Journal of Information Science and Technology
JIST Journal of Information Science and Technology
Volume 15, NO 2 | JUL – DEC 2025 | 15-23
ISSN: 2651-1053 (Online)
ISSN: 1906-9553 (Print)

Apply Monte Carlo Simulation to Synthesize Data

*Chaiyaporn Khemapatapan and Orawan Hensirisak **
Computer Engineering, College of Engineering and Technology,
Dhurakij Pundit University

**Corresponding Author: 66130226@dpu.ac.th*

Received: June 24, 2025; Revised: July 1, 2025; Accepted: August 5, 2025; Published: December 27, 2025

ABSTRACT – To address the problem of insufficient data for event analysis and simulation, especially in the AI era, data synthesization with accuracy and reliability is the solution. Data synthesization is the process of creating new data that mimics statistic of real data to increasing the amount of data available for further analysis and simulation. However, the existing methods like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) often struggle with categorical data. This research, therefore, presents data synthesis using Monte Carlo Simulation and compares with (GANs) and (VAEs). Seven datasets were used to synthesize 1,500 rows each then compared between real data and data synthesization using 1) The Kolmogorov-Smirnov Two-Sample Test, 2) T-test, 3) Cosine Similarity Test, 4) Multiple Linear Regression Analysis, and 5) direct data comparison. The results show that the Monte Carlo Simulation is highly effective in synthesizing data with numerous categorical variables, accurately simulating data distribution and means. The Monte Carlo simulation was 60.47% more efficient than GANs and 52.41% more efficient than VAEs. However, the Monte Carlo is not suitable for synthesizing data when there are many continuous variables or when the data distribution is not normal, as the synthetic data is generated from a normal distribution.

KEY WORDS -- Data Synthesization, Generative Adversarial Networks, Variational Autoencoders, Monte Carlo Simulation

1. Introduction

Data analysis using insufficient data can lead to various problems such as accuracy or reliability. Since the sample size cannot cover the entire population which will become a cause of data volatility, results might be biased towards one demographic group and risk of inaccurate decision-making due to difficulty in assessing potential risks which can be harmful to the organization and wasted time in analysis and produces results that cannot help in making the desired decisions. One of the solutions to the problem of insufficient data for analysis is using data synthetization which is very useful in case of insufficient data for analysis and it can avoid the risk of using real data containing personal information and save time and cost on collecting new data.

Currently, there are popular methods for data synthetization, such as Generative Adversarial Networks (GANs), which uses deep learning that focus on synthesizing data to closely resemble real data, or Variational Autoencoders (VAEs), which use the principle of dimensionality reduction through an encoder and reconstructing the data dimensions through a decoder.

However, these methods still have problems with inappropriateness for some variables, for example, being suitable for numeric variables but not for categorical variables, or being suitable for categorical variables but not for numeric variables. From the above problems, this research focuses on creating synthetic data using Monte Carlo Simulation method which uses random sampling from the distribution of real data to create a new data set of numerical variables with statistical properties close to the real data and uses the principle of Low of Change to create a new data set of categorical variables with a probability value of occurrence close to the real data.

2. Data Synthesization

Data synthesization is the method to create a new data. The concept for synthetization is for emulating statistically based on real data but without real data in result. The benefits of using data synthetization is reducing the risk of privacy

violations from using real data containing personal information, It can be customized in case of need different data set to cover situations that may not be found in real data and can reduce the cost and time of data collection. [1]

Currently, there are popular methods for data synthesization, such as Generative Adversarial Networks (GANs), which are a form of deep learning that focuses on synthesizing data to closely resemble real data, and Variational Autoencoders (VAEs), which use the principle of dimensionality reduction through an encoder and reverse the dimensionality transformation using a decoder. However, these methods may encounter issues with the generated data lacking diversity, being unsuitable for certain variables, and having statistical properties that do not closely match real data. This chapter will discuss the theories related to data synthesization using Generative Adversarial Networks, Variational Autoencoders (VAEs), and Monte Carlo Simulation methods, performance testing between real data and data synthesization, and related research.

Data Synthesization can be divided into 3 types:

1) Fully synthetic data. The synthetic data does not contain any real data.

2) Partially synthetic. The data from this synthesization contains both real and synthetic data.

