# Applying Deep Generative Methods to Generate Synthetic Data in Power Systems

Mohsen Kariman majd[1*] and Mohsen Niasati[1]

***Abstract*-- The lack of access to reliable databases, as well as the small number and imbalance of databases, is one of the main limitations of using machine learning methods in power systems, which can reduce efficiency and cause distrust in the results obtained from these methods. One of the solutions used to solve this problem is the use of Synthetic data generation. Two deep generative architectures, Generative Adversarial Network (GAN) and Variational Auto Encoder (VAE), are currently used to generate synthetic data. Due to the novelty and importance of the subject, until now, a comparative study has not been done on the research conducted in this field, in terms of subject classification, with an emphasis on validation methods of synthetic production databases. The purpose of this research is to review the studies done in this field up to now and examine the research trends for the future. In this regard, after introducing the principles of GAN and VAE deep architectures, the subject of synthetic data generation using the mentioned methods in power systems has been studied comparatively.**

***Index Terms*-- Synthetic data; Deep learning; Generative Adversarial Network; Variational Auto Encoder; Power systems.**

## I. INTRODUCTION

With the new approach in power systems and the transition to a smart grid, the use of artificial intelligence and its latest achievement, deep learning, in solving power system problems, has become more important. Among these cases, we can mention the prediction of random renewable production, the prediction of the amount and type of consumption in the smart distribution network based on the concept of smart homes, the evaluation of the condition and health measurement system of equipment based on continuous monitoring, etc. [1]. Due to the implementation of deep learning methods on the mentioned databases, problems arise in practice. These problems are divided into two main parts: Small databases and unbalanced databases. A small database will increase the risk of overfitting. In power systems, due to matters related to the rules of keeping confidential information and non-disclosure policies, the time-consuming or costly nature of data collection, or technical matters such as the low number of faults compared to normal work conditions and the lack of accurate documentation of abnormal situations. Performance, there is a problem of lack of data in public databases. According to the mentioned cases, the use of synthetic data generation methods is a necessary prerequisite for applying machine learning methods in general and especially deep learning in power systems databases. An unbalanced database is a database whose labeled data is unequally distributed among classes. Databases in power systems are often unbalanced in practice, and this problem will hurt the classification accuracy of deep learning-based methods so that the problem of overfitting and reducing the generalization of the obtained model is threatened If it is possible to add a set with a sufficient number of similar synthetic data to the primary database, both the problems of the small number of the database and the unbalanced distribution of data in the respective classes can be solved simultaneously.

The problems in the power systems that are solved by using deep generative methods are as follows:

- Rules related to the non-disclosure of specific data in public

- Lack of labeled data and sufficient numbers due to difficulty or time-consuming process

- The problem of imbalanced databases

Although a comprehensive study has been done in [2] in the field of classification of research conducted in the production of synthetic electrical networks in power systems, the said research is mostly based on network design based on graph theory, in line with systemic studies, and according to the authors, So far, there has not been a study in the field of reviewing the researches done in the field of synthetic databases in the power systems with thematic classification approach and stating the methods and limitations with an emphasis on identifying the validation methods of synthetic databases. With the introduction of smart grids and the use of new technologies for measuring system parameters such as electric current, voltage, and load databases have been provided for efficiency and analysis of system performance. The use of synthetic data is a solution to solve the problems of these databases for the optimal use of various methods of machine learning, especially deep learning in solving system problems. The purpose of the present research is to review and categorize the studies conducted on this emerging issue along with suggestions for further research in the future. The innovation of this article can be categorized as follows:

1- A comparative study on the research conducted using deep generative methods for synthetic data generation in power systems.

2- Subject classification emphasizes technical and statistical validation methods of synthetic databases.

1. Electrical and Computer Engineering Faculty, Semnan University, Semnan, Iran.

* Corresponding author Email:m_karimanmajd@semnan.ac.ir

## II. Principles of GAN

The GAN network is a method to create a set of synthetic data that is similar to the real data set in terms of statistical distribution. Fig.1 shows the working principles of the GAN network. The GAN network training algorithm is such that first the D network is trained by the real data set to solve a binary classification problem. On the other hand, block G is stimulated by a random input and creates a random output. The random output is fed to block D. The purpose of block G is to produce data that are so similar to real data that block D classifies them in the real set group. With an error backpropagation pattern, the weights of the G network are updated in this direction. According to the progress of the training of the D function in each step, the G function produces more accurate data in each step to mislead the D block. This competition goes so far that the data produced by block G will be very similar to the real data, and with the victory of block G over block D, the goal of GAN architecture is achieved. The term adversarial in GAN architecture refers to this process of competition. And finally, the victory of block G [3].

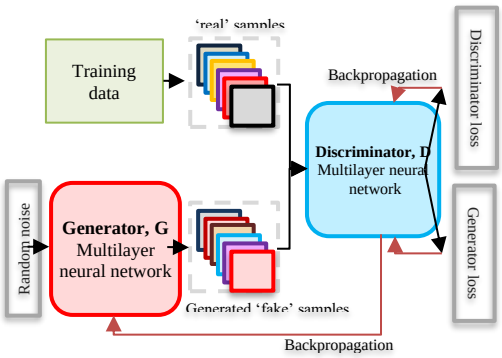


Fig. 1. Principles of GAN network [4]

Various types of GAN networks have been introduced during the research conducted to improve the performance of this architecture. The difference between some of these proposed models is in the architectural building used in the building blocks of the network. Deep Convolutional GAN (DCGAN), Conditional GAN (CGAN), and Semi-Supervised GAN (SSGAN) can be mentioned among these cases.

DCGAN uses deep convolutional architecture in the discriminator and generator part [5]. In CGAN, the input of the generating block is no longer a simple noise vector, and additional information is applied to it according to the nature of the problem for which the synthetic data is created [6]. Although GAN is classified under unsupervised machine learning, by converting the discriminator block into a multi-class classification, the learning type is changed to semi-supervised. SSGAN architecture is used to generate synthetic data in multi-class databases [7]. Some of the research has focused on improving the methods of optimizing the training process in GAN networks. One of the most important models proposed in this field is Wasserstein GAN (WGAN). In the WGAN algorithm, the Wasserstein distance criterion is used to measure the difference between the real data and the data generated by the network. This criterion is used in statistics to compare the difference between two different data distributions. Improving the gradient vanishing problem, which is one of the basic problems of deep learning, is one of the other achievements of the WGAN algorithm [8-9].

## III. Principles of VAE

The Variational Auto Encoder is one of the types of auto-encoder neural networks that can learn the statistical parameters of the domains in the database. In the VAE algorithm, in the middle layer, the data is developed as a normal distribution with a certain mean and variance as a statistical distribution, and instead of acting on the deterministic vector located in the middle layer, the decoder block operates on a Stochastic vector generated based on the normal distribution. As a result, the synthetic data, which is the purpose of using VAE, is produced in the output of the decoder block. Fig. 2 shows the principles of the VAE network [10].

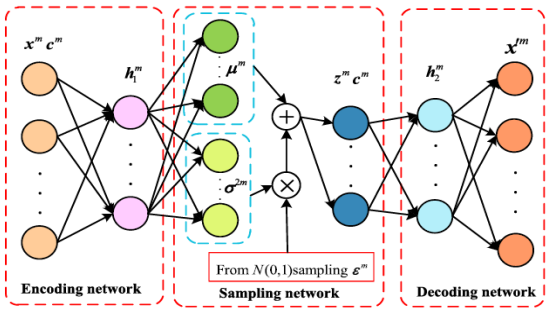


Fig. 2. Principles of VAE network [10]

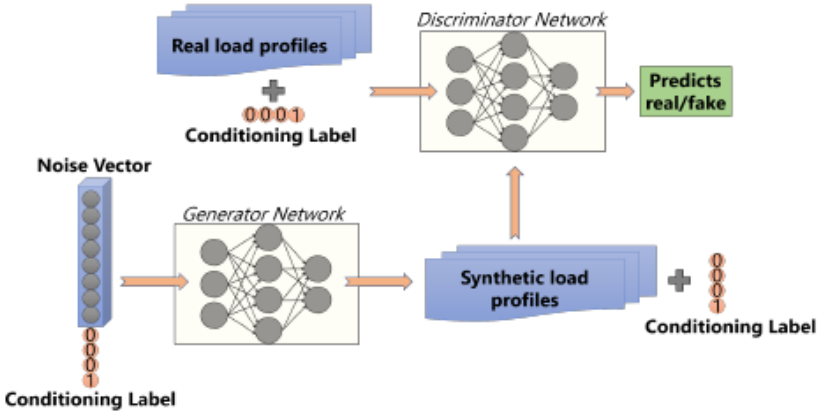


Fig. 3. The process of generating synthetic load profile data in the transmission section of the power systems [11]

## IV. DATABASES IN THE GENERATION, TRANSMISSION AND DISTRIBUTION SYSTEMS

In power systems, databases are formed in the generation, transmission and distribution sectors. Among these cases, we can mention voltage and current data recorded in phasor measurement units (PMU) and electric load information recorded by smart meters. The nature of the load in these studies is modeled as a time series. Using deep models to predict time series requires large amounts of recorded data. These data are recorded in the form of voltage and current in the transmission section by phasor measurement units. Deep generative methods can be used in the Data augmentation stage.

