



OPEN ACCESS

EDITED BY

Zening Li,
Taiyuan University of Technology, China

REVIEWED BY

Yixun Xue,
Taiyuan University of Technology, China
Tao Niu,
Chongqing University, China

*CORRESPONDENCE

Qinglai Guo,
✉ guoqinglai@mail.tsinghua.edu.cn

RECEIVED 16 November 2023

ACCEPTED 23 November 2023

PUBLISHED 29 December 2023

RETRACTED 12 November 2025

CITATION

Lan J, Zhou Y, Guo Q and Sun H (2023), A data-driven approach for generating load profiles based on InfoGAN and MKDE. *Front. Energy Res.* 11:1339543. doi: 10.3389/fenrg.2023.1339543

COPYRIGHT

© 2023 Lan, Zhou, Guo and Sun. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

RETRACTED: A data-driven approach for generating load profiles based on InfoGAN and MKDE

Jian Lan, Yanzhen Zhou, Qinglai Guo* and Hongbin Sun

State Key Laboratory of Power System and Generation Equipment, Department of Electrical Engineering, Tsinghua University, Beijing, China

High-quality demand-side management requires an abundance of load profiles to support decision-making processes. However, customer energy consumption data often contains sensitive personal information, and service providers face significant challenges in accessing a substantial amount of energy consumption data. To generate a large volume of customer data without compromising privacy, this study introduces a data-driven approach integrating Information Maximizing Generative Adversarial Networks (InfoGAN) with Multivariate Kernel Density Estimation (MKDE) for the generation of load profiles. InfoGAN is firstly trained based on existing customer load profiles, with the Q network disentangling the load into feature variables and the generator producing realistic profiles. Subsequently, MKDE is utilized to assess the distribution of these features, enabling the generation of new profiles by sampling new feature variables. The proposed method circumvents the need for intricate sampling or modeling processes and generates realistic data that represents the inherent uncertainties and fluctuations characterizing customers' electricity consumption. The generated data could be used as the substitution for real electricity consumption data, thereby facilitating further applications without compromising privacy concerns.

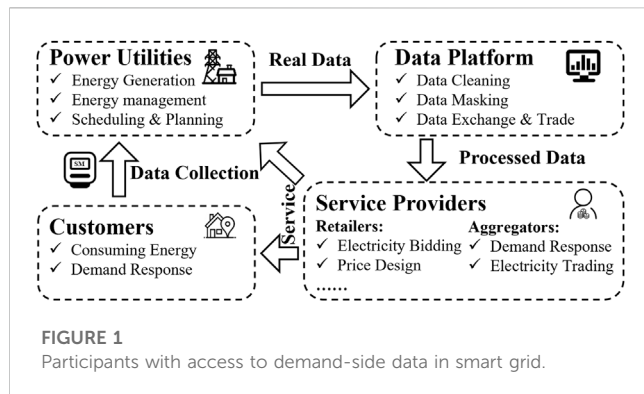
KEYWORDS

InfoGAN, MKDE, data generation, privacy, demand side management

1 Introduction

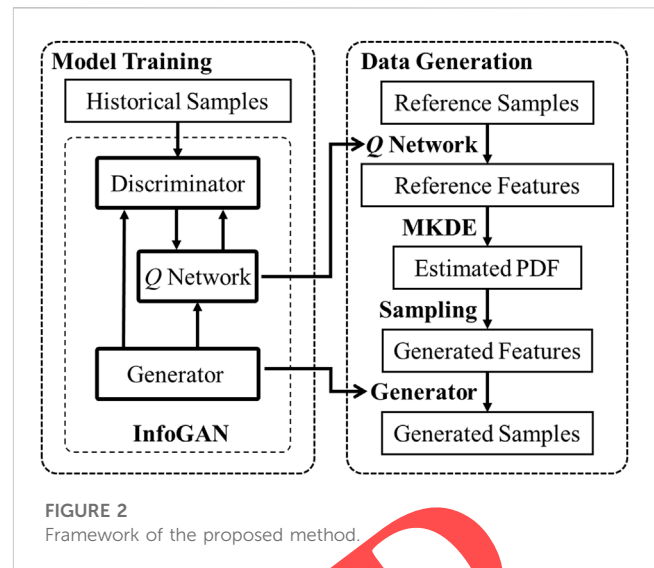
With the development of Advanced Metering Infrastructure (AMI) in smart grid, a large amount of fine-grained customers' power consumption data is collected by smart meters, leading to a better perception of the demand side for both power utilities and retailers and higher efficiency of all links in power system (Mohassel et al., 2014). However, these valuable data also carry inherent sensitive information risks, potentially revealing personal habits and lifestyle choices of customers, which poses great threats to customer privacy. Striking a balance between operational efficiency and privacy protection in the smart grid is an ongoing challenge, necessitating methods to model customer energy behavior while ensuring privacy.

The distribution of real-time customers' electricity demand can hardly be calculated because of the variation and acute fluctuating aspects between customers (Grandjean et al., 2012). Though many privacy issues are involved, load curves are still central to demand-side management rather than statistical indicators of consumption data, especially in the electricity market. For distribution network operators, smart meter data can help to realize better low-voltage network modeling and management (Haben et al., 2016). For



electricity retailers, the exact power consumption demand of their customers is vital to their marketing strategies, and high-resolution data guarantees demand forecasting ability, which may result in lower opportunity costs and higher profit (Da Silva et al., 2013). For aggregators, household load shapes have the potential to enhance the targeting and tailoring of demand response (DR) as well as improve energy reduction recommendations (Kwac et al., 2014), and when it comes to real-time demand response, potential evaluation load curves are indispensable. With the help of deep learning, smart meter data could also be used for customer characterization, where customers' sociodemographic characteristics could be inferred by their load profiles (Wang et al., 2018). Besides, Non-Intrusive Load Monitoring (NILM) has been used for classification and energy consumption estimation (Gillis et al., 2015). Besides, energy theft detection (Hu et al., 2019) and bad data detection (Li et al., 2009) are also conducted with demand-side energy consumption data. However, though better energy management could be achieved by analyzing the data collected from smart meters, there are also concerns about the abuse of these personal data (Hu and Vasiliakos, 2016), not only affecting the safe operation of the critical infrastructure but also violating customers' privacy. Energy consumption data collected from the demand side could expose customers' personal activities to anyone with access to these data and result in property damage and other undesirable outcomes.

