Synthetic Data Generation, Technological Aspects and Challenges

David Satseradze, Gocha Zedginidze

Georgian Technical University, Georgia, Tbilisi
dsatseradze@gtu.ge, g.zedginidze@gtu.ge

Abstract

The increasing complexity of machine learning models and stricter data protection regulations are driving a growing demand for large and diverse datasets. Synthetic data, artificially generated by algorithms, represents a powerful solution to this problem and are becoming an essential tool in modern data science and artificial intelligence. Their use addresses challenges related to data scarcity, privacy protection, and ensuring sample balance. This paper provides an overview of modern synthetic data generation techniques, their applications in artificial intelligence and computer science, and discusses key challenges and directions for future research.

Keywords: synthetic data, artificial intelligence, neural networks, GAN, VAE.

1. Introduction

In the era of rapid advancements in artificial intelligence (AI) and machine learning (ML), access to high-quality data is becoming a critical success factor. However, real data are often constrained by legal restrictions (GDPR, HIPAA) due to privacy concerns, high collection costs, and ethical considerations. The scarcity of high-quality data is becoming a bottleneck in machine learning development, as the use of real data containing personal information entails risks of leaks and privacy violations.

Synthetic data are not merely "random noise," but rather artificially created data generated by algorithms that preserve the statistical properties and patterns of the original real-world dataset, without directly containing confidential records or sensitive information. Such data can overcome the aforementioned barriers, accelerating the development and testing of machine learning models.

Synthetic data are widely used to train models where real data is scarce, to test systems, and to generate new insights. Recent studies indicate that synthetic data can significantly reduce software development costs and accelerate data-driven research and innovation.

The objective of this paper is to provide an overview of modern synthetic data generation methods, discuss key areas of application, and identify promising directions for further research.

2. Main part

Synthetic data refer to artificially created information generated through algorithms, models, or simulators that replicate the statistical characteristics of real datasets. Unlike anonymized data, synthetic data do not contain the original records, making it less susceptible to information leakage.
There are several types of synthetic data:

- **Tabular data**: Structured records that mimic databases, financial statements, medical records, etc.
- **Images**: Visual data generated by neural networks for tasks such as object or face recognition.
- **Text data**: Synthetic documents, messages, reviews generated using language models.
- **Audio and video**: Data used to train speech, gesture, and emotion recognition systems.

Classification may also account for the degree of realism, the level of noise (records in the data set that do not fit into one or another classification concept), the presence of labels, and other parameters.

Research on synthetic data has been actively developing since the early 2010s. Literature reviews indicate that statistical models, deep learning models, and hybrid approaches have been predominant paradigms among synthetic data generation (SDG) over the past decade [1]. Furthermore, research has focused on the use of machine learning for generating synthetic data, emphasizing the role of generative adversarial networks (GANs) and variation autoencoders (VAEs) [2].

According to published studies, synthetic data in healthcare is generated primarily using deep learning (72.6% of studies) [3] , followed by statistical methods (15.1%) [4] . Recent papers [5] presented on arXiv.org discuss the application of synthetic data in large language models (LLMs), including their roles in pretraining and fine - tuning. Furthermore, reviews highlight the ethical implications and potential biases in synthetic data.

Existing approaches to synthetic data generation can be divided into the following groups:

1. **Simple methods (Baseline):**
   - **Rule-based generation. Creating data based on predefined logical rules (e.g., generating simple 3D models).**
   - **Sampling from parametric distributions. We assume that the data follows a specific distribution (e.g., Gaussian) and generate data accordingly.**
2. **Generative neural network models (used in machine learning) - model complex data distributions using neural networks :**
   - **Generative adversarial networks (Generative Adversarial GANs (general adversarial networks) consist of a generator and a discriminator competing with each other. The generator creates "fake" data, and the discriminator learns to distinguish it from real data. GANs are effective for generating realistic images, audio, and tabular data, but are difficult to train (instability, mode collapse) and do not always guarantee coverage of the entire distribution. Examples:**
     - ❖ **DCGAN , StyleGAN (for images), WGAN , CTGAN / TVAE (for tabular data).**
   - **Variational autoencoders (Variational autoencoders (VAEs) are trained by compressing and decompressing data, creating a latent representation from which new examples can be generated. The encoder maps data to a latent space (latent variables), and the decoder reconstructs data from this space. Generation involves sampling from the latent space and passing it through the decoder. They learn more reliably than GANs and provide probabilistic interpretation, but sometimes produce less detailed examples than GANs.**
   - **Diffusion models are a modern class of generative models that demonstrate high-quality generation of images and other types of data by approximating an inverse stochastic process. Specifically, the first stage is " noising " the data (the forward process) and the subsequent "cleaning" of the noise to generate new data (the reverse process) . They are** characterized by stability and competitive quality compared to GANs. Examples:
     - ❖ DALL-E, Stable Diffusion (for images), their adaptations for tabular and time series.
   - **Autoregressive models (Autoregressive Models (**for text or sequences) predict the next element based on previous ones, meaning data is generated one element at a time, with each subsequent element dependent on the previous ones. This is well suited for text, code, and time series. It provides high quality for tasks involving sequences.

3. **Statistical and simulation methods** focus on formal modeling of known structures and laws. Classical approaches include the use of distributions, correlation models, and simulators. Examples include Monte Carlo and Gaussian models. Mixture Regression models are simple to implement but limited in their ability to reproduce complex relationships:
   - **Parametric statistical models** (multivariate normal, distribution mixtures, etc.) are simple and interpretable, but are inadequate for high-dimensional complex data with nonlinear dependencies.
   - **Simulation models** — a domain model (physics, transportation, medicine) that generates data according to specified rules; they provide precise control over scenarios and are suitable for generating rare/extreme cases.
   - **Agent-based and rule-based simulators** are used to model interacting agents (e.g. traffic, social systems).
4. **Hybrid and specialized approaches** are combinations of the above methods: a simulator plus a neural network to increase realism; statistical post-processing of synthetic samples; transfer learning to adapt the generator to the specifics of the data.
5. **Manual generation and simulators.** In some cases, synthetic data are created manually or using specialized simulators, for example, to model user behavior, network traffic, or biological processes.

In addition to real data, an increasing number of researchers are actively employing synthetic data across a wide range of fields:

- **Computer vision**: generation of images and annotations (semantic masks, bounding boxes) for detection and segmentation tasks; simulators (virtual environments) are widely used to create large-scale datasets.
- **Medicine and healthcare**: Generating synthetic electronic medical records and medical images helps protect patient privacy and increase training datasets.
- **Finance** : Modeling transactions and fraudulent schemes to train fraud detection systems.
- **AI model training**: Allows you to train models without the risk of personal data leakage, especially in education.
- **Systems testing**: used to check the functionality of software, databases, interfaces.
- **Dataset enrichment**: helps combat class imbalance by creating additional examples of rare categories.
- **Cybersecurity systems development**: used to simulate attacks, attacker behavior and testing defense mechanisms, as well as ISD (*Intrusion) training Detection System*) / IPS (Intrusion Prevention System).
- **Transport and robotics**: simulation of motion scenarios for autonomous systems.
- **Federated learning and confidential computing**: Synthetic data can be used to pre-train models without access to real data.

Most authors believe that using synthetic data gives them number of advantages, namely:

- **Confidentiality**: The absence of real records eliminates the risk of personal information leakage.
- **Scalability**: data of any volume and complexity can be generated.
- **Flexibility**: it is possible to model hypothetical scenarios, rare events, or extreme conditions.

