IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 1, November 2024

Comprehensive Review of Synthetic Data Generation Techniques and Their Applicationsin Healthcare, Finance, and Marketing

Prof. U. B. Bhadange[1], Swamini Jadhav[2], Bhagwati Jadhav[3], Sneha Ghatol[4], Piyush Kahale[5]

Guide, Department of Artificial Intelligence and Data Science[1]

Students, Department of Artificial Intelligence and Data Science[2,3,4,5]

Pune Vidyarthi Griha's College of Engineering and Shrikrushna S. Dhamankar Institute of Management, Nashik

Savitribai Phule Pune University (SPPU)

swamineejadhav@gmail.com, bhagwatijadhav271@gmail.com,

snehaghatol01@gmail.com, pyush671@gmail.com, urmila.bhadange@pvgcoe.org

Abstract: *The demand for privacy-preserving, high-quality data has driven the rapid development of synthetic data generation techniques. Data scarcity, privacy regulations, and the need for large-scale datasets are some of the challenges these methods aimto address. Key methodologies for synthetic data generation includeGenerative Adversarial Networks (GANs), Variational Autoen- coders (VAEs), and rule-based systems. This review highlights the strengths, limitations, and practical applications of these techniques across various fields. It also explores ethical considerations related to privacy and fairness, focusing on privacy-preserving models such as differential privacy and federated learning. Despite the potential of synthetic data to overcome major barriers in data-drivenindustries, issues around data fidelity, fairness, and utility remain unresolved. Future research should prioritize the responsible useof synthetic data.*

Keywords: Synthetic data generation, Privacy-preserving data, Generative Adversarial Networks (GANs), Variational Autoencoders(VAEs), Differential privacy

I. INTRODUCTION

Because of the digital era, data is the basis for advancesin machine learning and artificial intelligence. Data-driven technologies are revolutionizing various businesses, improv- ing decision-making procedures, and promoting progress in fields as disparate as marketing, finance, and healthcare. The obstacles that organizations confront as they increasingly use artificial intelligence to derive insights from data include data privacy, scarcity, and strict regulatory limits.

Strict rules on the gathering, use, and sharing of sensitive personal data are enforced by the Health Insurance Portability and Accountability Act (HIPAA) in the US and the General Data Protection Regulation (GDPR) in Europe. While the restrictions are intended to safeguard individuals' privacy, they also erect obstacles in the way of corporations that want access to vast and varied datasets in order to train their AI models. Patient data in the healthcare industry is frequently dispersed, lacking, or unavailable as a result oflegal constraints. The incapacity to leverage comprehensive datasets impedes the improvement of patient outcomes via tailored treatment regimens and enhanced clinical decision-making.

Techniques for creating synthetic data offer a possible way around the problems. Artificial data that imitates the statistical characteristics of real-world data without disclosing private or sensitive information is known as synthetic data. By applying sophisticated computational tools, organizations can produce synthetic datasets that are statistically equivalent to real data. This feature guarantees adherence to data protection regu- lations while enabling enterprises to build machine learning models.

Machine learning methods such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) are employed in the creation of synthetic data. These techniques enable the production of superior synthetic datasets that

DOI: 10.48175/IJARSCT-22066

may be customized to fulfill particular needs. GANs are well- suited for healthcare applications because they produce high- fidelity images, while VAEs are excellent at data imputation and augmentation, which is useful in clinical contexts.

The review's objective is to offer a thorough examinationof methods for creating synthetic data and the uses for them. A overview of generative artificial intelligence models is presented, emphasizing GANs and VAEs. Additionally, ethical issues and privacy-preserving methods that are crucial for guaranteeing the responsible use of synthetic data are discussed, and the usefulness of synthetic data is assessed.

This review offers a balanced perspective on synthetic data'srole in data-driven industries, emphasizing responsible use, ethics, and further research to address current limitations.

II. REVIEW OF LITERATURE

We emphasized the importance of our research issue and theobjectives of this review in the introduction. The literature on methods for creating synthetic data is examined in this section.We present an overview of the state of research now by evaluating a number of studies, highlighting the achievements of earlier efforts and pointing out importantgaps that require attention.

Reviews/Surveys Related to Synthetic Data GenerationTechniques.

A thorough analysis of methods for creating synthetic data has highlighted the pressing issues with real-world data, including its scarcity, quality issues, and privacy-related legal obstacles [1]. This comprehensive analysis covers a numberof industries, such as marketing, banking, and healthcare, and it shows how synthetic data may be used to solve problems with data scarcity and governance. However, the review pointsout certain shortcomings. The main reason is that it is difficult to assess the relative efficacy of the ways presented because there are no empirical benchmarks available to compare the performance of various synthetic data generation techniques. Furthermore, the emphasis on complex generative models mayobscure other straightforward, interpretable options that could be more suitable for particular uses.

Comparably, a different review concentrated on the potential applications of synthetic data generation in the medical field, particularly in the creation of medical images and patient records [2]. The study examined a variety of approaches, such as GANs and VAEs, but its insights are constrainedby a narrow focus that leaves out cutting-edge methods in the quickly developing field of synthetic data generation. Furthermore, the findings tend to concentrate on specific typesof health data, which may restrict their applicability across broader healthcare settings.

A recent study in the field of finance examined the applica- tion of generative models to create artificial financial datasets, highlighting its potential to improve prediction models [3]. Although the report points out important developments in this field, it also notes that issues with the reliability and qualityof the produced data still need to be resolved. The reliabilityof these models in real-world applications may be hampered by the persistent challenge of precisely assessing the qualityof synthetic data, even in spite of advances in predicting performance.

Large language models (LLMs) and their use in creating synthetic data, especially for text-based applications, were the subject of another study [4]. This research explores how LLMs can help with problems related to the quantity and quality of data, but it also points out that the main emphasis on textual data restricts the investigation of other data modalities, such photos and structured data. This restricted focus could limit the findings' wider usefulness in a variety of industries that need synthetic data solutions.

The development of synthetic data generation techniques has been extensively reviewed in other studies [5], moving from conventional probabilistic models to intricate deep learning techniques like GANs and VAEs. These studies do, however, frequently have drawbacks, such as the absence of standard evaluation metrics that make comparing various approaches more difficult. Additionally, they frequently fail to take into account the practical aspects of using these modelsin actual situations, which are crucial for directing future research.

Furthermore, a comprehensive study on the generation and use of synthetic datasets in diverse applications has emphasized the significance of both conventional statistical approaches and cutting-edge machine learning strategies [6]. Although this review offers a strong basis, it is deficient inits discussion of the long-term effects of employing synthetic data in practice as well as the scalability of synthetic data generation across various data types.

DOI: 10.48175/IJARSCT-22066

All of the study's limitations point to the necessity ofa more thorough literature review that fills in the gaps and resolves contradictions in the body of knowledge by synthesizing the body of research. This thorough analysis attempts to cover all important technologies and models used for synthetic data generation, including GANs, VAEs, and LLMs, in contrast to earlier evaluations that have only concentrated on particular techniques or applications. We hope to provide a detailed understanding of each model'sadvantages and disadvantages by contrasting them based on a range of factors, including performance, scalability, diversity of data, and practical implementation. This will help future researchers and applications in the fields of finance, marketing, healthcare, and artificial intelligence make well-informed decisions.

Reviews/Surveys Related to Applications of SyntheticData

In order to solve data scarcity and privacy problems, syn- thetic data has substantial potential in the healthcare and financial industries. Techniques like GANs and VAEs are frequently utilized in the healthcare industry to create artificialpatient records and medical images. These methods enhance diagnostic tools and predictive models while also providing solutions to privacy regulations [7]. But as the review in [5] makes clear, it mostly concentrates on these well-established models, ignoring more recent approaches and newly emerging data sources, such as genomic data, which restricts its wider use in the healthcare industry.

Synthetic data is also being utilized in finance to enhance prediction models, including those for fraud detection and credit scoring [8]. Sensitive data is kept hidden by creat- ing artificial financial datasets using GANs and VAEs that mimic real-world data. Nevertheless, [9] draws attention toan ongoing difficulty in assessing the validity and quality of this synthetic data, posing questions regarding bias and the reliability of the model, especially in crucial financial decision-making processes.

Synthetic data, especially that produced by large language models (LLMs), is used in marketing to provide product reviews, optimize advertising campaigns, and mimic user behavior [10]. LLMs provide a means of addressing the lackof data in text-based applications and guaranteeing adherence to privacy regulations such as GDPR. However, points outthat the limited application of synthetic data in marketing is due to its narrow concentration on textual data, particularly when it comes to other data types like photos or structured information.

GAN-generated synthetic data has been used in agriculture for applications such as disease diagnosis and crop monitoring[11]. In situations when real-world data is scarce, these meth- ods enable researchers to improve model training by adding existing agricultural datasets. But as mentioned in [11], this evaluation primarily concentrates on GANs, possibly ignoring other workable models such as VAEs or conventional statistical techniques. It also notes difficulties in implementing artificial intelligence (AI) models in actual agricultural environments, along with a lack of effective solutions.

Advanced machine learning techniques such as GANs, VAEs, and LLMs have replaced old statistical methods in the evolution of synthetic data creation techniques [12]. But as studies like as [10] highlight, there is a dearth of consistent evaluation metrics, which makes cross-domain model compar- isons challenging. Furthermore, these assessments frequently omit to address the difficulties in implementing synthetic data generation in real-world applications across industriesincluding finance, healthcare, and agriculture, as well as the approaches' practical scalability.

III. RESULTS

Synthetic data production techniques in marketing, finance, and healthcare have advanced significantly, yet there are still issues that need to be addressed. GANs and VAEs have demonstrated significant promise in the healthcare industry for creating artificial patient records and medical imaging. By avoiding sensitive data exposure, these techniques assist in addressing privacy concerns regulated by legislation such as HIPAA[5]. This permits the construction of prediction modelsand diagnostics. However, a large portion of the study ignores more recent approaches that may be able to handle newly emerging data types, such genomic and multi-modal healthcaredata, in favor of more well-established strategies like GANs and VAEs. This restriction limits the synthetic data's wider usefulness in different healthcare situations [5].

Synthetic data has been frequently utilized in finance to enhance models used for risk assessment, fraud detection,and credit scoring. Financial datasets that closely resemble real-world data can be produced with great success using

DOI: 10.48175/IJARSCT-22066

techniques like GANs and VAEs, all while protecting sensitive information. Ongoing difficulties, however, include evaluatingthe authenticity and quality of the synthetic data, which is still challenging and raises questions around bias and dependabil- ity in high-stakes financial applications [8]. Large language models (LLMs) have been used in marketing to produce arti- ficial customer behavior data that helps with market research and advertising optimization. Although this strategy assures compliance with privacy rules like as GDPR and helps with data scarcity, its limited use of text-based data restricts its usefulness in marketing, where other data kinds like photos and structured datasets also play important roles. [12].

By enhancing actual agricultural datasets, GANs have been used in agriculture to do tasks like disease diagnosis and crop monitoring. On the other hand, little is known aboutthe difficulties in implementing and expanding these modelsin practical settings [11]. The use of synthetic data has increased across all industries thanks to the progression from conventional statistical methods to sophisticated models like GANs, VAEs, and LLMs; however, reviews frequently high- light the absence of common evaluation metrics, which makesit challenging to compare the efficacy of various approaches. Moreover, scalability problems and practical deployment ob- stacles prevent synthetic data from reaching its full potentialin practical applications.

IV. CHALLENGES AND FUTURE DIRECTIONS

Bias and fairness in the creation of synthetic data present a substantial additional challenge. The synthetic data will inheritthe biases of the real-world datasets used to train generative models, which could result in skewed outcomes in downstreamapplications like financial predictions or healthcare diagnoses.As synthetic data is used more frequently in decision-making processes that have an immediate impact on society, such credit scoring and healthcare treatment plans, it is especially crucial to address bias [13]. More reliable methods must be developed by researchers to reduce bias and guarantee that artificial data is representative and fair.

Furthermore, assessing artificial intelligence data continues to be a critical concern. As of right now, there’s no acknowledged metric to evaluate the quality of synthetic data. While some research suggests utilizing criteria like model performance gains or resemblance to real data, these methods fall short of capturing the subtleties of data quality, such as privacy protection and usefulness in certain applications. The absence of uniform assessment frameworks makes it more difficult to compare various synthetic data generation methodsand assess their applicability in various sectors [14][15].

Privacy is still a major issue. While synthetic data is intended to conceal sensitive information, new research in- dicates that it may be possible to reverse-engineer synthetic data to reveal actual data in specific situations. To avoid privacy breaches, it is crucial to have strong privacy-preserving techniques in place while creating synthetic data, such as federated learning and differential privacy [16]. Furthermore, weighing the trade-off between data privacy and utility is still an important area for research, especially in industries like finance and healthcare where maintaining data privacy and accuracy is crucial[5].

Looking ahead, the creation of complex evaluation metrics for synthetic data quality that incorporate privacy, fairness, and practical applicability in addition to standard statistical mea- surements should be given top priority in future research[17]. Enhancing synthetic data models’ scalability and developing workable deployment techniques should also be research pri- orities, particularly in situations with limited resources. Lastly, interdisciplinary cooperation will be essential to creating legal frameworks that direct the ethical and responsible use of synthetic data, especially in highly regulated sectors like finance and healthcare[18].

V. CONCLUSION

In sectors including healthcare, banking, marketing, and agriculture, synthetic data production has become a potent instrument that may help with issues including data shortages, privacy concerns, and legal obstacles. Methods such as GANs, VAEs, and LLMs have demonstrated efficacy in generating synthetic datasets that promote creativity while protecting private data. Significant obstacles still need to be overcome, though, such as guaranteeing data fidelity, reducing preju- dice, and creating uniform evaluation standards. Further pri- vacy problems, such as the possibility of reverse-engineering synthetic data, underscore the need for enhanced privacy- preserving methods such as federated learning and differential privacy.