In [11], a CGAN model is proposed to generate synthetic load profile data in the transmission sector of power systems. Fig. 3 shows the process of running the algorithm. This research aims to produce a continuous load profile according to the season and type of load, and for effective forecasting, in addition to real raw data, additional information is needed such as season, type of load, and consumption pattern. This additional information requires that instead of using the simple GAN algorithm, the CGAN model is used to apply the mentioned conditions to the input of the network.

The Energy Density Spectrum (EDS) method has been used as a measurement criterion for two real and synthetic databases. This tool is used to measure random fluctuating components in time series.

In [12], another study has been done on creating a synthetic database using real data obtained from PMU. PMU's actual data collection is sometimes restricted by non-disclosure laws. The authors of the article created a synthetic database of voltage and current values of PMUs by using the GAN network to classify system events and improve the accuracy of the production database in the field of discovering system events such as load shedding, transmission line interruption, and busbar faults. The research has aimed to classify the events in the transmission sector more precisely. The modal analysis method has been used to check the efficiency of the generated database. Power system state variables in this analysis are the criteria for comparing two databases.

Another part of the research, in the field of synthetic load profile generation, is dedicated to the issue of non-technical losses in the distribution sector or electricity theft. Today, with the expansion of the use of cryptocurrency mining tools, the power grid has faced a serious threat, especially in countries that provide cheap electricity to consumers. In [13], variable autoencoder architecture is used to discover the electricity theft pattern. Considering that the definite cases detected in the field of discovering electricity theft patterns in the distribution system are very few, the need to use generative methods to generate synthetic electricity theft data is important. The model used for this purpose is the CVAE architecture. Load profile changes are applied to the input of a CVAE model in the few cases of power theft, and by implementing the variable auto encoder random mechanism, synthetic power theft load profiles are obtained at the output. The attack patterns have been used as a measurement criterion for the production database to more accurately discover the phenomenon of electricity theft. In addition, two original and synthetic databases have been compared with the Kullback-Leibler statistical distribution variance measurement parameter. In the science of statistics, the KL divergence criterion is mentioned as an index to measure the degree of divergence of a probability distribution from a secondary probability distribution. Values close to zero in the KL criterion indicate that we can expect similar behavior (not the same) from two distributions. The relation (1) describes the method of obtaining the value of this criterion [13]:

$$D_{KL}(p||q) = \sum_{i=1}^{N} p(x_i) \log(\frac{p(x_i)}{q(x_i)}) \quad (1)$$

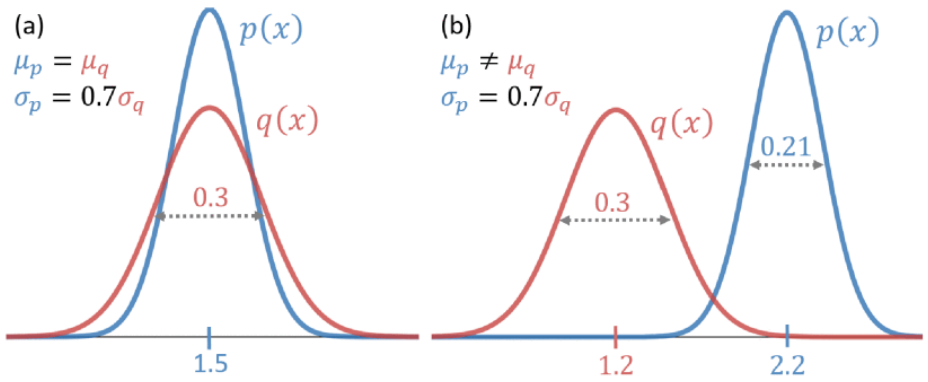


Fig. 4. Kullback-Leibler criterion to measure the degree of similarity of two sample statistical distributions a) KL=0.1, b) KL=5.7 [2]