To address privacy concerns in the provision of customer energy data to other entities in the electricity market, it is advisable to share the processed data instead of raw customer data. However, traditional data privacy protection methods, such as anonymization and adding random noise, have been found to be not always reliable based on existing studies (Armoogum and Bassoo, 2019). Moreover, the method of Differential Privacy (DP) is often used to conceal user information, but it may introduce excessive noise, particularly for high-dimensional time series data, which may compromise the utility of the data (Sangogboye et al., 2018). Utilizing models to synthesize new data is one of the approaches to address issues related to data insufficiency or data privacy concerns (Arif et al., 2017). systematically review existing load modeling techniques, but the biases are inherent in these model-based methods due to the assumptions made regarding the load operations. In recent years, data-driven generative models such as generative adversarial networks (GAN) (Goodfellow et al., 2014) have enabled the modeling of power systems without models. GAN



was first introduced to renewable scenario generation in (Chen et al., 2018), and has been used in load generation (Wang et al., 2021), reconstruction of high-temporal-resolution PV generation data (Zhang et al., 2021), etc. Besides, GAN has also been introduced to generating electroencephalographic data (Debie et al., 2020), spatial-temporal data (Qi et al., 2020), and sensitive data in IIoT operations (Hindistan and Yelken, 2023), etc. which could realize data privacy protection through data generation. Since the output-diversity characteristic of GAN could match the stochastic power consumption of demand side, it could also be used for generate new consumption data and replace real data sharing to avoid privacy leakage.

This paper introduces a novel approach using information maximizing generative adversarial networks (InfoGAN) combined with multivariate kernel density estimation (MKDE) for load profile generation. First, the InfoGAN model is utilized to learn from existing customer load profiles, where the Q network has the capability to decouple the load into feature variables, and the generator is capable of producing realistic load profiles. Subsequently, when new customer load profiles are needed, MKDE is employed to evaluate the distribution of existing feature variables, from which new feature variables can be sampled and corresponding load profiles can be generated. Notably, the proposed method is also applicable to inferring potential loads based on limited available usage data and generating load profiles for new customers. The key contribution of this paper can be summarized as follows.

- (1) A novel data-driven approach for generating load profiles is proposed. Information maximizing generative adversarial networks are first introduced to generate load profiles, which could achieve accurate modeling of customer energy demands through a data-driven approach. This allows for the rapid generation of extensive required customer load data, providing a robust data foundation for service providers.
- (2) The proposed method could extract the intrinsic features of the load profiles, which provides new insights for load modeling.

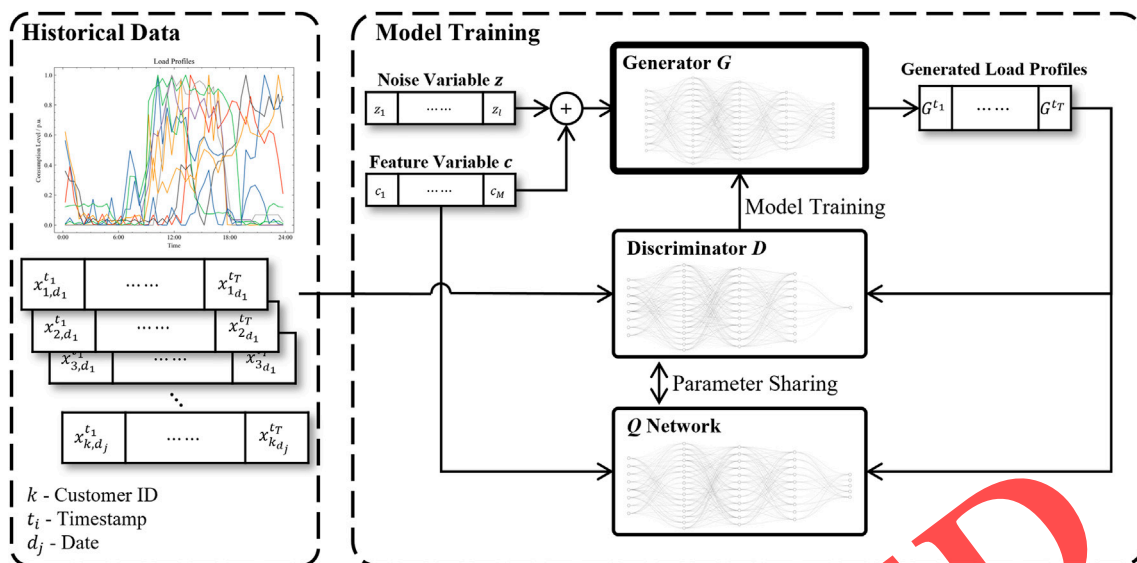


FIGURE 3
The structure of InfoGAN.

