IJSRA

International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(Review Article)

Check for updates

The spectrum of synthetic data generation: A comprehensive review

Satya Sudha S, Grishitha V *, Sai Rajeshwar V V N and Shiva Karthik P

Department of CSE (Artificial Intelligence and Machine Learning), ACE Engineering College, India.

International Journal of Science and Research Archive, 2025, 14(01), 1124-1128

Publication history: Received on 09 December 2024; revised on 17 January 2025; accepted on 20 January 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.14.1.0192

Abstract

Synthetic data, which is the data produced to mimic the characteristics of actual data without revealing any confidential information, is a much safer option than original data, especially when it comes to extreme instances such as personal data, financial data, or military intelligence. There are substantial dangers connected with the use of real-life data such as assault with the intent to commit identity theft, fraud, and hacking, but because synthetic data (SD) reproduces some of the elements of real data, without infringing on anyone's privacy, suffers from these risks. The project concentrates on the cutting-edge fields of Language Learning Models (LLM) and Deep Learning (DL) to generate synthetic data that mimics real-world data in its intricacy. Advances in LSTM networks and Generative Adversarial Networks (GAN) produce plausible and useful data in sequence forms for natural language processing and machine learning (ML) augmentation respectively. Applications of this technology include, but are not limited to, the use of augmented datasets to improve medical diagnosis, advanced finance fraud detection systems, and designing fictitious consumers in order to enhance AI-based system recommendations. The project which is implemented with Python programming language and also takes advantage of some open source packages such as SymPy, Pydbgen, Synthetic Data Vault (SDV), and Scikit-learn offers a solution to data scarcity and quality problems in order to improve the performance of the AI models in various sectors.

Keywords: Deep Learning; Large Language Models (LLM); Synthetic Data; Long - Short Term Memory (LSTM); Generative Adversarial Networks (GAN)

1. Introduction

The term synthetic data refers to algorithmically generated artificial data. Synthetic data is designed to mimic the characteristics of real data but contains no actual information itself. Used quite commonly in data science and machine learning, synthetic data enables testing algorithms and making improvements without risking the risk to privacy or confidentiality due to real-world data. It may also be employed as augmentation for an existing dataset, especially when the original data seems limited or biased.

Synthetic data offers several advantages, making it an attractive alternative to natural data. It addresses critical issues such as privacy concerns, data bias, and high acquisition costs, providing a viable solution for industries dependent on large- scale datasets. Synthetic data can be structured, unstructured, or semi-structured, depending on the domain and application. GANs use two neural networks—a generator and a discriminator—in a feedback loop to create data samples that mimic the real data distribution.

TGAN focuses on the generation of tabular data that uses both continuous and categorical variables. It learns the distribution of tabular datasets and generates data samples that maintain the properties of the original data set. It further relies on LSTM to model sequential dependencies in tabular data for column- wise data generation to capture feature relationships.

* Corresponding author: Grishitha V.

CGAN generates synthetic data for specific variables or categories. it useful for tabular data with imbalanced distributions. It is specifically used for specific scenarios and distributions. LSTMs are recurrent neural networks designed for sequential or time-series data generation. LSTMs are often integrated with TGANs or CGANs when generating tabular data such as financial transactions, stock price simulations, or patient monitoring records. CT GAN can leverage the output of LSTM for better conditioning on time series.

On the other hand, LLMs are utilized for text data generation but can be applied to other forms of data using prompting. They can be utilized for generating descriptive labels or narrative content for datasets. They can be especially utilized for GAN-based techniques by managing the unstructured data generation.

These techniques together form a harmonious framework for synthetic data generation across various domains to provide realistic datasets for healthcare and finance applications to autonomous systems and beyond.

Autonomous Vehicles - Companies like Waymo and Tesla use synthetic data to train self-driving algorithms. By creating virtual environments that replicate real-world scenarios, these companies enable their algorithms to learn how to respond effectively to diverse situations without the risks associated with real-world testing.

