Justin Morel

State Of The Art (SoTA) analysis report

Synthetic Data

I- What are synthetic data?

Synthetics data are artificially generated datasets that replicate or have similarity with real-world data. They are produced by capturing the statistical properties of real data to create new data points with similar characteristics [1]. Several methods can be used to generate synthetic data, including statistic-based methods such as the multivariate normal distribution (MVND) and bootstrapping, probabilistic-based methods like Stochastic Block Models (SBMs), machine learning-based methods such as tree ensembles, the Gaussian Mixture Models (GMMs), and deep learning-based methods including Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) [1]. But generating synthetic medical data is a challenge because of its inherent complexity and longitudinal nature [2].

Synthetic data offers the opportunity to protect and preserve the confidentiality of real data. They are used across various domains, with a particular emphasis on scientific research and applications in Artificial Intelligence (AI). In sensitive fields like healthcare, where patient data must remain confidential, synthetic data allow for anonymization while maintaining high fidelity to the original information. [3], [4], [5]

Not only do they protect patient privacy, but they also help researchers and medical professionals by increasing data availability, enabling broader experimentation and simulation without the risk of exposing personal data. However, a key challenge in generating synthetic data lies in maintaining a balance between realism and privacy. High-quality healthcare data is essential for high-quality research, better development initiatives, and outcomes, informed medical decisions, and better quality of life [6].

There exist different types of synthetic data:

- Full synthesis: Data entirely generated by models without using any actual records from the original dataset. It replaces all real data, ensuring maximum privacy but may risk some loss of statistical accuracy. [7]
- Partial synthesis: Only sensitive variables are replaced with synthetic values, while the rest of the dataset remains real. This strikes a balance between privacy protection and data utility. [7]

Depending on the original data type, synthetic data can also take specific forms [1]:

- Tabular Synthetic Data: Artificially generated data structured in a traditional table format, like spreadsheets or relational databases. Each row represents an individual sample, and each column corresponds to a variable or feature. [8]
- Radiomics Synthetic Data: Synthetic features derived from medical imaging (e.g., MRI, CT scans), representing quantitative attributes such as tumor shape, texture, and intensity. These are generated to simulate radiomic profiles without requiring real medical images. [9]
- Time-Series Synthetic Data: Data generated to mimic temporal sequences, preserving trends, seasonality, and time-dependent patterns found in real-world time series. [10]
- Omics Synthetic Data: Artificially generated data that replicates high-dimensional biological datasets from fields like genomics, transcriptomics, proteomics, and metabolomics. These datasets often include thousands of biological variables per sample. [11]
- Multimodal Synthetic Data: Synthetic datasets that integrate multiple data modalities, for example, combining text, images, and structured data to represent a single entity or scenario. [12]

II- <u>Why do we need synthetic data (their value)?</u>

Synthetic data has become increasingly essential with the rise of artificial intelligence (AI) and the implementation of regulations such as the General Data Protection Regulation (GDPR) in Europe [1]. Ethical concerns are now at the forefront of societal debates, making compliance with these norms and laws a necessity [13]. The use of data has become essential for decision making in public health at the local, national, and global level 14]. Unfortunately, data from healthcare has some challenges by their availability [14]. In the medical field in particular, data confidentiality is crucial to maintaining patient integrity. Health data has become an asset, with major tech companies like GAFAMs (Google, Apple, Facebook, Amazon, and Microsoft) seeking to exploit it [15]. As a result, synthetic data plays a critical role in preserving personal privacy. The value of medical, biological, and personal data has significantly increased in recent years. Their economic value is considerable, but it must be balanced with the need to protect privacy while supporting scientific research. These data also hold major strategic importance. This is why synthetic data have a primordial role.

