Received: 5 September 2023 | Revised: 1 February 2024 | Accepted: 20 February 2024

IET Smart Grid

DOI: 10.1049/stg2.12161

The Institution of Engineering and Technology WILEY

REVIEW

Smart grid public datasets: Characteristics and associated applications

Emran Altamimi[1] | Abdulaziz Al-Ali[2] | Qutaibah M. Malluhi[1,2] | Abdulla K. Al-Ali[1]

[1]Department of Computer Science and Engineering, Qatar University, Doha, Qatar

[2]KINDI Center for Computing Research, Qatar University, Doha, Qatar

Correspondence

Emran Altamimi.
Email: ea1510662@qu.edu.qa

Funding information

Qatar National Research Fund, Grant/Award Number: NPRP12C-33905-SP-66; Qatar National Library

Abstract

The development of smart grids, traditional power grids, and the integration of internet of things devices have resulted in a wealth of data crucial to advancing energy management and efficiency. Nevertheless, public datasets remain limited due to grid operators' and companies' reluctance to disclose proprietary information. The authors present a comprehensive analysis of more than 50 publicly available datasets, organised into three main categories: micro- and macro-consumption data, detailed in-home consumption data (often referred to as non-intrusive load monitoring datasets or building data) and grid data. Furthermore, the study underscores future research priorities, such as advancing synthetic data generation, improving data quality and standardisation, and enhancing big data management in smart grids. The aim of the authors is to enable researchers in the smart and power grid a comprehensive reference point to pick suitable and relevant public datasets to evaluate their proposed methods. The provided analysis highlights the importance of following a systematic and standardised approach in evaluating future methods and directs readers to future potential venues of research in the area of smart grid analytics.

KEYWORDS

building management systems, data analysis, power consumption, power grids, SCADA systems, smart metres, smart power grids

1 | INTRODUCTION

Smart grids (SGs) are intelligent electric network models that incorporate the actions of all connected end users, including internet of things (IoT) devices [1]. This infrastructure enables seamless communication between users and grid operators, supporting various applications, such as self-healing, automation of the power grid, and integration of distributed energy resources (DER) [2]. SGs generate a massive, constant stream of data from various sources, such as customer data, grid data, and external data [3]. The power system has become significantly more complex with the integration of DER, electric vehicles, and demand response (DR) techniques [4]. Advanced data analytics algorithms are required to process this data and derive valuable insights for SG operations and services.

IoT devices play an important role in the data generation process, as seen in the incorporation of advanced metering infrastructures (AMI) and the use of smart metres in the SG. The data generated from IoT devices in the SG is characterised by its enormous volume, wide varieties, varying sampling rate, veracity, and value range [5]. These data can be grouped into three categories: customer data, grid data, and external data.

Customer data refers to any type of information about a customer, such as energy consumption and other related data. Examples include non-intrusive load monitoring (NILM) datasets and smart metering data. Table 1 summarises the consumer data categories and their respective features.

Grid data include all information about the electricity grid, such as specifications for generation plants and DER, the distribution grid, the transmission grid, electrical substations, energy storage, and supervisory control and data acquisition (SCADA) system data, which refer to data coming from a wide range of sensor types (e.g. wide-area measurement systems, intelligent electronic devices, power quality analysers, and pole

TABLE 1 The categories and features related to consumer data.

Data categories for consumer data	Features
Energy consumption	Energy consumption in a particular time interval. Might be collected from appliances/plug loads/phase loads or the aggregated load at a house level/building level or neighbourhood level.
Electrical measurements	Voltage, current, power, power factor. Might be collected from appliances/plug loads/phase loads, or the aggregated load at a house level.
Metre data	Status, ID, circuit ID and section (metre's location within the grid), manufacturer, installation date, reprogramming messages, service points, GIS data
Customer account data	Contracted power (maximum power contracted), type and status billing information (e.g., late payments), pricing rates, fraud history, price, peak load, and load factor
Financial data	Market and billing data

mounted auto-recloser). SCADA data includes grid assessments, voltage, current, power factor, alarm data such as unsolicited openings, details pertaining to repaired faults, control commands, circuit outages, transmission loss, network quality data measurement flows data, set points, and event logs. The use of this comprehensive dataset extends to various applications, such as system control, monitoring, alarm processing, protection, and event management. Table 2 summarises the sensors, types of measurement, and purpose of the data.

External data sources include regional meteorological and weather data. Geographic information systems (GIS) and temperature data are often used in research. The integration of IoT devices in the SG enables various applications on both the grid side, such as DR, and the consumer side, including home energy management systems (EMSs), ambient assisted living (AAL), and appliance anomaly detection.

The vast variety and immense number of datasets and data sources, the diverse number of applications that can be based on them, and the scarce availability of studies conducted on publicly available datasets in comparison to studies conducted on private datasets stifles research in the field. This research work aims to address this issue by reviewing the characteristics, issues, and applications of existing public datasets in detail. Such work will help researchers identify research gaps and directions.

1.1 | Contributions

The rapid transition of the power sector towards more sustainable and efficient smart grid systems, enhanced by the integration of IoT technologies, has resulted in a complex and data-rich environment. This work is motivated by the pressing need to guide researchers through this intricate landscape. By offering a comprehensive review and comparative analysis of smart grid datasets, we seek to simplify dataset selection for specific applications. This paper makes several important contributions to the field of electrical grid research, data analysis, which are concisely outlined as follows:

- Offers an extensive and systematic review of data sources in the electrical grid domain, encompassing smart metre datasets, NILM datasets, and grid datasets. This review emphasises publicly available data and facilitates the identification of relevant datasets for specific research questions or analyses, addressing the challenge of selecting the most suitable data sources in a rapidly evolving field.
- Analyses the features and characteristics of SG datasets, elucidating their applications and relevance in various research contexts, including IoT-based energy management solutions. A comparative analysis of the features, strengths, and weaknesses of various datasets is presented, enabling researchers to make informed decisions when selecting appropriate data sources for their studies.
- Examines the preprocessing methodologies, feature engineering techniques, and evaluation procedures employed by researchers fostering a deeper understanding of best practices in the field. This also aims to mitigate potential pitfalls in the utilisation and handling of diverse datasets, promoting a more robust and rigorous approach to research in IoT-driven SG systems.

1.2 | Previous work

In this section, we examine notable literature reviews and surveys in the domain, providing a concise overview of the existing knowledge in this field.

The work in ref. [11] investigates the SG architecture for the study of software reliability engineering. The article cites and discusses the characteristics of 15 datasets, which can be used for reliability engineering, and divides them into three main categories: Loss of loading probability, power distribution, and hardware. However, the article does not offer a detailed analysis of these datasets and their characteristics.

The comprehensive study by ref. [12] presents 13 consumer datasets and their characteristics while exploring deep learning techniques applied to load analysis, forecasting and management systems. The challenges associated with implementing deep learning techniques are discussed as well as potential solutions to enhance performance. Furthermore, the authors identify five open research issues concerning the future of SGs. In a related review paper, ref. [13] focuses on data analytics applications of smart metre data, featuring 10 datasets

TABLE 2 The devices, measurements and applications relating to SCADA systems.

Source of data	Measurements and features	Applications
Wide area monitoring system and phasor measurement units [6]	Voltage, current, power, phase angle and harmonics	Validate models and identify parameters, ensure dynamic stability, state estimation, and system control and protection (e.g., controlled-island protection) [7]
Remote terminal unit	Local collection points for collecting sensor reports.	Delivering commands to control relays
Digital fault recorder	Records and classifies faults (ex. Power swings, frequency fluctuations, and time of fault)	Faults recording and classification
Fibre bragg grating sensor	Wavelength shift	Overheating, sag, vibration, and galloping prediction
Hall effect sensor	Voltage and magnetic field	Speed detection, current sensing, proximity switching, and positioning
Power quality analyser [8]	Voltage and current levels, power factor, frequency, waveform distortions, harmonics, flicker, phase imbalance, voltage sags and swells, transient events, outages, crest factor, energy consumption, load patterns, interharmonics and inrush current.	Record power parameters and power interruptions such as under/over voltage, sags, swells, and noise
Sagometer	Temperature	Line sagging
Transformer sensors	Voltage, current, temp, partial discharge, load tap changer values, oil pressure, tripped situations discharge ground and short circuit current, etc [9]	Preventive maintenance
High voltage line temperature	Temperature	Preventive maintenance
Intelligent electronic devices	Status changes in substation and outgoing feeders	Relay protection
Capacitor sensors	Voltage, current, volt-amps reactive, and harmonic monitoring	Capacitor's bank control and monitoring
Pole mounted auto reclosers	Pick up events details	Fault diagnosis and prognosis [10]
Magnetoresistive sensor	Modulation, frequency, I, P, and energy	Electromagnetic interference monitoring in substations

with general characteristics (e.g. number of records, frequency and duration) and corresponding references. Although both reviews contain useful information, neither delves into extensive detail about these public datasets, which would be beneficial for researchers seeking suitable datasets for their studies.

Iqbal et al. [14] provide a comprehensive review of 42 NILM datasets, detailing their characteristics and statistical information. However, the authors do not discuss NILM applications or reference research articles that utilised these datasets. In contrast, the study in ref. [15] reviews several NILM datasets and their characteristics, while also mentioning the types of NILM approaches they permit, such as event-based or event-less methods. Despite these insights, the review does not elaborate on how the datasets were used or the specific techniques that were applied.

In the work of ref. [16], the models of the tools and the datasets that can be used to operationalise local energy communities in practice were reviewed. The reviewed use cases are of interest to stakeholders but do not specify particular applications of the data. The mentioned datasets consist of demand-side data and climate-related data, with specified characteristics. However, the specific uses of these datasets were not referenced.

The review paper [17] discusses publicly available distribution and transmission grid datasets, detailing their characteristics and intended usage. However, the work does not provide examples of research efforts that demonstrate the practical application of these datasets. In contrast, the authors in ref. [18] focus on publicly available test distribution networks with features in the United States, characterising them and identifying their use cases. Although providing valuable information, its scope is limited to public grid datasets with US features, leaving a broader perspective unexplored.

A comparison of the review articles and the contribution of our article is provided in Table 3.

1.3 | Methodology

The methodology followed is to construct three different comprehensive search strings for each data types. We used six major search libraries namely IEEE Xplore, ScienceDirect, Wiley Online Library, SpringerLink, MDPI, and ACM digital library. The search string used for macro and micro-level consumption data is ((“smart metre” OR “energy consumption” OR “system level” OR “substation”) AND (“smart grid” OR “power grid”) AND (“public dataset” OR “publicly available”) AND (“dataset”)) and it returned a total of 275 articles. For the second type, which is, detailed in-home consumption data we used the search string ((“Buildings” OR “Non-intrusive load monitoring” OR “NILM”) AND (“public dataset” OR “publicly available”)) which returned 250 articles.

TABLE 3 Comparison of current work with existing survey papers.

Review article	Scope of datasets reviewed	Depth of analysis	Applications and research gaps highlighted
[11]	Focused on datasets for software reliability in SG	Limited analysis of dataset characteristics	No detailed discussion on applications or methodologies
[12]	Consumer datasets with a focus on deep learning applications	Moderate detail on dataset characteristics	Some discussion on challenges and solutions in deep learning for SGs
[13]	Smart metre data analytics applications	General characteristics of datasets reviewed	Lacks depth in dataset usage and application-specific details
[14]	42 NILM datasets	Detailed statistical information of datasets	Lacks discussion on NILM applications or specific research articles
[15]	NILM datasets with a focus on event-based or event-less methods	Moderate detail on dataset characteristics	Limited elaboration on dataset employment and techniques
[16]	Tools and datasets for local energy communities	Specifies dataset characteristics but not in-depth	Lacks detail on specific data applications
[17]	Public distribution and transmission grid datasets	Detailed characteristics and intended usage	The intended applications for each dataset are not discussed
[18]	US-featured public grid datasets	Characterisation of datasets and use cases	Limited to US datasets and lacks broader perspective
Our work	Comprehensive review covering smart metre, NILM and grid datasets	Extensive analysis of dataset features, characteristics, and applications in smart and power grids	Detailed discussion of preprocessing methodologies, feature engineering, evaluation procedures, and identification of research gaps

Finally, for grid datasets we used the following search string (("grid dataset" OR "test system" OR "benchmark grid" OR "Representative grid" OR "Generic grid") AND ("smart grid" OR "power grid")) which returned 501 articles. In addition to the public datasets published, the search strings return research articles that utilised public datasets to evaluate their proposed approaches. Since datasets are still relevant regardless of the time they were released we opted to keep all articles regardless of the year the articles were published. Lastly, irrelevant articles are excluded based on the title and abstract of the article. The new numbers of articles for Macro and Micro-level datasets, detailed in-home consumption datasets, and grid datasets were 192, 149 and 320 articles respectively. The datasets used for evaluation by the articles that remained in the final set were then extracted. The dataset characteristics and associated applications were then extracted from the datasets' meta data and articles that utilised the datasets for various applications.

The rest of this paper is organised as follows, Sections 2–4 discuss micro and macro consumption data (which includes both smart metre and system-level data), detailed in-home consumption data and grid data, respectively. Each section discusses the applications, public datasets, and reviews the literature on the most popular datasets in each category. Sections 5 and 6 highlight data issues and conclude the paper. A graphical representation of the structure of the paper is shown in Figure 1.

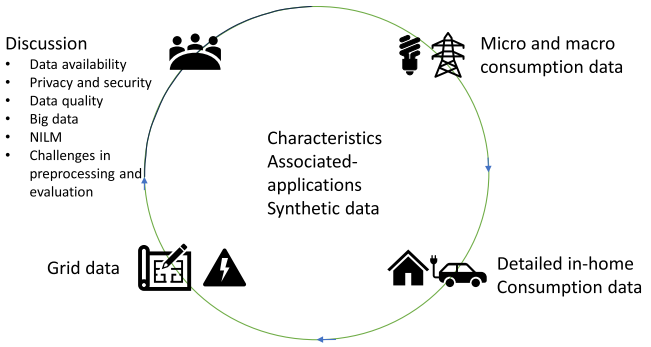

FIGURE 1 Road map of the paper.

2 | MACRO AND MICRO-LEVEL CONSUMPTION DATA

Smart metre data is the most commonly utilised in SG analytics, with a wide range of applications. They typically record data at intervals of 10–60 min. Smart metre data can significantly improve grid efficiency and long-term viability by providing valuable information on energy consumption, electrical measurements (e.g. voltage, current, power factors) [19, 20], metre-related issues, and outage data. Metre data management systems maintain records about each metre, such as its status, manufacturer, installation date, malicious behaviour, and reconfiguration data, as well as circuit installation locations and service point details. Service points represent the interface between the utility supply and a site's wiring system. Furthermore, customer account information, including contracted power, type, status, irregularities history, and billing data, can be leveraged for load forecasting by clustering similar customers [21]. A comprehensive summary of this information can be found in Table 1. Pablo et al. concluded in their work on 311K customers in Uruguay that complementary customer information and geo-localisation complement the consumption signal and are relevant features [22]. While smart metre data provides granular insights and allows for nuanced interventions and measures, system-level or macro-data carries its own significance. Macro

data provides a holistic view of consumption patterns in larger sections of the grid and are critical for high-level planning, management, and forecasting [23]. Conversely, micro or smart-metre data offers detailed load profiles of individual households, presenting opportunities for customised energy efficiency strategies and DR programs.

However, despite differences in scale and granularity, many applications, such as load forecasting, anomaly detection, and load management, incorporate an overlap between macro and micro data. For example, while load forecasting at the micro level informs individual household energy management strategies, at a macro level, it aids in power generation planning and grid stability measures. The methodologies developed for these applications can often be applied interchangeably between the two scales, although with adjustments to account for the inherent differences.

Therefore, given the considerable overlap in applications and in order to maintain coherence and efficiency in our presentation, we have elected to group both the smart metre (micro-level) and system-level (macro) datasets under the umbrella of "Macro and Micro-Level Consumption Data". This arrangement streamlines the discussion, eliminates redundancy, and underscores the interconnected nature of data analysis at different scales within the context of the smart grid.

This section discusses the most popular applications and public datasets and their characteristics.

2.1 | Consumption data applications

Yi Wang et al. categorised the applications of smart metre data into three main categories; load analysis, load forecasting [24] and load management [25–27]. Load analysis applications include bad data detection [28, 29], non-technical loss detection [30, 31] and load profiling [32–34]. Other applications include data compression [35], privacy [36–38] and outage management [39].

2.1.1 | Bad data detection

Bad data detection, or anomaly detection, is a crucial preprocessing technique that improves data quality and the accuracy of models and analytics by handling missing values or correcting/removing outlier data. Smart metre data are time series data, so existing techniques for time series data can be applied. However, traditional short-term load forecasting (STLF) methods for imputing bad data have limitations [40]. Probabilistic approaches also face challenges in determining optimal rejection thresholds, especially for large datasets [41–43].

