This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/UPEC57427.2023.10294666, 2023 58th International Universities Power Engineering Conference (UPEC)

Synthetic Electricity Consumption Data Generation using Tabular Generative Adversarial Networks

Thet Paing Tun
Department of Electronic and Electrical Engineering
Brunel University London
London, UK
thetpaing.tun@brunel.ac.uk

Ioana Pisica
Department of Electronic and Electrical Engineering
Brunel University London
London, UK
ioana.pisica@brunel.ac.uk

***Abstract*— Generating synthetic electricity consumption data is crucial for developing efficient energy systems in smart cities. In this paper, we propose the use of Tabular Generative Adversarial Networks (Tabular GAN) for generating synthetic data for residential electricity consumption. Tabular GANs have been used in various domains and have shown promising results in generating high-quality synthetic data. The performance of our proposed method was evaluated by comparing the probability density, mean, standard deviation, and variances of the synthetic data with the original data. The results showed that the Tabular GAN method generated synthetic data that closely match the statistical characteristics of the original data and the simulation outcome indicated that the synthetic data generated by Tabular GAN could effectively simulate the patterns and behaviors observed in the original data. Overall, the proposed method demonstrates the effectiveness of using Tabular GANs for generating synthetic electricity consumption data.**

Keywords*— *GAN, Tabular GAN, CTGAN, synthetic data, electricity consumption

I. INTRODUCTION

In the transition to net-zero energy scheme, the consumer data can enable decision-makers at the national level to shift from relying solely on market forces, which have guided past energy transitions, towards implementing psychological interventions by engaging consumers, influencing energy-related behaviours, and facilitating cross-sectoral transitions towards achieving Net-Zero [1].

The increasing connectivity and automation of today's power grid are being driven by the grid's pervasive communication and computational capabilities [2]. However, collecting and maintaining large-scale electricity consumption data is challenging due to various factors, such as privacy concerns, data accessibility, and data quality issues. Additionally, not every client currently possesses an advanced metering infrastructure (AMI) that permits the measurement and storage of load profiles [3].

These limitations pose significant challenges for researchers and practitioners who require accurate and diverse data to develop models, evaluate policies, and understand the behaviours of electricity consumers.

In addition to this, Oh et al. (2022) highlighted that data scarcity is a critical issue in the modern engineering industry [4]. To make matters worse, the obstacles in developing evidence-based economic growth policies for the energy sector of these economies arise from a noticeable lack of data [5]. Thus, the recent literature in [6] used synthetic data generation to facilitate the upcoming research and development of future smart grids.

The significance of synthetic data is increasing across several disciplines with its application as a replacement for real data to simulate alternative scenarios, and to facilitate the development and testing of AI models in fields where data scarcity or privacy concerns are critical factors [10].

Generative models have proven to be highly effective in generating synthetic data and among the most widely used and promising models are the Generative Adversarial Network (GAN) and Variational Autoencoder (VAE) models [11]. Between the two methods, GANs are generally considered to be superior to variational autoencoders [12].

Generative Adversarial Network (GAN) is a category of deep learning method, and it was first introduced by Goodfellow et al. (2014) in [13] with the objective of generating synthetic data by means of the adversarial process.

Despite the outperformances of GAN, it can be challenging to generate synthetic data that accurately reflects the underlying distribution, particularly in high-dimensional or complex data domains and thus it is necessary to develop more effective techniques for addressing the problem of uneven distribution in synthetic data generation.

To fulfil the abovementioned research gap, Ashrapov (2020) developed a Tabular Generative Adversarial Networks (GAN) method to reduce the probability of uneven distribution of data in the data synthesizing process [14].

The primary contribution of this study involves the advancement of a potential methodology for correlating the environmental and overall social behavior of energy users, enabling the synthesis of dynamic trends in electricity consumption. This is accomplished through the demonstration of a novel approach utilizing Tabular GAN to generate electricity consumption data, with the objective of supporting future energy management and grid management efforts in the face of uncertainty.

II. RELATED WORK

With the modernization of the electrical grid into a smart grid, the use of data science has become more crucial in operation monitoring grid activity [15].