Moreover, synthetic data can help reduce discrimination related to age, gender, race, and other biases that may be present in real datasets [1]. By eliminating such biases, it enables the development of fairer AI systems. Synthetic data also mitigates the risk of data leaks, thereby enhancing overall trust in AI systems [1]. Because synthetic datasets are not linked to real individuals, researchers can work with them freely without compromising privacy. This freedom allows researchers to bypass some of the constraints imposed by existing regulations. Synthetic data is especially necessary when real-world data is insufficient or of poor quality [16]. It offers a viable alternative that ensures the continuity and effectiveness of data-driven research and development.

III- What are the types of synthetic data generators?

There are several methods for generating synthetic data, with at least 77 different generators identified [1]. This reflects the wide variety of available tools, each tailored to specific needs and capable of producing optimal results depending on the type of medical data involved.

In the field of deep learning, the most used methods are Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). These models have demonstrated strong capabilities in generating privacy-preserving data, supporting applications such as clinical decision-making and predictive modeling. Variants like CycleGANs are particularly effective in image synthesis. There is machine learning–based tools such as the Synthetic Data Vault (SDV), which is designed for generating tabular data [17].

Synthetic Data Generator	**Usage**	**Problem to solve**	**Type of Data**	**Pros**	**Cons**
Generative Adversarial Networks (GAN) [18]	Deep learning	Generate realistic synthetic data from random noise	Radiomics, tabular, images, multimodal	Adaptable to various data	GANs are difficult to train due to several factors that include the loss function, hyperparameters, or a generator that can fool the discriminator.

Conditional GAN (CGAN) [18]	Deep Learning	synthesize images based on a chosen label, cannot sample an image of the desired class	Images, tabular	CGAN is trained using data instances and their respective labels	Dating from 2014, there are improvements possible
Deep Convolutional GAN (DCGAN) [18]	Deep learning	More precision to synthetize images	Images	Use of convolutional layers, detect edges, blur the images, remove noise	The features in the latent space had no semantic meaning. not possible to change the values of a feature in latent space and predict what that change would do to the image
Information Maximizing GAN (InfoGAN) [18]	Deep learning	to give semantic meaning to features in the latent space	Images	Capable of more precise recognition (handwritten or background numbers, glasses for example)	large number of hyperparameters and a large number of training samples, training process prohibitively expensive
Coupled GAN (CoGAN) [18]	Deep learning	Reduce the constraints of using a single GAN	Images	Paired of GAN, sharing weights requires fewer parameters than two individual GANs, less	Training intensive, large datasets

				memory consumption, less computational power, and fewer resources.	
Wasserstein GAN (WGAN) [18]	Deep learning	Improve the discriminator	images	avoids mode collapse and provides a meaningful loss metric that correlates with the generator's convergence and sample quality. Stable trainings and realistic samples	Require extensive computational resources
Cycle-Consistent GAN (CycleGAN)	Deep learning	image-to-image translation without paired data	Multimodal	Excellent for image-to-image translation tasks without paired data	Poor at maintaining consistency of synthesized images when large variation between input modalities.
Progressively growing GAN (ProGAN) [18]	Deep learning	Stabilize GAN training by progressively increasing generated	Images	accelerates and stabilizes training by producing images with	High resource consumption and complex trainings dynamics

		images resolution		few pixels. Layers corresponding to higher resolutions are added in the training process, allowing the creation of high-quality images.	
Style-Distribution GAN (SD-GAN) [19]	Deep learning	synthesize images of different styles based on several similar images	Images	Transferring and blending diverse style features.	Challenge to manage different variations styles
Bayesian models [20]	Machine learning	Synthetize tables from real data	Tabular	Flexible, non-linear data treated	Computationally intensive, needs substantial data.
multivariate normal distribution (MVND) [1]	Statistical	used for generating synthetic distributions that preserve statistical properties of the real data	Tabular, images	Robust and simple	Not usable with complex dependences in data
Vine Copula Models [21]	Statistical	Generate synthetic data with	Tabular	Excellent for modeling the dependencies	Complex to use and interpret