In relation (1), $p(x_i)$ is the statistical distribution function of the real database, and $q(x_i)$ is the statistical distribution function of the synthetic database. Typically, p(x) represents the "true" distribution of data, observations, or a precisely calculated theoretical distribution. The measure q(x) typically represents a theory, model, description, or approximation of p(x). $D_{KL}(p||q)$ is Kullback–Leibler Divergence, a measure of the difference or relative entropy between two probability density functions (pdfs). It quantifies the information lost or inefficiency when one PDF is used to approximate another. N is the total number of compared data in related

PDFs.. relation 1 is rewritten in continuous form as follows:

$$D_{KL(p||q)} = \int_{-\infty}^{\infty} p(x) log\left(\frac{p(x)}{q(x)}\right) dx \quad (2)$$

By placing the two normal distribution functions shown in Fig. 4 in the resulting integral relation and considering the ratio between the medians and the deviation from the criteria of the two normal distribution functions, the values related to the Kullback-Leibler criterion are obtained.

In Fig. 4, the Kullback-Leibler criterion is used to measure the similarity of two statistical distributions. The KL value in part (a) is equal to 0.1 and as can be seen in the figure, the two distributions are very similar to each other. In part (b) of Fig. 4, the divergence value is equal to 5.7 and the deviation from the uniformity of the two distributions is evident.

A hybrid model of GAN and VAE is proposed in [14] to generate a synthetic load profile in the field of smart grid, in the field of smart home. Fig. 5 shows the arrangement of generator blocks in the mentioned model. Photovoltaic unit generation and smart home demand have been modeled and the distribution of synthetic generation data with original data has been researched by statistical methods.

In [15], the generation of synthetic data to predict energy production in photovoltaic power plants has been independently researched.

In [16], about one of the features of a smart home called non-invasive load consumption monitoring, WGAN architecture has been used to produce an artificial characteristic sign for different types of loads to distinguish the type and amount of load consumption in an artificial intelligence classification system. To confirm the correct diagnosis of the size and type of loads on the demand side, the analysis of the load characteristic sign has been used.

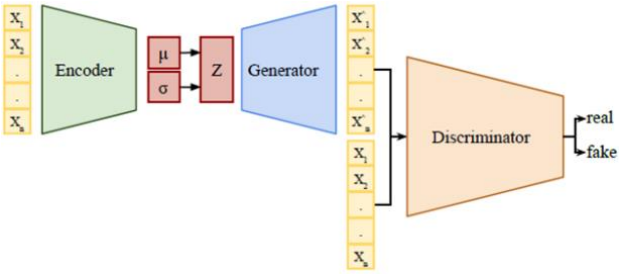


Fig. 5. VAE-GAN hybrid model to generate synthetic data of generation and consumption in the smart home [14]

## V. Databases in the condition assessment and internal fault diagnosis

Intelligent fault diagnosis is one of the interesting and up-to-date research trends in the field of power systems asset management. The purpose of this issue is early and accurate diagnosis of faults that are forming or have arisen in the equipment to evaluate the optimal condition and perform preventive maintenance. The information on the internal system of the desired equipment is sampled by measuring parameters such as vibration and sound signals and components soluble in oil. The sampled cases that lead to faults usually have a much smaller number than the total number of samples, and the obtained database is highly unbalanced. In addition, due to the lack of accurate documentation of abnormal performance modes, there is a problem of lack of sufficient data in public databases. In the following, the research was conducted in the direction of using synthetic data to solve the mentioned problems.

In [17] the problem of the lack of fault data in wind turbines has been solved by introducing a three-step algorithm. The first part of this algorithm uses an expert system as a database for fault analysis. In the second step, the data obtained from the first step is increased by a GAN network in a data augmentation process, and in the last step, the database obtained by an artificial intelligence model is used to detect various types of internal faults in wind turbines. Researchers in [18] studied the problem of imbalance between fault data and normal working mode data in industrial pumps used in power plants. The proposed method to solve the problem is a combination model of GAN and Stacked Auto Encoder (SAE) network. The GAN network is responsible for generating synthetic data and the SAE network is responsible for feature extraction in this model. The same problem has been researched in [19] for industrial chillers and by combining two architectures, WGAN and VAE, synthetic data necessary to solve the imbalance problem in the database has been provided. In [20], synthetic data generation is used to improve the battery condition assessment system. With the expansion of the use of electric vehicles and the need to manage the driving time of the car after each charge, monitoring the health of the batteries used in these cars is of great importance. Collecting real data related to the state of the charge of batteries is time-consuming and expensive. The databases of the working companies are also subject to non-publication restrictions. Therefore, generating synthetic data is a practical and available solution. The synthetic data produced with the standard databases have been statistically compared and the similarity in the statistical distribution of the two synthetic and real databases has been confirmed. In [21], for one of the most important issues in the field of asset management of equipment, determining the remaining useful life, the conventional method of synthetic data generation with GAN architecture has been used. The application of the proposed algorithm in predicting the remaining useful life of supercapacitors has been implemented. In Table I, the aforementioned studies have been summarized for comparison.