Synthetic data generation plays an important role in smart city planning by allowing for the prediction of energy demand from various sources. It facilitates the analysis of energy demand patterns, peak loads, and energy supply optimization.

Recent research has explored different approaches for generating synthetic load data. Pinceti et al. (2022) utilized a combination of GAN, conditional GAN, and singular value decomposition to synthesize time-series residential and industrial load data across varying timescales [16].

Similarly, Hosseini et al. (2017) developed a semi-synthetic dataset development tool using statistical methods to support house energy management systems [6].

Generating high-quality synthetic data is essential to monitor measurement error. Hazra et al. (2022) predicted smart meter measurement error using synthetic data generated through TGAN-skipped-WGAN-GP [17], while Moon et al. (2020) proposed a two-stage CTGAN method to generate

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/UPEC57427.2023.10294666, 2023 58th International Universities Power Engineering Conference (UPEC)

synthetic data for short-term load forecasting in distribution grids [18].

Other studies have focused on generating synthetic load consumption data for smart homes. Razghandi et al. (2022) integrated Variational Auto Encoder (VAE) and GAN for this purpose [9]. Chatterjee and Byun (2023) used an ensemble regression method to generate synthetic electric vehicle (EV) profiles in [8], and Ezhilarasi et al. (2023) introduced FBprophet as a tool to generate synthetic data for household electricity consumption based on the Low Carbon London project dataset [15].

Finally, Reis et al. (2020) proposed a queuing model implemented in Python for generating synthetic residential load data to support smart city energy management [7].

III. Methodology

The main programming language for the proposed work was Python and all the code implementation and simulation were done in PyCharm Community edition 2022.3 due to its flexibility in installing the essential library tools for the proposed work. The main library used for building Tabular GAN was 'tabgan'. Furthermore, the residential consumption data for Low Carbon London project available in [19] was used as a source data for generating new synthetic load data samples while the weather variables are based on the NASA data source available at [20].

A. Data Preparation

The overall process of data preparation is as described in Fig. 1.

To test and evaluate the effectiveness of Tabular GAN, the data for 8 consumption profiles were randomly selected from the meter data available for 'Low Carbon London' project. Each raw dataset comprises the electricity consumption of households in every 30-minute period (half-hourly) from November 2011 to February 2014.

As the date and time values were provided in the format of 'MM/DD/YYYY H:M:' in the original dataset, the proper date and time were split, and the day-of-week and month-of-year were excavated from the provided date by means of the 'datetime' function in Python. For this study, the variables for day type were carefully selected to include three categories: normal (working) day, day before holiday, and holiday. This is because consumer behavior may be particularly significant on the day before a holiday, especially in the evening. The UK Bank Holiday information were accessed via [21].

After cleaning or removing the outlier values that occurred due to measuring errors such as 'N/A', the last 1 year (from March 2013 to February 2014) from each of the selected datasets was reserved for testing while the rest data (from November 2011 to February 2013) were used for training with Tabular GAN.

According to the information provided in [19], the downloaded load dataset was based on the households located in East and South East London and thus, the latitude and

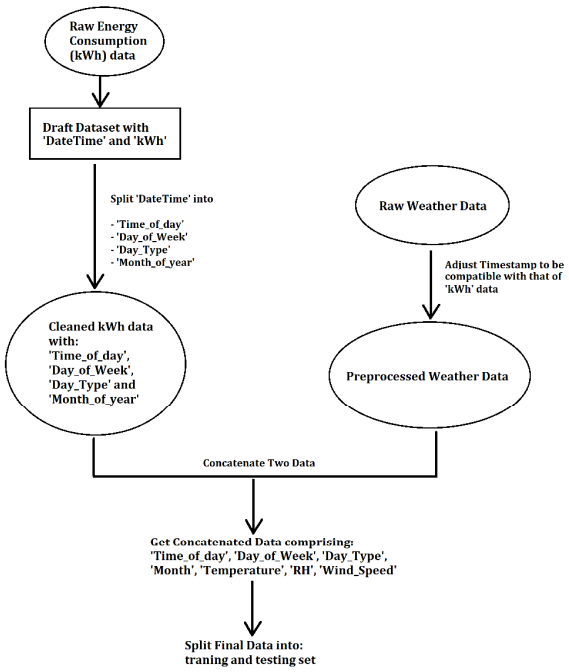


Fig. 1. Overall Process of Data Preparation

longitude of the selected location were 51.4941 and -0.0386 in degrees, respectively to access the historical weather information about average temperature, RH and wind speed.

