Generation of Synthetic 5G Network Dataset Using Generative Adversarial Network (GAN)

Muhammad Nur Aqmal Khatiman
Faculty of Engineering and Built Environment
Universiti Kebangsaan Malaysia
Bangi, Malaysia
aqmalkhatiman@gmail.com

Asma Abu-Samah
Wireless Research@UKM, Faculty of Engineering and Built Environment
Universiti Kebangsaan Malaysia
Bangi, Malaysia
asma@ukm.edu.my

Muhammad Amin Azman
Faculty of Engineering
Universiti Putra Malaysia
Serdang, Malaysia
amin.azman@upm.edu.my

Rosdiadee Nordin
Wireless Research@UKM, Faculty of Engineering and Built Environment
Universiti Kebangsaan Malaysia
Bangi, Malaysia
adee@ukm.edu.my

Nor Fadzilah Abdullah
Wireless Research@UKM, Faculty of Engineering and Built Environment
Universiti Kebangsaan Malaysia
Bangi, Malaysia
fadzilah.abdullah@ukm.edu.my

2023 IEEE 16th Malaysia International Conference on Communication (MICC) | 979-8-3503-0434-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/MICC59384.2023.10419563

***Abstract*—While the Fifth Generation (5G) network is actively being deployed in most countries to create new possibilities for better lifestyle and economic development, it is a technology that is currently being a focal point for researchers across the world along with 6G. Starting from 3GPP Release-18, Artificial Intelligent (AI) and Machine Learning (ML) are identified as enabler towards intelligent network in 5G and beyond. Nevertheless, the models based on AI/ML need a sufficient amount of data for learning patterns and relationships, enabling them to provide precise predictions for unfamiliar data and situations. The existence of Generative Adversarial Network (GAN) helps solve the issue by generating fake data from an existing dataset to resemble real-world data to be used in training and testing of different algorithms. In this paper, the process of generating synthetic data of 5G network was demonstrated from an extensive test drive results that will encourage innovation in mobile communication. Generation of data use two types of GAN which are the Conditional Tabular GAN (CTGAN) and Topological Variational Autoencoder (TVAE). The two algorithms were compared based on statistical analysis such as the distribution and Pearson Correlation analysis. TVAE showed a better overall performance score (94.14%) over CTGAN (89.66%) when compared with the original data, but the CTGAN produced more similar distribution for certain individual columns.**

Keywords— synthetic data generation, tabular data synthesis, 5G network, Generative Adversarial Network (GAN), Conditional Tabular Generative Adversarial Network (CTGAN)

I. INTRODUCTION

In the 3GPP standards that outline the technical aspects of mobile communication, machine learning (ML) and artificial intelligence (AI) have been increasingly integrated to enhance network performance, security, and user experience. ML is introduced for network optimization, traffic prediction, quality of service improvement, user experience enhancement and energy efficiency. Since its introduction as preliminary initiatives in the Release-15 to support 5G evolution, it is now proposed to facilitate the ubiquitous adoption of 5G-advance evolution [1].

However, to migrate from the use of conventional models to ML's, developers are often challenged and complicated not only by data privacy, but by the accuracy and efficacy of its modelling by the scarcity of real-world datasets that represents the various context of the users using the communication networks [2]. As machine learning is also an evolving field, many researchers and computer scientists relentlessly innovate the usage of synthetic data generation to help augment their dataset. Amongst the approaches, Generative Adversarial Network (GAN) deliver one of the highest performance solutions [3].

With the different types of data that exist nowadays, generating specific types of data can only be achieved using specific GAN. There are several variations of GAN that are available for use. Some of them are Conditional GAN [4], Progressive GAN [5], and Cycle GAN [6]. Most of these implementations however were demonstrated for image data generation and not of tabular or 2D data. Most people believe that synthesizing tabular data is easier than synthesizing other types of data such as images or texts as computers can crunch formatted numbers straightforwardly, but that is not the case in GAN application as the first development of GAN was to generate fake images. [7] explored various approaches to generate synthetic data, with a specific focus on tabular data and conclude that GAN still have lots of room for improvement in this direction.