## VI. . Conclusions

The use of deep generative methods in power systems is very new, and in this research, a systematic review with the separation of various specialized areas, along with the comparison of different validation methods was attempted. The most important reasons for using synthetic database production in power systems are the rules for maintaining confidential information and data non-disclosure policies, the time-consuming or costly nature of data collection, technical issues such as the low number of

faults compared to normal conditions, lack of Accurate documentation of abnormal performance modes and unbalanced databases. In the concept of synthetic data generation in power systems and the comparison of deep generator methods with traditional statistical methods, the most important challenge facing traditional methods is the low accuracy of synthetic database generation compared to the original database. This is especially important in big databases.

In big data, traditional methods suffer from overfitting problems in synthetic data generation. In deep generative methods, unlike traditional methods, the more original data there is, the more the possibility of overfitting in the production of synthetic data is reduced, and the generalizability of the model is increased.

The validation of the generated database is very important. Research that succeeds in proving the equivalence of two databases using both statistical measurement and system performance measurement methods is considered more successful. A new technical suggestion with the course of studies in this field is the use of deep generative methods to build a source database in the new subject of transfer learning. In transfer learning, the model is trained on a large initial database that is sufficiently similar to the target database and then fine-tuned on a small target database. According to the statistical and systematic evaluation methods, the synthetic database that is the subject of the present research is a database similar to the real database, and due to its artificial nature, it can be made large enough. Therefore, it can be used as a source for Transfer learning methods.

TABLE I

Research Conducted in the Field of Using Deep Generative Methods to Produce Synthetic Databases in Power Systems

| Specimens code | Reference article | Database | Architecture deep generator | System validation | Statistical validation |
|---|---|---|---|---|---|
| Generation, transmission and Consumption | Pinceti [11] | PMU | CGAN | Power Spectral Density | ------ |
| | Zheng [12] | PMU | GAN | Modal analysis | ------ |
| | Gong [13] | Smart meter | CVAE | Attack models | KL |
| | Razghandi [14] | Smart meter PV | VAE-GAN | ------ | KL MMD WD |
| | Rosa [11] | PV | VAE-GAN | ------ | MMD |
| Evaluating the condition of equipment and diagnosing internal faults | Harell [16] | Smart meter | WGAN | Power signature analysis | KL WD |
| | Liu [17] | Scada | GAN | Machine learning classification | CSP |
| | Han [18] | Sensors | GAN-SAE | Deep learning classification | ------ |
| | Yan [19] | Sensors | GAN-VAE | Deep learning classification | ------ |
| | Naaz [20] | Battery parameters | GAN | ------ | KL |
| | Udurume [21] | Capacitor parameters | GAN | Deep learning prediction | ------ |
| | Pinceti [11] | PMU | CGAN | Power Spectral Density | ------ |

## REFERENCES

[1] Mohsen Saffari, Mahdi Khodayar, "Spatiotemporal Deep Learning for Power System Applications: A Survey", IEEE Access, vol.12, pp.93623-93657, 2024.

[2] M. H. Mohammadi and K. Saleh, "Synthetic benchmarks for power systems", IEEE Access, vol. 9, pp. 162706-162730, 2021.

[3] Yang Zeng, Bolin Liao, Zhan Li, Cheng Hua, Shuai Li, "A Comprehensive Review of Recent Advances on Intelligence Algorithms and Information Engineering Applications", IEEE Access, vol.12, pp.135886-135912, 2024.

[4] C. Little, et al. "Generative Adversarial Networks for Synthetic Data Generation: A Comparative Study." arXiv preprint arXiv:2112.01925, 2021.

[5] S. Zeng, Y. Cai, R. Zhang, and X. Lyu, "Research on Human-Machine Collaborative Aesthetic Decision-Making and Evaluation Methods in Automotive Body Design: Based on DCGAN and ANN Models," in IEEE Access, vol. 12, pp. 91575-91589, 2024, doi: 10.1109/ACCESS.2024.