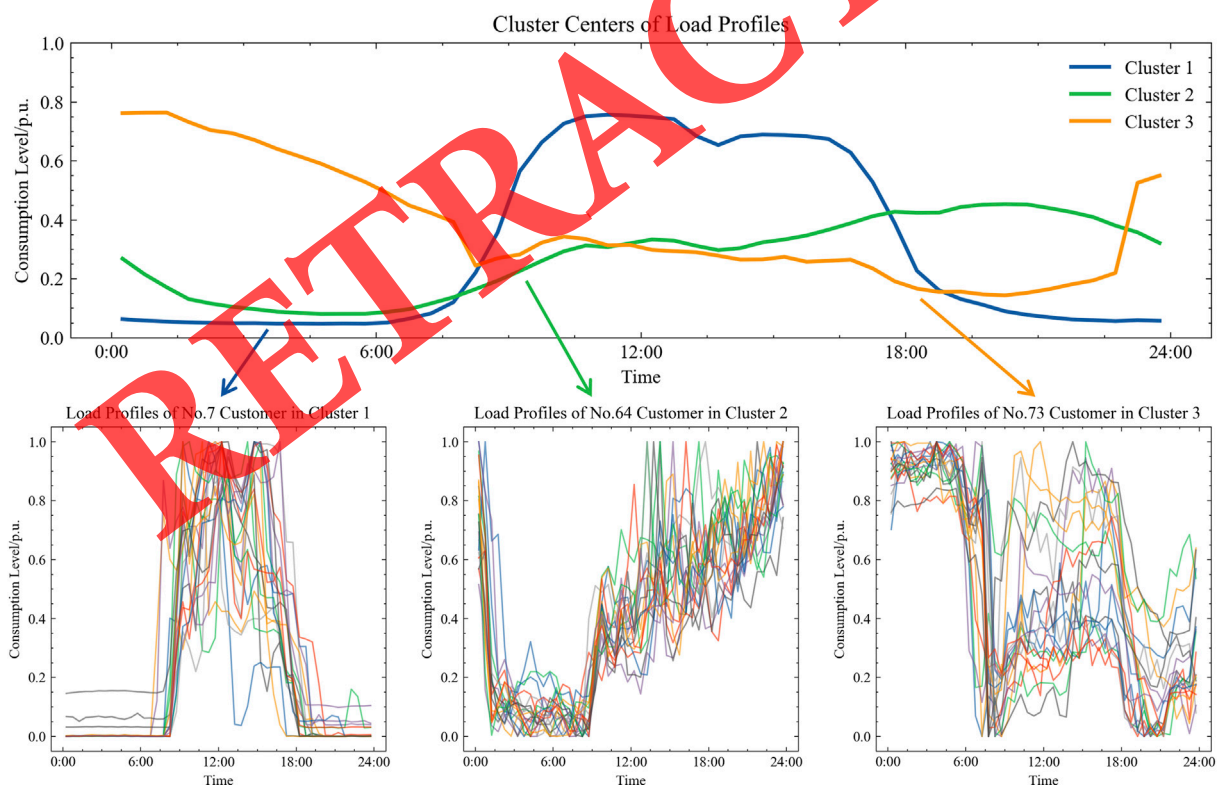
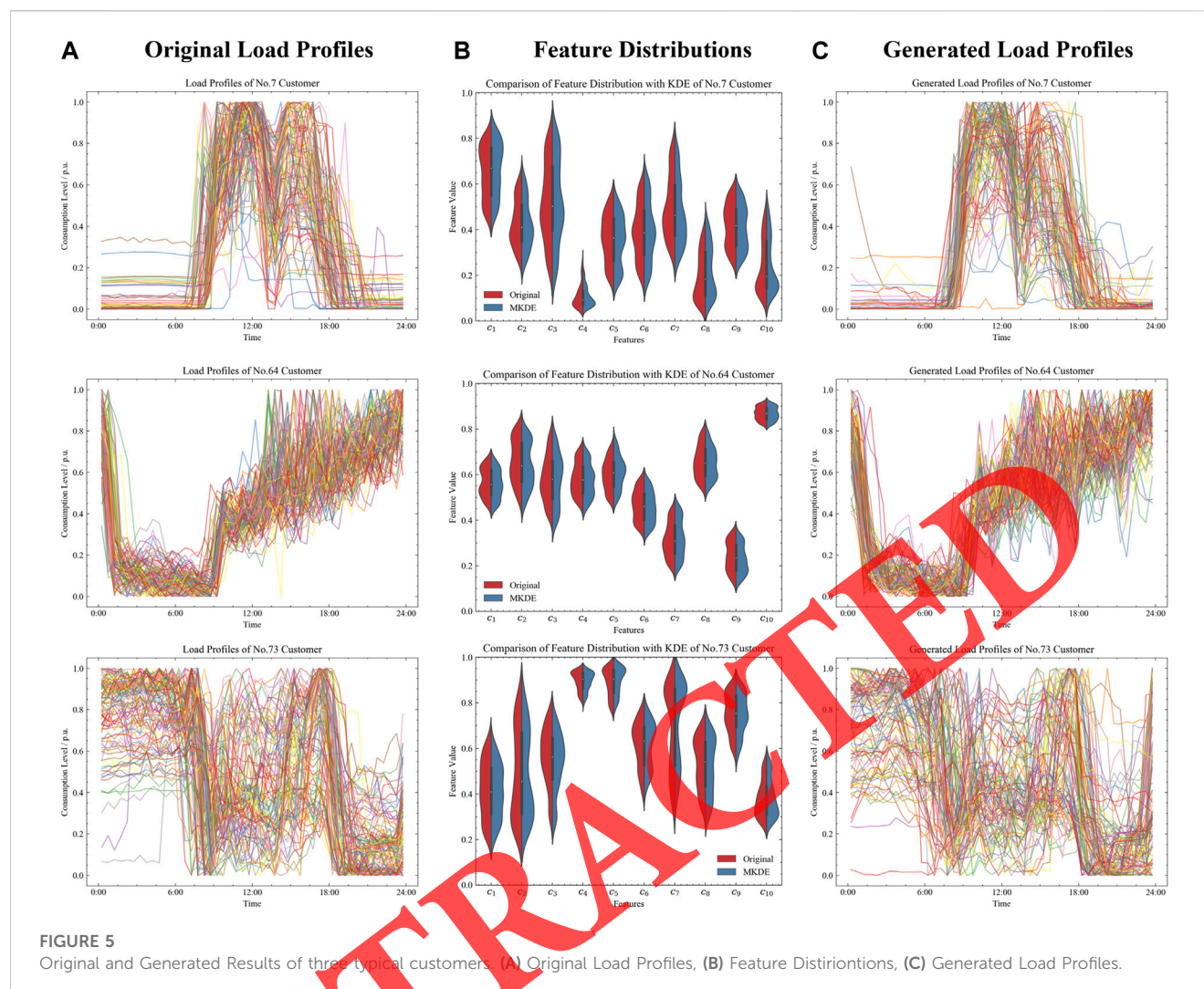


FIGURE 4
Clustering results of load profiles.

Combined with multivariate kernel density estimation, it enables the generation of any desired type of load profile, which has been validated through case studies.

(3) Based on the proposed method, potential utility curves can be efficiently and accurately generated from a limited sample of customer load data, which provides significant assistance in



researching potential customer load demands, offering valuable insights for further studies.

The paper is organized as follows: [Section 2](#) outlines our proposed method's framework; [Section 3](#) details the methodology, including algorithm introduction and model implementation; [Section 4](#) describes and evaluates the results; and [Section 5](#) concludes the paper.

2 Framework

As shown in [Figure 1](#), there are four major participants involved in data circulation in the smart grid: customers, power utilities, data platforms, and service providers. Power utilities collect data from customers through AML, as well as taking suggestions from service providers such as aggregators to achieve demand-side management. Service providers analyze data and provide suggestions to participants in the ancillary services market, which could be energy-saving advice for customers or efficient scheduling policies for power utilities, as well as getting involved in demand response as aggregators or participating in the electricity market as retailers.

As stated before, the original data collected from customers contains quality and privacy issues so that data platform is necessary to work as an information hub that gets real data from power utilities and provides service providers with cleaned and masked data. Moreover, with the popularization of the electricity market, data platforms may also be able to trade data someday since data itself is one of the most valuable assets in the electricity market.

To address these issues, we proposed a data-driven method for smart meter data generation based on InfoGAN and MKDE, which could capture the features of historical load profiles, and realistic data can be generated through generative models based on samples from feature space. The generated data maintains the characteristics of historical data as well as hiding detailed personal information, which is suitable for data circulation to service providers. The framework of the proposed method is shown in [Figure 2](#), encompassing both the model training and data generation phases.