3) Hybrid synthetic data. The data from this synthesization combines both real data and synthetic data. [2]

2.1 Data Synthesization Method

2.1.1 Generative Adversarial Networks (GANs) [3] are developed for data synthesization to be as realistic as possible. It uses a learning technique with a model consisting of two deep neural networks: 1) The generative model creates new data so that the discriminator model can detect whether the generated data is real or generated data. 2) The discriminator model which determines whether the samples coming from the generative model are real data or generated data.

Conditional Tabular Generative Adversarial Network (CTGAN) [4] is a method for generating synthetic tabular data based on GANs, with a key development principle being the use of non-Gaussian distributions.

2.1.2 Monte Carlo Simulation [5] is a synthesis method developed as a guideline for decision-making in problems that cannot be easily modeled or involve complex parameters. This method uses mathematical equations based on the principle known as the "Law of Chance".

2.1.3 Variational Autoencoders (VAEs) are a type of neural network that consists of two components: 1) Encoder, which creates encodings from data samples, and then this variable is fed into the generative network (Decoder). 2) Decoder network, which attempts to reconstruct the original data from the information encoded by the encoder network.

The Gaussian Variational Autoencoder (GVAE) [6] is based on the Variational Autoencoders (VAEs) and uses the Gaussian distribution equation derived from the distribution modeled by the Decoder network, allowing the basic distribution to be approximated.

2.2 Usable Synthetic Data

To assess whether synthetic data can be useful, this research employs various statistical testing methods [7] as listed in the below to evaluate the usable of synthetic data.

2.2.1 The Kolmogorov-Smirnov test is a test to determine whether two independent groups of real and synthetic data come from the same distribution.

2.2.2 The T-test is a test to compare the means of two independent sample data sets, with the preliminary agreement that the scores collected from both groups must be independent of each other.

2.2.3 Pearson's correlation is a test to find the correlation between datasets, used to explore whether two datasets are related or not.

2.2.4 Cosine Similarity Test is an examination of the similarity between two datasets using two vectors. It measures the size of the angle formed between the two vectors in an n-dimensional space, disregarding the length of the vectors and determining whether the two vectors point in the same direction.

2.2.5 Regression Analysis [8] is a statistical test for checking correlation between two or more variables, by dividing variable into independent variable and dependent variable then do a prediction equation.

2.3 Related Researches

Lei Xu and Kalyan Veeramachaneni [9] studied how to synthesis of tabular data using Generative Adversarial Networks. The researchers developed a model for data synthesization, focusing on creating tabular data with mixed variable types, meaning multiple variables with both continuous and discrete distributions. They compared the synthetic data obtained from the Generative Adversarial

Networks model in tabular form or used Macro-F1 to evaluate the model's performance in classifying data within each data group. The experiments found that Generative Adversarial Networks could effectively capture the relationships between variables and could be used with large datasets. However, they could only simulate data in numerical form.

Bauke Brenninkmeijer [4] conducted a study on modeling and developing tabular data using Generative Adversarial Networks with the aim of increasing the diversity of relationships among variables simulated from the Generative adversarial training for synthesizing tabular data (TGAN) model and to improve data synthesization methods using techniques such as TGAN (Generative adversarial training for synthesizing tabular data), WGAN-GP (Wasserstein GAN + Gradient Penalty), TGAN-SKIP, MedGAN, and TableGAN. From this research, it was found that the WGAN-GP model had the highest efficiency in data synthesization when compared to the three datasets.

Mathijs van Bree [6] conducted a study on data synthesization using the Variational Autoencoder (VAEs) method to test data synthesization and compare the suitability of variable types between categorical and continuous variables. The datasets used in this study consist of three sets, testing whether the synthetic data can replace the real data. The test models used for comparison are 7 models: 1. GVAE 2. ST-VAE 3. TGAN 4. WGAN 5. SKIP 6. MedGAN 7. TableGAN found that the dataset contains many categorical variables, making it unsuitable for using the Variational Autoencoder (VAEs) technique for data synthesization.

Kantana Loasirikul [5] studied the effectiveness of methods for handling imbalanced data for classification under different conditions, with the scope of the research being to examine the effectiveness of classifying educational data under imbalanced conditions. The data used in this study was synthetic using the Monte Carlo simulation method, with 8 independent variables, each following a normal distribution. The data was synthetic under 6 different conditions. From the experimental results, it was found that the data balancing method has a two-way interaction with the following conditions. (1) Sample size (2) Percentage of data between primary group data and secondary group data (3) Attrition rate and (4) data classification techniques, and found a three-way interaction with the following conditions (1) Sample size and the number of variables between categorical variables and continuous variables (2) Sample size and data classification techniques, and (3) Percentage of data between primary group data and secondary group data, and (4) Data classification techniques.