There is a lack of publicly available datasets with labelled ground-truth fine-grained anomalies for the SG context, except for EnerNOC [44], which has a limited number of anomalies. Anomaly detection work is typically divided into two steps: defining or injecting synthetic anomalies and implementing an anomaly detection technique. Imputation is important when dealing with missing values, considering the rate of missing values and the cause of failure.

In ref. [45], the authors used Prophet by Facebook to define anomalies and evaluated classification models in the Ausgrid residential dataset. The best performance was achieved using the Random Forest classifier. In ref. [46, 47], the authors used modified generative adversary networks (GANs) and removing variational autoencoder-based techniques to impute missing values and anomaly detection, respectively, on the GEFCom 2014 dataset [48].

2.1.2 | Non-technical loss detection

Non-technical loss detection focuses on identifying discrepancies between energy injected and electricity paid for. It is closely related to anomaly detection but evaluates users' load profiles in the same neighbourhood or previous benevolent profiles to detect anomalies [30]. The only publicly available known labelled dataset for this purpose is the State Grid Corporation of China (SGCC) dataset [49]. Another approach focuses on detecting abnormal behaviours in a private manner, such as the work in ref. [50].

2.1.3 | Load profiling

Load profiling aims to understand users' or groups' typical patterns of electricity use, which is valuable for DR programs and prospective load forecasting [32]. Load profiling helps to better comprehend socio-demographic factors and target potential consumers for DR programs. Another research direction explores the development of privacy-preserving techniques and integrity assurance mechanisms for load profiling in SGs, with the goal of safeguarding sensitive smart metre data and maintaining the accuracy of outsourced data analytics processes [51].

2.1.4 | Load forecasting

Load forecasting is essential in the electric power industry for operations, planning, pricing, procurement, and hedging decisions. Load forecasting can be long-term or short-term, with different use cases for each. Preprocessing techniques in load forecasting include smoothing and imputation, feature extraction and selection, and clustering [52]. Various techniques are used, such as artificial neural networks, time series analysis, bottom-up approaches, SVM, and regression.

2.1.5 | Load management

Load management can provide better and more personalised services by collecting sociodemographic information [53]. Key aspects include customer base load estimation and tariff design. Customer base load estimation evaluates the effectiveness of DR programs by estimating load profiles without the program.

The literature is categorised into similar-day methods, regression-based methods, and morning-consumption-adjustment methods [54]. New approaches using high-frequency data, such as clustering-based methods, improve performance. Tariff design, on the other hand, is essential in balancing consumer response and utility provider profits. Clustering consumers is an important first step, followed by solving optimisation problems based on each cluster's load profiles [55]. The real-time price determination problem aims to maximise profits for SG retailers [56]. Price bidding in the SG plays a crucial role in demand side management by allowing consumers to participate in electricity markets actively. By submitting price bids for electricity use, consumers can influence the market price of electricity, encouraging energy savings and peak load reduction. This interactive process not only empowers consumers but also helps stabilise the grid by aligning energy usage with real-time supply and demand conditions [57].

2.2 | Smart metre and system-level datasets

In this section, we present a comprehensive review of all public smart metre and/or region datasets identified in the literature to the best of our knowledge at the time of writing this article, examining their characteristics, features, associated applications, and privacy considerations. The information is summarised in Tables 4 and 5. Table 4 summarises the datasets commonly used in the literature for certain applications.

2.2.1 | Low Carbon London

Low Carbon London (LCL) dataset [58] is an open dataset that involved 5567 consumers. Dynamic time-of-use tariffs was applied on 1122 of the consumers as part of an experiment carried out over the year 2013. The data set consists of the following:

1) Energy consumption (in kWh) sampled from smart-metres at 30 min frequency for each consumer. Data were collected for a total of 12 months during the experiment (i.e. when the dynamic time-of-use tariffs was in effect), in addition to 6 months before and 2 months after the experiment period.
2) Appliance survey that includes information such as number of appliances, physical parameters of the household (e.g. insulation, number of rooms etc.) and basic details of the occupants (e.g. number of occupants, age categories etc.). Data include 990 records from the group that opted for the dynamic time-of-use tariffs, and 1870 from the group that did not.
3) Attitudes survey to assess the change in consumption behaviour of the group that opted for the experiment, such as the factors that made them more likely to change their behaviour. Seven hundred fourteen records were received.

The privacy of consumers was preserved by doing the following:

1) Identifying information such as names, locations, and addresses was omitted.
2) ID keys were generated randomly.
3) The surveys were manually checked for any inadvertent inclusion of personal details.

With the help of historical data prior to the implementation of DR programs, some baseline load estimation algorithms are developed on this dataset [54]. The high frequency data also led to works in long and STLF as well [71, 72].

TABLE 4 Applications and corresponding public datasets.

Application	Datasets
Load forecasting	Low Carbon London (LCL) [58], PecanStreet [59], UMass smart* [60], ausgrid distribution network [61], customer behaviour trials (CBT) [62], AEMO [63], ERCOT [64], building data genome project [65], energy market authority of Singapore [66], EnerNOC [44], GEFCom2012 [67]
Demand side management, price bidding, and power market design	Ausgrid distribution network [61], customer behaviour trials (CBT) [62], ISO new england [68], energy market authority of Singapore [66], ERCOT [64]
Solar panel generation and net demand forecasting	Ausgrid distribution network [61]
Equipment failure modelling and voltage regulation	Ausgrid distribution network [61], UMass smart* [60]
Descriptive analysis and building characteristics	Building data genome project [65], energy market authority of Singapore [66]
Energy storage research	EnerNOC [44], PecanStreet [59], AEMO [63]
Anomaly detection and concept drift aware algorithms	UCI ElectricityLoadDiagrams20112014 [69], customer behaviour trials (CBT) [62]
Web-of-things studies and appliances management	PecanStreet [59]
NILM, NIOM, and data compression research	PecanStreet [59], UMass smart* [60]
SGCC dataset [70]	Energy theft detection [70]
Renewable energy effects and solar panel simulation	AEMO [63]

2.2.2 | PecanStreet

The PecanStreet dataset [59] is supported by the Pecan Street experiment in Austin, TX. Pecan Street's research network is the world's only real-world electricity-gas-water testbed. It covers over 1000 homes without renewable energy sources, 250 homes with solar panels, and 65 owners of electric vehicles. The energy generated and used in each home is monitored in real time, down to the circuit level. Energy is used, generated, and stored in high resolution at a frequency of a reading per second to per minute (both at the whole-home level and the individually monitored appliance circuits level). The experiment aims to understand the effects of modern technologies such as EMSs and DR programs such as time-of-use pricing.

The studies done using this dataset are numerous and span across various applications; however, this dataset is best suited for: Web-of-Things studies (due to the diversity and sparsity of the collected data) [73], NILM [74], appliance scheduling and management (since individual appliances were monitored) [75], design controllers for solar panel energy storage devices [76], as well as customer baseline load estimation [77]; due to availability of solar panel loads and DR programs that affect consumption as in the LCL dataset. For privacy preservation, PecanStreet authors mentioned that they implement "commercially reasonable security measures" for privacy and data protection. However, no details were mentioned on their privacy policy websites.

2.2.3 | UMass smart*

The UMass Smart* [60] dataset is a collection consisting of the following 9 subsets:

- DeepRoof dataset: Satellite images of building roofs and the planar segmentation of each.
- Apartment dataset: Aggregated energy consumption of 114 single family apartments for the period of 2014–2016, together with their associated weather data. Readings were sampled once per minute.
- Home dataset (2017 release): The aggregated and individual circuit consumption of 7 households collected at a per minute interval over multiple years.
- NIOM (Non-intrusive-occupancy-monitoring) dataset: aggregated consumption at minute level for a 3 week period for two households of two occupants each, with the ground truth occupancy status.
- Home dataset (2013 release): This dataset focuses on depth instead of breadth. That is, only three houses were monitored. However, the data included information about consumption (per circuits and aggregate, individual metres, dimmable and non-dimmable switches), two electrical phase data (voltage and frequency), environmental data (indoor and outdoor), oven and door status, energy generation data (solar panels, wind, and battery voltage), and motion detector data. The dataset also includes micro-grid dataset of 443 homes over a single-day period.
- Solar-TK: contains solar energy generation data from 81 homes in the US.
- Solar panel: includes 50 rooftop solar panels energy generation data at a 1-min interval.
- SunDance: includes 1 year's data of 100 solar sites in North America in 2015. Net metre, solar generation, and weather data were collected at a frequency of 1 sample per hour.
- Physical-Black-Box Model: This dataset includes weather and normalised solar generation data to build the physical black-box model. The files also include code to model shading effects.

Applications of this dataset include a privacy-preserving architecture developed in ref. [78] while still meeting the utilities' needs to achieve a net metering goal. The solution uses the concept of Zero-Knowledge proofs and provides cryptographic guarantees for the integrity, authenticity, and accuracy of payments, while permitting changeable pricing without disclosing the power measurements acquired throughout a billing period. NILM and NIOM algorithms development can be done on this dataset, because of the availability of circuit and appliance level consumption data, as well as the ground truth for detecting occupancy status [79]. Solar panel data can also be used as auxiliary data for distribution grid management algorithms such as voltage regulation as in ref. [80]. The high frequency polling of data per minute also prompted some researchers to study data compression algorithms such as the work done in ref. [81].

2.2.4 | Ausgrid distribution network: residential and substations

The Ausgrid distribution network records and publishes four types of datasets [61]:

- Electricity consumption: Ausgrid has grouped the yearly residential and non-residential electricity consumption data by local government areas (LGA), total of 32 areas, in its distribution system. High-voltage customers and supply services such as public lighting and bus shelters are not included in these data.
- Solar panels and electricity consumption: A sample of 300 solar customers from Ausgrid's electricity network area was randomly selected, all of whom were billed on the domestic tariff and possessed a gross metered solar system throughout the duration from 1 July 2010 to 30 June 2013. To compile the data, metre reading processes were employed to obtain a comprehensive dataset of actual electricity consumption and production at half-hour intervals for the selected customers during the specified period. Customers who fell at the extremes of household consumption and solar generation performance during the first year of the study were excluded. Solar homes with rooftop solar systems connected to the grid through a gross metering configuration account for 2657 of the monthly

TABLE 5 Data-sets characteristics.

Dataset	Geographic location	Number of consumers	Data duration	Data frequency	Dataset characteristics	Privacy considerations
Low Carbon London (LCL)	London	5567	12 months during the experiment, 6 months before and 2 months after	30 min	Energy consumption from smart metres, appliance survey, attitudes survey	Omitted identifying information, randomly generated ID keys, manual check of surveys for personal details
PecanStreet	Austin, TX	Over 1000 homes without renewable energy, 250 homes with solar panels, and 65 electric vehicle owners	Real-time monitoring (1-min)	High-resolution, from a reading per second to per minute	Monitors energy usage, generation, and storage at the whole-home level and individually monitored appliance circuits level	No specific details provided
UMass smart	United States (specific locations vary depending on the subset)	Varies depending on the subset			Aggregated and individual energy consumption, circuit-level data, solar generation data, weather data, occupancy status, voltage and frequency data, environmental data, oven and door status, motion detector data, shading effect modelling code	Privacy-preserving architecture using zero-knowledge proofs.
Ausgrid distribution network: Residential and substations	Ausgrid's distribution system, Australia	32 LGAs, 300 solar customers, 2657 solar homes, 4064 non-solar homes, and 180 substations	Yearly data: 2010–2013; monthly data: 2007–2014; substation data: since 2005, regularly updated; past outages: Published every 3 months	Yearly, half-hourly intervals for solar customers, 15-min intervals for substations, and quarterly for past outages	Electricity consumption, solar generation, substation load profiles, and power outage records	Identifying information removed
Customer behaviour trials (CBT)	Ireland	5375 households	18 months	Electricity consumption every half an hour	Split into benchmarking and test period; four different groups with different time of use tariffs; includes a survey of 143 questions about household characteristics	Not mentioned
The State grid corporation of China (SGCC) dataset	China	42,372 (3615 thieves and 38,757 honest consumers)	1 January 2014–31 October 2016 (1035 days)	Daily consumption readings	Labelled for malicious activities (electricity theft)	Not mentioned
ISO new england[a]	New England, USA	Not specified	Since 2003	Hourly	System-level hourly load data, temperature data, location regional prices, market clearing prices, interchanges with other power systems for 9 zones	Not mentioned

TABLE 5 (Continued)

Dataset	Geographic location	Number of consumers	Data duration	Data frequency	Dataset characteristics	Privacy considerations
AEMO[a] [63]	5 Australian states	5 states	Since 1998	Half-hourly	Aggregated demand data, electricity price data	Select profile values may be manipulated if deemed sensitive.
ERCOT[a] [64]	Texas, USA	4 regions	Since 2001	Hourly	Market and grid information	Requires submission request for access
GEFCom2012 [67]	United States	20 zones	1 July 2003–30 June 2008	Hourly	Temperature data from 11 weather stations, holiday data	Not specified
Building data genome project [65]	University campuses (mostly)	507 whole (non-residential) buildings	February 2014—April 2016	Hourly	Electrical metre data, gross floor size, primary use type, meteorological information	Not specified
Energy market authority of Singapore[a] [66]	Singapore	Not specified	Since 2004	30-min intervals	System demand data, historical market prices	Not specified
EnerNOC [44]	Not specified	100 commercial/industrial sites	2012	5-min intervals	Energy consumption data	Anonymised measurement values, identifying information removed
UCI Electricity LoadDiagrams 20112014 [69]	Portugal	370 customers	2011–2014	4 readings per hour	Residential and commercial buildings and consumers	Not specified

[a]The dataset only includes system-level data.

electricity data points. The Ausgrid Distribution Network also provides monthly electrical data. Data are provided for the period from 1 January 2007 to 31 December 2014 and, as a result, it includes periods of household electricity consumption prior to the installation of the solar system. Furthermore, a data set of 4064 non-solar homes is provided for the same time period to compare electricity consumption patterns between the two datasets.

- Ausgrid substation data: Since 2005, Ausgrid has provided public access to the load profiles of approximately 180 zone substations through their website, with regular updates that ensure the data set remains current. Each entry in the dataset contains the year, zone substation name, date, and corresponding data unit, followed by a full day's worth of measurements at 15-min intervals.
- Past outages: Power supply interruptions that affect 50 or more customers and last for more than 5 min are recorded in the database and published quarterly. The dataset contains information on the start time, average duration of the outage in minutes, number of consumers affected, and its potential cause. The data are organised by LGA as is done for the electricity consumption subset.

Energy management solutions such as storage and DER scheduling and customer baseline load estimation can be implemented in the solar panels and electric consumption (residential) dataset. Electricity consumption, solar panel generation, and net demand forecasting can also be implemented in this dataset. For the Ausgrid substation and past outages datasets, aggregated load forecasting, and demand side management (e.g. planning of charging infrastructure [82] and electrical distribution system planning), and modelling equipment failure (e.g. power transform failures and retirement statistics [83]) are the most common applications. Customers in these datasets have been deidentified and do not represent a statistically significant sample of residential customers in the Ausgrid network area, nor have they been subjected to detailed occupancy checks.

2.2.5 | Customer behaviour trials

Customer behaviour trials (CBT) dataset [62] consists of 5375 households electricity consumption data recorded every half an hour for the span of 18 months. The data was collected by the Commission for Energy Regulation of Ireland. The objective of the CBT dataset is to evaluate smart-metre technology time-of-use tariffs and different demand side strategies. Therefore, the data was divided into two phases: the benchmark period (6 months) and the test period (12 months). In the trial, four different groups were assigned different time of use tariffs. A survey (of 143 questions) on household characteristics is also included. The survey aims to depict the socio-demographic characteristics of the household; employment status, household size, age, and the social class. Given that consumers were incentivised to change their behaviours through demand-side strategies, the authors believe that the dataset is a good benchmark to develop. Concept drift aware algorithms. As reported in the study, 82% reported making some changes in their consumption patterns and 74% reported drastic changes in their households. The trial reported noticeable drastic changes in 38% of the consumers.

2.2.6 | The state grid corporation of China

SGCC [70] released the daily electricity consumption of 42,372 consumers in the period from 1 January 2014 to 31 October 2016 for a total of 1035 days (with a consumption reading per day). The dataset is also labelled for malicious activities for a total of 3615 thieves. The other 38,757 consumers are labelled as honest consumers. Labelling electricity theft acts as the ground truth to evaluate models.

2.2.7 | Independent system operator New Englad

Every month since 2003, the independent system operator (ISO) New England publishes [68] system-level hourly load data, as well as corresponding temperature data, regional location prices, market clearing prices and interchanges with other power systems for 9 different zones. The market data allows for studies in power market design [84], price bidding [85], as well as price forecasting.