As the downloaded weather data were hourly sampled, the data values for the 30-minute ahead of each hour were assumed to be the same as that of the specific hour. For example, the data value of 15:30 was taken as the same value as that of 15:00. Then the electricity consumption and weather data were concatenated, and the specifications of the preprocessed data were described in Table 1.

TABLE I. Features and Response from the Preprocessed Data

Name	Feature/Response	Variables
Time_of_day	Feature	0 – 47 (represents 00:00 to 23:30 with 30-min timestamp)
Day	Feature	0 – 6 (represents Monday to Sunday)
Day_Type	Feature	0,1,2 (0 = normal working day 1= day before holiday 2= holiday)
Month	Feature	1 – 12 (represents January to December)
Temperature	Feature	+/- continuous variables
Relative Humidity (RH)	Feature	Continuous variables between 0 and 100
Wind Speed	Feature	Positive continuous variables
Electricity Consumption (kWh)	Response	Positive continuous variables

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI10.1109/UPEC57427.2023.10294666, 2023 58th International Universities Power Engineering Conference (UPEC)

B. Synthetic Data Generation

As it was described in the introduction section, the Tabular GAN was the main tool used for generating synthetic data for electricity consumption.

Generative Adversarial Network (GAN) is a deep learning framework comprising two neural networks, namely, generator and discriminator. The generator takes random noise as input and generates synthetic data that resembles real data. The discriminator takes real and fake data as input and distinguishes between them. The generator and the discriminator are trained in an adversarial manner, where the generator tries to generate more realistic data to pit the discriminator, and the discriminator learns to distinguish between real and fake data.

The generator and discriminator of GAN can be mathematically described as follows where 'D' represents the discriminator and 'G' the generator while 'x' is the sample drawn from the real dataset and 'z' is the noise data initially fed to the generator network.

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[log(D(x)] + \left(E_{z\sim p_z(z)}[log(1-D(G(z))]\right) \quad (1)$$

To learn the generator's distribution, a prior was defined on input noise variables, and a mapping to data space is represented as a differentiable function G(z). A second multilayer perceptron D(x) was also defined to output a single scalar representing the probability that x came from the data rather than the generator's distribution. D(x) was trained to maximize the probability of assigning the correct label to both training examples and samples from G(z), which was trained to minimize log(1 - D(G(z))). In other words, D and G play a two-player minimax game with a value function V(G, D).

Tabular GAN is a sub-category of GAN that is specifically designed for structured data such as tables and relational databases.

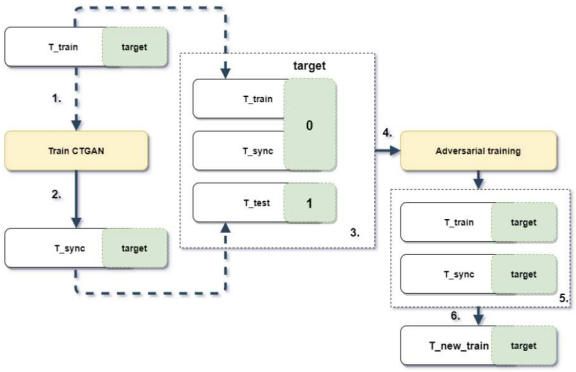


Fig. 2. Process flow of the proposed Tabular GAN [14]

For this particular scenario, the electricity consumption data from November 2011 to February 2013 was utilized as the training dataset (T-train). Meanwhile, the testing dataset (T-test) consisted of data from March 2013 to February 2014, as outlined in the 'Data Preparation' stage.

The goal was to improve the performance of a machine learning model trained on a dataset consisting of T train and tested on 'T-test'. To achieve this, the 'T-train' dataset was augmented by generating new data using the CTGAN method. This newly generated data, denoted as 'T-synth', was created in a manner that was similar to the distribution of 'T-test', without relying on ground truth labels.