[6] K. Kong, K. Kim, and S. -J. Kang, "Enhancing Stability in Training Conditional Generative Adversarial Networks via Selective Data Matching," in IEEE Access, vol. 12, pp. 119647-119659, 2024.

[7] C. Xu, T. Zhang, D. Zhang, D. Zhang, and J. Han, "Deep Generative Adversarial Reinforcement Learning for Semi-Supervised Segmentation of Low-Contrast and Small Objects in Medical Images," in IEEE Transactions on Medical Imaging, vol. 43, no. 9, pp. 3072-3084, Sept. 2024.

[8] X. Zhang, Z. Zhao, R. Shao, C. Li, and H. Tang, "Mechanical Anomaly Detection and Early Warning for Ultrahigh-Voltage Shunt Reactors via Adaptive Thresholds and WGAN-GP," in IEEE Sensors Journal, vol. 24, no. 12, pp. 20219-20230, 15 June 15, 2024.

[9] Q. Zhang, X. Wang, and C. Li, "SA-WGAN-Based Optimization Method for Network Traffic Feature Camouflage," in IEEE Access, vol. 12, pp. 111142-111157, 2024.

[10] Y. Wang, G. Sun and Q. Jin, "Imbalanced sample fault diagnosis of rotating machinery using conditional variational auto-encoder generative adversarial network", Applied Soft Computing, vol. 92, p. 106333, 2020.

[11] A. Pinceti, L. Sankar, and O. Kosut, "Synthetic Time-Series Load Data via Conditional Generative Adversarial Networks," 2021 IEEE Power & Energy Society General Meeting (PESGM), Jul. 2021.

[12] X. Zheng, B. Wang, D. Kalathil, and L. Xie, "Generative Adversarial Networks-Based Synthetic PMU Data Creation for Improved Event Classification," IEEE Open Access Journal of Power and Energy, vol. 8, pp. 68–76, 2021.

[13] X. Gong, B. Tang, R. Zhu, W. Liao, and L. Song, "Data Augmentation for Electricity Theft Detection Using Conditional Variational Auto-Encoder", Energies, vol. 13, no. 17, p. 4291, 2020.

[14] M. Razghandi, et al. "Variational Autoencoder Generative Adversarial Network for Synthetic Data Generation in Smart Home." arXiv preprint arXiv:2201.07387,2022.

[15] M. Razghandi, H. Zhou, M. Erol-Kantarci and D. Turgut, "Smart Home Energy Management: VAE-GAN Synthetic Dataset Generator and Q-Learning," in IEEE Transactions on Smart Grid, vol. 15, no. 2, pp. 1562-1573, March 2024.

[16] A. Harell, R. Jones, S. Makonin and I. Bajic, "TraceGAN: Synthesizing Appliance Power Signatures Using Generative Adversarial Networks", IEEE Transactions on Smart Grid, vol. 12, no. 5, pp. 4553-4563, 2021.

[17] J. Liu, F. Qu, X. Hong, and H. Zhang, "A Small-Sample Wind Turbine Fault Detection Method with Synthetic Fault Data Using Generative Adversarial Nets", IEEE Transactions on Industrial Informatics, vol. 15, no. 7, pp. 3877-3888, 2019.

[18] H. Han, L. Hao, D. Cheng, and H. Xu, "GAN-SAE based fault diagnosis method for electrically driven feed pumps", PLoS ONE, vol. 15, no. 10, Oct. 2020.

[19] K. Yan, J. Su, J. Huang, and Y. Mo, "Chiller fault diagnosis based on VAE-enabled generative adversarial networks", IEEE Transactions on Automation Science and Engineering: A Publication of the IEEE Robotics and Automation Society, pp. 1-9, 2020.

[20] F. Naaz, A. Herle, J. Channegowda, A. Raj and M. Lakshminarayanan, "A generative adversarial network-based synthetic data augmentation technique for battery condition evaluation", International Journal of Energy Research, vol. 45, no. 13, pp. 19120-19135, 2021.

[21] M. Udurume, C. Udeogu, AngelaC. Caliwag, and W. Lim, "Synthetic Data Generation Using GAN for RUL Prediction of Supercapacitors", The Journal of Korean Institute of Communications and Information Sciences, vol. 47, no. 3, pp. 492–500, Mar. 2022.