The adoption of synthetic data in the telecommunications industry are more relevant to tabular data and depends on the specific use cases, data privacy concerns, and regulatory requirements which can vary from different services and providers. This paper aims to show the possibility of GAN to synthesize 5G mobility management data from an opensource dataset using two algorithms namely CTGAN and TVAE. Section 2 elaborates the related works of GAN in telecommunication while Section 3 focuses on the methodology to generate data using the chosen algorithms and comparison of their performance. Section 4 shares the results and discussion, and finally Section 5 concludes the paper.

II. RELATED WORKS ON GAN IN TELECOMMUNICATIONS

A. Generative Adversarial Networks

As the field of synthetic data generation continues to evolve and improve, several GAN derived models were introduced recently to adapt the technique for tabular data. it is likely that telecommunications companies will increasingly use synthetic data to address data-related challenges while safeguarding privacy and security.

GAN has been a focus among researchers as they continue finding improved GAN techniques and new uses for GANs.

Following are several novel creations of GAN for specific applications. Kim et al., (2021) proposed a new variation of GAN based on Neural Ordinary Differential Equation (NODE) for tabular data synthesis [8]. In NODE, a neural network learns a system of ordinary differential equations to approximate the change of hidden vector at a time. Compared with other generative models through baseline method to synthesize tabular data, it is proven that the proposed OCT-GAN has the best performance in the classification, regression, and clustering experiments.

Other than that, [9] has suggested a new framework of GAN usage in synthesizing dataset that that has mixed data type variables, long tail distribution and skewed data but still maintain the statistical similarity. The model is tested on 5 different datasets and compared with several tabular data generation models (i.e., CTGAN, TableGAN, CWGAN, and MedGAN). From its evaluation, the proposed type of GAN outperforms all other models in terms of statistical similarity along with becoming the most reliable model as it provides the highest distance-based privacy guarantee compared to other models.

B. *Application of GAN in Real-World*

GAN has gained popularity in solving problems pertaining to limited public data and bridging the gap between building intelligent solutions using confidential data. Moon et al., 2020 demonstrate the usage of Conditional Tabular GAN (CTGAN) to generate synthetic data of electrical power usage for predicting load in various components [10]. It was found that forecasting models that are trained using synthetic data generated from Conditional Tabular Generative Adversarial Networks (CTGAN) showed the best performance compared to other models. It is believed that CTGAN has a superior ability to learn the overall distribution of the real data and effectively generate data following the distribution.

Similarly, Ullah & Mahmoud, 2021 in [11] believed that GAN implementation should be included in the framework for anomaly detection in an IoT network. Available anomaly detection models trained on datasets that are likely imbalanced have poor results. Different from usual implementation, this study uses Conditional GAN (CGAN) to augment certain variables to before feeding the dataset (with additional augmented column) for training a detector model. As a result, modelling a detection model using balanced dataset is more accurate compared to using a non-augmented dataset.

In much the same way, GAN is used in generating fake financial data. Cote et al., 2020 has compared performance of three types of GAN in generating synthetic insurance claims data [12]. This research demonstrates generating synthetic data using insurance-related dataset, French motor third-party liability policies that is available publicly. The dataset is synthesized using three different models which are MC-WGAN-GP, CTGAN and MNCDP-GAN. The comparison showed that the MC-WGAN-GP has the best performance followed by CTGAN and the worst performance is MC-WGAN.

Since its functionality is to generate synthetic data, engineers have benefited from GAN functionality that speeds up creating design according to specific conditions. Yılmaz & German, 2020 observe the ability of Conditional GAN (CGAN) to generate airfoil design given several performance specifications [13]. This study is a sample of an image-to-image translation application using GAN. The generated airfoil designs using GAN are compared with the training set that is taken from the database of aerodynamic analysis software, XFOIL and the result proved that both data show similar behavior.