The experimental design involves several steps, which are depicted in Figure 3. First, CTGAN is trained on 'T-train' using ground truth labels, and additional data 'T-synth' was generated (step 2). Then, an adversarial boosting technique was applied to concatenate 'T-train' and T synth (with the target set to 0) and 'T-test' (with the target set to 1), in order to train a new model (steps 3 and 4). The aim of this step was to use the newly trained adversarial boosting model to obtain rows that are more similar to 'T-test'. During the adversarial training, the original ground truth labels were not used.

The resulting rows were most similar to 'T-test' which were then selected from 'T-train' and 'T-synth', and sorted in correspondence to 'T-test' (steps 5 and 6).

Among the feature labels provided in Table 1, one-hot-encoding is used to train 'Time_of_day', 'Day', 'Day_Type' and 'Month' which were determined as categorical features.

The loss function used for GAN training is 'Wasserstein' distance function which could be mathematically described in general as follows. This function measures the distance between the joint distribution of the real and the generated synthetic to verify the generated data follow the trend of the original data.

$$W_p(P,Q) = (\sum_{i=1}^{n} \|R_i - T_i\|^p)^{\frac{1}{p}} \quad (2)$$

where P = joint distribution of real data R

Q = joint distribution of generated data T

n = number of samples randomly selected for testing

The simulation parameters for training the Tabular GAN is summarized in Table 2 as follows.

TABLE II. SIMULATION PARAMETERS

Parameter	Description	Value
gen_x_times	factor increasing the size of the generated data set	1.1
pregeneration_frac	fraction of data to generate before training GAN (amount of data generated = gen_x_times * pregeneration_frac)	2
bot_filter_quantile	0.001	Bottom quantile of data to keep (below which will be removed)
top_filter_quantile	0.999	Top quantile of data to keep (above which will be removed)
cat_cols	Categorical columns	Time_of_day, Day, Day_Type, Month
Batch_size	Batch_size	100
epochs	the number of iteration during training	120
Learning_rate	Learning_rate	0.01

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI10.1109/UPEC57427.2023.10294666, 2023 58th International Universities Power Engineering Conference (UPEC)

IV. Results

After running the simulation, the results are visualized in Table 3-4 and Fig. 3-4. 'M' in the provided tables means 'meter'.

TABLE III. STATISTICAL PARAMETERS FOR REAL AND SYNTHETIC KWH OUTPUT DATA

		Mean	STD	Var
M1	**Real**	0.076	0.063	0.004
	Synthetic	0.084	0.069	0.005
M2	**Real**	0.128	0.083	0.007
	Synthetic	0.149	0.097	0.009
M3	**Real**	0.28	0.231	0.054
	Synthetic	0.34	0.228	0.052
M4	**Real**	0.175	0.179	0.032
	Synthetic	0.16	0.163	0.027
M5	**Real**	0.079	0.084	0.007
	Synthetic	0.083	0.086	0.0073
M6	**Real**	0.117	0.103	0.011
	Synthetic	0.119	0.099	0.01
M7	**Real**	0.264	0.191	0.037
	Synthetic	0.298	0.194	0.038
M8	**Real**	0.13	0.122	0.015
	Synthetic	0.138	0.114	0.013

[a.] Mean=average, STD = standard deviation, VAR = variance

According to Table 3, the statistical variables of the synthetic data exhibited a close similarity to those of the real data, despite minor differences in the standard deviation of M2, M4, and M8 when compared to other meter units. Additionally, the probability density charts in Fig. 3 demonstrated that the synthetic data closely followed the trend of the real data, with small deviations observed in the distribution of the smallest kWh values, particularly in M3 and M7.