Leonardo Locowic et al. [10] conducted a study on data synthesization using the Monte Carlo Tree Search (MCTS) method and Large Language Models (LLMs). The researchers employed the data generation process from LLMs using temperature scaling and top-k sampling techniques to enhance the diversity of the generated data. Subsequently, they used Monte Carlo Tree Search to simulate evaluation outcomes, assessing the results using Kullback-Leibler Divergence (KL Divergence) and Jensen-Shannon Divergence (JS Divergence). The experiments revealed that the data synthesization method combining MCTS with LLMs can produce high-quality and realistic data synthesization, making it a good alternative for generating large-scale data synthesization.

Muhammad Nur Aqmal Khatiman et al. [11] conducted a study on the generation of synthetic data of 5G network performance metrics using Conditional Tabular Generative Adversarial Networks (CTGAN) and Topological Variational Autoencoder (TVAE) are trained and synthetic data is generated via the Synthetic Data Vault (SDV) module, a library for synthetic data generation in Python. From the experimental results, CTGAN gives data distribution values closer to the real data in some columns, while TVAE gives better overall results with a statistical similarity score of 94.14% compared to CTGAN's 89.66%.

Thi Thi Zin et al. [12] conducted a study on the generation of synthetic data of cow posture changes is helpful for reliability evaluation of indicators to predict when the calving event occurs using Markov Chain Monte Carlo (MCMC) to support the prediction of delivery time and reduce the burden of collecting real data which requires a lot of time and resources. The researchers limit the behavior patterns of a pregnant cow into four categories such as Lying State (L), Motion from Lying to Standing state (LS), Standing State (S) and Motion of Standing to Lying State (SL). Real data from 25 dairy cows were collected from the farm and used to construct a co-occurrence matrix and a transition matrix of their behaviors. Using these matrices, Monte Carlo Simulation was then applied to generate synthetic behavior sequences by randomly selecting each subsequent behavior based on the transition probabilities. Finally, it was found that synthesization data could be used to replace some of the real data to predict delivery time.

3. Methodology

3.1 Dataset

The datasets used in this study are listed as below:

3.1.1 The Delaware Births dataset [13]

3.1.2 Heart Rate Forecasting dataset [14]

3.1.3 Stock Market Data of USA dataset [15]

3.1.4 The Adult Census Income dataset [16]

3.1.5 The Car Evaluation dataset [17]

3.1.6 NIFTY-50 Stock Market (2000 - 2021) dataset [18]

3.1.7 Employee/HR (All in One) dataset [19]

3.2 Data Preparation

Data preparation is a step to make real data ready to synthesize new data. The process is as follows:

- to remove missing values
- to remove columns that cannot be synthetized
- to remove rows with abnormal values
- to randomly select data for 1500 rows
- to test the random data to see if the original data distribution and the original mean are equal to the random data using the Kolmogorov–Smirnov test and the T-test. If the original data distribution and the original mean are not equal to the random data, continue randomizing until the test results show that the data has the same distribution and same mean.

3.3 Data Synthesization

The procedures of data synthesization for various methods can be listed as follows:

3.3.1. Data synthesization using Monte Carlo Simulation as shown in Figure 1 and

3.3.2. Data synthesization using CTGAN and Variational Autoencoders (VAEs) as shown in Figure 2.

It can be noted that Monte Carlo Simulation method is more complex than the other methods because each type of variable has its own operation.

3.4 Performance Evaluation

Performance evaluation is a step to assess how well the synthetic data reflects the real data. The process is as follows:

3.4.1. Data Preparation Steps

Step 1. Convert categorical variables into numerical format using Label Encoding.

Step 2. Convert character-related variables into numerical format using Sentence Transformer.

Step 3. Arrange the columns between real and synthetic data to match.