2.2.8 | Australian energy market operator

The Australian Energy Market Operator (AEMO) [63] serves as the main entity responsible for overseeing the management and operation of electricity and gas networks, as well as price determination, in five states in Australia. This organisation maintains a comprehensive dataset that includes aggregated demand data and electricity price data for these states, with temporal granularity provided at a half-hourly rate. However, it should be noted that beginning in November 2021, the resolution of these data experienced a significant enhancement, with the frequency of data points increasing to every 5 min. Data have been updated and available since 1998. The research done on the datasets focused mostly on STLF. However, some descriptive analysis work was also done. For example, in ref. [86], the effects of wind and solar panel generation on wholesale electricity prices were studied. The authors in ref. [87] used the dataset to design the optimal battery capacity of solar panels. The authors also simulated the hourly generated solar panel power and made it public [88].

2.2.9 | Electric reliability council of Texas

The Electric Reliability Council of Texas (ERCOT) [64] is an ISO responsible for overseeing the state's electrical transmission and distribution network, serving over 25 million customers. Since its inception in 2001, ERCOT has managed the deregulated wholesale electricity market and has provided

various datasets to the public, including real-time and day-ahead market data, transmission and generation data, and renewable energy data. These datasets encompass energy prices, demand, and generation capacity for the entire ERCOT region, divided into four load zones. Access to ERCOT datasets requires a submission request through their website. These datasets have been utilised for various purposes, such as STLF and price forecasting (e.g. ref. [89]).

2.2.10 | Global energy forecasting competition 2012 (GEFCom2012)

The Global Energy Forecasting Competition 2012 (GEFCom2012) [67] is a hierarchical load forecasting contest for a utility located in the United States with 20 zones (from 1 July 2003 to 30 June 2008). The dataset includes temperature data from 11 weather stations and the holidays of that time period. The authors in ref. [67] reviewed the winning solutions.

2.2.11 | The building data genome project

The Building Data Genome Project [65] consists of 507 whole (non-residential) building electrical metres data from February 2014 to April 2016, most of which come from buildings on university campuses. The dataset also includes different distinctive meta-data such as gross floor size, primary use type, and meteorological information. The dataset was developed primary to test various algorithms and feature extraction techniques. Such use cases include load forecasting, load shape/profile clustering, and synthetic load data creation [90] and inference of buildings' characteristics [91].

2.2.12 | Energy market authority of Singapore

The Energy Market Authority of Singapore publishes numerous statistics pertaining to the grid operation [66], notably the system demand data polled at 30-min intervals since the beginning of 2004 and historical market prices. The datasets were used for reliability analysis [92, 93], descriptive analysis (for example, analysis of customers responding to socioeconomic determinants [94]) and demand-side bidding [95].

2.2.13 | EnerNOC

EnerNOC [44] collected 5-min energy consumption data for 2012, for 100 commercial/industrial sites. EnerNOC is the only dataset that we are aware of that has labelled anomaly data, which can be used as the ground truth for developing anomaly detection algorithms. However, further examination revealed that the number of anomalies in the dataset is arguably negligible (no more than 11 instances out of more than 100,000 readings). For privacy, the real measurement values and identifying information such as geolocations and floor area have been anonymised. However, the values were shifted on a linear scale to ensure consistency of the comparison over time and across sites. The data set has been used in the design of energy storage systems [96, 97] and load forecasting [98].

2.2.14 | UCI ElectricityLoadDiagrams20112014

The ElectricityLoadDiagrams20112014 is a real-world dataset from Portugal [69]. The dataset has a resolution of 4 readings per hour from 2011 to 2014 for 370 customers. The dataset includes residential and commercial buildings and consumers.

2.3 | Discussion

While datasets from the UK, US, Australia, Ireland, and Portugal provide valuable insights into energy consumption, the majority originate from developed nations in the Northern Hemisphere. This overrepresentation may limit the applicability of research outcomes to the diverse energy landscapes of developing countries, particularly those in the Southern Hemisphere, where different economic, infrastructural, and climatic conditions prevail. The geographic concentration of these datasets suggests potential limitations in the generalisability of research findings. The distinct energy consumption patterns, regulatory frameworks, and customer behaviours specific to the United States may not be directly transferrable to other global contexts. This limitation underscores the need for a more diverse compilation of datasets that encapsulate the variegated nature of energy systems across different regions and cultures to truly harness the universal applicability of smart grid analytics. The issue is particularly relevant in developing countries, where infrastructural, economic, and policy differences shape distinct energy dynamics. Emerging markets often prioritise expanding energy access, diverging from patterns found in more developed nations. Consequently, a more inclusive dataset collection is imperative for globally relevant smart grid analytics.

Additionally, the analysis of dataset utilisation across different applications as demonstrated in Figure 2 reveals a pronounced emphasis on load forecasting, particularly within system-level datasets that benefit from frequent updates. This trend aligns with the critical role that load forecasting plays in the operational planning and reliability of the electrical grid. Load forecasting's predominance in the literature is indicative of its foundational importance in grid management and the value placed on accurate and timely predictions.

Furthermore, the synthesis of the dataset characteristics and their respective applications into a coherent framework presents an opportunity for a targeted approach to dataset utilisation. Table 6, which delineates the most popular datasets for each application, serves as a practical guide for researchers and practitioners in the field. By identifying the datasets most suited to specific applications, this summary aids in the efficient allocation of analytical efforts and resources.

In conclusion, the analysis of smart metre and system-level datasets highlights the centrality of certain applications in

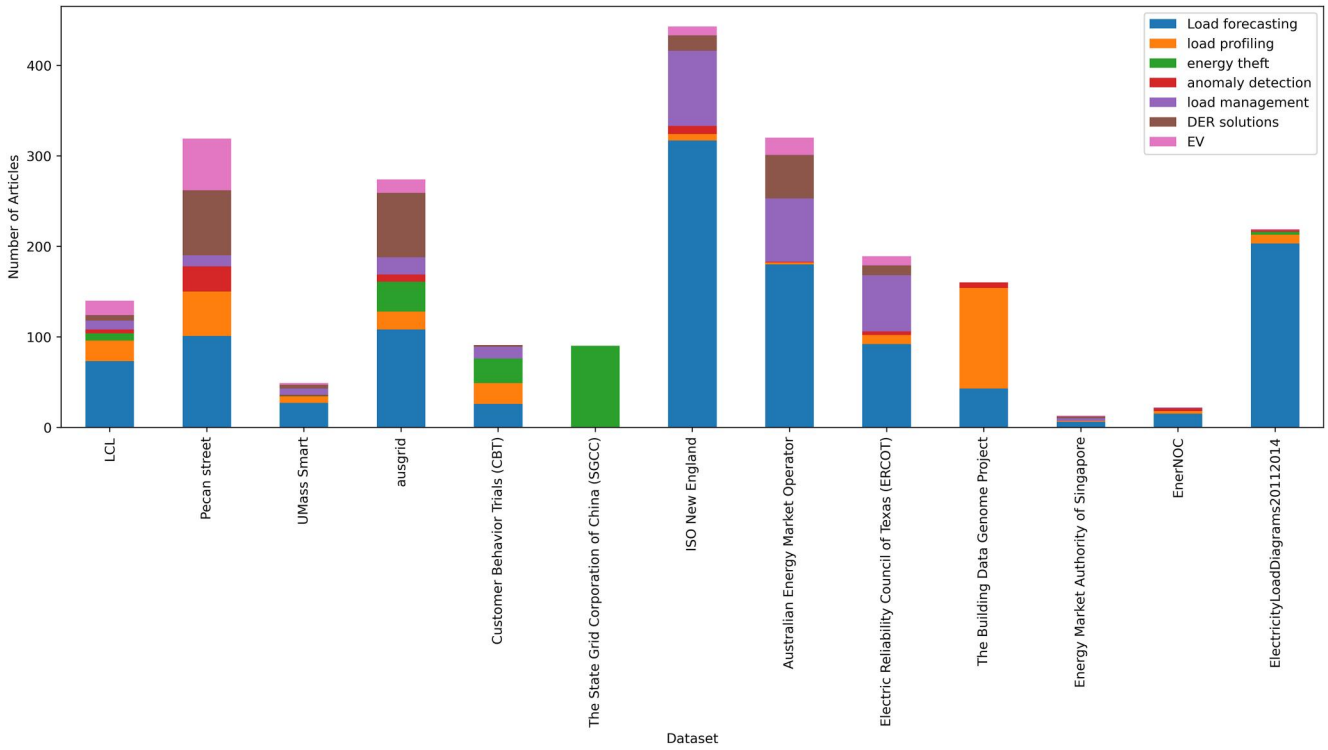

FIGURE 2 The count of articles that utilised the public datasets for particular applications.

TABLE 6 The most common public dataset for each application.

Application	Most common datasets
Load forecasting	ISO New England, ElectricityLoadDiagrams20112014 and Australian energy market operator
Load profiling	The building data genome project, pecan street and LCL
Energy theft	State grid corporation of China (SGCC), ausgrid and customer behaviour trials (CBT)
Anomaly detection	Pecan street, ISO New England and ausgrid
Load management	ISO New England, Australian energy market Operator and electric reliability council of Texas (ERCOT)
DER solutions	Pecan street, ausgrid and Australian energy market operator
EV	Pecan street, Australian energy market operator and LCL

smart grid analytics and the geographic concentration of dataset origins. The field stands to benefit from an expansion of data sources that better represent the global diversity of energy systems and from leveraging the specialised utility of each dataset. This dual approach can enhance both the breadth and depth of insights in smart grid analytics, fostering advancements that are both innovative and inclusive.

3 | DETAILED IN-HOME CONSUMPTION DATA

NILM systems provide an efficient way to monitor multiple appliances without the need for submonitoring, hence the name NILM. This section focuses on datasets that enable such systems which are commonly referred to as buildings' datasets [99] or NILM datasets [15]. NILM datasets contain data from electrical measurements taken at a very high sampling rate at the plug load, individual circuits in the house, and/or the main line. Data may also include environmental measurements (e.g. temperature), auxiliary data, and information about events such as occupancy status (i.e. how many occupants are inside at any given time) and switches. The availability of labelled power events allows for event-based approaches in energy disaggregation, in contrast to event-less approaches when power events are not labelled. The datasets might also include information on the weather both inside and outside the building. This high frequency is different from smart metre datasets, which typically take measurements every 10–30 min, with most of the commercial smart metre's sampling at less than 1 Hz. Gao et al. [100] suggested a 4 KHz threshold for a feasible and reliable classification of appliances in energy disaggregation.

Higher sampling frequencies of the electrical measurements enable features such as transient information, voltage-

current trajectories, electrical noise, and (power, reactive power, and distortion power) trajectories. Although low sampling frequencies can be used to achieve some NILM applications, transient analysis cannot be performed, limiting overall performance and the range of applications that can be used. Voltage, current, and power variables are the features that are most important in low-sampling-frequency NILM datasets, with reactive power being the distinctive feature that is most frequently used in research.

If the aggregate metering of electricity consumption is not available and only the measurements of individual appliances is available, the data lack a ground truth for evaluating and testing energy dis-aggregation models. Therefore, the dataset is used only for training, while the evaluation is performed on other datasets. The naive method of aggregating all the appliances does not serve as a ground truth since most of the appliances in the house are not monitored. If the appliances events are labelled, then such datasets might be used for event classification.

This section discusses the most popular applications and public datasets in NILM datasets.

3.1 | Detailed in-home consumption data applications

NILM datasets are primarily used for developing algorithms to disaggregate total consumption into individual appliances. The output of energy disaggregation systems can be used for purposes such as reducing energy consumption, preventing appliance failures, forecasting SG consumption peaks, and monitoring daily living activities [101].

3.1.1 | Energy disaggregation

The energy disaggregation process typically involves three stages: Event detection, feature extraction, and load identification [102]. Event detection captures appliance state transitions, while feature extraction uses steady-state, transient, and non-traditional event detection approaches to extract relevant features [103].

3.1.2 | Energy management system

EMS combines hardware and software to monitor and control energy consumption and generation within a home, helping consumers save on utility bills while maintaining comfort levels [104, 105]. DR solutions incentivise customers to actively control their energy demand based on market prices [106, 107].

3.1.3 | Condition-based maintenance

Condition-based maintenance monitors equipment conditions and performs maintenance tasks based on equipment status, allowing early detection of minor failures and more efficient maintenance strategies.

3.1.4 | Ambient assisted living

AAL focuses on products and services that improve the lives of elderly adults and promote their physical independence. NILM systems can facilitate AAL without the need for obtrusive monitoring [108].

3.1.5 | Appliance anomaly detection

Detecting anomalous appliances using NILM techniques is more cost-effective and practical than using individualised metres per appliance [109, 110]. However, further development is needed to improve the effectiveness of NILM-based anomaly detection [111, 112].

3.2 | Detailed in-home consumption public datasets

The NILM datasets have been extensively reviewed in the literature, for example, the authors in ref. [113] provided a comprehensive review of 29 existing open datasets, in terms of settings (residential or otherwise), measurement level (whole premises, individual appliances, and/or individual circuits), electrical and auxiliary measurements, time period, event labels availability and file format. The authors in ref. [16] reviewed 22 open datasets providing the country, the number of households/sites in the dataset and the sampling rate. In the work of ref. [15], 26 datasets were reviewed that provide the same information as the work of ref. [113], in addition to the country of origin. A critical review of all NILM datasets was published in ref. [14] in 2021, in which 42 datasets were comprehensively reviewed. The datasets were divided into high-frequency, low-frequency, and synthetic datasets. Providing the same characteristics mentioned in refs. [15, 16, 113] in addition to the name and number of appliances measured.

Table 7 reviews (24) NILM datasets with respect to their measurement levels and frequency, measured quantities and sampling rate, and the applications for the datasets.

4 | GRID DATA

Electrical grid data prove invaluable for examining typical grid operating conditions and analysing grid behaviour during failures and disturbances. Furthermore, it facilitates the investigation of microgrids in islanding conditions, where the microgrid is disconnected from the main grid, as well as the integration of renewable energy sources. The electrical grid encompasses power generation, transmission, and distribution components, and grid data in the literature enables the emulation of electrical measurements and sensors using various

TABLE 7 Measurement levels and frequency, measured quantities and the sampling rate, and the applications for 24 NILM datasets.

Dataset	Measurement levels	Measured quantities	Labelled events	Applications
REDD [114]	AGG (15 KHz), IC (0.5 Hz), IA (1 Hz)	P, V, I, S		STLF, load disaggregation
UK-DALE [115]	AGG (1 Hz), IC, IA (per 6 s)	P, V, I, S		Energy dissaggregation, behaviour analytics
BLUED [116]	AGG, IC, IA (all at 12 KHz)	P, V, I, Q	State transitions	Event detection
AMPds [117]	AGG, IC, IA (all at 1 sample per minute)	P, V, I, Q, f, phi, Pt		STLF, load disaggregation
ECO [118]	AGG, IA (both at 1 Hz)	P, V, I, Q	Occupancy status	Occupancy detection, energy dissaggregation
Tracebase [119]	IA (per 10 s)	P		Steady and transient state analysis of appliances, energy dissaggregation.
HES [120]	IA (per 2 min)	P		Behaviour and descriptive analytics
iAWE [121]	AGG, IC, IA (all at 1 Hz)	P, V, I, Q	Only 1 day, time-stamped data	Supervised energy dissaggregation techniques
GREEND [122]	IA (1 Hz)	P		Energy dissaggregation, EMS
PLAID-I [123]	IA (30 KHz)	V, I		Load identification at plug level
REFIT [124]	AGG, IA (both per 8 s)	P		Energy dissaggregation, behaviour and descriptive analytics
COMBED [125]	AGG, IC (both at per 30 s)	P, I		Energy dissaggregation for commercial building
DRED [126]	AGG (per 1 min.), IA (1 Hz)	P	Occupancy (room level per 1 min.)	Location aware energy dissaggregation
WHITED [127]	IA (44.1 KHz)	V, I		High frequency energy dissaggregation
HFED [128]	IA (10 and 5 MHz)	Electro-magnetic interference	Energy dissaggregation generalisation beyond lab settings	Developing energy dissaggregation models that are able to generalise beyond lab settings
SUSTDataED [129]	AGG (12.8 KHz), IA (50 Hz)	V, I	State transitions and occupancy measurements	energy dissaggregation and occupancy detection
ACS-fx [130]	IA (per 10 s)	P, V, I, phi		Low frequency load identification
COOLL [131]	IC (12 KHz)	V, I	20 variations of dictinct energy consumption profiles of 12 appliances.	Energy dissaggregation
Dataport [132]	AGG, IC (both at 1 Hz and per min.)	P, I		Energy dissaggregation
BLOND [133]	AGG (50 and 250 KHz), IA (6.4 and 50 KHz)	P, Q		High frequency energy dissaggregation
RAE [134]	AGG, IC (both at 1 Hz)	P, V, I, Q		Energy dissaggregation
BERDS [135]	AGG, IA (per 20 s)	P, Q, S		Energy dissaggregation
EEUD [136]	AGG (per 1 min.)	P		Simulating and analysing electricity consumption for residential building
I-BLEND [137]	AGG (per 1 min.)	P, V, I, pf, f	Occupancy status per 10 min	Occupancy detection, energy dissaggregation

Abbreviations: I, current; P, power; pf, the power factor; phi, the phase angle; Pt, total power; Q, reactive power; S, apparent power; V, voltage.

tools. It is worth noting that researchers often employ interchangeable terms when referring to grid datasets, such as network, case, system, and grid. There are several terms for grid datasets that are not always used consistently in the literature, due to a lack of standardisation [17]:

- Test systems: A simple grid built for the purpose of demonstrating a single problem or performing basic validation or testing. Synthetic [138, 139] or real grids [140, 141] are named test systems. IEEE case 9 [142] and ICPSs [143–146] are examples of test systems.
- Benchmark grids: Grids where the aim is to compare and evaluate different algorithms. For example, the CIGRE systems [147] and the authors of ref. [148] presented benchmark systems. However, it is worth noting that the IEEE test cases are typically used as a benchmark (e.g. for power flow analysis), which highlights the issue of the interchangeable use of grid terms.
- Representative grids: Are grids that represent real grids and/or a set of grids that share similar characteristic (e.g. rural grids). Such grids bridge the gap between technical findings and real-world grids [149].
- Generic grids: The work in ref. [150, 151] used the term generic to refer to a grid where different parameters can be tweaked to generate various grids. However, the term was synonymous with representative grid in the work of ref. [152].
- Synthetic grids: Grids that are neither models of real grids nor derived from a real-world grid.

This section discusses the most popular applications and public datasets for grid data.

4.1 | Applications

Grid datasets are used for various applications such as planning, stability analysis, reliability analysis, state estimation, and power flow analysis [153]. The SG paradigm has expanded research opportunities in the effective integration of DER and storage devices within the power grid, focusing on assessing the impact of incorporating these elements and evaluating their potential to reduce generation costs, smooth power generation curves, and maintain sustainable service reliability for users [154].

4.1.1 | Planning

Power system planning faces challenges such as generation expansion planning (GEP) and transmission expansion planning (TEP), which involve determining the ideal combination of technology, location, and building time for new generation units and power lines [155]. Both GEP and TEP are formulated as optimisation problems with constraints such as the electricity market, congestion, uncertainties, and other considerations [156–163].

4.1.2 | State estimation

State estimation determines the state of the power grid from imperfect measurements, used for online applications like security analysis, anomaly detection, and fault diagnosis, or offline purposes like planning [164]. With the advent of the SG, state estimation is increasingly important for distribution grids [165].

4.1.3 | Power flow analysis

Power flow analysis examines the flow of power in a networked system, analysing steady-state operations of power systems and optimising power flow for efficiency [166].

4.1.4 | Reliability and stability analysis

Reliability analysis studies the life cycle of components and the system level, while stability studies examine the steady state and transient stability of power grids [167].

4.2 | Transmission and distribution grids

The transmission grid is responsible for delivering the load over long distances from a generating site to electrical substations, while the distribution grid is responsible for delivering energy to consumers. The authors of ref. [168, 169] classify the data collected from the grid into:

- Standard equipment (e.g. transformers, switch gears, circuit breakers, storage batteries, transmission cables, and ccables)
- Technical parameters (e.g. transformers and capacitor ratings, voltage levels, and number of buses)
- Cost and maintenance data
- GIS data of the power lines, service points, and buildings
- Substations data and locations
- Parcel use category (e.g. residential)

There are several works that reviewed available grid datasets. The work in ref. [17] has reviewed steady-state distribution grid datasets highlighting the intended use case. The authors in ref. [153] reviewed the IEEE and CIGRE benchmark test systems, highlighting the applications done on each. A review of distribution test systems in the United States is presented in ref. [18]. The authors analysed IEEE test systems, Pacific Northwest National Laboratory test systems, Electric Power Research Institute representative systems, and the Pacific Gas and Electric Company (PG&E) grids. The IEEE PES Working Group on Cascading Failure [170] provides a comprehensive review of test systems providing the intended use case and technical details on the test grids. This section provides a comprehensive concise summary of the most popular grid datasets along side the intended use cases and popular applications for these datasets.

The Test Feeder Working Group originally released five test feeders: IEEE 4, 13, 34, 37, and 123 bus test feeders. Test feeders are synonymous with test systems with the exception that test feeders have only one power source while test systems incorporate multiple power sources. They were intended to benchmark power flow algorithms, however, various analysis and research was conducted on the five test feeders originally released [171]. The test feeders are not representative of large and complex distribution grids and were small to medium radial feeders. In 2010, a sixth test feeder, called the IEEE Comprehensive Test Feeder, was added to model various components of the grid and transformers in particular [172]. The feeders are comprised of overhead lines and underground cables, voltage regulators, shunt capacitors, and various degrees of load unbalance [171]. Table 8 summarises the intended use cases of the original feeders and other prominent applications in the datasets.

Since then, several benchmark test systems have been made public to serve as a standardised dataset to test various methods and algorithms [186]. All IEEE 9, 14, 30, 39, 57, 118, 300, and Reliability Test Systems (RTS)-24 and RTS-73 test systems allow for power flow, state estimation, and planning studies. However, only IEEE RTS-24/73 allows for reliability analysis and IEEE 39 for stability analysis and development of control schemes. Modifying the test systems to allow for different analyses is possible [153].

In 2010 a **8500-bus test feeder** was published to represent a full-size distribution system [187], still allowing for the same intended use cases in Table 8. The test system was also used in time series load modelling [188] and DER integration in the SG [189].

Three test feeders and systems were published to tackle specific scenarios and to subvert common assumptions. Table 9 summarises the test feeders and systems with the intended use case and common applications.

Texas A&M university hosts several datasets on their website [198] for electric grid test cases that cover a variety of systems and scenarios, and are crucial in different power system analyses. These datasets do not contain Critical Energy Infrastructure Information (CEII), making them widely accessible for research purposes.

Among the datasets are the latest synthetic electric grid cases of 2023, which include a smaller self-contained island test case for the Hawaiian island of Oahu with a synthetic 138/69 kV transmission network. For larger-scale scenarios, there are datasets such as the Texas Synthetic Grid, which covers the ERCOT portion of Texas with a 6717-bus transmission network, and the Combined East-West US Grid, representing a

TABLE 8 The original IEEE test systems and their respective intended use case and common applications.

IEEE original test systems	Intended use case and common applications
4-bus	Transformer modelling testing. State estimation [173] and step-voltage regulators [174].
13-bus	Testing power flow convergance in unbalanced systems. Optimal capacitor placement [175], control of renewable energy batteries in microgrid [176], and islanding detection in microgrids [177].
34-bus	A test system that requires voltage regulators to comply with ANSI voltage standards. Optimal distributed generator placement [178] and optimal placement of storage systems [179].
37-bus	Capability of software to solve for the less common three-wire delta systems. Power flow analysis with DER [180], distributed generators for providing reactive power [181], and micro-grid small signal analysis [182]
123-bus	Minimising voltage drops with voltage regulators and shunt capacitor. Power flow analysis in unbalanced systems, operational planning for self-healing action [183], stochastic reactive power management in microgrids with renewable energy [184]
CTF	Capability of software to solve for a variety of components in one system. Distributed generation applications [185]

TABLE 9 Test feeders and systems, highlighted characteristic, and the intended use case and common applications.

Test feeders and systems	Highlighted characteristic	Intended use case and applications
Neutral-earth-voltage test feeder	The neutral conductor is not reduced by Kron reduction [190] because the neutral voltage is above zero.	Study neutral voltages in case of connection failures. Harmonic analysis [191] and load modelling [192]
Low voltage network test system	A low voltage highly meshed system that represents typical urban areas. The system is also referred to as 342-bus LVNTS.	Tests software capability to handle highly meshed systems. Economic dispatch with DER integration [193] and planning of communications systems [194].
European low voltage test feeder [195]	Represents a typical feeder in Europe and the first feeder to operate at 50 Hz.	Tests software capability to solve for various test feeders. State estimation with DER integration [196] and optimal sizing and placement of renewable energy batteries [197].

synchronously intertied model of the US portion of the eastern and western interconnects.

Datasets also exist for the ARPA-E Performance-based Energy Resource Feedback, Optimisation, and Risk Management (PERFORM) program. This program aims to optimise grid management as the penetration of variable renewable resources continues to increase. For this program, specific cases such as the 6717-bus Texas Case and 24,000-bus Midwest Case were created.

Additional datasets from 2021 to 2023 include synthetic transmission and distribution test cases, such as the Full Texas Synthetic Transmission and Distribution Test Case, and the 150-bus Synthetic Transmission and Distribution Test Case based on Travis County, Texas. These datasets also include restoration data and scenarios, as well as an associated natural gas pipeline network.

Other notable datasets come from the ARPA-E GridData program, which offers synthetic electric grid models. These models are designed to be statistically and functionally similar to actual electric grids, ensuring the confidentiality of CEII. Examples include a 200-bus synthetic grid on the footprint of Central Illinois, a 500-bus synthetic grid on the footprint of South Carolina, and a 2000-bus synthetic grid on the footprint of Texas, among others.

Datasets for competitions like the GO Competition Challenge 1, and literature-based power flow test cases such as IEEE bus systems and the Kundur Two-Area System, are also included. Moreover, for stability analysis and control, the dataset provides small signal stability test cases, such as the Three Machines Infinite Bus Benchmark System, the Brazilian Seven Bus System, and the New England 68-Bus Test System.

All of these cases include feasible AC power flow solutions, and some have additional parameters or models for analyses such as transient stability, geomagnetic disturbance analysis, energy economic study, and more. They have been developed to improve the situational awareness of current system operating conditions and to support various studies and research in power systems.

4.3 | Power generation

Electricity generation can be divided into two categories: centralised and distributed generation. Centralised generation refers to the generation of electricity through large-scale production plants and the distribution of that electricity to consumers, whereas distributed generation refers to the generation of electricity on a much smaller scale, typically by individuals using renewable energy sources.

In both centralised and distributed generation, the data collected is identical. It includes load demand, historical power measurements, capacity, generating unit, cost, performance, ramp-rate limit, operating zones, and carbon dioxide emissions data. These data are used in the management of both the power generation side and the microgrid side. Power generators are modelled on both transmission and distribution grids (in the case of DER).

5 | DISCUSSION

After evaluating a comprehensive number of the most popular public datasets and the work done on them, several aspects of the discussion were identified. This section covers these aspects, as well as research gaps and future research directions. For example, for data availability and synthetic data, we identified that the generation of synthetic data oriented to privacy preserving data could solve the problem of data availability, allowing realistic data analysis on otherwise private data [199]. In terms of privacy preservation, we have identified two main categories of techniques, which are either 'consumer-oriented' [200] or 'utility-oriented' [201]. Analysing the impact of 'consumer-oriented' privacy techniques on the utility of the datasets is an interesting future work. Moreover, to the best of our knowledge, there is no work that aims at identifying consumers that practice such privacy preserving techniques. Regarding data quality, since the number of popular public datasets is fairly low, an interesting research direction is to develop at this early stage a toolkit to unify consumer datasets in terms of format, data exploration, preprocessing techniques, and feature engineering techniques similar to the toolkit developed for NILM datasets [202].

Moreover after analysing the public datasets and the literature, we noticed two relevant and prominent issues:

1. Energy theft detection (and more broadly anomaly detection) and EV detection and load forecasting predominantly rely on synthetic or private datasets. The preference for non-public datasets is largely due to privacy concerns and the proprietary nature of the data. In energy theft detection, the data involves sensitive user information and operational details from utility companies, which are legally protected and competitively sensitive. Similarly, for EV forecasting, private companies hold detailed charging data that is commercially valuable and often kept confidential.
2. Newly emerging challenges, such as the detection of unauthorised crypto mining, suffer from a lack of public datasets. The surge in cryptocurrency mining poses challenges to smart grid management. Unauthorised mining operations and rapid technological advancements in this field hinder the collection of accurate and up-to-date datasets. One study estimates the energy consumption to be between 120 and 240 billion kilowatt-hours yearly [203]. This level of consumption suggests a significant impact on grid resources, yet the lack of detailed data impedes comprehensive analysis and grid optimisation efforts. Notably, during Texas's energy conservation periods, such as the 2022 summer heatwave, mining activities demonstrated demand flexibility [204]. This behaviour indicates a potential adaptive load management strategy, but a detailed dataset is critical for evaluating the feasibility and reliability of such an approach.

These trends highlight a gap in available resources for researchers, emphasising the need for a collaborative effort to establish data-sharing protocols that can balance privacy,

commercial value, and research needs to support advancements in smart grid technologies.

5.1 | Data availability and synthetic data

One of the major issues facing SG data analytics is the lack of public datasets available, which can be attributed to the reluctance of energy providers to publish their data. Privacy, security, and political issues all contribute to this issue [205]. Aside from the privacy concerns posed by energy disaggregation discussed in previous sections, geographical location of consumers can be compromised by solar panels generation data as in ref. [206]. The lack of data availability and a standard benchmark is more prevalent in the findings of a 2019 systematic mapping study of 358 articles in SG data analytics [207]. Their findings revealed that 70% of the articles were conducted on private datasets, 26% used publicly available datasets, 15% synthetically generated the data, and the remaining 4% used a combination of public and private datasets. Without a standardised large set of public datasets, the issue of reproducibility is expected to persist. As a result, there has been an interest in developing sophisticated techniques to synthesise SG data, and, in particular, energy consumption data, either at an aggregate level or appliance level (i.e. the case of NILM data). There is a lack of focus on synthesising other categories of data such as market data. These types of data are abundant and made public by grid operators because they are necessary information for ISO and consumers. Grid data, on the other hand, is mostly synthetic since they are considered critical information for grid operators. Synthetic data generation, especially using data-driven approaches, also gives rise to opportunities for grid operators to allow realistic data analytics without sacrificing their customer's privacy. GANs were first introduced to generate synthetic data in the work of ref. [208] in 2018 and since then several other works have utilised GANs to generate time series data [90, 209–211]. The results of these efforts suggest that GANs are a promising research direction. However, simply using GANs is not enough to conceal privacy, as they are susceptible to membership inference attacks [199].

5.2 | Privacy and security

With higher sampling rate readings, the analysis of smart metre data on energy consumption patterns can be used to determine household occupancy and other more detailed sensitive information about the household. The serious nature of the privacy issues that smart metres raise has been shown to be a barrier to the widespread implementation of smart metres in some countries [212–214].

The work in ref. [215] reviewed the existing literature on smart metres privacy and categorised the techniques into two broad techniques:

- Data manipulation: In this category, the high-resolution data is manipulated from the consumer's end before being communicated. Data aggregation, quantisation, and differential privacy techniques [37, 216–218] all fall into this category. For example, the effect of data granularity on privacy was studied in ref. [219]. However, more sophisticated privacy-aware techniques are required to ensure the aggregation of private data [220]. Secure multi-party computation coupled with homomorphic encryption [221], and secret sharing [222] are considered powerful candidates to achieve privacy aware data aggregation.
- Demand shaping and scheduling: In this category, smart-metre values are not modified or obfuscated. Instead, batteries, appliance scheduling, and renewable sources hide energy usage within the house and hinder privacy-intrusive attacks, such as NILM. In these cases, smart metres measure perturbed usage after using the battery and renewable sources. As such, locally installed batteries and renewable sources could provide total household demand and privacy is absolutely ensured. Table 10 illustrates the four main categories and exemplary articles.

Security is another critical issue in the SG. Recent published work in ref. [233] provides a comprehensive review of AMI security vulnerabilities in SG in the three layers: hardware, data and communication layers. The identified countermeasures fall into three main categories:

- Data encryption: Encryption is critical to preserving confidentiality and privacy at the data layer. The techniques here focus on encrypting the data before communicating them to the utility with minimal computational and communication overhead [234, 235].
- Authentication mechanisms: Authentication is critical to verify the sources of messages in the SG and to prevent impersonation attacks [236, 237].
- Intrusion detection systems (IDS): IDSs are a critical second line of defence for detecting security breaches in critical infrastructure. Recent works in IDS for AMI include [238–240].