		complex dependencies		between variables	
Bayesian (hierarchical) generalized linear models (hGLM) [1]	Machine learning	Synthetic generation with hierarchic structure	Tabular		Require extensive computational resources
Tree Ensembles [1]	Machine learning	Data generation by learning of complex structures	Tabular	Combine multiple decision tree to make realistic data	Large Training so can be expensive, high number of trees can increase time of process and resources usage
Gaussian Mixture Models (GMM) [22]	Machine Learning	represent a dataset as a mixture of Gaussian distributions.	Tabular	Fast, easy to train	Data have to follow a Gaussian distribution
Hidden Markov Model (HMM) and regression algorithm [1] [23]	Machine Learning	describe the evolution of observable events, which themselves, are dependent on internal factors that can't be directly observed	Time-series, tabular	Effective for capturing sequences and transitions in time series data.	Complex integration of multiple modeling techniques
Variational AutoEncoders (VAE) [24] [1]	Deep learning	compress data into a lower-	Time-series	Good for modeling distribution of	Less precise, quality not

		dimensional latent space, a dre-construct data from this space		data for simulation.	perfect for samples
Diffusion Model [25]	Deep learning	Generate highly realistic data by denoising random noise	Images, Time-series, multimodal	Extremely high sample quality, stable training, good control over output	very computationally expensive

IV- **References**

[1] Synthetic data generation methods in healthcare: A review on open-source tools and methods. Vasileios C Pezoulas, Dimitrios I Zaridi, Eugenia Mylona, Christos Androutsos, Kosmas Apostolidis, Nikolaos S Tachos, Dimitrios I Fotiadis (https://pmc.ncbi.nlm.nih.gov/articles/PMC11301073/)

[2] Big Healthcare Data Analytics: Challenges and Applications. Chonho Lee, Zhaojing Luo, Kee Yuan Ngiam, Meihui Zhang, Kaiping Zheng, Gang Chen, Beng Chin Ooi & Wei Luen James Yip (https://link.springer.com/chapter/10.1007/978-3-319-58280-1_2)

[3] Between Access and Privacy: Challenges in Sharing Health Data. Bradley Malin, Kenneth Goodman (https://www.thieme-connect.de/products/ejournals/pdf/10.1055/s-0038-1641216.pdf?utm_source=sciencedirect_contenthosting&getft_integrator=sciencedirect_contenthosting)

[4] Privacy Law, Data Sharing Policies, and Medical Data: A Comparative Perspective. Edward S. Dove & Mark Phillips (https://link.springer.com/chapter/10.1007/978-3-319-23633-9_24?utm_source=getftr&utm_medium=getftr&utm_campaign=getftr_pilot&getft_integrator=sciencedirect_contenthosting)

[5] Synthetic data generation: a privacy-preserving approach to accelerate rare disease research. Jorge M Mendes, Aziz Barbar, Marwa Refaie (https://pmc.ncbi.nlm.nih.gov/articles/PMC11958975/)

[6] Electronic health records to facilitate clinical research. Martin R. Cowie, Juuso I. Blomster, Lesley H. Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, Alexander Michel, Seleen Ong, Jill P. Pell, Mary Ross Southworth, Wendy Gattis Stough, Martin Thoenes, Faiez Zannad & Andrew Zalewski (https://link.springer.com/article/10.1007/s00392-016-1025-6#Sec15)

[7] Synthetic Data. Trivellore E. Raghunathan (https://www.annualreviews.org/content/journals/10.1146/annurev-statistics-040720-031848)

[8] Tabular and latent space synthetic data generation: a literature review. Joao Fonseca & Fernando Bacao (https://doi.org/10.1186/s40537-023-00792-7)

[9] Overcoming data scarcity in radiomics/radiogenomics using synthetic radiomic features. Milad Ahmadian, Zuhir Bodalal, Hedda J. van der Hulst, Conchita Vens, Luc H.E. Karssemakers, Nino Bogveradze, Francesca Castagnoli, Federica Landolfi, Eun Kyoung Hong, Nicolo Gennaro, Andrea Delli Pizzi, Regina G.H. Beets-Tan, Michiel W.M. van den Brekel, Jonas A. Castelijns (https://doi.org/10.1016/j.compbiomed.2024.108389.)