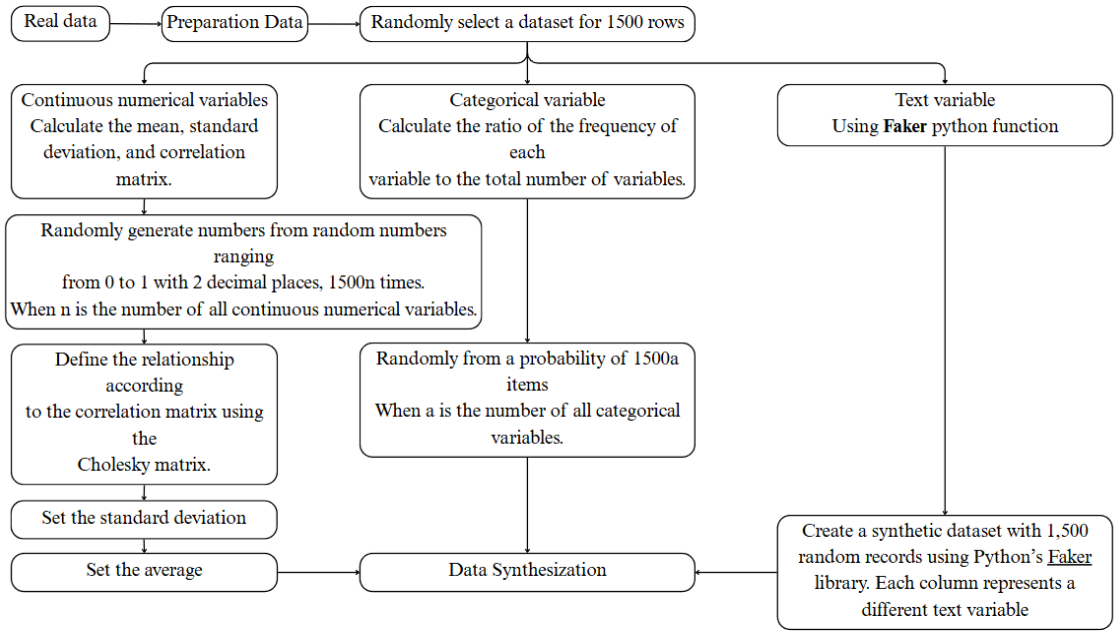

Figure 1 The process of data synthesization using Monte Carlo Simulation

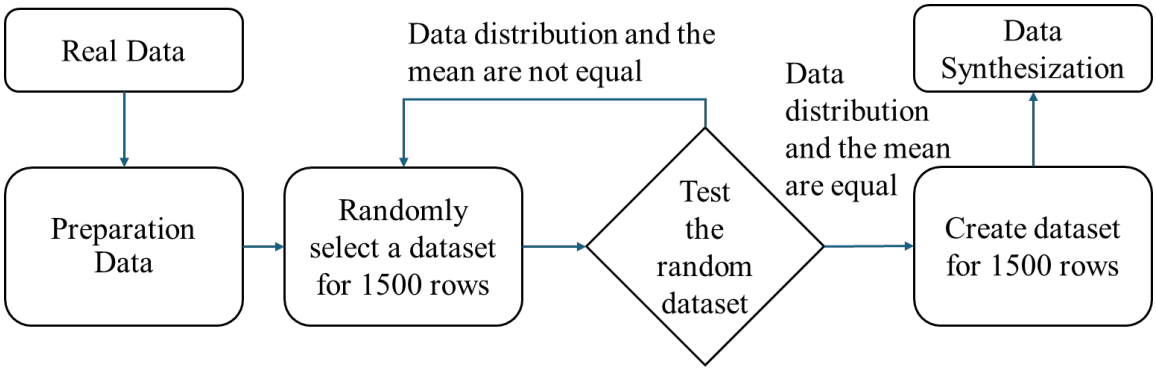

Figure 2 The process of data synthesization using CTGAN and VAEs

Table 1 the total number of variables obtained from data synthesization.

Dataset name	Total number of variables	Number of categorical variables	Number of numerical variables	Number of character-related variables
Delaware Births [13]	16	16	0	0
Heart Rate Forecasting [14]	32	1	31	0
Stock Market Data of USA [15]	7	1	6	0
Adult Census Income [16]	15	9	6	0
Car Evaluation [17]	7	7	0	0
NIFTY-50 Stock Market [18]	14	2	12	0
Employee/HR [19]	17	3	2	12
Total	108	39	57	12

3.4.2. Test the specified statistics using testing methods as follows:

- Kolmogorov-Smirnov Two-Sample test is used to determine whether real and synthetic data have the same distribution,
- T-test is used to test whether the means obtained from real and synthetic data are same,
- Similarity Score test is used to measure the similarity between real and synthetic data, and
- Regression analysis is used to predict the value of the dependent variable when the values of the independent variables are specified.