III. Methodology

A. *GAN Philosophy*

GAN composes of two deep networks, the generator and the discriminator. Both networks are trained simultaneously, and they are adversarial in nature. In other words, each is designed to achieve the opposite goal of the other. The generator is a network that generates new data, while the discriminator is a network that tries its best to differentiate the fake and the real data. One of them is trying to fool the other, and the other is trying to avoid being fooled. The generator takes random variables from latent space and produces fake data samples. The discriminator takes real data from the training sample and fake data from the generated sample and classifies which one is the real data and which one is fake [14]. Fig. 1 shows the overview of the simplest GAN structure, sometimes called Vanilla GAN.

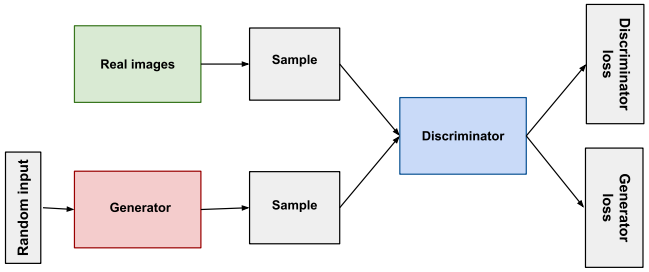

Fig. 1. Structure of GAN [15]

From Fig. 1, the discriminator input is connected directly to the output of the generator. The discriminator penalizes the generator for producing fake data. The generator will use the feedback from the discriminator to update its weight to generate fake data that is more similar to real data. As training progresses, the generator gets closer to generating data that can fool the discriminator. Eventually, after a sufficient time of training, the generator can generate data that mimics the real data, and the discriminator is no longer able to classify between the real data and the fake data. At that time, the discriminator's accuracy decreases as it starts to classify fake data as real data. Here, the Generator and the Discriminator are simple multi-layer perceptron. In vanilla GAN, the simplest form of GAN, the algorithm is really simple, it tries to optimize the mathematical equation using stochastic gradient descent.

CTGAN's is a derived architecture based on the architecture of Conditional GAN (CGAN) with T stands for Tabular. As the name suggests, CGAN takes a conditional vector as an input when training the generator and the discriminator. In other words, CGAN is trained to generate data following the specified conditions. The generator and the discriminator have more information about the ideal output they should produce. CGAN has two advantages to offer. Firstly, the training will not take longer than vanilla GAN as it will reach a convergence state faster. Secondly, the generator's output can be determined during test time by giving the label of data that wants to be generated. CGAN is the basic architecture of any conditional-based GAN and CTGAN is the modified version that is implemented to

generate tabular format data. Figure 2 shows the architecture of CGAN model.

To generate synthetic data of tabular format, it is recommended to use Conditional Tabular GAN (CTGAN) or Topological Variational Autoencoder (TVAE) because they are tailored to the specific characteristics and challenges of tabular data. CTGAN focuses on preserving relationships and constraints, while TVAE captures the underlying topology, both of which are crucial for generating high-quality synthetic tabular data. The choice between these models might depend on the specific requirements of the application and the desired characteristics of the generated data. TVAE is a type of deep learning algorithm that combines concepts from topology and variational autoencoders to learn a compressed representation of complex high-dimensional data. TVAE aims to capture the topological structure of the data in a low-dimensional space, allowing for more efficient and accurate analysis, clustering, and visualization of the data.

This study chose to demonstrate generating synthetic data of tabular 5G Network dataset using CTGAN as it is said to have the best performance for synthesizing data. We will later compare the performance of using CTGAN with TVAE. The goal of TVAE is to capture the topological structure of the data in a low-dimensional space, allowing for more efficient and accurate analysis, clustering, and visualization of the data.