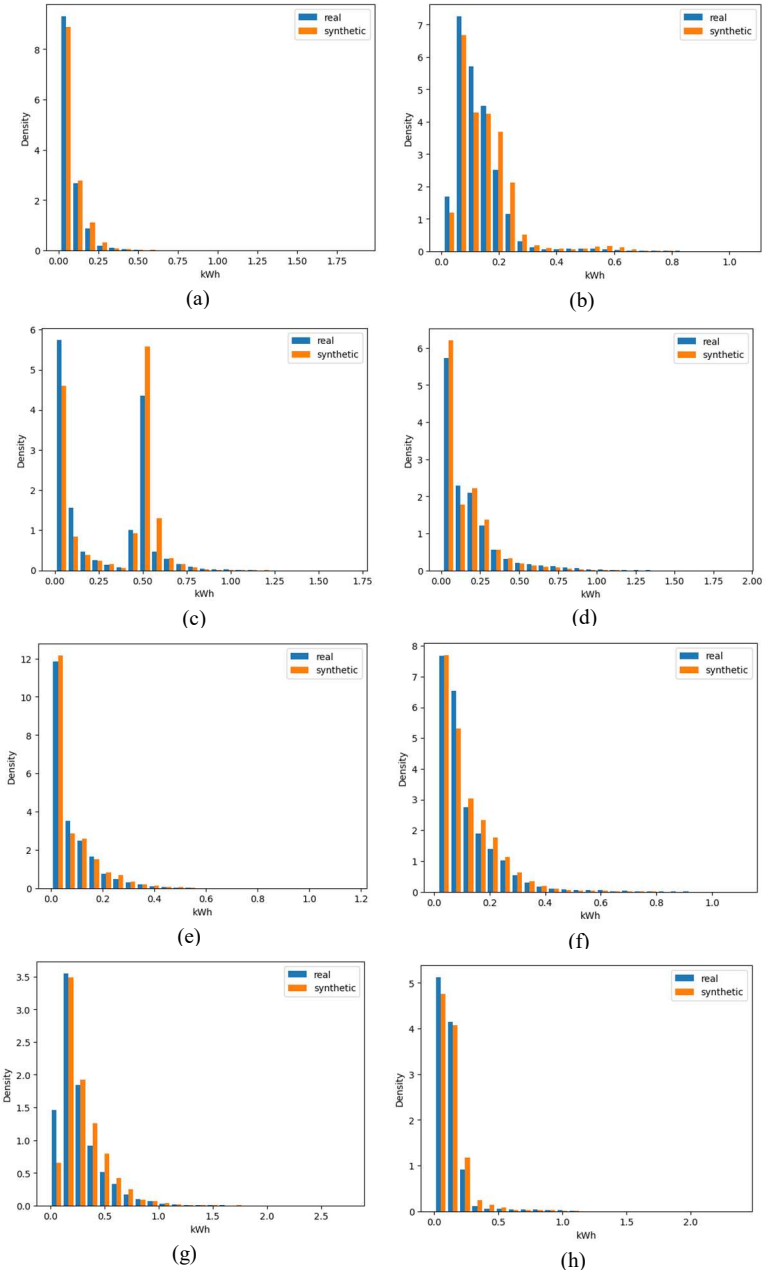


Fig. 3. Comparison of Probability Density of kWh between real and synthetic data for (a) Meter 1, (b) Meter 2, (c) Meter 3, (d) Meter 4, (e) Meter 5, (f) Meter 6, (g) Meter 7 and (h) Meter 8 (blue = real data, orange = synthetic data)

However, the synthetic kWh data were properly correlated with the individual generated features, which closely followed the correlation trend between actual kWh data and its respective features as depicted in Table 4.

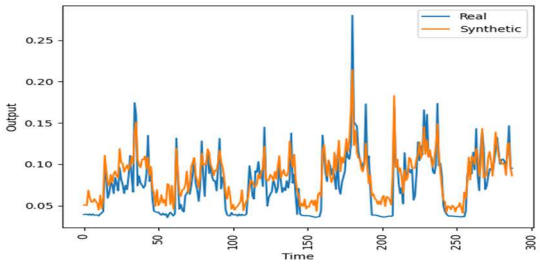
Moreover, Fig. 4 visually represented the average half-hourly load profiles with seasonal variation. It should be noted that the Autumn and Spring patterns were omitted from the figure, as they followed a similar trend to that of Winter and Summer.

The graphs in Fig. 4, from a to h, depicted the average half-hourly consumption for six scenarios, starting with Summer: normal day, day-before-holiday, holiday, and Winter: normal day, day-before-holiday, and holiday, each containing 48 samples (24-hour profile with each 30-minute sample). The graphical results confirmed that the synthetic kWh data closely followed the trend of the real kWh data with small deviations.