Table 1 shows the total number of variables that can be used in data synthesization by comparing the result of the Kolmogorov-Smirnov two-sample test and the T-test

4. Experimental Results

4.1 The Kolmogorov-Smirnov test

From the Figure 3, the results found that the Monte Carlo technique was the most effective to accurately synthesize categorical variable and character-related data because It is able to simulate the distribution of data for 38 out of 39 variables while it can simulate the character-related data for 8 out of 12 variables. However, data synthesization for numerical variables of all three techniques did not perform well, with CTGAN, Monte Carlo, and VAEs being able to simulate 13, 3, and 9 variables, respectively, out of a total of 57 variables.

4.2 The T-test

From the Figure 4, it was found that the data synthesization of categorical variables using the Monte Carlo technique is more efficient. In additon, it can accurately simulate the mean for 38 variables out of 39 variables. Character-related variables were able to simulate the mean for 8 out of 12, meaning it can accurately simulate the mean of the data for datasets with many categorical variables due to the data generation method based on probability principles for random data. The data synthesization of numerical variables using the CTGAN technique was more efficient, accurately

simulating the mean of the data for 54 variables, next is the synthesis using the VAEs technique, which can accurately simulate the mean of the data for 46 variables out of 57 variables.

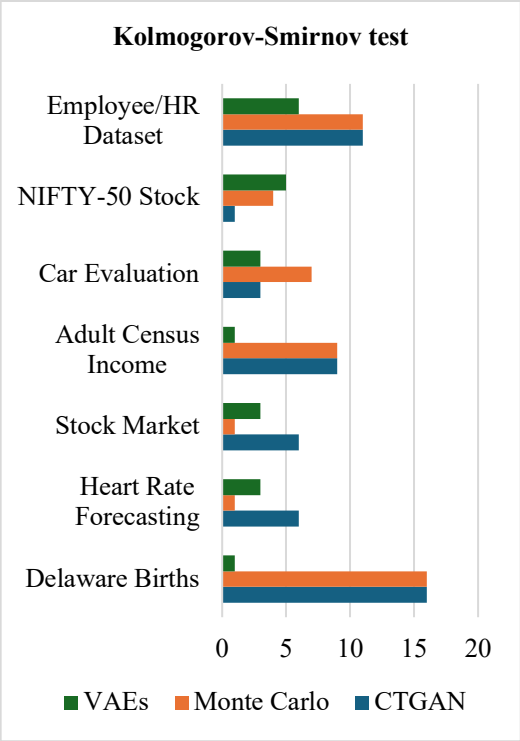

Figure 3 Results of the Kolmogorov-Smirnov test

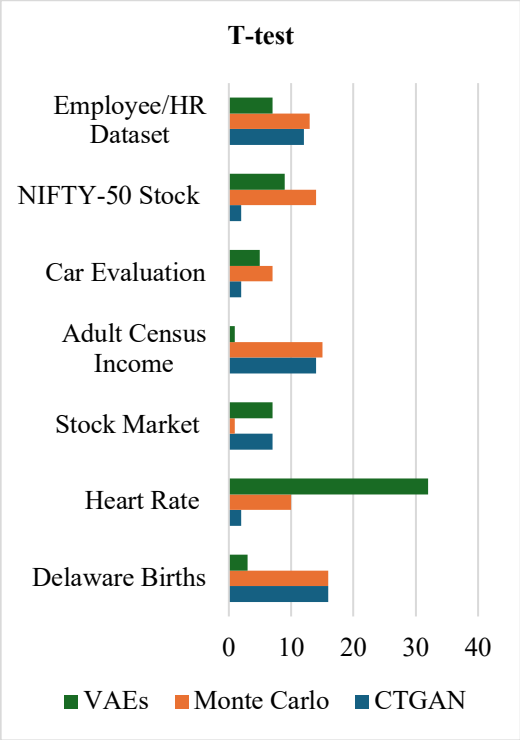

Figure 4 Results of the T-test

4.3 The Similarity Score

The analysis from the Table 2 found that all three techniques can synthesize well in cases where the dataset includes numerical variables. However, the data synthesize using the VAEs technique had a Similarity Score of 0.5248 for the STOCK MARKET DATA OF USA dataset because it could not synthesize data with high value, e.g., 1.394626e+09. The data synthesization using the GANs technique has a Similarity Score of 0.4413 for the NIFTY-50 Stock Market Data (2000 - 2021) dataset due to its complex numerical relationships. All three data synthesization techniques have different Similarity Scores for the Car Evaluation dataset because it consists entirely of categorical numerical data. The average Similarity Score of the Monte Carlo method is the highest at 0.9296, indicating the consistency of the synthetic data.

