Synthetic Data: A Solution for Privacy Challenges in Industries with Sensitive Data?

Novalytics Gibraltar

Abstract—The growing demand for privacy-preserving solutions in industries handling sensitive information, such as healthcare and finance, has accelerated the exploration of synthetic data generation. Synthetic data offers a promising approach by replicating the statistical properties of real datasets without exposing individual records. This paper discusses the potential of synthetic data to facilitate privacy protection, support data sharing, enhance AI model performance, and mitigate biases in real-world datasets. It further examines key generation methods, such as Generative Adversarial Networks (GANs), while highlighting challenges related to fidelity, bias, and regulatory compliance. Although synthetic data remains in its early adoption phase, it has the potential to transform data-driven research and operational practices across high-risk sectors.

Keywords—*Synthetic Data, Privacy, Healthcare, Finance, Machine Learning, Data Sharing, Bias Mitigation, GANs, Data Protection, Data Science*

1. Introduction

The ability to derive insights from data is a key driver of innovation in the healthcare and finance sectors. However, organisations must also balance this need with stringent data privacy requirements. Synthetic data generation offers a powerful solution that allows access to realistic datasets without compromising individual privacy. When applied effectively, synthetic data can accelerate research and development, support regulatory compliance, enhance operational resilience, and reduce the risks associated with handling sensitive personal information.

Data privacy remains a critical concern within the healthcare and finance sectors due to the sensitive nature of personal health and financial information. Any misuse or breach of such data can have severe implications for businesses and public trust. Consequently, access to data is often restricted or withheld to mitigate privacy risks. However, this practice can limit the insights obtainable from machine learning models, as reliable and accurate predictions require access to large-scale datasets.

Synthetic data can help organisations meet regulatory obligations under frameworks such as the General Data Protection Regulation (GDPR) [8]. In particular, it supports the principle of data minimisation (Article 5(1)(c)) and privacy by design and by default (Article 25), whilst also reducing risks identified during Data Protection Impact Assessments (DPIAs). As synthetic data sets do not directly identify individuals, their use can mitigate many compliance burdens associated with real-world data handling.

One potential solution is the generation of synthetic data that mimics real-world data without compromising individual privacy [6], [7]. Synthetic data are created by transforming the statistical properties of real datasets, thereby retaining representativeness and simulating real-world processes, while avoiding the risks associated with privacy breaches (Figure 1).

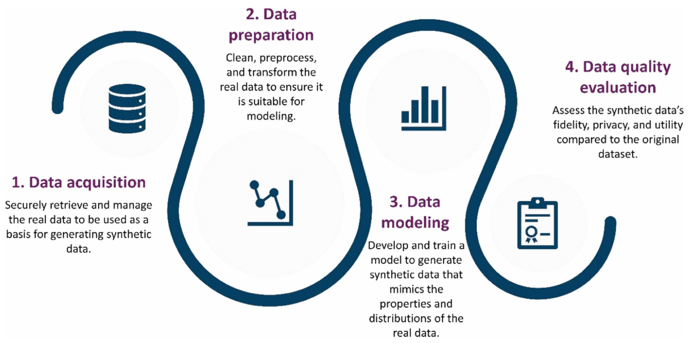
2. Data preparation
Clean, preprocess, and transform the real data to ensure it is suitable for modeling.
4. Data quality evaluation
Assess the synthetic data's fidelity, privacy, and utility compared to the original dataset.
1. Data acquisition
Securely retrieve and manage the real data to be used as a basis for generating synthetic data.
3. Data modeling
Develop and train a model to generate synthetic data that mimics the properties and distributions of the real data.

Figure 1. The four stages of the synthetic data generation workflow.

2. Applications

- **Privacy protection** — Synthetic data enable the creation of datasets that closely resemble real-world data, supporting innovation whilst maintaining compliance with privacy regulations such as the GDPR [8]. This approach allows organisations to develop AI models and analytical tools without exposing sensitive personal information.
- **Data sharing and collaboration** - Synthetic datasets facilitate safer data sharing between internal teams, partners, and regulators. By generating realistic yet anonymised datasets, organisations can collaborate more effectively, detect subtler signals, and gain broader insights into market and operational dynamics.
- **Enhanced AI model performance** — Access to large, diverse synthetic datasets strengthens the performance of AI models, particularly for applications such as fraud detection, anti-money laundering (AML), and predictive risk analysis. Synthetic data also enable robust testing of extreme outlier scenarios, for example, crisis situations, market fluctuations, or emerging fraud techniques, thereby improving model resilience.
- **Bias reduction and generalisability** — Real-world datasets often reflect systemic biases, under-representing rare events or certain demographic groups. Synthetic data generation can address these gaps by creating more balanced datasets, improving the fairness, generalisability, and accuracy of predictive models across a wider population.

3. Methods of synthetic data generation

Several methods have been developed for synthetic data generation; however, one of the most widely adopted approaches is the use of Generative Adversarial Networks (GANs), which are based on deep learning and were originally proposed by Ian Goodfellow et al. in 2014 [1] (Figure 2). GANs have shown particular success in generating synthetic images, supporting applications such as the detection of medical pathologies [2]. Beyond imaging, GANs have also been applied to the generation of artificial genomes [4], longitudinal electronic health records, and financial time series data [3], enabling organisations to simulate complex real-world processes without exposing sensitive information.