[10] Comparative assessment of synthetic time series generation approaches in healthcare: leveraging patient metadata for accurate data synthesis. Isasa, I., Hernandez, M., Epelde, G. et al. (https://doi.org/10.1186/s12911-024-02427-0)

[11] Multi-omics data integration by generative adversarial network. Khandakar Tanvir Ahmed, Jiao Sun, Sze Cheng, Jeongsik Yong, Wei Zhang. (https://academic.oup.com/bioinformatics/article/38/1/179/6355579)

[12] Generation of realistic synthetic data using Multimodal Neural Ordinary Differential Equations. Wendland, P., Birkenbihl, C., Gomez-Freixa, M. et al. (https://doi.org/10.1038/s41746-022-00666-x)

[13] Synthetic Data in AI: Challenges, Applications, and Ethical Implications. Shuang Hao, Wenfeng Han, Tao Jiang, Yiping Li, Haonan Wu, Chunlin Zhong, Zhangjun Zhou, He Tang (https://arxiv.org/pdf/2401.01629)

[14] A systematic review of barriers to data sharing in public health. Willem G van Panhuis, Proma Paul, Claudia Emerson, John Grefenstette, Richard Wilder, Abraham J Herbst, David Heymann & Donald S Burke (https://link.springer.com/article/10.1186/1471-2458-14-1144?utm_source=getftr&utm_medium=getftr&utm_campaign=getftr_pilot&getft_integrator=sciencedirect_contenthosting)

[15] Votre Santé, Un Trésor Convoité : Arte Reportage. David Carr-Brown (https://www.youtube.com/watch?v=v9DfKROPV78&t=9s) (Studied at my university)

[16] Survey on Synthetic Data Generation, Evaluation Methods and GANs (https://www.mdpi.com/2227-7390/10/15/2733)

[17] GitHub Synthetic Data Vault (https://github.com/sdv-dev/SDV)

[18] Survey on Synthetic Data Generation, Evaluation Methods and GANs. Figueira A, Vaz B. (https://doi.org/10.3390/math10152733)

[19] SD-GAN: A Style Distribution Transfer Generative Adversarial Network for Covid-19 Detection Through X-Ray Images. T. Kausar, Y. Lu, A. Kausar, M. Ali and A. Yousaf (https://ieeexplore.ieee.org/document/10061161)

[20] Synthetic data generation with probabilistic Bayesian Networks. Gogoshin G, Branciamore S, Rodin AS (https://pmc.ncbi.nlm.nih.gov/articles/PMC8848551/)

[21] Learning Vine Copula Models For Synthetic Data Generation. Yi Sun, Alfredo Cuesta-Infante, Kalyan Veeramachaneni (https://arxiv.org/pdf/1812.01226)

[22] Synthetic data generation with Gaussian Mixture Models. May 24, 2023 (https://ydata.ai/resources/synthetic-data-generation-with-gaussian-mixture-models)

[23] Hidden Markov Models and their Applications in Biological Sequence Analysis. Curr Genomics. Yoon BJ. (https://pmc.ncbi.nlm.nih.gov/articles/PMC2766791/)

[24] An Introduction to Variational Autoencoders. Diederik P. Kingma, Max Welling (https://arxiv.org/pdf/1906.02691)

[25] An overview of diffusion models for generative artificial intelligence. Davide Gallon, Arnulf Jentzen, Philippe von Wurstemberger (https://arxiv.org/pdf/2412.01371)

ChatGPT (for reformulation and translation)