For data encryption and authentication mechanisms, the work is typically evaluated using simulations on any energy consumption dataset to measure the computational and communication overheads. On the other hand, IDS are evaluated on popular datasets that are not specific to the SG. An unpopular solution is to develop testbeds and simulations such as in ref. [240]. Developing an IDS dataset in the context of the SG or evaluating the effectiveness of IDS trained on typical IDS datasets in the context of the SG is a necessary research direction.

On the basis of the above, we argue that more focus should be put forward on understanding the impact of demand shaping and load scheduling approaches to preserve privacy on the electrical utility. From a management perspective. These techniques might, for example, induce uncertainties similar to NTL leading to poor utilisation of resources and poor tariff design [241]. From a data analytics perspective, such techniques could potentially disrupt the efficacy of load forecasting

TABLE 10 Demand shaping and load scheduling categories.

Categories	Explanation	Ref.
Demand shaping: Batteries	A battery (physical or virtual) used for energy consumption can be charged and discharged to obfuscate the fine grain consumption data of the house, thus preserving privacy.	[200, 223–225]
Demand shaping: Renewable energy	These techniques obfuscate energy consumption with batteries; however, renewable energy generation must also be modelled.	[223, 226–228]
Demand shaping: Heating and cooling	Since cooling and heating have high consumption, scheduling them in a specific way would be able to obfuscate the consumption of smaller appliances and provide more privacy	[229–231]
Load scheduling	Scheduling appliances to make non-intrusive load monitoring more difficult	[232]

or energy theft detection models. Another research direction is to consider techniques that identify consumers that practice such privacy-preserving practices, to limit their possible problematic impact on energy management and data analytics.

5.3 | Data quality

In the SG context, missing values, outliers, and noisy data (i.e. logical errors or inconsistent data) are the three most common data quality issues [242]. Several solutions to each of these problems were suggested by existing work.

Regarding missing values, most datasets do not report missing data, forcing data analysts to manually detect and manage them. For time-series forecasting applications, data replacement (also called imputation) is typically required to preserve the integrity and pattern of the data. In general, the approaches to replace missing data are categorised as interpolation-based and prediction-model-based algorithms. The former being used for a few missing data points, while the latter for longer periods. However, since accurate time series data are necessary to train forecasting models, researchers mostly opt to omit a certain portion or timeframe in the data (e.g. the whole day or similar omission criteria). Although this is a common and straightforward way to deal with missing data, it omits a portion of the available data, which may lead to bias in common statistical analysis (e.g. linear regression) [243]. The work in ref. [244] outlines an industry-recognised recommended practice for imputing faulty or missing smart metre data. Periods less than 2 h are often imputed by using linear interpolation to the adjacent data. For times longer than 2 hours, the standard technique is to develop daily load profiles based on previously verified historical data of 'like weekdays' and 'like days'. Holidays or other exceptional cases are often addressed individually. It is important to note that dealing with missing values is not always necessary. For example, in ref. [244] when creating a representative load pattern of a cluster of consumers, the average of the available data points in a given point stamp is taken.

In outliers detection (anomalous data detection), most work utilise the two-standard-deviations rule as a preprocessing step for their respective application. According to ref. [245], there are two types of outliers that should be taken into account when dealing with time series data: isolated anomalies or events where the error is local to a certain set of data points and innovative anomalies where the errors are propagated throughout the time series in the system.

Real power systems also suffer significantly from noise [246], especially after the introduction of powerline communication technologies (PLCs) that support higher data rate transmission (also called high data rate narrowband 3–500 kHz PLC systems). These new technologies are desirable because they can be built on the existing power systems, however they are designed for one-way communication and not the two-way communication necessary for SG applications [247]. The noise present in these systems affect very high-frequency electrical measurement devices such as PMU devices. The noise of voltage and current measurements of the phases (e.g. in NILM datasets) at 60 or 120 Hz is negligible and can be ignored.

Data quality issues can extend to several other dimensions, namely contextual, representational, and accessibility. Contextual quality are several characteristics of the data that must be present in certain applications but not others. Such qualities are record time (the time it took for the data to be available after it happened in actual time), sampling rate, and quantity of the data. The representational qualities simply refer to how well a dataset follows the format and structure of similar datasets, as well as interpretability of notations. This issue was found to be a significant hurdle for data analytics [248]. The last dimension is accessibility, in particular availability, which is one of the most prevalent issues in the SG context, as some datasets are more readily available to researchers than others. For example, some datasets require extra procedures such as login credentials and/or licencing.

In light of these issues, we argue that more effort should be put to develop toolkits to standardise the datasets as a future research direction. In terms of formatting, for example, the Ausgrid dataset [61] for electricity consumption combines three consumption categories in the same Excel data sheet, while the LCL contains only one. A toolkit with a unified API would make the repeatability of studies much more feasible. Another issue that could be addressed by the toolkits is data preprocessing, since most work on energy consumption utilises similar preprocessing techniques. Feature engineering is another possible extension of such toolkits. For example, a toolkit can facilitate the extraction of time-related features (e.g. peak hours) or apply simple clustering techniques to help with data exploration; clustering daily consumption profiles helps identify common, uncommon, and anomalous consumption habits [249].

5.4 | Big data in the smart grid

The key steps to handle and use big data are data acquisition, storage, analysis, and operational integration. The work in ref. [250] reviewed data management for SGs and its technical requirements, the tools, and the necessary steps to integrate big data solutions in the SG context. The authors highlighted three main issues: standards and interoperability; lack of infrastructure to be able to fully utilise the big data; and privacy, integrity, authentication, and security. Furthermore, the authors of ref. [251] discussed several challenges in the area of big data analytics, including data indexing and time synchronisation. The two broad categories of applications in big data are smart metre big data and PMU big data. Smart metre big data applications are related to energy management such as load forecasting, profiling, DR, baseline estimation. The CBT dataset is a common public dataset used for this area of research due to its large volume (167 million data rows) [252]. PMU big data are used for state estimation, transmission grid visualisation, and SG reliability and stability. Simulations are commonly used to generate PMU data [253].

5.5 | Detailed in-home consumption datasets

Currently available detailed in-home consumption datasets, or as commonly referred to as buildings datasets or NILM datasets, fall into two categories: laboratory measurements and data from the actual environment. Available laboratory measurements include data from individual devices, although these data are of very little use for overall benchmark tests because real-world datasets contain measurements where multiple devices are active concurrently. However, assigning reference data in real-world scenarios presents difficulties:

1) The synchronisation of references and measured data; that is, a label should correspond to a pattern shift in the data that corresponds to the labelled pattern. A further requirement is that all data streams must be in sync with one another.
2) The absence or excess of events, and the number of "on" and "off" cycles for each device.
3) The probability distribution of the devices, as well as the lengthy measurement cycles containing a correspondingly large volume of data that contain a small number of events.

While NILM datasets face trade-offs between covering a large number of houses or focusing on a more extensive set of appliances and measurements, an equally important aspect that is often overlooked in the literature is the preprocessing of data. Ensuring data quality and dealing with missing values is a crucial step in the development of effective energy disaggregation models, as it can significantly improve their performance. Unfortunately, the lack of transparency regarding preprocessing procedures in many studies makes it difficult to replicate results and assess the true impact of any potential bias or domain-specific knowledge that may have been introduced during this stage. By addressing both the trade-offs inherent in dataset design and the need for clear documentation of preprocessing techniques, researchers can work towards developing more robust and generalisable energy disaggregation models.

5.6 | Challenges in preprocessing and evaluation

In this subsection, we discuss the challenges and limitations faced in current approaches to data pre-processing, post-processing, model evaluation, and generalisability in the context of electrical grid data analysis.

Most literature does not mention data preprocessing steps such as data cleaning and dealing with missing values despite being a crucial step known to boost performance. These steps are presumably taken, but not mentioned. Not explicitly stating the preprocessing procedure harms replicability of the work as there are several preprocessing procedures that can be followed. The authors could have introduced bias and/or domain knowledge in the data, which may have enhanced the performance of their models.

We have also observed a lack of post-processing techniques, which we believe is a potential future work to explore due to its promise to enhance performance (especially reducing false positives [254]) and mitigate common typical biases especially in energy disaggregation. For example, the authors in ref. [255] discovered that disaggregation techniques typically overestimate or underestimate disaggregated loads and proposed a technique that ensures that the disaggregated loads sum up to approximately the true aggregate consumption. Similarly, the authors of ref. [256] discovered bias when dealing with appliances that operate on multi-states (e.g. dishwashers and washing machines). Models typically produce several sporadic activations for such appliances.

Another preprocessing issue observed in the literature is the arbitrary exclusion of some data and without justification, which threatens the validity of the models. For example, some houses in the REDD datasets include very few events. These houses were mostly excluded due to the effect they have on training. The issue is not specific to the REDD dataset, as each model has its own setbacks that can be revealed if tested on more houses. To this end, we recommend using techniques such as leave-one-house-out cross-validation for a more complete evaluation in future work. Different authors also select the appliances and a number of appliances that they will train and test on without justification.

There is no clear justification and/or consensus for the selection of the training and testing split. Some train their model for 5 days and test only on one, while others follow a different evaluation strategy. This makes it difficult to compare and evaluate models, not to mention that models will be more likely to overfit the test data and perform better but have lower generalisability. Some models also train and test on the same house, while others train on a house and test on another, which means the former has lower generalisability.

The authors also define steady states and transient states differently. For example, a steady-state power signal must not fluctuate more than a certain threshold and must last for a period of time. In probabilistic models, such assumptions extend to the appliances (average, maximum, minimum, and duration of power consumed). While necessary, this poses a trade-off as follows: a more strict (i.e. high threshold) definition will eliminate noise; however, this may lead to not being able to detect small appliances consumption (they will still be considered in the steady state). To better illustrate this point, imagine a kettle that consumes 10 W, if the steady-state threshold was, for example, 20 W then the kettle's consumption will be considered noise and will not be detected. A more lenient definition (or a lower threshold) will allow for small appliances to be detected, however, poorer performance becomes inevitable. In probabilistic models, a more "diverse" assumption on the appliances (e.g., picking appliances that have a high difference in their average consumption) will allow for better distinction between the appliances and better performance overall. However, this will require handpicking of appliances and is thus not practical.

We believe that more attention must be paid to developing models with genralisability and transferability in mind [257]. This can be evaluated by training and testing on different datasets or on different houses. Comparing the same model with different datasets poses several challenges. First is the different percentage of missing data in the datasets; some loads that are not sub-metered and their consumption data become missing. The second is the scarcity of fully labelled NILM datasets. The last challenge is the different characteristics of the datasets, such as the type and sampling rate of the measurements, and the different formats.

The learned models are affected by the sampling rate of their associated dataset. Data preprocessing techniques that can capture most of the features at lower sampling rates while still maintaining high performance are a promising future research direction.

Another notable issue is associated with the use of metrics that favour classifying high-power consumption devices. Such metrics do not capture information about how well the model performs in low-power devices. It is argued, however, that such information is valuable since low-power appliances are typically what the user has the greatest control over.

6 | FUTURE WORK

This section provides future research directions highlighting key areas that require further exploration and development in the field of SG data analytics.

- **Synthetic Data Generation**: Future research can focus on developing privacy-preserving synthetic data generation techniques for SG, particularly for market and grid data. Another potential avenue is investigating advanced synthetic data generation methods such as the use of GANs while addressing privacy concerns like membership inference attacks.
- **Advancing Privacy Preservation and Security**: Future research can focus on exploring the impact of privacy preservation techniques, particularly demand shaping and load scheduling, on SG data utility and energy management. Another avenue is to develop methods to identify consumers using privacy-preserving techniques because these consumers may affect utilities data analytics.
- **Improving Data Quality and Standardisation**: Researchers in the future may address SG data quality issues by creating comprehensive toolkits for data standardisation and preprocessing, including feature engineering and clustering. A key focus may be on unifying dataset formats and structures for better data exploration and better analytics accuracy.
- **Big Data Management and Analytics in SG**: Investigate the integration of big data solutions in SG, addressing challenges in data management, standards, interoperability, and infrastructure development. Emphasise improving data acquisition, storage, analysis, and operational integration.
- **Detailed In-Home Consumption Datasets and Preprocessing Techniques**: Future research may aim at improving in-home consumption datasets by refining preprocessing methods for handling data synchronisation, event detection, and large data volumes. Standardising preprocessing steps is crucial for enhancing study replicability and minimising biases. This effort includes better strategies for data cleaning and handling missing values. There's also a need for more inclusive datasets covering diverse appliances and conditions to foster robust, generalisable energy disaggregation models.

By addressing these areas, future research can significantly contribute to the advancement of SG data analytics, ensuring more efficient, secure, and reliable data management systems.

7 | CONCLUSION

Power grids generate huge volumes of data and specifically in the SG context, where various types of data originate from several sources and typically at higher sampling rates. In addition to enabling safe operation of the grid itself, such data enable a wide variety of applications. Despite their high utility, the availability of public real-world smart grid datasets is very limited. In this work we reviewed over 50 public datasets in the smart grid context, categorising them into three main categories; Consumers' data, NILM data, and Grid data. Each category can enable for a distinct set of applications. After considering the characteristics of the individual datasets, 14 of their most popular applications were discussed, as well as numerous other less popular applications. Several findings are discussed and highlighted throughout this contribution. In the end, we present a discussion of some prevalent issues that motivate potential future research and development directions.

Ultimately, this review provides a comprehensive survey of public datasets in smart and power grid research, with the aim of improving reproducibility and serving as a key reference for researchers developing applications in this domain.

AUTHOR CONTRIBUTIONS

The initial manuscript draft was prepared by the first author, who was also responsible for the primary data collection and analysis. All co-authors provided intellectual oversight and contributed significantly to the research design and methodology. They played an active role in refining the manuscript through rigorous critique, constructive feedback, and substantive editing. Additionally, they provided expert guidance in the framing and structuring of the manuscript, ensuring its academic rigour and integrity.

ACKNOWLEDGEMENTS

This publication is supported in part by grant NPRP12C-33905-SP-66 from the Qatar National Research Fund. The findings achieved herein are solely the responsibility of the authors.

Open Access funding provided by the Qatar National Library.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no financial or personal relationships with other people or organisations that could inappropriately influence or bias their work.

DATA AVAILABILITY STATEMENT

No new data were generated or analysed in this study. All datasets reviewed are publicly available and are cited within the article.

PERMISSION TO REPRODUCE MATERIALS

None.