B. Evaluation Metrics

In this study, two state-of-the-art synthesizers, CTGAN and TVAE were employed and using the original 5G network performance dataset. These synthesizers were implemented using the Synthetic Data Vault (SDV) module, an extensive library of synthetic data generation techniques available freely using Phyton programming. The module creates datasets by maintaining their statistical characteristics of the original data. To evaluate the quality and accuracy of the generated synthetic data, statistical tests and correlation analysis were employed. These metrics are suitable for tabular data, and they assess the degree to which the synthetic datasets preserve the statistical properties of the original dataset such as the distribution of columns, and the overall structure.

By comparing the performance of the CTGAN and TVAE synthesizers, this study aims to identify the most effective method for generating high-quality synthetic data that accurately represents the nuances of real-world 5G mobile management data while maintaining the privacy of the original data. This evaluation will provide valuable insights for researchers and industry professionals seeking to use synthetic data in their analyses and decision-making processes.

C. Dataset

The dataset used in this study offers a comprehensive collection of 5G network performance metrics, extracted from Hassan's et al. (2022) research work [16]. The dataset mainly consisted of over 600GB collected logs that consist of more than 47,000 handovers and that span multiple dimensions, (i) 3 anonymous carriers/operators denoted as OpX, Y and Z, (ii) radio technologies (5G vs. 4G), (iii) 5G architectures (NSA vs. SA), and (4) 5G bands–low-band, mid-band, mmWave (high-band). The dataset was generated from a 6,200 km drive-tests across major cities and interstate freeways in the U.S.

The mobile applications on test were real-time volumetric video streaming, cloud gaming and zoom live video conferencing. This study employs data from Operator X with Mid-band LTE, Low-band NSA and mmWave NSA. However, the frequency range for each band were not specified in their work. The data structure with 10 variables is given in Table 1.

TABLE I. DATA STRUCTURE OF THE DATASET FOR 1 OPERATOR

latitude	**longitude**	**handoffType**	**run_number**	**Carrier**
Coordinates	Coordinates	[0…12]	[1…9]	OpX
sw_power	**pci**	**nr_pci**	**Location**	**network**
[min,max] = [932.34, 11053.392]	[49,50, 210, 216, 222,401]	[49,50, 206, 210, 216, 222, 401]	Home, MPA, MRCHMS, MRCHMSS	NSA, LTE, mmWAVE

IV. RESULTS AND DISCUSSION

The previous sections have detailed the methods for generating synthetic 5G network datasets using the CTGAN algorithm. In this section, the results of the study is presented, which aims to evaluate the effectiveness of the CTGAN algorithm in generating synthetic datasets that accurately reflect the characteristics of real-world 5G networks. Specifically, the extent to which the synthetic datasets generated by CTGAN replicate the statistical properties of the original dataset was assessed. It provides a sense of the distribution of network traffic volume in the synthetic 5G network dataset generated using CTGAN. A series of experiments was conducted to evaluate the performance of network algorithms and protocols when trained on synthetic datasets. The results provide valuable insights into the potential uses of synthetic datasets for testing and evaluating network technologies and contribute to the growing body of research on the use of GANs for generating synthetic data.

The mean and standard deviation of network traffic volume were 70.79 and 27.51, respectively, indicating that the traffic volume was moderately variable across the dataset. The minimum and maximum values were 20.14 and 159.28, respectively, indicating a wide range of traffic volumes. The distribution of traffic volume was approximately normal, with a skewness of 0.18 and a kurtosis of -0.73. Overall, these descriptive statistics suggest that the synthetic dataset generated using CTGAN provides a reasonable approximation of the statistical properties of real-world 5G network data in terms of network traffic volume.

To investigate the relationship between network traffic volume and the number of handovers in the synthetic 5G network dataset generated using CTGAN, Pearson correlation analysis was conducted on the 5G handover dataset. The results showed a strong positive correlation between network traffic volume and the number of handovers ($r = 0.81$, $p < 0.001$), indicating that higher traffic volumes were associated with more frequent handovers. The effect size, calculated using Cohen's d, was large. These findings provide valuable insights into the relationship between network traffic volume and the number of handovers in 5G networks and highlight the potential of synthetic datasets generated using CTGAN for studying the performance of handover algorithms and protocols under varying traffic conditions.