In the model training process, historical customer energy consumption data are used as the training for the generative model. The training process is meticulously designed to enable the generator within InfoGAN to learn the distribution of historical data through adversarial training, and the Q network is capable of extracting key features without labeled data. When new

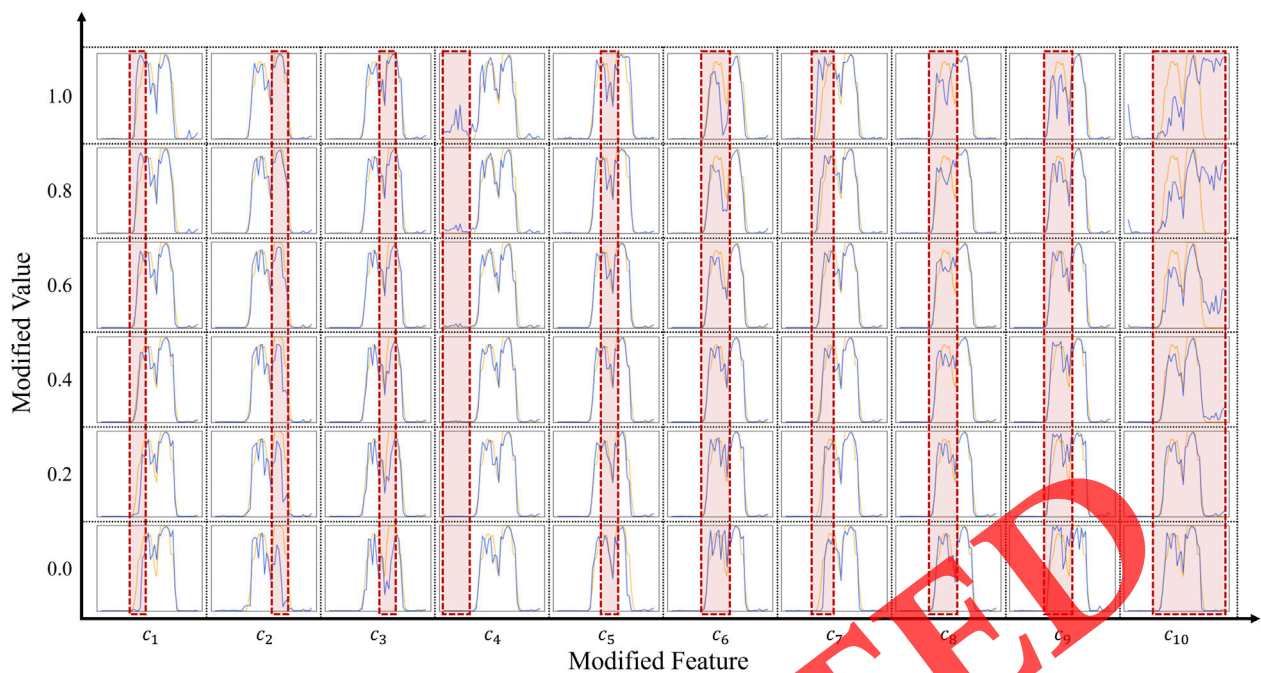


FIGURE 6

The impact of feature value on the generated results.

data is required, historical data or specified types of customer load profiles can be used as reference samples. The Q network is then utilized to obtain reference features from these samples. Subsequently, Multivariate Kernel Density Estimation (MKDE) is used to estimate the probability density functions (PDFs) of features to decipher their distribution in a multi-dimensional space non-parametrically, allowing for the generation of feature variables through sampling. Finally, these features are used as input for the generator, which could generate the required load profiles.

It should be mentioned that the feature variables for generation are sampled from the distribution of existing data, ensuring that the generated profiles do not directly correspond to specific customers, which could also preserve privacy. Moreover, this approach is applicable even with limited historical data. Sampling in the feature space reduces computational complexity and leverages the attributes of historical data, yielding more realistic load profiles.

3 Methodology

In this section, information maximizing generative adversarial networks and multivariate kernel density estimators are introduced to load profile generation, and detailed implementation are described.

3.1 Load profile generation based on InfoGAN

Generative Adversarial Networks (GANs) were first proposed as an unsupervised generative model, which has the ability to generate high-quality, realistic data. The aim of GAN is to capture the potential

distribution of input data and generate new identically distributed data samples, which also corresponds to some data issues in the power system. There are two deep neural networks known as generative model G and discriminative model D being trained simultaneously which corresponds to a minimax two-player game. The goal of GAN is to train generative model G to capture exactly the real distribution of the input data x assisted by constantly optimized discriminative model D . The competition in this game pushes both of these two models to improve their performance until Nash equilibria are achieved that the samples generated by G can't be distinguished from the original data x by D . The distribution of the input data x is defined as P_{data} and noise variable z under a known prior distribution P_z such as Gaussian distribution is used as the input of G . The target of G is to present a mapping from prior distribution to the data space denoted as $G(z)$. The output of D denoted as $D(x)$ is a single scalar representing the estimation that x comes from P_{data} rather than the generated distribution P_g . As a result, the objective function of training G is maximizing $D(G(z))$, and the object function of training D is minimizing $D(G(z))$ as well as maximizing $D(x)_{x \sim P_{data}}$, and the value function of GAN can be written as follows:

$$\min_G \max_D V(D, G) = E_{x \sim P_{data}(x)} [\log D(x)] + E_{z \sim P_z(z)} [\log (1 - D(G(z)))]$$

Despite its effectiveness, the vanilla GAN encounters challenges in learning interpretable and disentangled representations of data, which is critical for understanding and controlling the generative process. To put it into practical usage, a method to extract the features of data is needed. As an innovative iteration of GANs, InfoGAN (Chen et al., 2016) addresses this limitation by learning to disentangle representations of the data in an unsupervised manner.

The core innovation of InfoGAN lies in the introduction of an auxiliary network, the Q network, which maximizes the mutual information between the feature variables c and the observations $G(z, c)$, effectively inducing the generator to learn meaningful and interpretable representations. Here, feature variable c represents the conditional variable, encoding interpretable and meaningful attributes of generated samples, while z is the noise variables introducing randomness to ensure diversity in generation.

Figure 3 illustrates the structure of InfoGAN. Unlike conventional GANs, where the generator only receives a noise variable z as input, InfoGAN augments this by incorporating feature variables c . These variables are designed to represent distinct and interpretable attributes of the generated samples, and the Q network is trained to predict these feature variables. Besides, the Q network often shares parameters with the discriminator to enhance training efficiency and model compactness in practice, which leverages the feature-discriminating capabilities of D , facilitating more effective inference of the conditional variables. This training encourages the generator to produce outputs where variations in the features correspond to variations in specific, interpretable aspects of the generated data. $Q(c|x)$ is computed to approximate the posterior $P(c|x)$, and it is proved that mutual information $I(c, G(z, c))$ can be quantified by $E_{x \sim G(z, c)}[E_{c' \sim P(c|x)}[\log Q(c' | x)]]$. Consequently, the minimax game of InfoGAN can be described as follows, where λ is a regularization coefficient that balances the conventional GAN objective with mutual information maximization.

$$\begin{aligned} & \min_G \max_D V_{\text{InfoGAN}}(D, G, Q) \\ & = V(D, G) - \lambda E_{x \sim G(z, c)}[E_{c' \sim P(c|x)}[\log Q(c' | x)]] \end{aligned}$$

In this paper, the training of InfoGAN is described as follows.