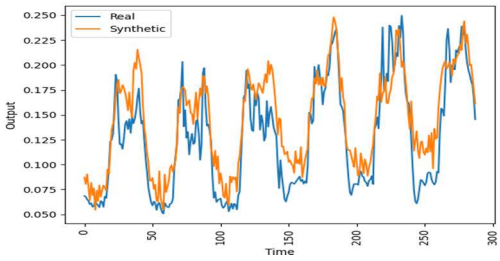
This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/UPEC57427.2023.10294666, 2023 58th International Universities Power Engineering Conference (UPEC)

TABLE IV. COMPARISON OF KWH CORRELATION WITH FEATURES IN BOTH REAL AND SYNTHETIC DATA

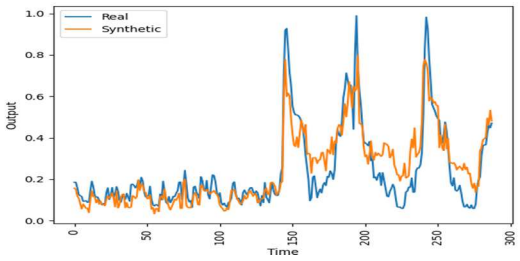
		Time_of_day	Day	Day_Type	Month	Temperature	RH	Wind_Speed
kWh (M1)	**Real**	0.35	-0.022	-0.031	-0.025	0.01	-0.1	0.1
	Synthetic	0.29	-0.027	-0.033	-0.031	0.014	-0.12	0.07
kWh (M2)	**Real**	0.45	0.0086	0.027	-0.037	-0.028	-0.15	0.128
	Synthetic	0.37	0.0112	0.038	-0.047	-0.019	-0.10	0.092
kWh (M3)	**Real**	0.016	-0.029	-0.027	-0.39	-0.55	0.13	0.16
	Synthetic	0.022	-0.036	-0.018	-0.29	-0.51	0.18	0.16
kWh (M4)	**Real**	0.35	0.012	0.026	0.033	-0.037	0.0010	0.074
	Synthetic	0.27	0.014	0.031	0.024	-0.049	0.0007	0.059
kWh (M5)	**Real**	0.33	0.037	0.041	0.038	-0.18	0.13	0.085
	Synthetic	0.26	0.032	0.039	0.042	-0.24	0.15	0.058
kWh (M6)	**Real**	0.38	0.022	0.037	-0.0037	-0.084	0.017	0.074
	Synthetic	0.31	0.023	0.039	-0.0052	-0.119	0.014	0.089
kWh (M7)	**Real**	0.12	-0.034	-0.032	-0.022	0.0030	-0.017	0.059
	Synthetic	0.09	-0.026	-0.029	-0.029	0.0022	-0.023	0.065
kWh (M8)	**Real**	0.25	0.065	0.074	-0.008	0.115	-0.16	0.056
	Synthetic	0.21	0.063	0.072	-0.013	0.122	-0.19	0.087



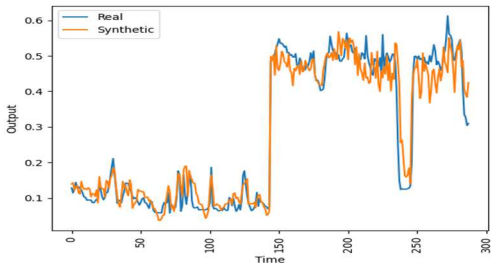
(a)



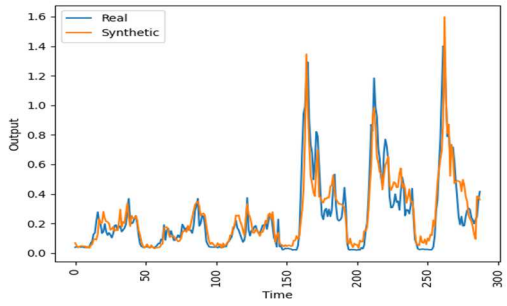
(b)