Healthcare - Synthetic data is instrumental in generating realistic health records for research purposes. It allows researchers to work with data that mirrors the statistical properties of actual patient data while preserving patient privacy. For instance, synthetic images of organs or tissues can train algorithms to recognize patterns and detect abnormalities in real patient images. This approach facilitates accurate diagnoses and treatment planning without requiring vast amounts of sensitive real-world data.

Finance - Synthetic data can simulate financial markets, enabling businesses to test and refine trading strategies without relying on historical market data.

As synthetic data continues to become more scalable, private, and high-quality datasets prove to be a foundation of innovation. Synthetic data bridges gaps in the traditionally available and limited quality data. With AI and machine learning, it can empower businesses to break complex challenges with confidence and efficiency.

2. Why Synthetic Data

The exponential growth of technologies based on data has brought to the forefront the significance of data in many industries. Synthetic data, derived through artificial means that include algorithms, machine learning models, or simulations, holds some promise and is a good replacement for real-world data. Its ability to address a number of critical challenges related to privacy, data scarcity, and scalability makes it an indispensable tool for applications in Data Analytics and Artificial Intelligence (AI). Its necessity must, therefore be understood in order to advance technological developments in a manner that is ethically and practically sound.

Probably the most significant catalyst for synthetic data usage is the increasing concern about data privacy and governmental compliance with regulations such as Central Consumer Protection Authority (CCPA). The generated data is very advantageous for various Machine Learning (ML) algorithms. Real data often contains sensitive information and can be a source of privacy breaches if not used in a proper manner. Synthetic data, since it is artificially generated, can mimic some statistical properties of real data without replicating specific personal details, therefore to comply with the privacy of data. This makes it highly valuable in industries such as health care and finance, where privacy is paramount but access to high-quality data is indispensable. The European law, General Data Protection Regulation (GDPR) ensures that their personal data is being protected in stake of their privacy concerns.

Hence, synthetic data serves more purposes than being an almost real alternates to replace real-world data. It then becomes a solution to the most sensitive pain of the modern data-driven industries. It is a revolutionary technology that will find solutions to the issues of data privacy, availability, scalability, and fairness, making it an indispensable tool within the technological boundaries. Synthetic data will ensure that organizations operate within tight privacy compliances, and thus protect the rights of individuals and bring about trust and transparency in using the data. Synthetic data fills the gaps in accessibility to data as it generates diverse and representative datasets to overcome the constraints of scarcity or imbalance in data availability. This flexibility allows the machine learning models and algorithms to work reproducibly and inclusively over all scenarios. It accelerates experimentation and prototype development by reducing reliance on the expensive and time- consuming need to collect real-world data. This ability is very critical in advanced fields such as autonomous systems, health care, and financial technology, wherein fast growth depends on having good, scalable-quality data. In addition, synthetic data allows researchers and developers to practice infrequently or extreme scenarios, which leads to the generation of very robust models capable of managing edge cases well. It builds not only the performance of systems but also their dependability [1] in real-world applications.

3. Generating Synthetic Data

3.1. Generative Adversarial Networks (GAN)

In Neural Networks, Generative adversarial networks are another powerful networks that are applied for unsupervised learning. They are of two components, A) Discriminator B) Generator. They apply the principle of adversarial training to produce Synthetic Data. Generative Adversarial Networks are very effective in generating realistic synthetic data by learning the underlying distribution of real-world datasets. They find [2] widespread applications when the real data cannot be practically or expensively collected or even when it's sensitive. GANs generate synthetic data using two networks, namely the generator, which generates samples of data, and the discriminator, which assesses them for authenticity. Ap- plications include image synthesis, such as creating medical imaging datasets (e.g., MRI scans) for healthcare applications, generating realistic text or audio, or producing synthetic video sequences. GANs can also balance out data imbalances in the training dataset by generating more samples for under- represented classes.