•Hypermeter: α , learning rate, m , batch size, n_{critic} , the number of D updates per G updates.

•Require: θ_{g0} , initial G parameters, θ_{d0} , initial D parameters, θ_{q0} , initial Q parameters.

```

1: while  $\theta$  has not converged, do
2:   for  $t = 0, \dots, n_{\text{critic}}$  do
3:     Sample  $x^{(i)}_{i=1}^m \sim p_{\text{data}}$  # a batch from the training data
4:     Sample  $z^{(i)}_{i=1}^m \sim p_z$  and  $c^{(i)}_{i=1}^m \sim p_c$  # a batch from latent distribution
5:     # Update discriminator  $D$ :
6:      $g_{\theta_d} \leftarrow \nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^m [\log D(x^{(i)}) + \log(1 - D(G(z^{(i)}, c^{(i)})))] - \lambda \log Q(c^{(i)} | G(z^{(i)}, c^{(i)}))$ 
7:      $\theta_d \leftarrow \theta_d + \alpha \cdot \text{RMSProp}(\theta_d, g_{\theta_d})$ 
8:     # Update  $Q$  network:
9:      $g_{\theta_q} \leftarrow \nabla_{\theta_q} \frac{1}{m} \sum_{i=1}^m [\lambda \log Q(c^{(i)} | G(z^{(i)}, c^{(i)}))]$ 
10:     $\theta_q \leftarrow \theta_q + \alpha \cdot \text{RMSProp}(\theta_q, g_{\theta_q})$ 
11:   end for
12:   # Update generator  $G$ :
13:    $g_{\theta_g} \leftarrow \nabla_{\theta_g} [-\frac{1}{m} \sum_{i=1}^m \log(1 - D(G(z^{(i)}, c^{(i)})))] + \lambda \log Q(c^{(i)} | G(z^{(i)}, c^{(i)}))$ 
14:    $\theta_g \leftarrow \theta_g - \alpha \cdot \text{RMSProp}(\theta_g, g_{\theta_g})$ 
15: end while

```

Algorithm 1. Information Maximizing Generative Adversarial Networks.

In practice, generator G and discriminator D are both deep neural networks composed of multilayer perceptron, normalization, and leaky rectified linear units (leaky ReLU) with RMSProp algorithm for weight updates. n_{critic} denotes the number of discriminator updates per generator update, balancing their training pace and ensuring the discriminator's effectiveness in guiding the generator's learning process. Besides, 1D convolutional layers are also adopted in the proposed models.

In summary, InfoGAN is introduced to load profile generation, aiming at disentangling the features of power consumption data, which not only maintains the generative strengths of traditional GANs but also significantly improves the model's utility in understanding and manipulating complex load data distributions.

3.2 Multivariate kernel density estimation for feature modeling

To generate data similar to that of a specific customer, the feature variables corresponding to the customer's historical data can be used as a reference. As multivariate kernel density estimation (MKDE) extends the concept of kernel density estimation (KDE) to multiple dimensions, it could estimate the probability density functions of a vector of variables, enabling the sampling of new feature variables for the generation of new samples.

The typical formula of the MKDE can be expressed as follows, where $\hat{f}(x)$ denotes the estimated probability density function at point x , n represents the number of data points, x_i is the i -th data points, K_H is the kernel function measuring the similarity between the point x and the data point x_i .

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K_H(x - x_i)$$

However, MKDE presents increased computational complexity and challenges in bandwidth selection. The choice of bandwidth, critical in density estimation, becomes more complex as it often requires a matrix to appropriately scale the kernel in each dimension, considering inter-variable correlations. In this paper, we adopted the commonly used Silverman's rule (Zhang et al., 2006), which suggests the use of a diagonal bandwidth matrix $H = \text{diagonal}(h_1, \dots, h_d)$ where each diagonal element is derived from the corresponding univariate bandwidth estimate for each dimension. Each diagonal element can be expressed as follows, where d is the dimensionality of the data, n is the sample size, and σ_i is the standard deviation of the i -th dimension.

$$h_i = \left(\frac{4}{(d+2)n} \right)^{1/(d+4)} \sigma_i$$

Besides, the kernel function is pivotal as it determines the manner and extent of smoothing applied to the data in MKDE. Compared to other kernels, the Gaussian kernel facilitates the handling of the tails in the data distribution more effectively. As a result, the Gaussian kernel is employed as kernel function K to ensure a continuous, smooth density estimate.

4 Case study

In this section, the proposed method is trained with historical consumption data, aiming at extracting key features of load profiles