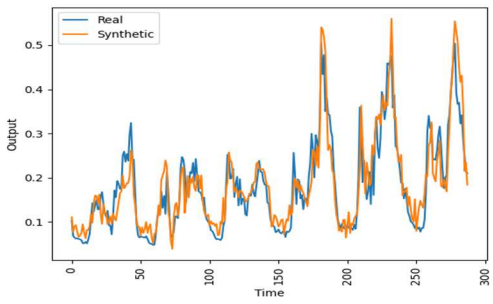
(c)



(d)

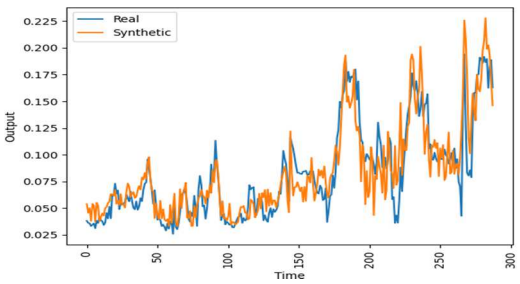


(e)

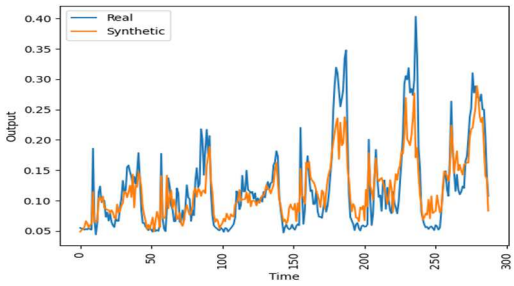


(f)

This article has been accepted for publication in a future proceedings of this conference, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI10.1109/UPEC57427.2023.10294666, 2023 58th International Universities Power Engineering Conference (UPEC)



(g)



(h)

Fig.4. Average Half-hourly kWh Profile of (a) Meter 1, (b) Meter 2, (c) Meter 3, (d) Meter 4, (e) Meter 5, (f) Meter 6, (g) Meter 7 and (h) Meter 8 (In x-axis, 0-47 = normal working day in summer, 48-95 = day before holiday in summer, 96-143 = holiday in summer, 144-191 = normal working day in winter, 192-239 = day before holiday in winter and 240-287 = holiday in winter)

V. Conclusion

To summarize, this study showcased the significant potential of Tabular Generative Adversarial Networks (Tabular GAN) in bolstering energy management systems and facilitating load forecasting in smart cities. Moreover, it delved into the exploration of the diverse electricity consumption patterns of residential consumers, considering factors such as environmental conditions and holidays, thus laying the foundation for subsequent tasks involving consumer behavior modeling. Future endeavors should entail rigorous testing and validation of the proposed method using extended load data sets and additional sources of energy-related data. Such efforts will contribute to the advancement of this research field and enhance its practical applicability in real-world scenarios.

References

[1] L. Liu, M. Workman, and S. Hayes, “Net Zero and the potential of consumer data - United Kingdom energy sector case study: The need for cross-sectoral best data practice principles,” in Energy Policy, vol. 163, pp. 112803, February 2022, doi: https://doi.org/10.1016/j.enpol.2022.112803

[2] S. E. Kababji and P. Srikantha, "A Data-Driven Approach for Generating Synthetic Load Patterns and Usage Habits," in IEEE Transactions on Smart Grid, vol. 11, no. 6, pp. 4984-4995, Nov. 2020, doi: 10.1109/TSG.2020.3007984.

[3] P. Wiest, D. Contreras, D. Gross and K. Rudion, "Synthetic Load Profiles of Various Customer Types for Smart Grid Simulations," NEIS 2018; Conference on Sustainable Energy Supply and Energy Storage Systems, Hamburg, Germany, 2018, pp. 1-6.

[4] K. Oh, E. J. Kim, and C. Y. Park, “A Physical Model-Based Data-Driven Approach to Overcome Data Scarcity and Predict Building Energy Consumption,” in Sustainability, vol. 14, no. 15, pp. 9464, August 2022, doi: https://doi.org/10.3390/su14159464

[5] R. Nepal, B. Sharma and M. I. a. Irsyad, “Scarce data and energy research: Estimating regional energy consumption in complex economies,” in Economic Analysis and Policy, vol. 65, pp. 139-152, December 2019, doi: https://doi.org/10.1016/j.eap.2019.12.002

[6] S. Hosseini, S. Kelouwani, K. Agbossou, A. Cardenas and N. Henao, "A semi-synthetic dataset development tool for household energy consumption analysis," 2017 IEEE International Conference on Industrial Technology (ICIT), Toronto, ON, Canada, 2017, pp. 564-569, doi: 10.1109/ICIT.2017.7915420.

[7] F. B. d. Reis, R. Tonkoski and T. M. Hansen, " Synthetic residential load models for smart city energy management simulations," in IET Smart Grid, vol. 3, no. 3, pp. 342-354, May. 2020, doi: https://doi.org/10.1049/iet-stg.2019.0296

[8] S. Chatterjee and Y. C. Byun, " A Synthetic Data Generation Technique for Enhancement of Prediction Accuracy of Electric Vehicles Demand," in Sensors, vol. 23, pp. 594, January. 2023, doi: https://doi.org/10.3390/s23020594

[9] M. Razghandi, H. Zhou, M. Erol-Kantarci and D. Turgut, "Variational Autoencoder Generative Adversarial Network for Synthetic Data Generation in Smart Home," ICC 2022 - IEEE International Conference on Communications, Seoul, Korea, Republic of, 2022, pp. 4781-4786, doi: 10.1109/ICC45855.2022.9839249.

[10] S. Thorve, A. Vullikanti, H. S. Mortveit, S. Swarup and M. V. Marathe, " Fidelity and diversity metrics for validating hierarchical synthetic data: Application to residential energy demand," 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022, pp. 1377-1382, doi: 10.1109/BigData55660.2022.10020837.

[11] P. Pandey, 2018. *Deep Generative Models*. [online] towardsdatascience.com. Available at: < https://towardsdatascience.com/deep-generative-models-25ab2821afd3> [Accessed 1 May 2023].

[12] M. Stewart, 2019. *GANs vs. Autoencoders: Comparison of Deep Generative Models*. [online] towardsdatascience.com. Available at: <https://towardsdatascience.com/gans-vs-autoencoders-comparison-of-deep-generative-models-985cf15936ea > [Accessed 1 May 2023].

[13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, “Generative Adversarial Networks,” in arXiv (Cornell University), Jun. 2014, doi: https://doi.org/10.48550/arxiv.1406.2661

[14] I. Ashrapov, “Tabular GANs for uneven distribution,” in arXiv (Cornell University), Oct. 2020, doi: https://doi.org/10.48550/arXiv.2010.00638

[15] P. Ezhilarasi, L. Ramesh, X. Liu and J. B. Holm-Nielsen, “Smart Meter Synthetic Data Generator development in python using FBProphet,” in Software Impacts, vol. 15, p. 100468, Mar. 2023, doi: https://doi.org/10.1016/j.simpa.2023.100468

[16] A. Pinceti, L. Sankar, and O. Kosut, “Generation of Synthetic Multi-Resolution Time Series Load Data,” in arXiv (Cornell University), July 2022. Available: https://doi.org/10.48550/arXiv.2107.03547

[17] D. Hazra, W. Shafqat and Y. Byun, "Generating synthetic data to reduce prediction error of energy consumption," Computers, Materials & Continua, vol. 70, no.2, pp. 3151–3167, 2022, doi: https://doi.org/10.32604/cmc.2022.020143

[18] J. Moon, S. Jung, S. Park and E. Hwang, "Conditional Tabular GAN-Based Two-Stage Data Generation Scheme for Short-Term Load Forecasting," in *IEEE Access*, vol. 8, pp. 205327-205339, 2020, doi: 10.1109/ACCESS.2020.3037063.

[19] “SmartMeter Energy Consumption Data in London Households – London Datastore.” https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households

[20] NASA, “ArcGIS Web Application,” Nasa.gov, 2018. https://power.larc.nasa.gov/data-access-viewer/

[21] “United Kingdom Bank Holidays,” www.ukbankholidays.co.uk. https://www.ukbankholidays.co.uk/ (accessed March 1, 2023).