ORCID

Emran Altamimi https://orcid.org/0000-0002-2902-663X

REFERENCES

1. Mattioli, R., Moulinos, K.: Communication Network Interdependencies in Smart Grids. EUA FNAI Security, Ed., EU: ENISA, (2015)
2. Geisler, K.: The Relationship Between Smart Grids and Smart Cities. IEEE Smart Grid Newsletter (2013)
3. Marinakis, V.: Big data for energy management and energy-efficient buildings. Energies 13(7), 1555 (2020). https://doi.org/10.3390/en13071555
4. Siebert, L.C., et al.: An agent-based approach for the planning of distribution grids as a socio-technical system. Energies 13(18), 4837 (2020). https://doi.org/10.3390/en13184837
5. Zhou, K., Fu, C., Yang, S.: Big data driven smart energy management: from big data to big insights. Renew. Sustain. Energy Rev. 56, 215–225 (2016). https://doi.org/10.1016/j.rser.2015.11.050
6. Stewart, E., Liao, A., Roberts, C.: Open μpmu: A Real World Reference Distribution Micro-phasor Measurement Unit Data Set for Research and Application Development (2016)
7. Chintakindi, R., Mitra, A.: Wams challenges and limitations in load modeling, voltage stability improvement, and controlled island protection—a review. Energy Rep. 8, 699–709 (2022). https://doi.org/10.1016/j.egyr.2021.11.217
8. Florencias-Oliveros, O., et al.: Real-life power quality sags. IEEE Dataport (2017)
9. Wang, Y., et al.: Power system disaster-mitigating dispatch platform based on big data. In: 2014 International Conference on Power System Technology, pp. 1014–1019. IEEE (2014)
10. Wang, X., et al.: Automatic analysis of pole mounted auto-recloser data for fault prognosis to mitigate customer supply interruptions. In: 2014 49th International Universities Power Engineering Conference (UPEC), pp. 1–6. IEEE (2014)
11. Chren, S., et al.: Reliability data for smart grids: where the real data can be found. In: 2018 Smart City symposium prague (scsp), pp. 1–6. IEEE (2018)
12. Elahe, M.F., Jin, M., Zeng, P.: Review of load data analytics using deep learning in smart grids: open load datasets, methodologies, and application challenges. Int. J. Energy Res. 45(10), 14274–14305 (2021). https://doi.org/10.1002/er.6745
13. Wang, Y., et al.: Review of smart meter data analytics: applications, methodologies, and challenges. IEEE Trans. Smart Grid 10(3), 3125–3148 (2018). https://doi.org/10.1109/tsg.2018.2818167
14. Iqbal, H.K., et al.: A critical review of state-of-the-art non-intrusive load monitoring datasets. Elec. Power Syst. Res. 192, 106921 (2021). https://doi.org/10.1016/j.epsr.2020.106921
15. Pereira, L., Nunes, N.: Performance evaluation in non-intrusive load monitoring: datasets, metrics, and tools—a review. Wiley Interdiscip. Rev. 8(6), e1265 (2018). https://doi.org/10.1002/widm.1265
16. Kazmi, H., et al.: Towards data-driven energy communities: a review of open-source datasets, models and tools. Renew. Sustain. Energy Rev. 148, 111290 (2021). https://doi.org/10.1016/j.rser.2021.111290
17. Meinecke, S., Thurner, L., Braun, M.: Review of steady-state electric power distribution system datasets. Energies 13(18), 4826 (2020). https://doi.org/10.3390/en13184826
18. Postigo Marcos, F.E., et al.: A review of power distribution test feeders in the United States and the need for synthetic representative networks. Energies 10(11), 1896 (2017). https://doi.org/10.3390/en10111896
19. Kabalci, E., Kabalci, Y.: From Smart Grid to Internet of Energy. Academic Press (2019)
20. Sun, Q., et al.: A comprehensive review of smart energy meters in intelligent energy networks. IEEE Internet Things J. 3(4), 464–479 (2015). https://doi.org/10.1109/jiot.2015.2512325
21. Humeau, S., et al.: Electricity load forecasting for residential customers: exploiting aggregation and correlation between households. In: 2013 Sustainable Internet and ICT for Sustainability (SustainIT), pp. 1–6. IEEE (2013)
22. Massaferro, P., Di Martino, J.M., Fernández, A.: Ntl detection: overview of classic and dnn-based approaches on a labeled dataset of 311k customers. In: 2021 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), pp. 1–5. IEEE (2021)
23. Saleh, S., Pijnenburg, P., Castillo-Guerra, E.: Load aggregation from generation-follows-load to load-follows-generation: residential loads. IEEE Trans. Ind. Appl. 53(2), 833–842 (2016). https://doi.org/10.1109/tia.2016.2626261
24. Edwards, R.E., New, J., Parker, L.E.: Predicting future hourly residential electrical consumption: a machine learning case study. Energy Build. 49, 591–603 (2012). https://doi.org/10.1016/j.enbuild.2012.03.010
25. Kavousian, A., Rajagopal, R., Fischer, M.: Determinants of residential electricity consumption: using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. Energy 55, 184–194 (2013). https://doi.org/10.1016/j.energy.2013.03.086
26. Bai, Y., Zhong, H., Xia, Q.: 2016 IEEE Power and Energy Society General Meeting (PESGM). IEEE. pp. 1–5 (2016)
27. Mahmoudi-Kohan, N., et al.: A three-stage strategy for optimal price offering by a retailer based on clustering techniques. Int. J. Electr. Power Energy Syst. 32(10), 1135–1142 (2010). https://doi.org/10.1016/j.ijepes.2010.06.011

28. Huang, H., et al.: False data separation for data security in smart grids. Knowl. Inf. Syst. 52(3), 815–834 (2017). https://doi.org/10.1007/s10115-016-1019-8
29. Mateos, G., Giannakis, G.B.: Load curve data cleansing and imputation via sparsity and low rank. IEEE Trans. Smart Grid 4(4), 2347–2355 (2013). https://doi.org/10.1109/tsg.2013.2259853
30. Chuwa, M.G., Wang, F.: A review of non-technical loss attack models and detection methods in the smart grid. Elec. Power Syst. Res. 199, 107415 (2021). https://doi.org/10.1016/j.epsr.2021.107415
31. Zhao, Q., Chang, Z., Min, G.: Anomaly detection and classification of household electricity data: a time window and multilayer hierarchical network approach. IEEE Internet Things J. 9(5), 3704–3716 (2021). https://doi.org/10.1109/jiot.2021.3098735
32. Yang, S.L., Shen, C.: A review of electric load classification in smart grid environment. Renew. Sustain. Energy Rev. 24, 103–110 (2013). https://doi.org/10.1016/j.rser.2013.03.023
33. Chicco, G.: Overview and performance assessment of the clustering methods for electrical load pattern grouping. Energy 42(1), 68–80 (2012). https://doi.org/10.1016/j.energy.2011.12.031
34. Wang, Y., et al.: Load profiling and its application to demand response: a review. Tsinghua Sci. Technol. 20(2), 117–129 (2015). https://doi.org/10.1109/tst.2015.7085625
35. Unterweger, A., Engel, D., Ringwelski, M.: The effect of data granularity on load data compression. In: DA-CH Conference on Energy Informatics, pp. 69–80. Springer (2015)
36. Rottondi, C., Verticale, G., Krauss, C.: Distributed privacy-preserving aggregation of metering data in smart grids. IEEE J. Sel. Area. Commun. 31(7), 1342–1354 (2013). https://doi.org/10.1109/jsac.2013.130716
37. Khwaja, A.S., et al.: Smart meter data obfuscation using correlated noise. IEEE Internet Things J. 7(8), 7250–7264 (2020). https://doi.org/10.1109/jiot.2020.2983213
38. Shateri, M., et al.: Privacy-cost management in smart meters with mutual-information-based reinforcement learning. IEEE Internet Things J. 9(22), 22389–22398 (2022). https://doi.org/10.1109/jiot.2021.3128488
39. Wang, X., et al.: Detection and isolation of false data injection attacks in smart grids via nonlinear interval observer. IEEE Internet Things J. 6(4), 6498–6512 (2019). https://doi.org/10.1109/jiot.2019.2916670
40. Peppanen, J., et al.: Handling bad or missing smart meter data through advanced data imputation. In: 2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), pp. 1–5. IEEE (2016)
41. Barnett, V., Lewis, T.: Evolution by gene duplication. In: Outliers in Statistical Data, 3rd ed Wiley, Hoboken, NJ, USA (1994)
42. Buzzi-Ferraris, G., Manenti, F.: Outlier detection in large data sets. Comput. Chem. Eng. 35(2), 388–390 (2011). https://doi.org/10.1016/j.compchemeng.2010.11.004
43. Markou, M., Singh, S.: Novelty detection: a review—part 1: statistical approaches. Signal Process. 83(12), 2481–2497 (2003). https://doi.org/10.1016/j.sigpro.2003.07.018
44. Enernoc open data. [Online]. https://open-enernoc-data.s3.amazonaws.com/anon/index.html
45. Jaiswal, R., et al.: Anomaly detection in smart meter data for preventing potential smart grid imbalance. In: 2021 4th Artificial Intelligence and Cloud Computing Conference, pp. 150–159 (2021)
46. Zhang, W., et al.: Solargan: multivariate solar data imputation using generative adversarial network. IEEE Trans. Sustain. Energy 12(1), 743–746 (2020). https://doi.org/10.1109/tste.2020.3004751
47. Zheng, R., et al.: Load forecasting under data corruption based on anomaly detection and combined robust regression. Int. Trans. Electr. Energy Syst. 30(7), e12103 (2020). https://doi.org/10.1002/2050-7038.12103
48. Hong, T., et al.: Probabilistic Energy Forecasting: Global Energy Forecasting Competition 2014 and beyond, pp. 896–913 (2016)
49. State grid corporation of China: electricity theft dataset. [Online]. https://github.com/henryRDlab/ElectricityTheftDetection
50. Yao, D., et al.: Energy theft detection with energy privacy preservation in the smart grid. IEEE Internet Things J. 6(5), 7659–7669 (2019). https://doi.org/10.1109/jiot.2019.2903312
51. Yang, H., et al.: Pipc: privacy- and integrity-preserving clustering analysis for load profiling in smart grids. IEEE Internet Things J. 9(13), 10851–10861 (2022). https://doi.org/10.1109/jiot.2021.3125674
52. Kuster, C., Rezgui, Y., Mourshed, M.: Electrical load forecasting models: a critical systematic review. Sustain. Cities Soc. 35, 257–270 (2017). https://doi.org/10.1016/j.scs.2017.08.009
53. Ahmad, S., Naeem, M., Ahmad, A.: Unified optimization model for energy management in sustainable smart power systems. Int. Trans. Electr. Energy Syst. 30(4), e12144 (2020). https://doi.org/10.1002/2050-7038.12144
54. Sun, M., et al.: Clustering-based residential baseline estimation: a probabilistic perspective. IEEE Trans. Smart Grid 10(6), 6014–6028 (2019). https://doi.org/10.1109/tsg.2019.2895333
55. Yusta, J., et al.: Optimal electricity price calculation model for retailers in a deregulated market. Int. J. Electr. Power Energy Syst. 27(5-6), 437–447 (2005). https://doi.org/10.1016/j.ijepes.2005.03.002
56. Ahmad, S., et al.: A compendium of performance metrics, pricing schemes, optimization objectives, and solution methodologies of demand side management for the smart grid. Energies 11(10), 2801 (2018). https://doi.org/10.3390/en11102801
57. Zhang, R., et al.: A wind energy supplier bidding strategy using combined ega-inspired hpsoifa optimizer and deep learning predictor. Energies 14(11), 3059 (2021). https://doi.org/10.3390/en14113059
58. Low carbon london (lcl) dataset. [Online]. https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london%20-households
59. Dataport. (2021). [Online]. https://www.pecanstreet.org/dataport/
60. Weibel, T.: Smart* Data Set for Sustainability. [Online]. https://traces.cs.umass.edu/index.php/smart/smart
61. Ausgrid datasets. [Online]. https://www.ausgrid.com.au/Industry/Our-Research/Data-to-share
62. C. for Energy Regulation (CER): Cer smart metering project - electricity customer behaviour trial, 2009-2010. Irish Soc. Sci. Data Arch. (2012). sN: 0012-00. [Online]. https://www.ucd.ie/issda/data/commissionforenergyregulationcer/
63. Australian energy market operator (aemo) dataset. [Online]. https://aemo.com.au/en/energy-systems/electricity/national-electricity-%20market-nem/data-nem
64. Electric reliability council of Texas (ercot) dataset. [Online]. http://www.ercot.com/
65. Miller, C., Meggers, F.: The building data genome project: an open, public data set from non-residential building electrical meters. Energy Proc. 122, 439–444 (2017). https://doi.org/10.1016/j.egypro.2017.07.400
66. Energy Market Authority of singapore. (2022). [Online]. https://www.ema.gov.sg/Statistics.aspx
67. Hong, T., Pinson, P., Fan, S.: Global Energy Forecasting Competition 2012, pp. 357–363 (2014)
68. Energy, Load, and Demand Reports. ISO New England, (2012)-present. [Online]. https://www.iso-ne.com/isoexpress/web/reports/load-and-demand/-/tree/dmnd-rt-hourly-sys
69. Trindade, A.: Electricityloaddiagrams20112014. [Online]. https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014
70. HenryRDlab: The State Grid Corporation of china (Sgcc) Dataset. https://github.com/henryRDlab/ElectricityTheftDetection
71. Laurinec, P., et al.: Density-based unsupervised ensemble learning methods for time series forecasting of aggregated or clustered electricity consumption. J. Intell. Inf. Syst. 53(2), 219–239 (2019). https://doi.org/10.1007/s10844-019-00550-3
72. Laurinec, P., Lucká, M.: Interpretable multiple data streams clustering with clipped streams representation for the improvement of electricity consumption forecasting. Data Min. Knowl. Discov. 33(2), 413–445 (2019). https://doi.org/10.1007/s10618-018-0598-2
73. Venkatesh, J., et al.: Scalable-application design for the iot. IEEE Softw. 34(1), 62–70 (2017). https://doi.org/10.1109/ms.2017.4

74. Singh, S., Majumdar, A.: Deep sparse coding for non–intrusive load monitoring. IEEE Trans. Smart Grid 9(5), 4669–4678 (2017). https://doi.org/10.1109/tsg.2017.2666220
75. Mocanu, E., et al.: On-line building energy optimization using deep reinforcement learning. IEEE Trans. Smart Grid 10(4), 3698–3708 (2018). https://doi.org/10.1109/tsg.2018.2834219
76. Henri, G., Lu, N.: A supervised machine learning approach to control energy storage devices. IEEE Trans. Smart Grid 10(6), 5910–5919 (2019). https://doi.org/10.1109/tsg.2019.2892586
77. Xuan, Z., et al.: Pv-load decoupling based demand response baseline load estimation approach for residential customer with distributed pv system. IEEE Trans. Ind. Appl. 56(6), 6128–6137 (2020). https://doi.org/10.1109/tia.2020.3014575
78. Molina-Markham, A., et al.: Private memoirs of a smart meter. In: Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building, pp. 61–66 (2010)
79. Jin, M., Jia, R., Spanos, C.J.: Virtual occupancy sensing: using smart meters to indicate your presence. IEEE Trans. Mobile Comput. 16(11), 3264–3277 (2017). https://doi.org/10.1109/tmc.2017.2684806
80. Kekatos, V., et al.: Voltage regulation algorithms for multiphase power distribution grids. IEEE Trans. Power Syst. 31(5), 3913–3923 (2015). https://doi.org/10.1109/tpwrs.2015.2493520
81. Eichinger, F., et al.: A time-series compression technique and its application to the smart grid. VLDB J. 24(2), 193–218 (2015). https://doi.org/10.1007/s00778-014-0368-8
82. Li, C., et al.: Data-driven planning of electric vehicle charging infrastructure: a case study of Sydney, Australia. IEEE Trans. Smart Grid 12(4), 3289–3304 (2021). https://doi.org/10.1109/tsg.2021.3054763
83. Martin, D., et al.: Investigation into modeling Australian power transformer failure and retirement statistics. IEEE Trans. Power Deliv. 33(4), 2011–2019 (2018). https://doi.org/10.1109/tpwrd.2018.2814588
84. Krishnamurthy, D., Li, W., Tesfatsion, L.: An 8-zone test system based on iso new england data: development and application. IEEE Trans. Power Syst. 31(1), 234–246 (2015). https://doi.org/10.1109/tpwrs.2015.2399171
85. Wang, W., Yu, N.: A machine learning framework for algorithmic trading with virtual bids in electricity markets. In: 2019 IEEE Power & Energy Society General Meeting (PESGM), pp. 1–5. IEEE (2019)
86. Csereklyei, Z., Qu, S., Ancev, T.: The effect of wind and solar power generation on wholesale electricity prices in Australia. Energy Pol. 131, 358–369 (2019). https://doi.org/10.1016/j.enpol.2019.04.007
87. Sharma, V., Haque, M.H., Aziz, S.M.: Energy cost minimization for net zero energy homes through optimal sizing of battery storage system. Renew. Energy 141, 278–286 (2019). https://doi.org/10.1016/j.renene.2019.03.144
88. Sharma, V., Haque, M.H., Aziz, S.M.: Pv generation and load profile data of net zero energy homes in south Australia. Data Brief 25, 104235 (2019). https://doi.org/10.1016/j.dib.2019.104235
89. Luo, S., Weng, Y.: A two-stage supervised learning approach for electricity price forecasting by leveraging different data sources. Appl. Energy 242, 1497–1512 (2019). https://doi.org/10.1016/j.apenergy.2019.03.129
90. Fekri, M.N., Ghosh, A.M., Grolinger, K.: Generating energy data for machine learning with recurrent generative adversarial networks. Energies 13(1), 130 (2019). https://doi.org/10.3390/en13010130
91. Miller, C., Meggers, F.: Mining electrical meter data to predict principal building use, performance class, and operations strategy for hundreds of non-residential buildings. Energy Build. 156, 360–373 (2017). https://doi.org/10.1016/j.enbuild.2017.09.056
92. Koh, L., et al.: Impact of energy storage and variability of pv on power system reliability. Energy Proc. 33, 302–310 (2013). https://doi.org/10.1016/j.egypro.2013.05.071
93. Shu, Z., Jirutitijaroen, P.: Non-sequential simulation methods for reliability analysis of power systems with photovoltaic generation. In: 2010 IEEE 11th International Conference on Probabilistic Methods Applied to Power Systems, pp. 703–709. IEEE (2010)
94. Loi, T.S.A., Le Ng, J.: Analysing households' responsiveness towards socio-economic determinants of residential electricity consumption in Singapore. Energy Pol. 112, 415–426 (2018). https://doi.org/10.1016/j.enpol.2017.09.052
95. Srinivasan, D., et al.: Game-theory based dynamic pricing strategies for demand side management in smart grids. Energy 126, 132–143 (2017). https://doi.org/10.1016/j.energy.2016.11.142
96. Yoshida, Y., Figueroa, H.P., Dougal, R.A.: Use of time series load data to size energy storage systems. In: 2018 IEEE Green Energy and Smart Systems Conference (IGESSC), pp. 1–6. IEEE (2018)
97. Günter, N., Marinopoulos, A.: Energy storage for grid services and applications: classification, market review, metrics, and methodology for evaluation of deployment cases. J. Energy Storage 8, 226–234 (2016). https://doi.org/10.1016/j.est.2016.08.011
98. Tajeuna, E.G., Bouguessa, M., Wang, S.: A network-based approach to enhance electricity load forecasting. In: 2018 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 266–275. IEEE (2018)
99. Himeur, Y., et al.: Building power consumption datasets: survey, taxonomy and future directions. Energy Build. 227, 110404 (2020). https://doi.org/10.1016/j.enbuild.2020.110404
100. Gao, J., et al.: A feasibility study of automated plug-load identification from high-frequency measurements. In: 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 220–224. IEEE (2015)
101. He, J., et al.: An efficient and accurate nonintrusive load monitoring scheme for power consumption. IEEE Internet Things J. 6(5), 9054–9063 (2019). https://doi.org/10.1109/jiot.2019.2926815
102. Hart, G.: Nonintrusive appliance load monitoring. Proc. IEEE 80(12), 1870–1891 (1992). https://doi.org/10.1109/5.192069
103. Zoha, A., et al.: Non-intrusive load monitoring approaches for disaggregated energy sensing: a survey. Sensors 12(12), 16838–16866 (2012). https://doi.org/10.3390/s121216838
104. Ahmad, S., et al.: Joint energy management and energy trading in residential microgrid system. IEEE Access 8, 123334–123346 (2020). https://doi.org/10.1109/access.2020.3007154
105. Yaqub, R., et al.: Smart energy-consumption management system considering consumers' spending goals (sems-ccsg). Int. Trans. Electr. Energy Syst. 26(7), 1570–1584 (2016). https://doi.org/10.1002/etep.2167
106. Ahmad, H., Ahmad, A., Ahmad, S.: Efficient energy management in a microgrid. In: 2018 International Conference on Power Generation Systems and Renewable Energy Technologies (PGSRET), pp. 1–5. IEEE (2018)
107. Ahmad, S., Naeem, M., Ahmad, A.: Low complexity approach for energy management in residential buildings. Int. Trans. Electr. Energy Syst. 29(1), e2680 (2019). https://doi.org/10.1002/etep.2680
108. Ruano, A., et al.: Nilm techniques for intelligent home energy management and ambient assisted living: a review. Energies 12(11), 2203 (2019). https://doi.org/10.3390/en12112203
109. Ganu, T., et al.: Socketwatch: an autonomous appliance monitoring system. In: 2014 IEEE International Conference on Pervasive Computing and Communications (PerCom), pp. 38–43. IEEE (2014)
110. Pereira, W., Ferscha, A., Weigl, K.: Unsupervised detection of unusual behaviors from smart home energy data. In: International Conference on Artificial Intelligence and Soft Computing, pp. 523–534. Springer (2016)
111. Rashid, H., et al.: Can non-intrusive load monitoring be used for identifying an appliance's anomalous behaviour? Appl. Energy 238, 796–805 (2019). https://doi.org/10.1016/j.apenergy.2019.01.061
112. Rashid, H., et al.: Evaluation of non-intrusive load monitoring algorithms for appliance-level anomaly detection. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8325–8329. IEEE (2019)
113. Ahajjam, M.A., et al.: Mored: a moroccan buildings' electricity consumption dataset. Energies 13(24), 6737 (2020). https://doi.org/10.3390/en13246737
114. Kolter, J.Z., Johnson, M.J.: Redd: a public data set for energy disaggregation research. In: Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA, vol. 25, pp. 59–62 (2011). Citeseer