4.4 The Regression Analysis

From the results as shown in Table 3, it can be found that data synthesization using the Monte Carlo technique is highly effective and mostly close to the Adj. R-squared values. The differences from the four datasets—Stock Market Data of the USA, Adult Census Income, NIFTY-50 Stock Market Data and Employee/HR Dataset were the smallest compared to other synthesis techniques. However, the Delaware Births dataset still show substantially lower than those of the real data, dropping from 0.718 to 0.000. This indicates that while Monte Carlo outperforms other methods in this specific metric, it still lacks the ability to fully replicate complex statistical relationships.

4.5 The comparison between real data and synthetic data

Table 4 shows the number of identical rows between real and synthetic data. The Monte Carlo technique is most effective. The GAN technique has duplicates in two datasets: Delaware Births and Car Evaluation. The dataset with the most duplicates among the three synthesis techniques is Car Evaluation, which is a dataset with all categorical variables. Therefore, it exceeds the number of

duplicates because the number of categorical data is less than using numerical randomization.

Table 2 The Similarity Score

Dataset Name	Simulation method		
	GANs	Monte Carlo	VAEs
Delaware Births [13]	0.9998	0.9999	0.9999
Heart Rate Forecasting [14]	0.8357	0.8567	0.9872
Stock Market Data of USA [15]	0.8627	0.9970	0.5248
Adult Census Income [16]	0.8441	0.9279	0.9983
Car Evaluation [17]	0.6613	0.7302	0.7561
NIFTY-50 Stock Market [18]	0.4413	0.9999	0.9999
Employee/HR [19]	0.9999	0.9960	0.9999

Table 3 The ADJ. R-SQUARED value from the regression analysis

Dataset name	Adj R -Squared from real data	GANs	Monte Carlo	VAEs
Delaware Births [13]	0.718	0.000	-0.000	0.005
Heart Rate Forecasting [14]	0.190	0.016	0.024	0.279
Stock Market Data of USA [15]	0.121	0.002	0.077	0.070
Adult Census Income [16]	0.093	0.006	0.046	0.155
Car Evaluation [17]	0.107	-0.000	-0.001	0.324
NIFTY-50 Stock Market [18]	0.437	-0.586	0.485	0.259
Employee/HR [19]	0.011	0.033	0.012	0.088

Table 4 The number of rows that are the same between real data and data synthesization.

Dataset name	The same amount		
	GANs	Monte Carlo	VAEs
Delaware Births [13]	17	0	0
Heart Rate Forecasting [14]	0	0	0
Stock Market Data of USA [15]	0	0	0
Adult Census Income [16]	0	0	0
Car Evaluation [17]	484	706	1143
NIFTY-50 Stock Market [18]	0	0	0

5. Conclusion

Three data synthesization techniques were identified: GANs, Monte Carlo technique and VAEs. The comparison in terms of Kolmogorov-Smirnov test, T-test, Regression Analysis, means, similarity score, and the number of duplicates are applied. The results found that data synthesization using the Monte Carlo method performed well in scenarios where the number of categorical variables is more than the number of continuous variables but comparison of correlation coefficients found that data synthesization using the Monte Carlo method is not suitable for applications where predictive modeling or inferential analysis is essential. This is because the Monte Carlo method uses the law of large numbers, making the random sampling of categorical variables more realistic than other data synthesization methods. However, it is not suitable for synthesizing data when data contain many continuous variables or when the data distribution is not normal, as the synthetic data is generated from a normal distribution.

From the directly comparison of real and synthetic data to determine whether the data synthesization can conceal personal information, the datasets used include the Heart Rate Forecasting dataset, the Delaware Births dataset, and the Adult dataset. The data synthesization generated using the Monte Carlo method and VAEs performed the best, with no duplicate data found between the real data and data synthesization.