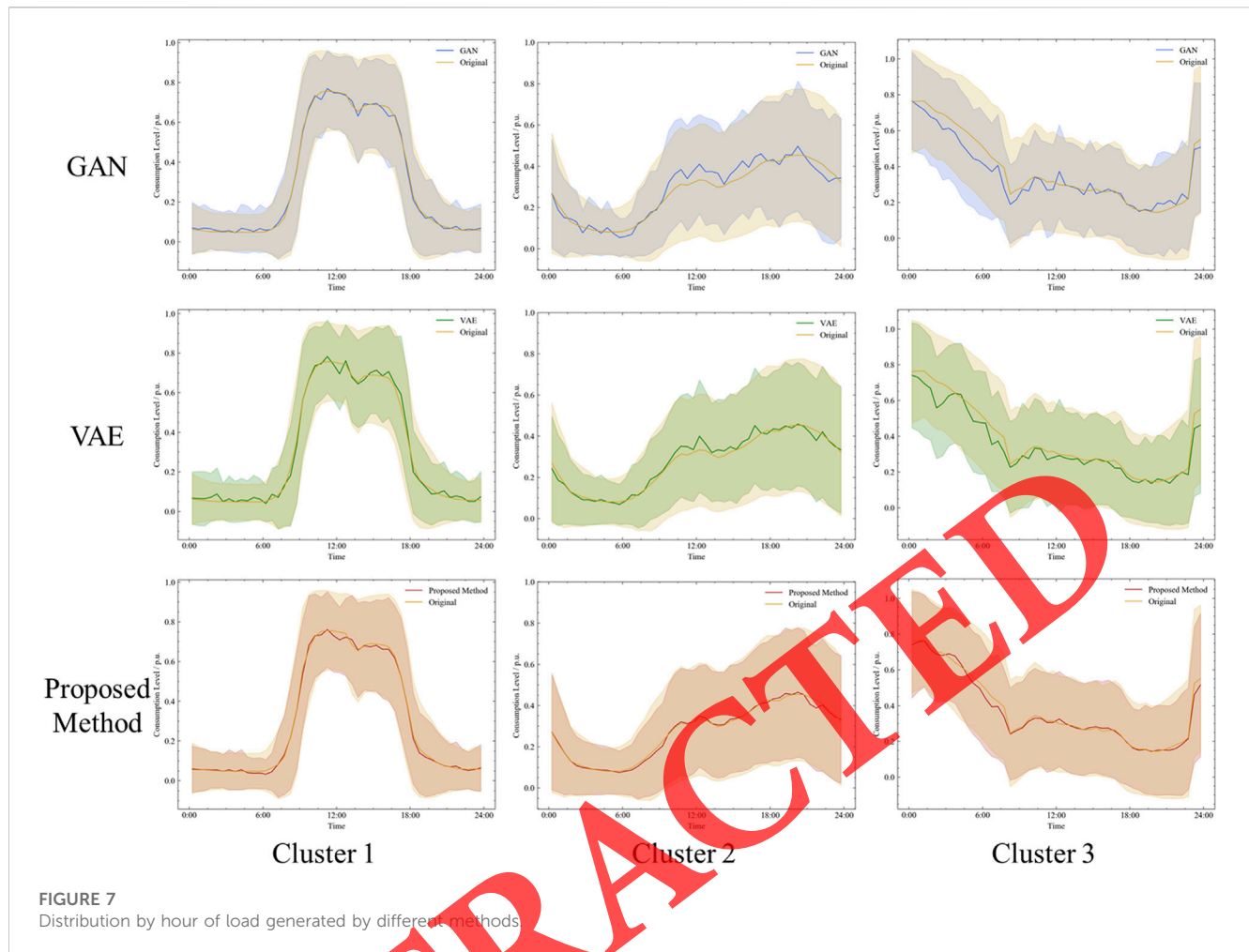


FIGURE 7
Distribution by hour of load generated by different methods

and generating realistic load profiles conformed to regular patterns of customers' electricity consumption.

4.1 Data description

The proposed method was validated on a dataset of small and medium enterprises (SME) in Ireland, which includes electricity consumption data collected every 30 min over a period of 1.5 years. Notably, the missing or abnormal data in this dataset, which can be the result of faulty data collection instruments, is fully removed instead of extrapolating the missing values to protect the original features. To address potential challenges arising from absolute consumption values, which could negatively impact the training approximation and generalization, all features were normalized through min-max scaling, resulting in a standardized range of [0,1]. This normalization process facilitated the comparison of load profiles across diverse customers and served to enhance the disclosure of dynamic data characteristics. Consequently, 90,155 days' load profiles from 319 customers were selected, and the data from 256 customers were used to train the generative model, while the rest was used for the testing.

To further understand the characteristics of load profiles, the K-means clustering method with $K = 3$ was employed to cluster all

load data, which also facilitates a more structured comparison within homogenous groups and effectively showcases the proposed method's performance. The clustering results and the consumption profiles of typical customers within each cluster are shown in Figure 4. It is observed that cluster 1 primarily exhibits consumption peaks at noon with lower loads in the morning and evening, resembling the electricity usage pattern of commercial office buildings. Cluster 2 demonstrates a gradual increase in load from noon to past midnight, likely representing businesses such as restaurants that operate into the evening. Cluster 3 shows higher loads during the early morning hours, suggesting enterprises that operate at night. The clustering results also reveal significant variability and uncertainty in the load profiles, making them difficult to describe with mathematical models. However, similarities could also be found among load profiles, which suggests that a data-driven approach could be effectively used for modeling.

4.2 Load profile generation for existing customers

In this paper, the feature variable c is a continuous variable containing ten features with a range of values between 0 and 1, and the dimension of noise variable z takes 50 which is sampled from a

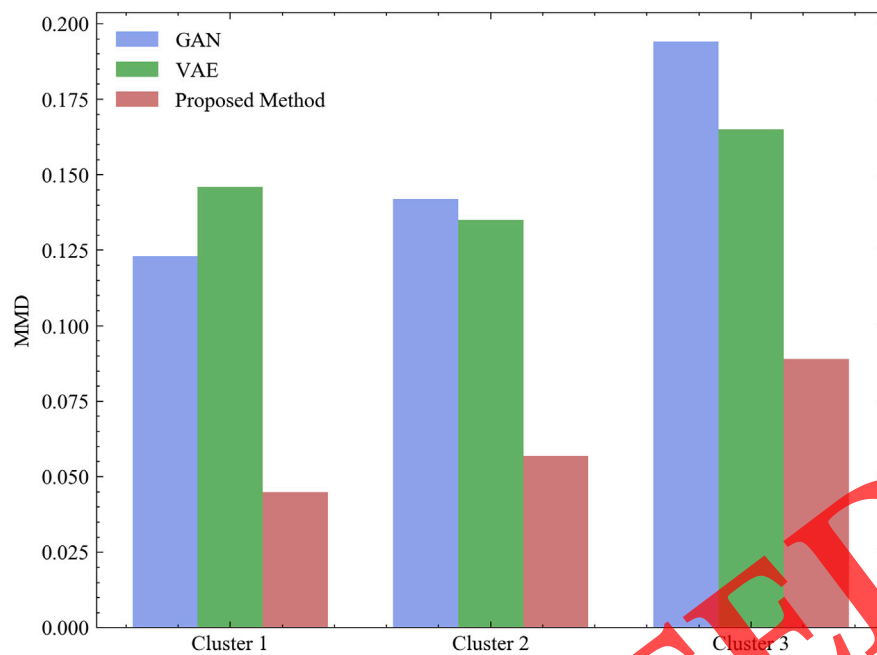


FIGURE 8
Comparison of MMD of different methods.

Gaussian distribution. All weights of G , D , and Q are initialized from a centered Normal distribution with a standard deviation of 0.02, and batch normalization is adopted.

Figure 5 provides a visual representation of the efficacy of the proposed method in capturing and replicating the inherent variability in electricity consumption patterns across different customers. Figure 5A delineates the actual load profiles for customers 7, 64, and 73, aggregating over a period of 100 days. These profiles are characterized by distinctive usage patterns, underscoring the individualized nature of electricity consumption. In Figure 5B, the red sections encapsulate the distribution of original feature variables extracted by the Q network. The blue section presents the distribution formed by randomly sampling 100 feature vectors after analyzing the original feature distribution with MKDE, illustrating the close approximation to the original feature distribution. Figure 5C showcases the generated load profiles based on the aforementioned sampled features. Remarkably, these synthesized profiles exhibit a high degree of resemblance to the actual load profiles, which is indicative of the model's ability to learn and simulate complex, real-world data distributions. It should be noted that since the generated data is randomly sampled from the feature space, there is no deterministic mapping between the original and the generated load profiles.