115. Kelly, J., Knottenbelt, W.: The UK-dale dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. Sci. Data 2(1), 1–14 (2015). https://doi.org/10.1038/sdata.2015.7
116. Anderson, K., et al.: Blued: a fully labeled public dataset for event-based non-intrusive load monitoring research. In: Proceedings of the 2nd KDD Workshop on Data Mining Applications in Sustainability (SustKDD), vol. 7, pp. 1–5. ACM New York (2012)
117. Makonin, S., et al.: Ampds: a public dataset for load disaggregation and eco-feedback research. In: 2013 IEEE Electrical Power & Energy Conference, pp. 1–6. IEEE (2013)
118. Beckel, C., et al.: The eco data set and the performance of non-intrusive load monitoring algorithms. In: Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings, pp. 80–89 (2014)
119. Reinhardt, A., et al.: On the accuracy of appliance identification based on distributed load metering data. In: 2012 Sustainable Internet and ICT for Sustainability (SustainIT), pp. 1–9. IEEE (2012)
120. Zimmermann, J.-P., et al.: Household Electricity Survey: A Study of Domestic Electrical Product Usage, pp. 213–214. Intertek Testing & Certification Ltd (2012)
121. Batra, N., et al.: It's different: insights into home energy consumption in India. In: Proceedings of the 5th ACM Workshop on Embedded Systems for Energy-Efficient Buildings, pp. 1–8 (2013)
122. Monacchi, A., et al.: Greend: an energy consumption dataset of households in Italy and Austria. In: 2014 IEEE International Conference on Smart Grid Communications (SmartGridComm), pp. 511–516. IEEE (2014)
123. Gao, J., et al.: Plaid: a public dataset of high-resoultion electrical appliance measurements for load identification research: demo abstract. In: Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings, pp. 198–199 (2014)
124. Murray, D., et al.: A Data Management Platform for Personalised Real-Time Energy Feedback (2015)
125. Batra, N., et al.: A comparison of non-intrusive load monitoring methods for commercial and residential buildings. arXiv preprint arXiv:1408.6595 (2014)
126. Uttama Nambi, A.S., Reyes Lua, A., Prasad, V.R.: Loced: location-aware energy disaggregation framework. In: Proceedings of the 2nd Acm International Conference on Embedded Systems for Energy-Efficient Built Environments, pp. 45–54 (2015)
127. Kahl, M., et al.: Whited-a worldwide household and industry transient energy data set. In: 3rd International Workshop on Non-intrusive Load Monitoring, pp. 1–4 (2016)
128. Gulati, M., Ram, S.S., Singh, A.: An in depth study into using emi signatures for appliance identification. In: Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings, pp. 70–79 (2014)
129. Ribeiro, M., et al.: Sustdataed: a public dataset for electric energy disaggregation research. In: Proceedings of the ICT for Sustainability, pp. 14–16 (2016)
130. Gisler, C., et al.: Appliance consumption signature database and recognition test protocols. In: 2013 8th International Workshop on Systems, Signal Processing and their Applications (WoSSPA), pp. 336–341. IEEE (2013)
131. Picon, T., et al.: Cooll: controlled on/off loads library, a public dataset of high-sampled electrical signals for appliance identification. arXiv preprint arXiv:1611.05803 (2016)
132. Holcomb, C.: Pecan street inc.: a test-bed for nilm. In: International Workshop on Non-intrusive Load Monitoring, Pittsburgh, PA, USA (2012)
133. Kriechbaumer, T., Jacobsen, H.-A.: Blond, a building-level office environment dataset of typical electrical appliances. Sci. Data 5(1), 1–14 (2018). https://doi.org/10.1038/sdata.2018.48
134. Makonin, S., Wang, Z.J., Tumpach, C.: Rae: the rainforest automation energy dataset for smart grid meter data analysis. Data 3(1), 8 (2018). https://doi.org/10.3390/data3010008
135. Maasoumy, M., et al.: Berds-berkeley energy disaggregation data set. In: Proceedings of the Workshop on Big Learning at the Conference on Neural Information Processing Systems (NIPS), pp. 1–6 (2013)
136. Johnson, G., Beausoleil-Morrison, I.: Electrical-end-use data from 23 houses sampled each minute for simulating micro-generation systems. Appl. Therm. Eng. 114, 1449–1456 (2017). https://doi.org/10.1016/j.applthermaleng.2016.07.133
137. Rashid, H., Singh, P., Singh, A.: I-blend, a campus-scale commercial and residential buildings electrical energy dataset. Sci. Data 6(1), 1–12 (2019). https://doi.org/10.1038/sdata.2019.15
138. Rui, H., Arnold, M., Wellssow, W.H.: Synthetic medium voltage grids for the assessment of smart grid techniques. In: 2012 3rd IEEE PES Innovative Smart Grid Technologies Europe (ISGT Europe), pp. 1–8. IEEE (2012)
139. Seack, A., Kays, J., Rehtanz, C.: Generating low voltage grids on the basis of public available map data. CIRED Workshop: Rome, Italy. 11, (2014)
140. Han, X., et al.: Real-time measurements and their effects on state estimation of distribution power system. In: IEEE PES ISGT Europe 2013, pp. 1–5. IEEE (2013)
141. Bracale, A., et al.: Active Management of Distribution Networks with the Atlantide Models (2012)
142. Anderson, P.M., Fouad, A.A.: Power System Control and Stability. John Wiley & Sons (2008)
143. Tripathy, S., et al.: Load-flow solutions for ill-conditioned power systems by a Newton-like method. IEEE Trans. Power Apparatus Syst. 10, 3648–3657 (1982). https://doi.org/10.1109/tpas.1982.317050
144. Iwamoto, S., Tamura, Y.: A load flow calculation method for ill-conditioned power systems. IEEE Trans. Power Apparatus Syst. 4, 1736–1743 (1981). https://doi.org/10.1109/tpas.1981.316511
145. Zollenkopf, K.: Load-flow calculation using loss-minimisation techniques. Proc. Inst. Electr. Eng. 115(1), 121–127 (1968). IET. https://doi.org/10.1049/piee.1968.0019
146. Stott, B., Alsac, O.: Fast decoupled load flow. IEEE Trans. Power Apparatus Syst. 3, 859–869 (1974). https://doi.org/10.1109/tpas.1974.293985
147. Barsali, S., et al.: Benchmark Systems for Network Integration of Renewable and Distributed Energy Resources (2014)
148. Dickert, J., Domagk, M., Schegner, P.: Benchmark low voltage distribution networks based on cluster analysis of actual grid properties. In: 2013 IEEE Grenoble Conference, pp. 1–6. IEEE (2013)
149. Scheidler, A., et al.: Der Integration Study for the German State of Hesse–Methodology and Key Results (2019)
150. Garske, S., Schloemer, G., Hofmann, L.: Evaluation of reactive power management strategies and grid loss characteristics based on generic distribution grid models. In: NEIS 2017; Conference on Sustainable Energy Supply and Energy Storage Systems, pp. 1–6. VDE (2017)
151. Adamczyk, A., et al.: Generic 12-bus test system for wind power integration studies. In: 2013 15th European Conference on Power Electronics and Applications (EPE), pp. 1–6. IEEE (2013)
152. Kraiczy, M., Stetz, T., Braun, M.: Parallel operation of transformers with on load tap changer and photovoltaic systems with reactive power control. IEEE Trans. Smart Grid 9(6), 6419–6428 (2017). https://doi.org/10.1109/tsg.2017.2712633
153. Peyghami, S., et al.: Standard test systems for modern power system analysis: an overview. IEEE Ind. Electron. Mag. 13(4), 86–105 (2019). https://doi.org/10.1109/mie.2019.2942376
154. Xu, G., et al.: Toward integrating distributed energy resources and storage devices in smart grid. IEEE Internet Things J. 4(1), 192–204 (2016)
155. Hemmati, R., Hooshmand, R.-A., Khodabakhshian, A.: Comprehensive review of generation and transmission expansion planning. IET Gener. Transm.Distrib. 7(9), 955–964 (2013). https://doi.org/10.1049/iet-gtd.2013.0031
156. Leou, R.-C.: A multi-year transmission planning under a deregulated market. Int. J. Electr. Power Energy Syst. 33(3), 708–714 (2011). https://doi.org/10.1016/j.ijepes.2010.11.020
157. Akbari, T., et al.: Towards integrated planning: simultaneous transmission and substation expansion planning. Elec. Power Syst. Res. 86, 131–139 (2012). https://doi.org/10.1016/j.epsr.2011.12.012
158. Gu, Y., McCalley, J.D., Ni, M.: Coordinating large-scale wind integration and transmission planning. IEEE Trans. Sustain. Energy 3(4), 652–659 (2012). https://doi.org/10.1109/tste.2012.2204069

159. Rahmani, M., et al.: Efficient method for ac transmission network expansion planning. Elec. Power Syst. Res. 80(9), 1056–1064 (2010). https://doi.org/10.1016/j.epsr.2010.01.012
160. Zhao, J.H., et al.: Flexible transmission network planning considering distributed generation impacts. IEEE Trans. Power Syst. 26(3), 1434–1443 (2010). https://doi.org/10.1109/tpwrs.2010.2089994
161. Cedeño, E.B., Arora, S.: Performance comparison of transmission network expansion planning under deterministic and uncertain conditions. Int. J. Electr. Power Energy Syst. 33(7), 1288–1295 (2011). https://doi.org/10.1016/j.ijepes.2011.05.005
162. Feng, Y., Ryan, S.M.: Scenario construction and reduction applied to stochastic power generation expansion planning. Comput. Oper. Res. 40(1), 9–23 (2013). https://doi.org/10.1016/j.cor.2012.05.005
163. Pereira, A.J., Saraiva, J.T.: Generation expansion planning (gep)–a long-term approach using system dynamics and genetic algorithms (gas). Energy 36(8), 5180–5199 (2011). https://doi.org/10.1016/j.energy.2011.06.021
164. Huang, Y.-F., et al.: State estimation in electric power grids: meeting new challenges presented by the requirements of the future grid. IEEE Signal Process. Mag. 29(5), 33–43 (2012). https://doi.org/10.1109/msp.2012.2187037
165. Ahmad, F., et al.: Distribution system state estimation-a step towards smart grid. Renew. Sustain. Energy Rev. 81, 2659–2671 (2018). https://doi.org/10.1016/j.rser.2017.06.071
166. Ghaddar, B., Marecek, J., Mevissen, M.: Optimal power flow as a polynomial optimization problem. IEEE Trans. Power Syst. 31(1), 539–546 (2015). https://doi.org/10.1109/tpwrs.2015.2390037
167. Zhang, B., Wang, M., Su, W.: Reliability analysis of power systems integrated with high-penetration of power converters. IEEE Trans. Power Syst. 36(3), 1998–2009 (2020). https://doi.org/10.1109/tpwrs.2020.3032579
168. Krishnan, V., et al.: Validation of synthetic us electric power distribution system data sets. IEEE Trans. Smart Grid 11(5), 4477–4489 (2020). https://doi.org/10.1109/tsg.2020.2981077
169. Mateo, C., et al.: Building large-scale us synthetic electric distribution system models. IEEE Trans. Smart Grid 11(6), 5301–5313 (2020). https://doi.org/10.1109/tsg.2020.3001495
170. Ieee pes cascading failure working group. [Online]. https://site.ieee.org/pes-cascading/publications/
171. Kersting, W.H.: Radial distribution test feeders. IEEE Trans. Power Syst. 6(3), 975–985 (1991). https://doi.org/10.1109/59.119237
172. Kersting, W.: A comprehensive distribution test feeder. In: IEEE PES T&D 2010, pp. 1–4. IEEE (2010)
173. De Oliveira-De Jesus, P.M., Quintana, A.R.: New formulation for distribution system state estimation. In: 2012 VI Andean Region International Conference, pp. 3–6. IEEE (2012)
174. Mojumdar, R.R., Arboleya, P., González-Morán, C.: Step-voltage regulator model test system. In: 2015 IEEE Power & Energy Society General Meeting, pp. 1–5. IEEE (2015)
175. Eajal, A.A., El-Hawary, M.: Optimal capacitor placement and sizing in unbalanced distribution systems with harmonics consideration using particle swarm optimization. IEEE Trans. Power Deliv. 25(3), 1734–1741 (2010). https://doi.org/10.1109/tpwrd.2009.2035425
176. Gottwalt, S., et al.: Modeling and valuation of residential demand flexibility for renewable energy integration. IEEE Trans. Smart Grid 8(6), 2565–2574 (2016). https://doi.org/10.1109/tsg.2016.2529424
177. Matic-Cuka, B., Kezunovic, M.: Islanding detection for inverter-based distributed generation using support vector machine method. IEEE Trans. Smart Grid 5(6), 2676–2686 (2014). https://doi.org/10.1109/tsg.2014.2338736
178. Hedayati, H., Nabaviniaki, S.A., Akbarimajd, A.: A method for placement of dg units in distribution networks. IEEE Trans. Power Deliv. 23(3), 1620–1628 (2008). https://doi.org/10.1109/tpwrd.2007.916106
179. Nick, M., Cherkaoui, R., Paolone, M.: Optimal allocation of dispersed energy storage systems in active distribution networks for energy balance and grid support. IEEE Trans. Power Syst. 29(5), 2300–2310 (2014). https://doi.org/10.1109/tpwrs.2014.2302020
180. Dall'Anese, E., Zhu, H., Giannakis, G.B.: Distributed optimal power flow for smart microgrids. IEEE Trans. Smart Grid 4(3), 1464–1475 (2013). https://doi.org/10.1109/tsg.2013.2248175
181. Zin, A.A.M., et al.: Two circular-updating hybrid heuristic methods for minimum-loss reconfiguration of electrical distribution network. IEEE Trans. Power Syst. 28(2), 1318–1323 (2012)
182. Rasheduzzaman, M., Mueller, J.A., Kimball, J.W.: Reduced-order small-signal model of microgrid systems. IEEE Trans. Sustain. Energy 6(4), 1292–1305 (2015). https://doi.org/10.1109/tste.2015.2433177
183. Meng, F., et al.: Dual passive harmonic reduction at dc link of the double-star uncontrolled rectifier. IEEE Trans. Ind. Electron. 66(4), 3303–3309 (2018). https://doi.org/10.1109/tie.2018.2844840
184. Hobbs, B.F., Rijkers, F.A.: Strategic generation with conjectured transmission price responses in a mixed transmission pricing system-part i: formulation. IEEE Trans. Power Syst. 19(2), 707–717 (2004). https://doi.org/10.1109/tpwrs.2003.821628
185. Zhao, S., Meng, X., Song, X.: Increasing maximum penetration of distributed generation by voltage regulation in smart distribution grid. In: 2015 5th International Conference on Electric Utility Deregulation and Restructuring and Power Technologies (DRPT), pp. 1894–1898. IEEE (2015)
186. Dugan, R.C., et al.: Roadmap for the ieee pes test feeders. In: 2009 IEEE/PES Power Systems Conference and Exposition, pp. 1–4 (2009)
187. Dugan, R., Arritt, R.: The Ieee 8500-node Test Feeder. Electric Power Research Institute, Palo Alto (2010)
188. Schneider, K.P., Fuller, J.C., Chassin, D.P.: Multi-state load models for distribution system analysis. IEEE Trans. Power Syst. 26(4), 2425–2433 (2011). https://doi.org/10.1109/tpwrs.2011.2132154
189. Arritt, R.F., Dugan, R.C.: Distribution system analysis and the future smart grid. IEEE Trans. Ind. Appl. 47(6), 2343–2350 (2011). https://doi.org/10.1109/tia.2011.2168932
190. Penido, D., et al.: Solving the single-circuit nev test case using the current injection full-Newton power flow. In: 2008 IEEE Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century, pp. 1–7. IEEE (2008)
191. Variz, A., et al.: Harmonic analysis of the power distribution neutral-to-earth voltage (nev) test case using four-wire three-phase harmonic current injection method. In: 2009 IEEE Power & Energy Society General Meeting, pp. 1–7. IEEE (2009)
192. Horton, R., et al.: Effect of line modeling methods on neutral-to-earth voltage analysis of multi-grounded distribution feeders. In: 2011 IEEE/PES Power Systems Conference and Exposition, pp. 1–6. IEEE (2011)
193. Yuan, Z., Hesamzadeh, M.R.: Hierarchical coordination of tso-dso economic dispatch considering large-scale integration of distributed energy resources. Appl. Energy 195, 600–615 (2017). https://doi.org/10.1016/j.apenergy.2017.03.042
194. Yang, T., et al.: Optimal planning of communication system of cps for distribution network. J. Sens. 2017, 1–10 (2017). https://doi.org/10.1155/2017/9303989
195. [Online]. https://cmte.ieee.org/pes-testfeeders/resources/
196. Ni, F., et al.: Uncertainty analysis of aggregated smart meter data for state estimation. In: 2016 IEEE International Workshop on Applied Measurements for Power Systems (AMPS), pp. 1–6. IEEE (2016)
197. Khalid Mehmood, K., et al.: Optimal sizing and allocation of battery energy storage systems with wind and solar power dgs in a distribution network for voltage regulation considering the lifespan of batteries. IET Renew. Power Gener. 11(10), 1305–1315 (2017). https://doi.org/10.1049/iet-rpg.2016.0938
198. Electric grid test cases. (2023), [Online] https://electricgrids.engr.tamu.edu/electric-grid-test-cases/ Accessed 19 July 2023
199. Del Grosso, G., Pichler, G., Piantanida, P.: Privacy-preserving synthetic smart meters data. In: 2021 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), pp. 1–5. IEEE (2021)
200. Farokhi, F., Sandberg, H.: Fisher information as a measure of privacy: preserving privacy of households with smart meters using batteries. IEEE Trans. Smart Grid 9(5), 4726–4734 (2017). https://doi.org/10.1109/tsg.2017.2667702

201. Buescher, N., et al.: Two is not enough: privacy assessment of aggregation schemes in smart metering. Proc. Priv. Enhancing Technol. 2017(4), 198–214 (2017). https://doi.org/10.1515/popets-2017-0045
202. Batra, N., et al.: Nilmtk: an open source toolkit for non-intrusive load monitoring. In: Proceedings of the 5th International Conference on Future Energy Systems, pp. 265–276 (2014)
203. Menati, A., Lee, K., Xie, L.: Modeling and analysis of utilizing cryptocurrency mining for demand flexibility in electric energy systems: a synthetic Texas grid case study. IEEE Trans. Energy Mark. Pol. Regul. 1(1), 1–10 (2023). https://doi.org/10.1109/tempr.2022.3230953
204. Menati, A., et al.: High resolution modeling and analysis of cryptocurrency mining's impact on power grids: carbon footprint, reliability, and electricity price. Adv. Appl. Energy 10, 100136 (2023). [Online]. https://doi.org/10.1016/j.adapen.2023.100136
205. Hu, J., Vasilakos, A.V.: Energy big data analytics and security: challenges and opportunities. IEEE Trans. Smart Grid 7(5), 2423–2436 (2016). https://doi.org/10.1109/tsg.2016.2563461
206. Chen, D., et al.: Sunspot: exposing the location of anonymous solar-powered homes. In: Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments, pp. 85–94 (2016)
207. Rossi, B., Chren, S.: Smart grids data analysis: a systematic mapping study. IEEE Trans. Ind. Inf. 16(6), 3619–3639 (2019). https://doi.org/10.1109/tii.2019.2954098
208. Zhang, C., et al.: Generative adversarial network for synthetic time series data generation in smart grids. In: 2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), pp. 1–6. IEEE (2018)
209. Li, J., et al.: Energy data generation with wasserstein deep convolutional generative adversarial networks. Energy 257, 124694 (2022). https://doi.org/10.1016/j.energy.2022.124694
210. El Kababji, S., Srikantha, P.: A data-driven approach for generating synthetic load patterns and usage habits. IEEE Trans. Smart Grid 11(6), 4984–4995 (2020). https://doi.org/10.1109/tsg.2020.3007984
211. Chen, Y., et al.: Model-free renewable scenario generation using generative adversarial networks. IEEE Trans. Power Syst. 33(3), 3265–3275 (2018). https://doi.org/10.1109/tpwrs.2018.2794541
212. Erdemir, E., Gündüz, D., Dragotti, P.L.: Smart meter privacy. In: Privacy in Dynamical Systems, pp. 19–41. Springer (2020)
213. Naperville Smart Meter V. City of Naperville. p. 521, (2018)
214. Cuijpers, C., Koops, B.-J.: Smart metering and privacy in europe: lessons from the Dutch case. In: European Data Protection: Coming of Age, pp. 269–293. Springer (2013)
215. Farokhi, F.: Review of results on smart-meter privacy by data manipulation, demand shaping, and load scheduling. IET Smart Grid 3(5), 605–613 (2020). https://doi.org/10.1049/iet-stg.2020.0129
216. Ács, G., Castelluccia, C.: I have a dream!(differentially private smart metering). In: International Workshop on Information Hiding, pp. 118–132. Springer (2011)
217. Sandberg, H., Dán, G., Thobaben, R.: Differentially private state estimation in distribution networks with smart meters. In: 2015 54th IEEE Conference on Decision and Control (CDC), pp. 4492–4498. IEEE (2015)
218. Kserawi, F., Al-Marri, S., Malluhi, Q.: Privacy-preserving fog aggregation of smart grid data using dynamic differentially-private data perturbation. IEEE Access 10, 43159–43174 (2022). https://doi.org/10.1109/access.2022.3167015
219. Eibl, G., Engel, D.: Influence of data granularity on smart meter privacy. IEEE Trans. Smart Grid 6(2), 930–939 (2014). https://doi.org/10.1109/tsg.2014.2376613
220. Merad-Boudia, O.R., Senouci, S.M.: An efficient and secure multidimensional data aggregation for fog-computing-based smart grid. IEEE Internet Things J. 8(8), 6143–6153 (2020). https://doi.org/10.1109/jiot.2020.3040982
221. Keoh, S.L., Ang, Y.H., Tang, Z.: A lightweight privacy-preserved spatial and temporal aggregation of energy data. In: 2015 11th International Conference on Information Assurance and Security (IAS), pp. 1–6. IEEE (2015)
222. Danezis, G., et al.: Smart meter aggregation via secret-sharing. In: Proceedings of the First ACM Workshop on Smart Energy Grid Security, pp. 75–80 (2013)
223. Giaconi, G., Gündüz, D., Poor, H.V.: Smart meter privacy with renewable energy and an energy storage device. IEEE Trans. Inf. Forensics Secur. 13(1), 129–142 (2017). https://doi.org/10.1109/tifs.2017.2744601
224. Li, S., Khisti, A., Mahajan, A.: Information-theoretic privacy for smart metering systems with a rechargeable battery. IEEE Trans. Inf. Theor. 64(5), 3679–3695 (2018). https://doi.org/10.1109/tit.2018.2809005
225. Kserawi, F., Malluhi, Q.M.: Privacy preservation of aggregated data using virtual battery in the smart grid. In: 2020 IEEE 6th International Conference on Dependability in Sensor, Cloud and Big Data Systems and Application (DependSys), pp. 106–111 (2020)
226. Giaconi, G., Gündüz, D.: Smart meter privacy with renewable energy and a finite capacity battery. In: 2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pp. 1–5. IEEE (2016)
227. Erdemir, E., Dragotti, P.L., Gündüz, D.: Privacy-cost trade-off in a smart meter system with a renewable energy source and a rechargeable battery. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2687–2691. IEEE (2019)
228. Isikman, A.O., et al.: Power scheduling in privacy enhanced microgrid networks with renewables and storage. In: 2016 13th IEEE Annual Consumer Communications & Networking Conference (CCNC), pp. 405–410. IEEE (2016)
229. Farokhi, F., Sandberg, H.: Fisher information privacy with application to smart meter privacy using hvac units. In: Privacy in Dynamical Systems, pp. 3–17. Springer (2020)
230. Sun, Y., Lampe, L., Wong, V.W.: Smart meter privacy: exploiting the potential of household energy storage units. IEEE Internet Things J. 5(1), 69–78 (2017). https://doi.org/10.1109/jiot.2017.2771370
231. Chen, D., et al.: Combined heat and privacy: preventing occupancy detection from smart meters. In: 2014 IEEE International Conference on Pervasive Computing and Communications (PerCom), pp. 208–215. IEEE (2014)
232. Liu, E., Cheng, P.: Achieving privacy protection using distributed load scheduling: a randomized approach. IEEE Trans. Smart Grid 8(5), 2460–2473 (2017). https://doi.org/10.1109/tsg.2017.2703400
233. Shokry, M., et al.: Systematic survey of advanced metering infrastructure security: vulnerabilities, attacks, countermeasures, and future vision. Future Generat. Comput. Syst. 136, 358–377 (2022). https://doi.org/10.1016/j.future.2022.06.013
234. Alsharif, A., et al.: Epda: efficient and privacy-preserving data collection and access control scheme for multi-recipient ami networks. IEEE Access 7, 27829–27845 (2019). https://doi.org/10.1109/access.2019.2900934
235. Ibrahem, M.I., et al.: Pmbfe: efficient and privacy-preserving monitoring and billing using functional encryption for ami networks. In: 2020 International Symposium on Networks, Computers and Communications (ISNCC), pp. 1–7. IEEE (2020)
236. Naseer, H., Bhutta, M.N.M., Alojail, M.A.: A key transport protocol for advance metering infrastructure (ami) based on public key cryptography. In: 2020 International Conference on Cyber Warfare and Security (ICCWS), pp. 1–5. IEEE (2020)
237. Lee, Y., Hwang, E., Choi, J.: A unified approach for compression and authentication of smart meter reading in ami. IEEE Access 7, 34383–34394 (2019). https://doi.org/10.1109/access.2019.2903574
238. Huang, C.: Forest management and resource monitoring based on ami intrusion detection algorithm and artificial intelligence. J. Ambient Intell. Hum. Comput., 1–15 (2021)
239. Yao, R., et al.: Intrusion detection system in the advanced metering infrastructure: a cross-layer feature-fusion cnn-lstm-based approach. Sensors 21(2), 626 (2021). https://doi.org/10.3390/s21020626
240. Sun, C.-C., et al.: Intrusion detection for cybersecurity of smart meters. IEEE Trans. Smart Grid 12(1), 612–622 (2020). https://doi.org/10.1109/tsg.2020.3010230

241. Sajid, Z., Javaid, A.: A stochastic approach to energy policy and management: a case study of the Pakistan energy crisis. Energies 11(9), 2424 (2018). https://doi.org/10.3390/en11092424
242. Chen, W., et al.: Data quality of electricity consumption data in a smart grid environment. Renew. Sustain. Energy Rev. 75, 98–105 (2017). https://doi.org/10.1016/j.rser.2016.10.054
243. Allison, P.D.: Missing Data. Sage publications (2001)
244. Uniform Business Practices for the Retail Energy Market. Edison Electric Institute, (2000)
245. Choy, K.: Outlier detection for stationary time series. J. Stat. Plann. Inference 99(2), 111–127 (2001). https://doi.org/10.1016/s0378-3758(01)00081-7
246. Korki, M., Hosseinzadeh, N., Moazzeni, T.: Performance evaluation of a narrowband power line communication for smart grid with noise reduction technique. IEEE Trans. Consum. Electron. 57(4), 1598–1606 (2011). https://doi.org/10.1109/tce.2011.6131131
247. Nassar, M., et al.: Cyclostationary noise modeling in narrowband powerline communication for smart grid applications. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3089–3092. IEEE (2012)
248. Radhakrishnan, A., Das, S.: Quality assessment of smart grid data. In: 2018 20th National Power Systems Conference (NPSC), pp. 1–6. IEEE (2018)
249. Alhussein, M., Aurangzeb, K., Haider, S.I.: Hybrid cnn-lstm model for short-term individual household load forecasting. IEEE Access 8, 180544–180557 (2020). https://doi.org/10.1109/access.2020.3028281
250. Daki, H., et al.: Big data management in smart grid: concepts, requirements and implementation. J. Big Data 4(1), 1–19 (2017). https://doi.org/10.1186/s40537-017-0070-y
251. Bhattarai, B.P., et al.: Big data analytics in smart grids: state-of-the-art, challenges, opportunities, and future directions. IET Smart Grid 2(2), 141–154 (2019). https://doi.org/10.1049/iet-stg.2018.0261
252. Wang, Y., et al.: Clustering of electricity consumption behavior dynamics toward big data applications. IEEE Trans. Smart Grid 7(5), 2437–2447 (2016). https://doi.org/10.1109/tsg.2016.2548565
253. Shyam, R., et al.: Apache spark a big data analytics platform for smart grid. Proc. Technol. 21, 171–178 (2015). https://doi.org/10.1016/j.protcy.2015.10.085
254. Khazaei, M., Stankovic, L., Stankovic, V.: Evaluation of low-complexity supervised and unsupervised nilm methods and pre-processing for detection of multistate white goods. In: Proceedings of the 5th International Workshop on Non-intrusive Load Monitoring, pp. 34–38 (2020)
255. He, K., et al.: Post-processing for event-based non-intrusive load monitoring. In: 4th International Workshop on Non-intrusive Load Monitoring, pp. 1–4 (2018)
256. Kong, W., et al.: A practical solution for non-intrusive type ii load monitoring based on deep learning and post-processing. IEEE Trans. Smart Grid 11(1), 148–160 (2019). https://doi.org/10.1109/tsg.2019.2918330
257. Pinto, G., et al.: Transfer learning for smart buildings: a critical review of algorithms, applications, and future perspectives. Adv. Appl. Energy 5, 100084 (2022). https://doi.org/10.1016/j.adapen.2022.100084

How to cite this article: Altamimi, E., et al.: Smart grid public datasets: characteristics and associated applications. IET Smart Grid. 1–28 (2024). https://doi.org/10.1049/stg2.12161