Data synthesization using the Monte Carlo simulation technique is the most efficient, especially for categorical variables. The Monte Carlo simulation technique can be done quickly as it does not require the model to learn all the previous data first. The Monte carlo simulation can manually adjust the parameters to introduce greater diversity in the data, ensuring alignment with specific data characteristics and analytical requirements. Moreover, under privacy constraints, the Monte Carlo simulation technique can still generate data synthesization without retaining any real data while maintaining the statistical properties of the data.

References

[1] Turing, "Synthetic Data Generation: Definition, Types, Techniques, and Tools," TURING [online]. Available: https://www.turing.com/kb/synthetic-data-generation-techniques#what-is-synthetic-data? (Accessed Feb. 12, 2025).

[2] A. Beduschi, "Synthetic data protection: Towards a paradigm change in data regulation", Law School, University of Exeter, Exeter, UK, 2024.

[3] W. Phusomsai, "Extending GANs' Latent Space for Diverse Image Generation from Sketches," M.S. thesis, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand, 2019.

[4] B. Brenninkmeijer, "On the Generation and Evaluation of Tabular Data using GANs," Radboud University, Houtlaan 4, 6525 XZ Nijmegen, Netherlands, 2019.

[5] K. Laosirikul, "The Performance of Imbalanced Data Handling Methods for Classification under Different Conditions," M.S. thesis, Dept. of Statistics, Chulalongkorn University, Bangkok, Thailand, 2022.

[6] M. van Brer, "Unlocking the Potential of Synthetic Tabular Data Generation with Variational Autoencoders," Radboud University, Nijmegen, Netherlands, 2019.

[7] C. Noikhamyang, "A Comparison of Parametric and Nonparametric Statistical Tests for Identifying Differences between Two Independent Populations," M.S. project, Dept. of Statistics, Srinakharinwirot University, Bangkok, Thailand, 2009.

[8] Chiang Mai University, "Statistics and Health Data Analysis," Chiang Mai, Thailand, 2020.

[9] L. Xu and K. Veeramachaneni, "Synthesizing Tabular Data using Generative Adversarial Networks," LIDS, MIT, Cambridge, MA, USA, 2018.

[10] L. Locowic, Alessandro Monteverdi, "Synthetic Data Generation from Real Data Sources using Monte Carlo Tree Search and Large Language Models," arXiv preprint arXiv:2401.12345, 2024. Available: https://d197for5662m48.cloudfront.net/documents/publicationstatus/224165/preprint_pdf/3c3ef1837551b4cf3bb7cfd68385de99.pdf

[11] M. N. A. Khatiman, et.al, "Generation of Synthetic 5G Network Dataset Using Generative Adversarial Network (GAN)," 2023 IEEE 16th Malaysia International Conference on Communication (MICC). IEEE, pp. 141–145, Dec. 10, 2023. doi: 10.1109/micc59384.2023.10419563.

[12] T. T. Zin, et.al, "Markov Chain Monte Carlo Method for the Modeling of Posture Changes Prior to Calving," The IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech). IEEE, pp. 291–292, Mar. 09, 2021, doi: 10.1109/lifetech52111.2021. 9391822.

[13] Noey, *Delaware Births* [Online]. Available: https://www.kaggle.com/datasets/noeyislearning/delaware-births/data (Accessed Feb. 12, 2025).

[14] G. Dutta, *Heart Rate Forecasting* [Online]. Available: https://www.kaggle.com/datasets/gauravduttakiit/heart-rate-forecasting (Accessed Feb. 12, 2025).

[15] A. Rafiee, *Stock Market Data of USA* [Online]. Available: https://www.kaggle.com/datasets/ahmadrafiee/stock-market (Accessed Feb. 12, 2025).

[16] R. Sandiani, *Census Income* [Online]. Available: https://www.kaggle.com/datasets/uciml/adult-census-income/data (Accessed Feb. 12, 2025).

[17] U. Zia, *Car Classification Dataset* [Online]. Available: https://www.kaggle.com/datasets/stealthtechnologies/car-evaluation-classification (Accessed Feb. 12, 2025).

[18] Vopani, *NIFTY-50 Stock Market Data (2000 - 2021)* [Online]. Available: https://www.kaggle.com/datasets/rohanrao/nifty50-stock-

market-data (Accessed Feb. 12, 2025).
[19] R. S. Rana, *Employee/HR Dataset (All in One)* [Online]. Available: https://www.kaggle.com/datasets/ravindrasinghrana/employeedataset/data?select=recruitment_data.csv (Accessed May 31, 2025).