International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 6, Nov-Dec-2026, ISSN (Online): 2395-566X

Synthetic Data Generation for Machine Learning

Mantasha Najir Jamadar
Department of Artificial Intelligence and Data Science (AI&DS)
Sahyadri Valley College of Engineering and Technology, Rajuri, Pune, India

Abstract — Machine learning (ML) systems depend on large, representative, and reliable datasets. In many practical situations, real data may be limited, expensive to collect, imbalanced, or restricted by privacy and confidentiality requirements. Synthetic data generation addresses these challenges by creating artificial records that reproduce useful statistical patterns and relationships of an original dataset without directly copying every real observation. This paper presents a structured study of synthetic data generation for machine learning. It reviews statistical methods, resampling techniques, Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and diffusion-based approaches. A general framework covering data collection, preprocessing, model selection, synthetic sample generation, quality evaluation, and downstream ML testing is described. The paper discusses data fidelity, utility, diversity, and privacy as important evaluation dimensions, along with applications in healthcare, finance, cybersecurity, and computer vision. Challenges such as mode collapse, bias reproduction, privacy leakage, and computational cost are also examined. The proposed framework is intended to support responsible dataset development; synthetic data should be validated before being used as a replacement or supplement for real-world data.

Keywords— Synthetic Data, Machine Learning, Data Augmentation, Generative Adversarial Networks, Variational Autoencoder, Data Privacy, Diffusion Models.

I. INTRODUCTION

Machine learning has become an important part of modern applications such as medical diagnosis, fraud detection, recommendation systems, autonomous vehicles, and intelligent monitoring. The performance of an ML model is strongly influenced by the quantity, quality, and diversity of its training data. However, collecting real-world data can be difficult because of high acquisition costs, limited access, class imbalance, confidentiality requirements, and the presence of sensitive personal information.

Synthetic data is artificially generated information designed to represent selected characteristics of real data. It may be created using statistical models, simulation, interpolation, or generative neural networks. Unlike simple duplication, a well-designed generator attempts to learn the structure and distribution of the source data and produce new samples. Synthetic data can support model training, software testing, data sharing, and experimentation when real records are scarce or restricted [1], [2].

The o
bjective of this paper is to study the role of synthetic data generation in machine learning and describe a general framework for producing and evaluating synthetic datasets. The study emphasizes that synthetic data is not automatically private, unbiased, or useful; these properties must be assessed for each application.

Objectives and Scope

The main objectives are: (1) to understand the need for synthetic data in ML; (2) to study common data generation techniques; (3) to describe a general synthetic data generation workflow; (4) to identify suitable evaluation measures for fidelity, utility, diversity, and privacy; and (5) to discuss applications, limitations, and future improvements.

The scope includes tabular, image, and sequential data, with emphasis on structured/tabular datasets commonly used in supervised learning. The framework can be adapted to classification and regression tasks. Synthetic data may supplement real training records, support testing, or enable controlled sharing, but it should not be assumed to preserve every rare event or causal relationship.

II. LITERATURE REVIEW

Figueira and Vaz survey synthetic data generation and evaluation methods, with particular attention to Generative Adversarial Networks. Their review describes how generative models learn data patterns and produce artificial samples for different use cases [1].

Fonseca and Bacao review tabular and latent-space synthetic data generation. They discuss applications such as anonymization, regularization, oversampling, and semi-supervised learning, highlighting the need for methods suited to mixed numerical and categorical data [2].

© 2026 IJSRET

Goyal and Mahmoud provide a systematic review of generative-AI-based synthetic data methods, including GANs, VAEs, and large language models. They identify data scarcity and privacy as major motivations while noting challenges in computational requirements, training stability, and privacy protection [3].

Recent work also compares tabular generators such as CTGAN and TVAE and evaluates them using statistical similarity and downstream predictive performance [4]. These studies indicate that no single generator is best for every dataset; selection depends on data type, sample size, task, and evaluation requirements.

III. PROPOSED SYSTEM

1. System Overview

The proposed framework accepts an original dataset as input. The data is inspected, cleaned, and divided into training and validation subsets. A suitable generator is selected according to the data format and task. The generator learns the patterns of the training data and produces artificial samples. The generated data is then checked for statistical similarity, diversity, privacy risk, and usefulness in a downstream ML model. Only samples that satisfy the application requirements are retained.

2. System Architecture

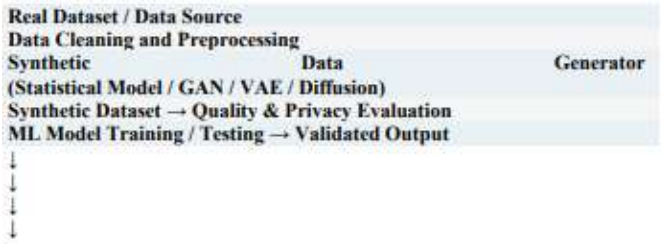

Fig. 1. General architecture of the proposed synthetic data generation framework.