The findings of this study suggest that the synthetic 5G network dataset generated using CTGAN and TVAE provides a reasonable approximation of the statistical properties of the original dataset. Statistical tests comparing the synthetic dataset to a real-world dataset showed no significant

differences in these variables. Comparing the performance of CTGAN and TVAE synthesizing the original dataset, TVAE showed a better performance score over CTGAN, with its performance scores are 94.14% and 89.66%.

On the other hand, even though TVAE outperformed CTGAN on the overall level, CTGAN produced a more similar distribution for individual columns compared to TVAE which cannot mimic every individual column's distribution efficiently. Figure 2 shows the frequency of appearance of the 3 types of networks. The CTGAN and TVAE produce opposite results in the over/under generation of each network's appearance hinting there is a pattern in the network's category that can influence how the data is discriminated.

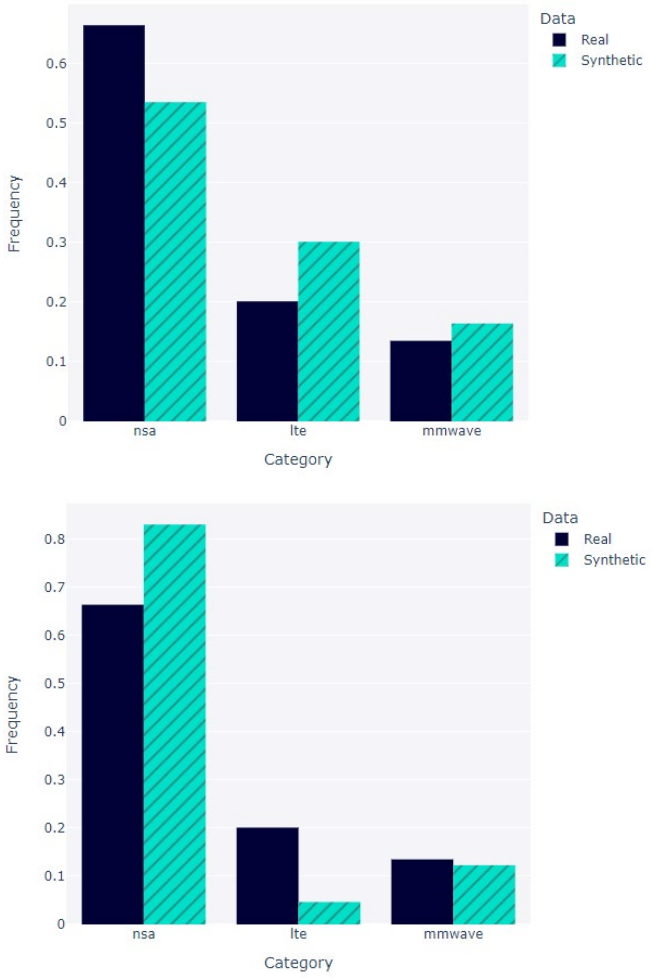

Fig. 2. Real vs. Synthetic data for column network representing the frequency vs the category in both the CTGAN (top) and TVAE (bottom)

Meanwhile, Fig. 3 shows the column shape results that is used to generate a visualization to provide insights into the shapes or distributions of data within each column of the synthetic data, focusing on characteristics such as data types, value ranges, and distributions. KScomplement refers to the Kolmogorov-Smirnov analysis and TVcomplement is the Total Variation analysis. The graph shows only the best complement score between the two for each variable generated using the CTGAN and TVAE methods. In both analysis, lower values indicate better similarity with the original dataset.