Use of CT-GAN for Tabular Data Conditional Tabular GAN (CT-GAN) is a sub - class of GANs tailored specifically for tabular data generation. Contrary to the traditional GAN, which are really great at generating image and audio data, CT-GAN is designed for numerical, categorical, or mixed tabular datasets. It generates synthetic records that can retain the statistical relationships and patterns found in the original data set. This is particularly [3] valuable in business spheres such as finance, health care, and customer analytics, as tabular datasets usually contain sensitive information. CT-GAN upholds data privacy since it generates synthetic datasets that appear similar to real ones without unveiling the original sensitive records, thus allowing safe data sharing and robust machine learning model training.

Use of TGAN for Sequential Data: Temporal GAN (TGAN) is designed to generate synthetic time-series data while preserving temporal dependencies. It is widely applied in domains such as finance (e.g., stock market simulations), health care (e.g., patient monitoring data), and IoT (e.g., sensor readings). TGAN models the sequential nature of time-series data and captures patterns such as trends, seasonality, and periodicity. This capability makes TGAN highly valuable for training Machine Learning models in scenarios where historical data is incomplete or unavailable. [4] Enables better forecasting, anomaly detection, and simulation of rare events, providing a robust foundation for time-series-based AI applications.

3.2. Large Language Models (LLM)

LLMs can process textual data, handle numerical data, and do value prediction work. For example, it showed that LLMs can better classify tasks based on tabular data than the traditional deep learning-based tabular classification. [5] Studied the structural understanding abilities of LLMs and discovered that using prompting techniques to inform structural completeness in tabular data may, in turn, enhance the performance of LLMs for several tabular tasks. These advanced Artificial Intelligence systems are primarily designed to process and generate humanized text by leveraging various deep learning techniques and typically based on transformer architectures. These models, such as GPT, BERT, or PaLM, are processed with on enormous [6] masses of textual data and can therefore understand very complicated linguistic patterns, contexts, and semantics. LLMs excel in a variety of NLP tasks such as text generation, translation, summarization, question answering, and code generation. Their generalization capabilities across domains make them highly adaptable for applications in health, educational, and financial sectors. This is because LLMs are widely used for synthetic data generation to better replicate and mimic complex data patterns.

Well, trained on vast amounts of diverse and extensive datasets, LLMs can create realistic synthetic data across a wide range of domains, in text, in structures of data, and in code. Their ability to simulate complex dependencies among variables in the data makes them applicable for creating rich and quality datasets that recreate few of the statistical properties of original data. Furthermore, LLMs enable privacy-preserving synthetic data generation [7] as they produce distributions that mimic the real world while not revealing any sensitive information. This capability is also very valuable in areas like health and finance, where data privacy and regulatory compliance are important. With LLMs, researchers and organizations can augment otherwise limiting datasets, improve model training, and enable robust data-driven innovation.

3.3. Long - Short Term Memory (LSTM)

LSTMs are a sub - class of RNNs (Recurrent Neural Networks) that is formulated to learn and process sequential data over long time intervals to bypass the vanishing and exploding gradient problems common in standard RNNs.

Why Use LSTMs for Data Generation?

It retains the sequential nature, and dependencies present in the data, hence, outputting synthetic which are coherent to real outputs and is flexible to generate data from many domains such as finance, health, and activity recognition. While handling Data Scarcity - Synthetic sequences enrich the datasets so that models are trained without gathering fresh data from reality.

LSTMs create synthetic data by learning patterns and dependencies of time from the source dataset and then repeating similar sequences namely A) Training Phase B) Synthetic Data Generation phase. Using the VAE-LSTM-based augmentation approach time-series data sets are synthetically generated. The VAE component captured latent representations of the time-series data, and the LSTM network generated realistic temporal sequences. This method addressed the data scarcity challenge effectively and improved the training process. Incorporation VAE-LSTM generated data with raw data in a 1:1 ratio greatly

This research employs a recursive LSTM network to project the time courses of intrinsic connectivity networks. Deep learning models such as multi-channel CNNs, CNNs with attention mechanisms, and time-attention LSTMs were used for chronological age prediction tasks on the comparison between augmented data sets empirically and the original one. It was found that the augmentation resulted in a marked improvement in prediction accuracy.