An empirical analysis of the impact of individual feature variables on the generated load profile is also conducted. A day's load data of a customer was chosen as the baseline, and its corresponding features were extracted through the Q network. Then, different features were modified separately, which were then used as input of the generator along with a fixed noise variable z to produce new load profiles. As shown in Figure 6, the abscissa of the figure corresponds to ten features, while the ordinate reflects the modified value of these features. The

baseline load data, represented in orange, serves as a reference against which the impact of feature adjustments can be measured, while the generated load profiles are depicted in blue. Notably, regions with significant deviations from the baseline are encapsulated within red dashed boxes. It can be observed that each feature has a noticeable impact on the load profiles. For example, feature c_1 predominantly impacts the onset of the morning load peak and a c_2 appears to determine the peak load value in the afternoon. Besides, the interaction between features is not isolated, as the morning load variation is collectively influenced by features c_6 through c_9 , which suggests a complex and interconnected feature space contributes to shaping the load profile. The figure demonstrates the capacity of the proposed method to learn and generate the dynamics of load profiles, which also proves its utility in learning disentangled and interpretable representations of the load profiles.

4.3 Comparative analysis

In this paper, the two most commonly used generative models, Variational Autoencoders (VAE) (Kingma et al., 2019) and Vanilla GAN, are used for comparison. Due to the extensive variety of load profiles in the original dataset, evaluating the effectiveness of the proposed method using the complete dataset is challenging. Therefore, a comparative approach using different clusters is adopted. Both VAE and GAN are trained with corresponding data of the test clusters and generate 1000 samples. In the proposed method, MKDE is used to estimate the feature distribution under different clusters and feature variables are then sampled based on these distributions to generate samples.

Figure 7 presents a schematic of the samples generated by different methods compared to the samples of the corresponding clusters. The

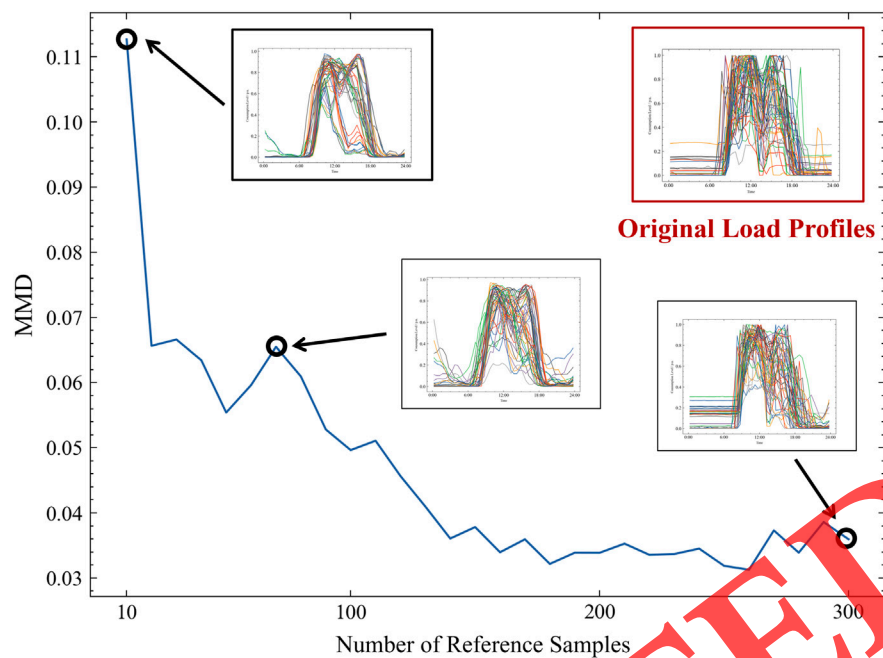


FIGURE 9

The relationship between generated sample quality and reference sample quantity.

central line graph in each cluster illustrates the average load profile generated by the respective method, with the surrounding shaded area denoting the standard deviation from this mean, encapsulating the variability and dispersion of the samples. It is apparent that the proposed method closely approximates the original data, as indicated by both mean and standard deviation. The results demonstrate the proposed method maintains good statistical characteristics when generating a large volume of samples.

Furthermore, the Maximum Mean Discrepancy (MMD) metric (Gretton et al., 2012) is also used to estimate the effectiveness of the proposed method. This metric quantifies the disparity between the distributions of real and generated data, with a smaller MMD value signifying a greater similarity between the two data sets. As shown in Figure 8, the results demonstrate that the data generated by the proposed method exhibit higher similarity across various scenarios compared to traditional generative methods, thereby validating the effectiveness of the proposed approach.

4.4 Load profile generation for new customers

In demand-side management, there also arises a need to generate additional user load data from a limited dataset to infer probable energy usage scenarios, and the enhanced load data could facilitate comprehensive demand forecasting, load balancing, and tailoring energy efficiency measures. Compared to traditional models like the vanilla Generative Adversarial Networks (GAN), which have uncontrollable output results, the proposed method allows for the generation of customer load profiles that align with known information. Besides, the proposed method does not require prior knowledge of a customer's specific category, as is necessary for

models like conditional GANs (cGANs), and leverages existing samples as a reference. When limited real data of a customer is available, the proposed method can also rapidly generate potential load profiles for research and analysis.

Different quantities of samples were selected from the test set as reference samples, and the feature distribution was first obtained using MKDE, followed by generating corresponding samples. As shown in Figure 9, the proposed method is capable of generating similar samples even with as few as ten reference samples. As the quantity of reference samples increases, the Maximum Mean Discrepancy (MMD) between the generated samples and the real samples progressively decreases, which indicates that the generated dataset increasingly resembles the real sample set, demonstrating the method's efficacy in accurately replicating real-world data. Additionally, the diversity of the generated samples also increases with the number of reference samples, suggesting that more reference samples can better depict the feature space corresponding to the real samples.

5 Conclusion

This paper presents a novel methodology combining InfoGAN and MKDE for generating customer load profiles. The approach leverages InfoGAN to learn from existing load data, with its Q network effectively disentangling feature variables and the generator producing realistic load profiles. MKDE is then used to assess the distribution of these features for new profile generation. Through this procedure, the privacy of customers is well protected because the real data are separated from third parties by generative models. The case studies have demonstrated the quality of generated samples compared to the real load profiles, which could be proof of the effectiveness of the proposed method. The proposed method provides an effective tool for load data analysis in

power systems, offering significant support for the planning and management of power systems.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <http://www.ucd.ie/issda/data/commissionforenergyregulationcer/>.

Author contributions

JL: Writing—original draft, Writing—review and editing. YZ: Writing—original draft, Writing—review and editing. QG: Writing—review and editing. HS: Writing—review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by National Key R&D Program of China (2018AAA0101503) and by the Science and Technology Project of State Grid Corporation of China: Fundamental Theory of

Human-in-the-loop Hybrid-augmented Intelligence for Power Grid Dispatch and Control.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The authors declare that this study received funding from the Science and Technology Project of State Grid Corporation of China: Fundamental Theory of Human-in-the-loop Hybrid-augmented Intelligence for Power Grid Dispatch and Control. The funder had the following involvement in the study: the decision to submit it for publication.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Arif, A., Wang, Z., Wang, J., Mather, B., Bashualdo, H., and Zhao, D. (2017). Load modeling—a review. *IEEE Trans. Smart Grid* 9, 5986–5999. doi:10.1109/TSG.2017.2700436
- Armoogum, S., and Bassoo, V. (2019). "Privacy of energy consumption data of a household in a smart grid," in *Smart power distribution systems* (Amsterdam, Netherlands: Elsevier), 163–177.
- Chen, X., Duan, Y., Houthoofd, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016). Infogan: interpretable representation learning by information maximizing generative adversarial nets. *Adv. neural Inf. Process. Syst.* 29. Available at: https://proceedings.neurips.cc/paper_files/paper/2016/hash/7c9d0b1f96aebd7b5eca8c3edaa19ebb-Abstract.html.
- Chen, Y., Wang, Y., Kirschen, D., and Zhang, B. (2018). Model-free renewable scenario generation using generative adversarial networks. *IEEE Trans. Power Syst.* 33, 3265–3275. doi:10.1109/TPWRS.2018.2794541
- Da Silva, P. G., Ilić, D., and Karnouskos, S. (2013). The impact of smart grid prosumer grouping on forecasting accuracy and its benefits for local electricity market trading. *IEEE Trans. Smart Grid* 5, 402–410. doi:10.1109/TSG.2013.2278868
- Debie, E., Moustafa, N., and Whitty, M. T. (2020). "A privacy-preserving generative adversarial network method for securing EEG brain signals," in *Proceedings of the 2020 international joint conference on neural networks (IJCNN)*, Glasgow, UK, July 2020 (IEEE), 1–8. doi:10.1109/IJCNN48605.2020.9206683
- Gillis, J. M., Alshareef, S. M., and Morsi, W. G. (2015). Nonintrusive load monitoring using wavelet design and machine learning. *IEEE Trans. Smart Grid* 7, 320–328. doi:10.1109/TSG.2015.2428706
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. *Adv. neural Inf. Process. Syst.* 27. Available at: https://proceedings.neurips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.
- Grandjean, A., Adnot, J., and Binet, G. (2012). A review and an analysis of the residential electric load curve models. *Renew. Sustain. energy Rev.* 16, 6539–6565. doi:10.1016/j.rser.2012.08.013
- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* 13, 723–773. Available at: <http://jmlr.org/papers/v13/gretton12a.html>.
- Haben, S., Singleton, C., and Grindrod, P. (2016). Analysis and clustering of residential customers energy behavioral demand using smart meter data. *IEEE Trans. Smart Grid* 7, 136–144. doi:10.1109/TSG.2015.2409786
- Hindistan, Y. S., and Yetkin, E. F. (2023). A hybrid approach with GAN and DP for privacy preservation of IIoT data. *IEEE Access* 11, 5837–5849. doi:10.1109/access.2023.3235969
- Hu, J., and Vasilakos, A. V. (2016). Energy big data analytics and security: challenges and opportunities. *IEEE Trans. Smart Grid* 7, 2423–2436. doi:10.1109/TSG.2016.2563461
- Hu, T., Guo, Q., Shen, X., Sun, H., Wu, R., and Xi, H. (2019). Utilizing unlabeled data to detect electricity fraud in AMI: a semisupervised deep learning approach. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 3287–3299. doi:10.1109/TNNLS.2018.2890663
- Kingma, D. P., and Welling, M. (2019). An introduction to variational autoencoders. *Found. Trends® Mach. Learn.* 12, 307–392. doi:10.1561/22000000056
- Kwac, J., Flora, J., and Rajagopal, R. (2014). Household energy consumption segmentation using hourly data. *IEEE Trans. Smart Grid* 5, 420–430. doi:10.1109/TSG.2013.2278477
- Li, X., Bowers, C. P., and Schnier, T. (2009). Classification of energy consumption in buildings with outlier detection. *IEEE Trans. Industrial Electron.* 57, 3639–3644. doi:10.1109/TIE.2009.2027926
- Mohassel, R. R., Fung, A., Mohammadi, F., and Raahemifar, K. (2014). A survey on advanced metering infrastructure. *Int. J. Electr. Power and Energy Syst.* 63, 473–484. doi:10.1016/j.jepes.2014.06.025
- Qu, Y., Yu, S., Zhou, W., and Tian, Y. (2020). Gan-driven personalized spatial-temporal private data sharing in cyber-physical social systems. *IEEE Trans. Netw. Sci. Eng.* 7, 2576–2586. doi:10.1109/TNSE.2020.3001061
- Sangogboye, F. C., Jia, R., Hong, T., Spanos, C., and Kjærgaard, M. B. (2018). A framework for privacy-preserving data publishing with enhanced utility for cyber-physical systems. *ACM Trans. Sens. Netw. (TOSN)* 14, 1–22. doi:10.1145/3275520
- Wang, H., Qi, L., Yan, L., and Li, Z. (2021). Load photo: a novel analysis method for load data. *IEEE Trans. Smart Grid* 12, 1394–1404. doi:10.1109/TSG.2020.3025936
- Wang, Y., Chen, Q., Gan, D., Yang, J., Kirschen, D. S., and Kang, C. (2018). Deep learning-based socio-demographic information identification from smart meter data. *IEEE Trans. Smart Grid* 10, 2593–2602. doi:10.1109/TSG.2018.2805723
- Zhang, C., Shao, Z., Jiang, C., and Chen, F. (2021). A PV generation data reconstruction method based on improved super-resolution generative adversarial network. *Int. J. Electr. Power and Energy Syst.* 132, 107129. doi:10.1016/j.jepes.2021.107129
- Zhang, X., King, M. L., and Hyndman, R. J. (2006). A Bayesian approach to bandwidth selection for multivariate kernel density estimation. *Comput. Statistics Data Analysis* 50, 3009–3031. doi:10.1016/j.csda.2005.06.019