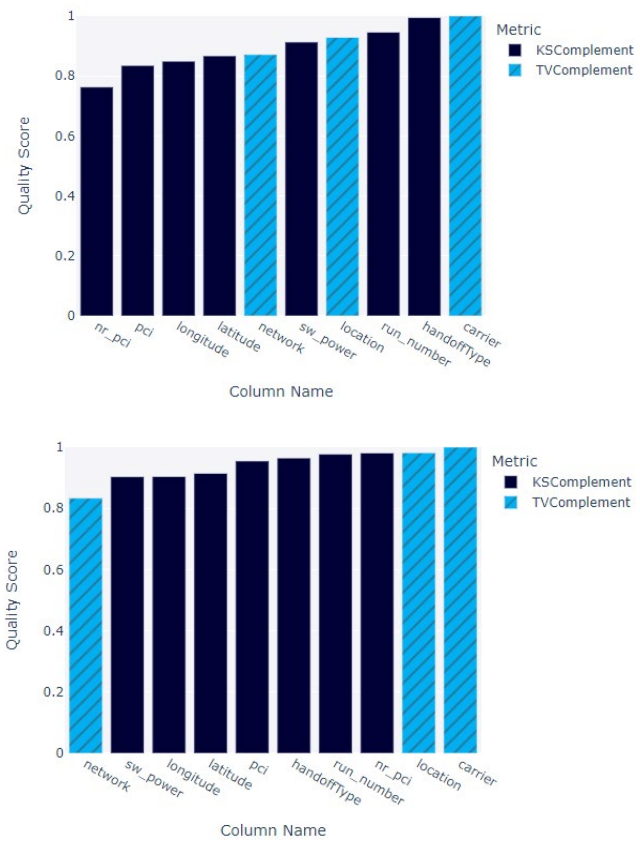

Fig. 3. The data quality between the CTGAN (top) and TVAE (bottom)

The findings suggest that both CTGAN and TVAE are promising approaches for generating synthetic 5G network datasets that replicate the statistical properties of real-world data. The synthetic datasets generated using CTGAN and TVAE were similar to real-world data in terms of network traffic volume, network topology, and other variables of interest.

However, the study is limited by several factors. First, only one dataset (Operator X) was used for evaluation, and therefore, the generalization of the results may be limited. Second, the quality of the synthetic datasets was evaluated for a limited set of variables, and future research should explore the effectiveness of both approaches for generating synthetic datasets for a wider range of variables in the 5G network domain.

Despite these limitations, our study contributes to the growing research on synthetic data generation and its applications for network technology testing. Future research should continue to explore the effectiveness of both CTGAN and TVAE in generating synthetic datasets for various domains and variables and evaluate the quality of the generated data against real-world data from different sources. Additionally, the ethical implications of synthetic data generation should be carefully considered, particularly in sensitive domains such as healthcare or finance, to ensure that the use of synthetic data does not have any unintended negative consequences. experience.

V. Conclusion

In conclusion, the study demonstrates the potential of CTGAN and TVAE for generating synthetic 5G network datasets that replicate the statistical properties of real-world

data. While this study is limited to using a single dataset for evaluation and evaluation of a limited set of variables, we believe that our findings contribute to the growing body of research on synthetic data generation and its applications for network technology testing especially in the communication field. Future research should explore the effectiveness of these approaches for generating synthetic datasets for a wider range of variables and domains and evaluate the quality of the generated data against real-world data from different sources. Additionally, the ethical implications of synthetic data generation should be carefully considered, particularly in sensitive domains such as healthcare or finance, to ensure that synthetic data does not have any unintended negative consequences.

Overall, the study highlights the potential of synthetic data generation as a valuable tool for network technology testing and contributes to advancing the field toward more efficient and cost-effective". Secondly, while Pearson correlation could potentially be used to analyse specific aspects of the generated data, it's not a comprehensive metric for evaluating GANs. Metrics like Fréchet Inception Distance (FID), Inception Score (IS), and others mentioned earlier are more commonly used in the context of generative models, as they are designed to assess factors like diversity, quality, and distribution similarity, which are critical in evaluating GAN performance.