The stateless LSTM employed a single layer with hidden size 50 and reset its hidden states at the end of each batch to avoid interference. The recursive LSTM predicts one time point per instance, using it iteratively to forecast subsequent points. This adaptive multi-step strategy captures temporal dependencies efficiently, which is one of the reasons why the augmented dataset outperformed.

3.4. Synthetic Data Vault (SDV)

SDVs are a framework for generating and using synthetic data in machine learning, data science, and research, with the purpose of providing modeling, synthesis, and evaluation tools and libraries specifically for datasets, especially in cases of restricted access to real-world data due to privacy or security concerns. Some of the key features: A) Data Modeling B) Data Generation C) Data Assessment.

The Synthetic Data Vault (SDV) framework improves syn- thetic data generation by providing the solution to missing values in a real-world data set. When SDV detects any column containing missing values, it generates two new columns: A) Input Data Column B) Missingness Indicator Column. It allows SDV to create complete and representative synthetic data sets while maintaining the integrity of missing data patterns. The imputation step ensures realistic data, and the indicator column enables models to account for missingness, facilitating better handling of incomplete data in the context of machine learning tasks. The essence of handling missing values while generating synthetic data improves the data completeness of the generated synthetic data, supports privacy- preserving practices, and especially allows more robust model training as missing data patterns are explicitly managed.

4. Conclusion

Generating data that mimics the real - world data is truly a work that is emphasized in Artificial Intelligence (AI) world. This review paper thoroughly interprets various ways of creating the synthetic data, importance, and endorse various Machine Learning (ML) and AI engineers about the significant use in training and testing experimental models for solving complex algorithms. This paper discusses various ways to create SD by using various LLMs and GANs. Synthetic Data can also be generated by Synthetic Data Vault (SDV), for creating a structured data (Tabular Data). SD is also created Long Short Term Memory (LSTM) which are mostly used for analysing sequential data like time - series.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] G. Albuquerque, T. Lowe, and M. Magnor, “Synthetic generation of high- dimensional datasets,” IEEE Transactions on Visualization and Computer Graphics, vol. 17, no. 12, pp. 2317–2324, 2011.

[2] C. Esteban, S. L. Hyland, and G. Ra¨tsch, “Real-valued (medical) time series generation with recurrent conditional gans,” 2017. [Online]. Available: https://arxiv.org/abs/1706.02633

[3] Y. E. Sagduyu, A. Grushin, and Y. Shi, “Synthetic social media data generation,” IEEE Transactions on Computational Social Systems, vol. 5, no. 3, pp. 605–620, 2018.

[4] A. J. Rodriguez-Almeida, H. Fabelo, S. Ortega, A. Deniz, F. J. Balea- Fernandez, E. Quevedo, C. Soguero-Ruiz, A. M. Wa¨gner, and G. M. Callico, “Synthetic patient data generation and evaluation in disease pre- diction using small and imbalanced datasets,” IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 6, pp. 2670–2680, 2023.

[5] P. Eigenschink, T. Reutterer, S. Vamosi, R. Vamosi, C. Sun, and K. Kalcher, “Deep generative models for synthetic data: A survey,” IEEE Access, vol. 11, pp. 47 304–47 320, 2023.

[6] T. Dahmen, P. Trampert, F. Boughorbel, J. Sprenger, M. Klusch, K. Fis- cher, C. Ku¨bel, and P. Slusallek, “Digital reality: a model-based approach to supervised learning from synthetic data,” AI Perspectives, vol. 1, no. 1, p. Article No.2, 2019, 43.22.02; LK 01.

[7] Q. Liu, M. Khalil, J. Jovanovic, and R. Shakya, “Scaling while privacy preserving: A comprehensive synthetic tabular data generation and evaluation in learning analytics,” in Proceedings of the 14th Learning Analytics and Knowledge Conference, ser. LAK ’24. New York, NY, USA: Association for Computing Machinery, 2024, p. 620–631. [Online]. Available: https://doi.org/10.1145/3636555.3636921.