- **Reduced costs**: Reduces the need for expensive real-world data collection and labeling. A critical question when using synthetic data is how to measure its quality, i.e., how accurately it simulates real data. It is clear that quality cannot be measured by a single metric, as it has multiple dimensions. A review of the existing literature suggests the following metrics for answering this question:

1. **Sample quality metrics (fidelity / realism)**. How similar are synthetic data to real data?
   - Statistical discrepancies (e.g., **Distance Correlation** , KS-test).
   - Visual assessment (for images).
   - Classification reports (**Classification Report**) — a model trained on synthetic data should show similar metrics on real test data. The following are used:
   - ❖ For images: **FID (Fréchet Inception Distance)**, **IS (Inception Score)**.
   - ❖ For tabular data: comparison of feature distributions (KS test, MMD) and analysis of correlation structures.
2. **Utility metrics.** How well does synthetic data solve the target problem?
   - Model-in-model (**Train on Synthetic, Test on Real - TSTR**). Models **' accuracy** (measures the proportion of correct model predictions among all predictions, used when classes in the data are balanced) and/or **F1-score** (combines precision and recall , providing a more comprehensive assessment of model performance, used when there is class imbalance or different error costs) trained on synthetic data are compared with a model trained on real data.
   - **Downstream task performance**: training a model on synthetic data and testing it on real data (**Accuracy, AUC, F1**, etc.). This is one of the most practical metrics— synthetic data are useful, if models trained on it perform well on real data.
3. **Privacy and security metrics.** How secure is the underlying data ?
   - **Differential Privacy (DP) is** a formal guarantee of privacy during generation. It includes mechanisms for noise injection and DP training of generators.
   - **Membership inference/disclosure risk** — practical tests of the risk of recovering or recognizing specific records from synthetic datasets.
4. **Diversity and coverage metrics:**
   - Assessing how well synthetic data covers the real data space (precision / recall for generative models), mode identification collapse.

It should be noted that, in practice, most specialists combine several types of metrics: **Fidelity + Utility + Privacy**.

Despite the great advantages, synthetic data still has number of limitations, namely:

- **Lack of realism**: especially when generating complex dependencies or rare events.
- **Risk of information leakage**: If models are configured incorrectly, data close to the original may be generated.
- **Limited applicability**: Not all tasks allow the use of synthetic data without loss of quality.
- **Ethical aspects:** Potential for generating disinformation (**deepfakes**), amplification of bias (if the original data is biased, synthetic data will amplify this bias).
- **Legal aspects**: It is necessary to consider the potential consequences of using synthetic data, especially in medicine and law.

Generating synthetic data, especially using modern methods such as generative adversarial networks (GANs), diffusion models, or large language models (LLMs), requires substantial computational resources. The appropriate technology depends on the scale of the problem, the type of data (images, text, tabular data, time series), and the algorithms used. A key principle is that the problem determines the technology.

- **Simple data (tables, logical rules):** A computer with a powerful CPU and a large amount of RAM is required.
- **Images, audio, text (Medium) Scale):** Entry-level/mid-range GPU required .
- **Video, complex 3D scenes, large-scale projects (Large Scale):** Requires **powerful GPUs (often multiple)** and/or the use of cloud clusters.

Below are recommendations for hardware, software tools, and other aspects to consider when generating synthetic data.

Table. 1. Recommendations for hardware and software tools

Data type	Hardware requirements	Software	Example of a problem
Images	GPU (NVIDIA A100, RTX 3090), 32–64 GB RAM, 2 TB SSD +	PyTorch , Diffusers, StyleGAN	Generation medical MRI
Tabular data	CPU (16 cores), GPU (12 GB VRAM), 32 GB RAM	SDV, CTGAN, Synthcity	Synthetic financial records
Text	GPU (24 GB + VRAM), 64 GB RAM	Hugging Face Transformers	Generating Synthetic Texts for NLP
Time series	GPU (A100), 64 GB RAM, SSD 4 TB +	TimeGAN , PyTorch	Synthetic sensory data