ACKNOWLEDGMENT

The authors would like to acknowledge the support provided by the Air Force Office of Scientific Research: FA2386–20-1–4045 (UKM Ref: KK-2021–013) throughout the period of the study and the future of it.

REFERENCES

[1] W. Chen, X. Lin, J. Lee, A. Toskala, S. Shu, C. -F. Chiasserini and L. Liu, "5G-Advanced Towards 6G: Past, Present, and Future", IEEE Journal On Selected Areas In Communications, 2023.

[2] A. Tosun, B. Turhan and A. Bener, "Practical considerations in deploying AI for defect prediction: a case study within the turkish telecommunication industry" In Proceedings of the 5th International Conference on Predictor Models in Software Engineering, pp. 1-9. 2009.

[3] A. Sajeeda and B. M Mainul Hossain, "Exploring generative adversarial networks and adversarial training", International Journal of Cognitive Computing in Engineering 3 (2022): 78-89.

[4] M. Mirza and S. Osindero, " Conditional Generative Adversarial Nets", 2014. http://arxiv.org/abs/1411.1784

[5] T. Karras, T. Aila, S. Laine and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation", (2017). http://arxiv.org/abs/1710.10196

[6] J. Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks Monet Photos", In Proceedings of the IEEE international conference on computer vision, pp. 2223-2232, 2017. https://github.com/junyanz/CycleGAN

[7] A. Figueira and B. Vaz, "Survey on synthetic data generation, evaluation methods and GANs", Mathematics 10, no. 15 (2022): 2733.

[8] J. Kim, J. Jeon, J. Lee, J. Hyeong and N. Park. "Oct-gan: Neural ode-based conditional tabular gans". In Proceedings of the Web Conference 2021, pp. 1506-1515, 2021. https://doi.org/10.1145/3442381.3449999

[9] Z. Zhao, A. Kunar, R. Birke and L. Y. Chen. "Ctab-gan: Effective table data synthesizing", In Asian Conference on Machine Learning, vol. 157, pp. 97-112, PMLR, 2021. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-l

[10] J. Moon, S. Jung, S.Park and Eenjun Hwang, "Conditional tabular GAN-based two-stage data generation scheme for short-term load forecasting", IEEE Access 8, 205327-205339, (2020). https://doi.org/10.1109/ACCESS.2020.3037063

[11] I. Ullah and Q. H. Mahmoud, "A framework for anomaly detection in IoT networks using conditional generative adversarial networks", IEEE Access 9 (2021): 165907-165931. https://doi.org/10.1109/ACCESS.2021.3132127

[12] M. -P.Cote, B. Hartman, O. Mercier, J. Meyers, J. Cummings and E. Harmon, "Synthesizing property & casualty ratemaking datasets using generative adversarial networks", arXiv preprint arXiv:2008.06110 (2020)

[13] E. Yilmaz and B. German, "Conditional generative adversarial network framework for airfoil inverse design", In AIAA aviation 2020 forum, p. 3185. 2020. https://doi.org/10.2514/6.2020-3185

[14] J. Hui. "GAN – What is Generative Adversarial Networks GAN?" (2018). https://jonathan-hui.medium.com/gan-whats-generative-adversarial-networks-and-its-application-f39ed278ef09

[15] "Overview of GAN Structure | Machine Learning | Google Developers." Google Developers, https://developers.google.com/machine-learning/gan/gan_structure. Accessed 14 Apr. 2023

[16] Hassan, Ahmad, Arvind Narayanan, A. Zhang, W. Ye, R. Zhu, S. Jin, J. Carpenter, Z. M. Mao, F. Qian and Z. -L. Zhang, "Vivisecting mobility management in 5G cellular networks", In Proceedings of the ACM SIGCOMM 2022 Conference, pp. 86-100. 2022. https://doi.org/10.1145/3544216.3544217