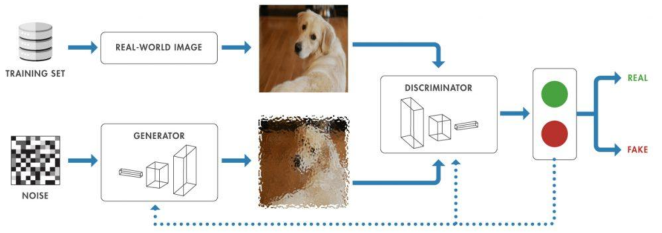

Figure 2. Typical Generative Adversarial Network (GAN) workflow using image discrimination as an example [5]. A GAN comprises two neural network components: the generator and the discriminator. The generator creates synthetic data by starting with random noise (typically a vector of random numbers) and attempting to produce data that resemble the real dataset. The discriminator's role is to evaluate whether a given sample originates from the real training dataset or has been synthetically generated. Through adversarial training, the generator progressively improves its outputs until the discriminator can no longer reliably distinguish between real and synthetic data. This process enables the generator to learn and replicate the statistical characteristics and patterns inherent in real-world data.

4. Challenges

Although synthetic data offer significant potential, several key challenges remain. A critical issue is maintaining a balance between generating highly realistic data (high fidelity) and preserving individual privacy. Ensuring that synthetic data sets cover the full range of real-world scenarios without introducing bias is another major challenge. Furthermore, in sectors such as healthcare and finance, synthetic data must comply with strict data privacy regulations, such as the GDPR in Europe, which may restrict the methods available for data generation and use.

5. Conclusion

Although the use of synthetic data in healthcare and finance remains at an early stage, its potential is considerable. Synthetic data can enable data sharing and improve the accuracy of statistical models whilst maintaining the protection of individual data. However, challenges related to data quality, privacy, and regulatory compliance must be addressed before synthetic data can be deployed at scale in industries handling highly sensitive information.

6. Opportunities for Financial Businesses

Financial organisations are particularly well positioned to benefit from synthetic data solutions. By generating large and representative datasets that reflect real-world market behaviours and fraud patterns, firms can enhance the performance of fraud detection systems, anti-money laundering (AML) models, and risk prediction tools without exposing sensitive client data. Synthetic data also supports regulatory compliance by enabling data-driven innovation while preserving privacy and meeting the requirements of the GDPR and similar frameworks. Furthermore, synthetic data allow robust model testing against rare or hypothetical market events, strengthening the resilience of predictive analytics against financial shocks. By mitigating demographic and geographic biases often embedded in real-world financial data, synthetic datasets also promote the development of more inclusive and generalisable financial services and risk models.

7. Contact Novalytics for More Information

Novalytics specialises in advising on privacy-preserving analytics and data governance solutions tailored for SMEs operating within high-risk and regulated sectors. We help organisations drive innovation whilst ensuring regulatory compliance and maintaining the highest standards of data protection.

For additional details on your data strategy, data protection compliance, or data analytics, please contact us via:

- Website: https://www.novalytics.co.uk
- Email: contact@novalytics.co.uk

References

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, *et al.*, *Generative adversarial networks*, arXiv preprint arXiv:1406.2661, Accessed March 9, 2025, 2014. [Online]. Available: http://arxiv.org/abs/1406.2661.

[2] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, p. 101 552, 2019. DOI: 10.1016/j.media.2019.101552.

[3] H. Buehler, B. Horvath, T. Lyons, I. Perez Arribas, and B. Wood, *Generating financial markets with signatures*, SSRN Electronic Journal, Accessed March 9, 2025, 2020. [Online]. Available: https://www.ssrn.com/abstract=3657366.

[4] B. Yelmen, A. Decelle, L. Ongaro, D. Marnetto, C. Tallec, F. Montinaro, *et al.*, "Creating artificial human genomes using generative neural networks," *PLoS Genetics*, vol. 17, no. 2, e1009303, 2021. DOI: 10.1371/journal.pgen.1009303.

[5] H. Al-Obaidi and S. Kurnaz, "Divergent cnn architectures for novel image generation: A dual-pipeline approach," *Wireless Personal Communications*, 2023, Accessed March 9, 2025. DOI: 10.1007/s11277-023-10758-w.

[6] V. Pezoulas, D. Zaridis, E. Mylona, C. Androutsos, K. Apostolidis, N. Tachos, *et al.*, "Synthetic data generation methods in healthcare: A review on open-source tools and methods," *Computational and Structural Biotechnology Journal*, vol. 23, pp. 2892–2910, 2024. DOI: 10.1016/j.csbj.2024.07.005.

[7] V. Potluru, D. Borrajo, A. Coletta, N. Dalmasso, Y. El-Laham, E. Fons, *et al.*, *Synthetic data applications in finance*, arXiv preprint arXiv:2401.00081, Accessed March 9, 2025, 2024. [Online]. Available: http://arxiv.org/abs/2401.00081.

[8] *Regulation - 2016/679 - EN - gdpr - EUR-Lex*, [Online; accessed 28. Apr. 2025], Apr. 2025. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng.