IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

Synthetic Data Generation for Privacy-Preserving

Prof. Gujjala Usha[1], Chinthana P[2], Rakshitha G[3], V Haritha[4] & Taqiya Taskeen[5]
Department of AI & ML[1-5]
Proudhadevaraya Institute of Technology, Hospet, India
[1] gujjal.usha@gmail.com[2] pchinthana85@gmail.com
[3] rakshithag765@gmail.com, hh6575206@gmail.com, taqiyataskeen7@gmail.com

Abstract: *The growing dependence on data for machine learning and analytics has increased concerns about privacy and secure data sharing. Many real-world datasets contain sensitive information that cannot be openly distributed. The proposed study examines how artificially generated datasets can serve as an effective approach for overcoming these limitations. The proposed system creates artificial datasets that reflect the important characteristics of the original data while ensuring that confidential information is not revealed. Different generative techniques are used to learn patterns from existing data and produce realistic synthetic records. These datasets can be utilized for model training, system validation, analytical studies, and experimental tasks while safeguarding sensitive information. The work demonstrates that synthetic data can support data-driven applications while maintaining security and compliance with privacy requirements.*

Keywords: Synthetic Data Generation for Privacy-Preserving

I. INTRODUCTION

The continuous advancement of modern computing and digital systems has made data an essential resource for innovation and decision-making. Organizations rely on data to improve services, develop intelligent systems, and support decision-making processes. However, much of this information contains personal or confidential details that require protection. The sharing or processing of confidential information without appropriate security controls may result in privacy breaches and legal challenges. To overcome these challenges, researchers have developed techniques that enable data-driven operations while protecting confidential records. A promising method is the creation of synthetic datasets. Rather than sharing real-world information, computer-generated data is produced to replicate the patterns and characteristics of the original dataset. This approach enables organizations to conduct analytical tasks, build predictive systems, and evaluate machine learning solutions while minimizing privacy-related risks. The primary goal of this project is to design a secure synthetic data generation framework capable of producing realistic, high-quality, and privacy-preserving datasets.

II. LITERATURE SURVEY

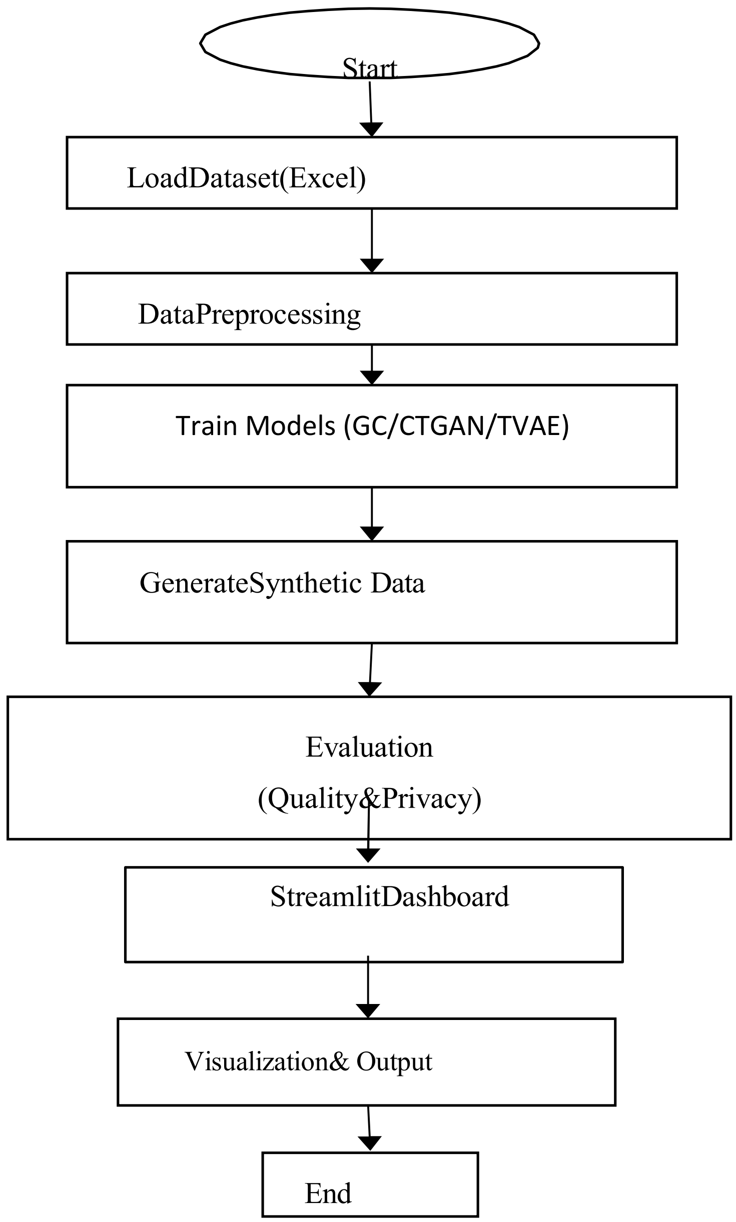
Researchers have proposed several techniques to protect sensitive information while maintaining data usability. Traditional approaches such as data masking, encryption, and anonymization help reduce privacy risks but often affect data quality and may still allow re-identification under certain conditions. Recent studies have explored synthetic data generation as an alternative solution. Statistical models were initially used to create artificial datasets, but they struggled to capture complex relationships. With advancements inmachine learning, models such as Generative Adversarial Networks (GANs), Variational Autoencoder architectures (VAEs) together with CTGAN are capable of producing synthetic datasets that closely reflect the patterns and characteristics of real-world data. Current research indicates that synthetic data can assist in the creation and evaluation of machine learning solutions testing, and research activities while preserving privacy. These findings encourage the design of highly effective and secure synthetic data generation systems.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

III. RESEARCH OBJECTIVES

This project aims to study how synthetic data can be generated from real-world datasets while safeguarding sensitive and private information.The study focuses on creating artificial records that preserve important data characteristics while protecting individual privacy. Another objective is to examine different synthetic data generation models and compare their ability to produce realistic outputs. The project also evaluates whether the generated data can be effectively used for machine learning applications, testing activities, and research purposes. special attention is given to achieving an effective trade-off between data protection and data quality of synthetic data in real-world environments.

IV. FLOW CHART



IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

V. FUTURE SCOPE

Synthetic data generation is expected to play an increasingly important role as organizations continue to rely on large volumes of information for decision-making and innovation. Future advancements may lead to the development of more sophisticated models capable of creating datasets that closely resemble real-world data while maintaining strong privacy standards. Researchers can focus on improving the quality, consistency, and practical usefulness of synthetic datasets so

that they can be applied with greater confidence in sectors such as healthcare, finance, and cybersecurity. Another area of future research involves developing automated techniques for assessing privacy risks and evaluating data quality before synthetic datasets are released. The combination of synthetic data generation with cloud-based platforms and real-time processing technologies may further enhance scalability and efficiency. As data protection regulations continue to evolve, synthetic data is likely to support secure information exchange, collaborative innovation, and responsible AI development. Ongoing research in this field will contribute to better privacy preservation while ensuring that generated data remains valuable for analytical and machine learning applications.

VI. REQUIREMENTS

1. Programming Language:
    - Python

2. Python Libraries:
    - Pandas
    - NumPy
    - Matplotlib
    - Seaborn
    - Scikit-learn

3. Synthetic Data Generation Tools:
    - SDV (Synthetic Data Vault)
    - Gaussian Copula Synthesizer
    - CTGAN Synthesizer
    - TVAE Synthesizer
    - CTGAN library (fallback support)

4. Evaluation Metrics:
    - SD Metrics

5. Machine Learning Testing:
    - XG Boost Classifier
    - Label Encoding for Categorical Features

6. Output and Visualization:
    - Excel Reports
    - CSV Files
    - Statistical Summaries
    - KDE (Kernel Density Estimation) Plots
    - Downloadable Output Files

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

7. Development Environment:
   - Jupyter Notebook
   - Anaconda Distribution

VII. APPLICATIONS

1.Healthcare Sector: Synthetic data can be used for medical research, disease prediction, healthcare analytics, and diagnostic model development without exposing sensitive patient information.

2.Financeand Banking:It supports financial fraud identification, assessment of lending risks, and monitoring of banking activities, along with financial forecasting while protecting customer privacy.

3.E-Commerce and Retail: Synthetic datasets help analyze customer behavior, improve recommendation systems, and support demand forecasting without revealing personal details of consumers.

4.Education and Research: Researchers and students can access realistic datasets for experiments, academic studies, and artificial intelligence training while maintaining data confidentiality.

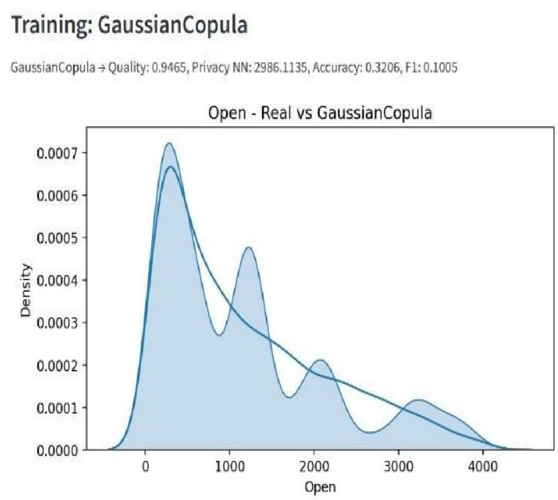
5. Government and Public Policy: Synthetic data enables secure data sharing, policy evaluation, demographic analysis, and decision-making without compromising citizen privacy.

6. Information Technology: It is useful for software testing, database validation, cybersecurity research, and machine learning model development when real data cannot be used directly.

VIII. OUTPUT

1. Gaussain Copula

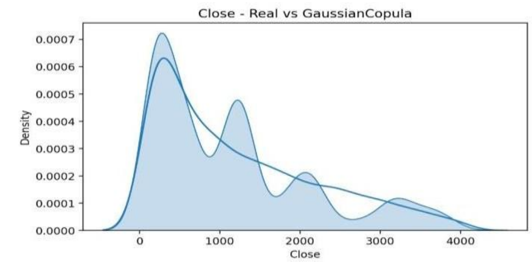
Open:

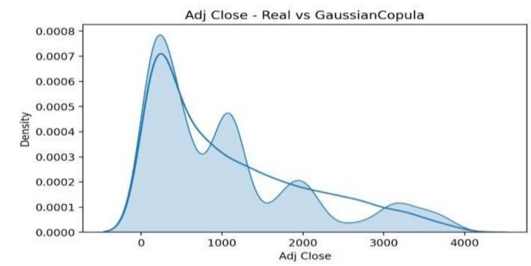


DOI: 10.48175/IJARSCT-37359

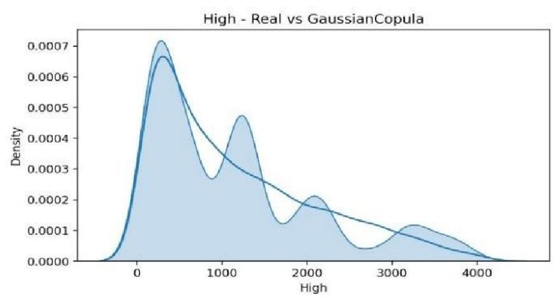
IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

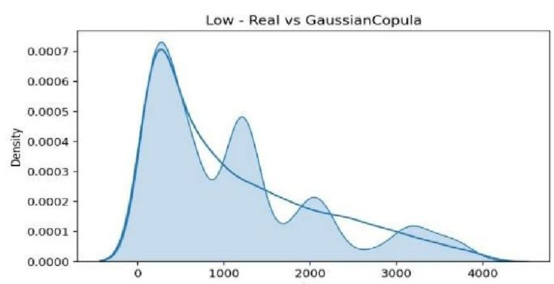
Close & Adjust:





High and low:

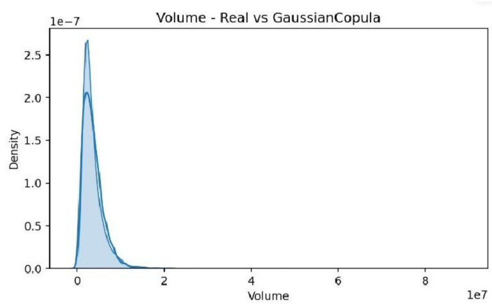




DOI: 10.48175/IJARSCT-37359

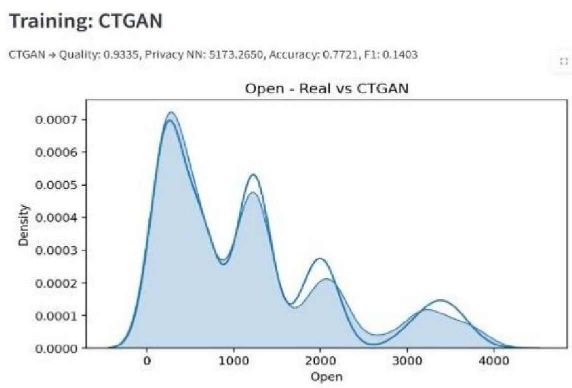
IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

Volume:

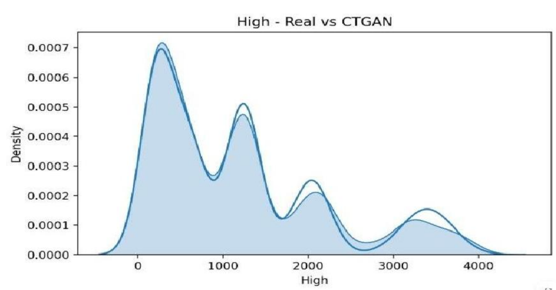


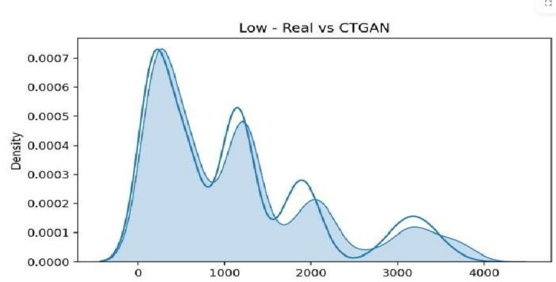
2. CTGAN

Open:



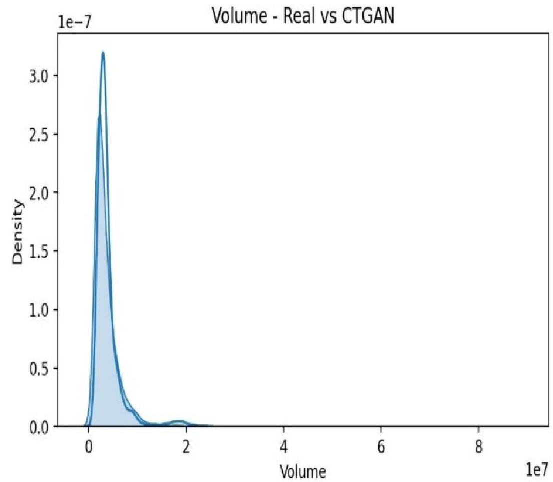
High and low:



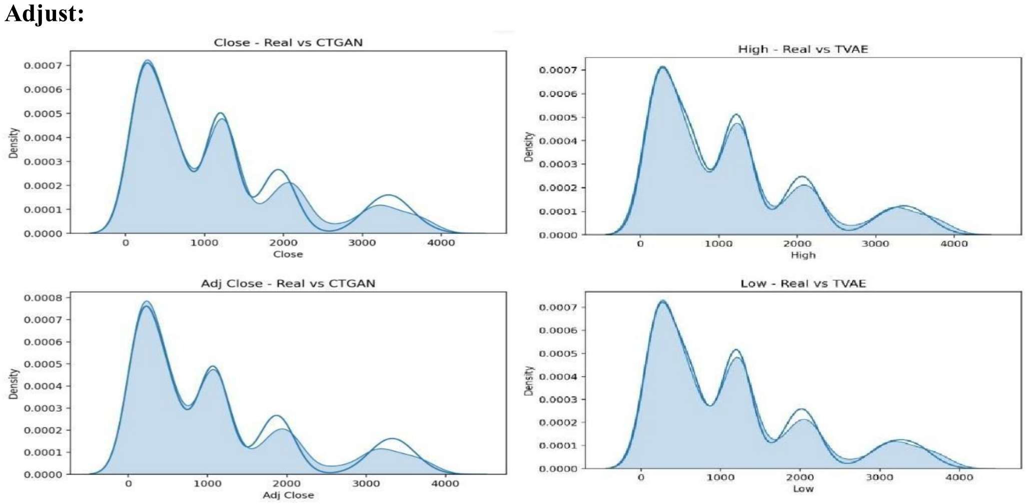


IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

Volume:



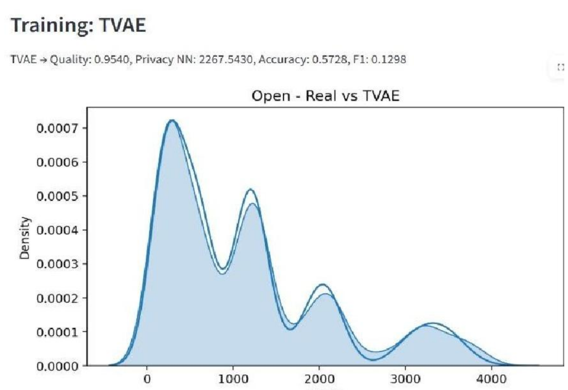
Close & Adjust:



Open:

3. TVAE

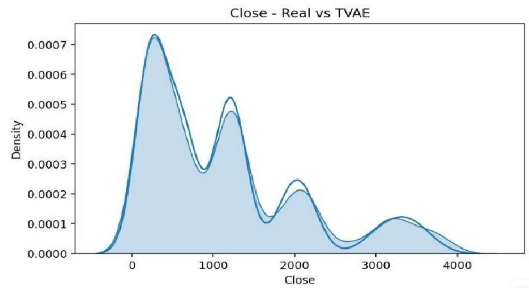
Volume:

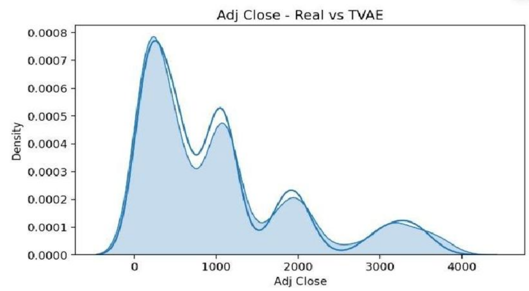


DOI: 10.48175/IJARSCT-37359

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

Close & Adjust:





High and low:

	A	B	C	D	E	F	G
1	Date	Open	High	Low	Close	Adj Close	Volume
2	01-01-2007	243.8	227.227036	204.545718	235.180584	174.633051	3927478
3	01-01-2008	215	172	172.174589	198.414666	121.311243	5055117
4	01-01-2011	252	297.44	294.85	269.092615	217.612572	1609019
5	01-01-2011	575	540	240.9	498.95	424.84311	5688194
6	01-01-2012	608	620.219709	507.907274	790.015039	665.09429	2074964
7	01-01-2012	563.603761	563.746998	265.147875	597.981461	157.281966	2799770
8	01-01-2013	326	289	221.152313	565.304313	521.426096	2013371
9	01-01-2017	1287.5	1270	1226.366192	1194.261053	1138.410773	1569121
10	01-01-2017	1271.5	1240.663499	1246.649124	1213.491017	1119.392244	2135073
11	01-01-2018	1237.494292	1258	1214.040534	1249.207716	1111.456487	1764305
12	01-01-2019	3037	3035	2075.522196	1933.98	2818.499375	4545420
13	01-01-2020	3306	3270.236269	3379.38	3307.317239	1869.53645	3669515
14	01-01-2021	2211.790204	2240.634	2207.559196	2248.655918	1960.555558	3769675
15	01-01-2021	1928.8	2003.8	2113.2	1963.963858	1816.515401	4528259

Real Data:

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 11, June 2026
Impact Factor: 8.2

9. Result Data

Synthetic Data:

	A	B	C	D	E	F	G
1	Date	Open	High	Low	Close	Adj Close	Volume
2	27-08-2004	122.800003	122.800003	119.82	120.332497	88.088272	30646000
3	30-08-2004	121.237503	123.75	120.625	123.345001	90.293549	24465208
4	31-08-2004	123.3125	123.75	122	123.512497	90.416122	21194656
5	01-09-2004	123.75	124.375	122.949997	123.487503	90.39782	19935544
6	02-09-2004	123.737503	125.574997	123.25	124.207497	90.924896	21356352
7	03-09-2004	125.75	137.5	123.794998	124.732498	91.309212	9869856
8	06-09-2004	129.987503	129.987503	124.112503	124.357498	91.034721	9038672
9	07-09-2004	129.375	129.375	124.375	124.449997	91.10244	5772232
10	08-09-2004	124.5	125.099998	123.887497	124.212502	90.928566	6593984
11	09-09-2004	124.625	124.737503	122.307503	122.495003	89.67131	7947184
12	10-09-2004	123.75	123.75	122	123.599998	90.480194	6415072
13	13-09-2004	123.875	126.9375	123.875	125.4375	91.825348	20904912
14	14-09-2004	125.625	127.349998	125.370003	126.957497	92.938019	15335472
15	15-09-2004	127.25	127.5	125.150002	125.762497	92.063225	11988288

IX. CONCLUSION

This project explored the use of synthetic data generation as a privacy-preserving solution for modern data-driven applications. The study showed that synthetic datasets can capture the important patterns of original data without revealing sensitive information. Such datasets can be safely used for machine learning, research, testing, and analysis.

The implementation demonstrated that it is possible to achieve a balance between privacy protection and data usability. Although synthetic data generation still faces challenges related to quality and realism, continuous improvements in machine learning techniques are making it more effective and reliable.

Overall, synthetic data generation represents an important step toward secure data sharing and responsible use of information in various domains. Future developments in this field are expected to further enhance both privacy and data quality.

REFERENCES

1. 1.N. Patki, R. Wedge, and K. Veeramachaneni, "The Synthetic Data Vault," in 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada, 2016, pp. 399–410.
2. 2.L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling Tabular Data Using Conditional GAN," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019.
3. D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," arXiv preprint arXiv:1312.6114, 2013.
4. R. Xu, N. Baracaldo, Y. Zhou, A. Anwar, J. Joshi, and H. Ludwig, "Federated Learning for Healthcare Informatics," Journal of Healthcare Informatics Research, vol. 5, no. 1, pp. 1–19, 2021.
5. C. Dwork, "Differential Privacy," in Automata, Languages and Programming, Berlin, Germany: Springer, 2006, pp. 1–12.
6. A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, and K. Bennett, "PrivacyPreserving Synthetic Health Data," in ESANN 2020 Proceedings, Bruges, Belgium, 2020.

7. SDV Development Team, "Synthetic Data Vault (SDV)," 2025. [Online]. Available:https://github.com/sdvdev/SDV. [Accessed: Jun. 24, 2026].
8. S. Bond-Taylor, P. Leach, Y. Long, and C. G. Willcocks, "Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 11, pp. 7327–7347, 20.