3. Conclusions

Given the rapid evolution of AI, it is reasonable to assume that the future of synthetic data will be linked to number of areas, such as:

- **Improving generative models**: developing architectures that can reproduce complex dependencies and rare cases.
- **Quality assessment**: development of unified metrics to objectively assess the realism, usefulness, and privacy of synthetic data.
- **Standardization**: creating common protocols, formats and requirements for synthetic data.
- **Integration with Privacy-Preserving Machine Learning**: Synergy with federated learning, differential privacy, and other approaches.
- **Research into guaranteed coverage of rare events** - ensuring adequate representation of rare events through improved sampling strategies.
- **Automatic validation of synthetic kits** - automating validation pipelines to evaluate dataset suitability for specific tasks.
- **Integration of simulators with neural network generators** to obtain controlled and realistic data synthesis.
- **Ethical and regulatory aspects** – defining ethical and regulatory frameworks governing the use of synthetic data in sensitive domains such as healthcare and finance.

Reference:

[1] André Bauer, “Comprehensive Exploration of Synthetic Data Generation: A Survey,” 02 2024. Internet resurs: https://arxiv.org/pdf/2401.02524. [Accessed: 22/08/2025].

[2] Yingzhou Lu, “Machine Learning for Synthetic Data Generation: A Review,” JOURNAL OF LATEX CLASS FILES, VOL. 14, NO. 8, T, AUGUS 2021. Internet resources: https :// arxiv. org / pdf /2302.04062 v 8. [Accessed: 01/06/2025].

[3] Vasileios C Pezoulas, “Synthetic data generation methods in healthcare: A review on open-source tools and methods,” Comput Struct Biotechnol J. 23:2892–2910., 23 Jul 9 2024. Internet resources: https://www.sciencedirect.com/science/article/pii/S2001037024002393. [Accessed: 23/06/2025].

[4] Shuang Hao, “Synthetic Data in AI: Challenges, Applications, and Ethical Implications,” 01 Jan 2024. Internet resources: https :// arxiv . org / html /2401.01629 v 1. [Accessed: 12/06/2025].

[5] Lin Long, “On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey,” 14 Jul 2024. Internet resources: https :// arxiv . org / pdf /2406.15126 v 1. [Accessed: 12/06/2025].

სინთეტიკური მონაცემების გენერაცია, ტექნოლოგიური ასპექტები და გამოწვევები

დავითი საცერაძე, გოჩა ზედგინიძე

საქართველოს ტექნიკური უნივერსიტეტი, თბილისი

dsatseradze@gtu.ge, g.zedginidze@gtu.ge

რეზიუმე

მანქანური სწავლების მოდელების მზარდი სირთულე და მონაცემთა დაცვის უფრო მკაცრი რეგულაციები ზრდის დიდი და მრავალფეროვანი მონაცემთა ნაკრებების მოთხოვნას. ალგორითმების მიერ ხელოვნურად გენერირებული სინთეზური მონაცემები წარმოადგენს ამ პრობლემის ძლიერ გადაწყვეტას და ხდება აუცილებელი ინსტრუმენტი თანამედროვე მონაცემთა მეცნიერებასა და ხელოვნურ ინტელექტში. მათი გამოყენება წყვეტს რეალური მონაცემების სიმწირესთან, კონფიდენციალურობის დაცვასთან და ნიმუშის ბალანსის უზრუნველყოფასთან დაკავშირებულ გამოწვევებს. ეს სტატია იძლევა თანამედროვე სინთეზური მონაცემების გენერირების ტექნიკის მიმოხილვას, მათ გამოყენებას ხელოვნურ ინტელექტსა და კომპიუტერულ მეცნიერებაში და განიხილავს ძირითად გამოწვევებსა და სამომავლო კვლევის მიმართულებებს.

საკვანძო სიტყვები: სინთეზური მონაცემები, ხელოვნური ინტელექტი, ნეირონული ქსელები, **GAN,**