Future studies need to close these gaps by creating more complex models for creating synthetic data that faithfully capture the intricacy of real-world data. Moreover, realizing the full potential of synthetic data will depend on building strong assessment frameworks, reducing bias, and improving scalability. Establishing ethical principles and legal frame- works that guarantee the proper use of synthetic data across businesses—especially in industries where accuracy and fair- ness are critical, like healthcare and finance—will require interdisciplinary collaboration. In the end, new applications enabled by sustained advancement in synthetic data techniques promise to revolutionize data-driven fields while upholding privacy and regulatory compliance.

REFERENCES

[1] Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., Fu, T., Wei, W. (2024). Machine Learning for Synthetic Data Generation: A Review. Journal of Artificial Intelligence Research.

[2] Skandarani, Y., Jodoin, P.-M., Lalande, A. (2024). GANs for Medical Image Synthesis: An Empirical Study. Journal of Medical Imaging and Artificial Intelligence Research.

[3] Assefa, S., Dervovic, D., Mahfouz, M., Reddy, P., Veloso, M. (2019). Generating Synthetic Data in Finance: Opportunities, Challenges, and Pitfalls. Journal of Financial Data Science.

[4] Guo, X., Chen, Y. (2024). Generative AI for Synthetic Data Generation: Methods, Challenges, and the Future. Journal of Artificial Intelligence Research.

[5] Jadon, A., Kumar, S. (2023). Leveraging Generative AI Models for Syn- thetic Data Generation in Healthcare: Balancing Research and Privacy. Journal of Healthcare Data Science and Privacy.

[6] D'Amico, S., Dall'Olio, D., Sala, C., Dall'Olio, L., Sauta, E., Zampini, M., Asti, G., Lanino, L., Maggioni, G., Campagna, A., Ubezio, M., Russo, A., Bicchieri, M. E., Riva, E., Tentori, C. A., Travaglino, E., Morandini, P., Savevski, V., Santoro, A., Prada-Luengo, I., Krogh, A., Santini, V., Kordasti, S., Platzbecker, U., Diez-Campelo, M., Fenaux, P., Haferlach, T., Castellani, G., Della Porta, M. G. (2024). Synthetic Data Generation by Artificial Intelligence to Accelerate Research and Precision Medicine in Hematology. Journal of Hematology and Precision Medicine Research.

[7] Potluru, V. K., Borrajo, D., Coletta, A., Dalmasso, N., El-Laham, Y., Fons, E., Ghassemi, M., Gopalakrishnan, S., Gosai, V., Kreacˇic´, E., Mani, G., Obitayo, S., Paramanand, D., Raman, N., Solonin, M., Sood, S., Vyetrenko, S., Zhu, H., Veloso, M., Balch, T. (2024). Synthetic Data Applications in Finance. Journal of Financial Artificial Intelligence Research.

[8] Xuereb, S. (2023). Generation of Synthetic Data to Improve Financial Prediction Models. Journal of Financial Data Science and Prediction.

[9] Mistry, H. M. (2024). Optimizing Fraud Detection Models with Synthetic Data: Advancements and Challenges. International Journal of Advanced Research, 1(2), 249-264.

[10] Brand, J., Israeli, A., Ngwe, D. (2023). Using LLMs for Market Research. Journal of Marketing Research and Data Science.

[11] Akkem, Y., Biswas, S. K., Varanasi, A. (2024). A Comprehensive Review of Synthetic Data Generation in Smart Farming by Using Variational Autoencoder and Generative Adversarial Network. Journal of Smart Agriculture and Data Science.

[12] Goyal, M., Mahmoud, Q. H. (2024). A Systematic Review of Synthetic Data Generation Techniques Using Generative AI. Journal of Artificial Intelligence Research and Data Science.

[13] Gianfrancesco, M. A., Tamang, S., Yazdany, J., Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. JAMA Internal Medicine, 178(11), 1544–1547.

[14] Anonymous Authors (2004). Really Useful Synthetic Data. A Framework to Evaluate the Quality of Differentially Private Synthetic Data.

[15] Pereira, M., Kshirsagar, M., Mukherjee, S., Dodhia, R., Lavista Ferres, J., de Sousa, R. (2024). Assessment of differentially private synthetic data for utility and fairness in end-to-end machine learning pipelines for tabular data.

[16] Hasan, J. (2024). Security and Privacy Issues of Federated Learning.

[17] Hyrup, T., Zimek, A., Schneider-Kamp, P. (2024). SynthEval: A Frame- work for Detailed Utility and Privacy Evaluation of Tabular Synthetic Data. Journal of Artificial Intelligence Research.

[18] Bhanot, K., Qi, M., Erickson, J. S., Guyon, I., Bennett, K. P. (2021). The Problem of Fairness in Synthetic Healthcare Data. Entropy, 23(9), 1165. https://doi.org/10.3390/e23091165