3. Working Procedure

First, a relevant dataset is collected from an authorized source. Second, preprocessing handles missing values, duplicates, inconsistent labels, and categorical encoding. Third, the dataset is split before generator training to prevent information leakage. Fourth, a generation method is selected and trained using only the training portion. Fifth, new synthetic records are sampled and checked for invalid values and unrealistic combinations. Finally, quality and privacy tests are performed, and a downstream ML model is trained or evaluated using the accepted synthetic data.

IV. SYNTHETIC DATA GENERATION TECHNIQUES

1. Statistical Methods

Statistical generators estimate probability distributions and relationships among variables, then sample new records from the fitted model. Methods such as Gaussian models, Bayesian networks, and copulas can be interpretable and relatively efficient. Their limitation is that complex dependencies may be difficult to represent when the data is high-dimensional or contains mixed variable types.

2. Data Augmentation and SMOTE

Data augmentation creates modified examples from existing samples. In image tasks, transformations may include rotation, cropping, and flipping. For imbalanced tabular classification, Synthetic Minority Over-sampling Technique (SMOTE) generates minority-class samples by interpolation between nearby observations. It is useful for balancing classes, but interpolation can create unrealistic records when feature relationships are complex.

3. Generative Adversarial Networks

A Generative Adversarial Network (GAN) consists of a generator and a discriminator. The generator creates synthetic samples, while the discriminator attempts to distinguish generated samples from real ones. Both networks improve through adversarial training. Conditional GANs, including CTGAN, can generate tabular records conditioned on selected categories. GANs can model complex distributions, but training may be unstable and can suffer from mode collapse, where output diversity is limited [1], [4].

4. Variational Autoencoders

A Variational Autoencoder (VAE) learns a compressed latent representation of the input data. An encoder maps records to a probability distribution in latent space, and a decoder reconstructs samples from that representation. New synthetic records can be generated by sampling latent vectors and decoding them. TVAE adapts this approach to tabular data. VAEs offer a structured probabilistic framework, although generated samples may be smoother or less precise than real observations.

5. Diffusion and Language-Model Approaches

Diffusion models learn to reverse a gradual noise-adding process and can generate high-quality samples through iterative denoising. They have been explored for images and tabular data, but generation can require substantial computation. Language-model-based approaches can generate structured or

© 2026 IJSRET

text data when prompts and schemas are carefully designed. Their output must be checked for factual consistency, schema validity, and possible memorization.

V. DATASET AND PREPROCESSING

The choice of dataset depends on the intended ML task. Public datasets from repositories such as UCI Machine Learning Repository can be used for tabular experiments, while image datasets may be selected for computer-vision tasks. Healthcare, finance, and institutional datasets may contain sensitive information and should be used only with appropriate authorization and safeguards.

Preprocessing may include removing duplicates, handling missing and infinite values, correcting data types, encoding categorical fields, scaling numerical features, and examining class imbalance. The train-test split should be performed before fitting the generator. Synthetic samples used for model training should be generated from the training data only; the real test set must remain separate to provide an unbiased evaluation. For time-series data, chronological splitting is important to avoid future information leaking into training.

VI. PERFORMANCE EVALUATION

A synthetic dataset should be assessed from multiple perspectives. A single similarity score is not sufficient because a dataset can look statistically similar while failing to improve the target ML task or exposing private information.

1. Statistical Fidelity
Fidelity measures how closely synthetic data resembles the real data distribution. Comparisons may include means, standard deviations, category frequencies, correlation matrices, and distributional tests such as the Kolmogorov–Smirnov test for numerical features.

2. Downstream Utility
Utility measures whether synthetic data supports the intended task. A common approach is Train on Synthetic, Test on Real (TSTR): an ML model is trained using synthetic records and evaluated on a held-out real test set. Accuracy, precision, recall, F1-score, and area under the ROC curve may be reported for classification.

3. Diversity and Coverage
Diversity evaluates whether the generator produces varied samples rather than repeated or near-duplicate records. Coverage checks whether important subgroups, rare categories, and minority patterns are represented. Low diversity may indicate mode collapse or poor learning of the original distribution.

4. Privacy Risk
Privacy evaluation examines whether synthetic records reveal information about training individuals. Tests may include membership-inference and nearest-neighbour disclosure-risk analyses. Differential privacy can provide formal protection when correctly implemented, but it may reduce data utility and requires careful selection of privacy parameters [5].

Applications
Synthetic data generation can support healthcare research by creating realistic training records when access to patient data is restricted. In finance, it can help test fraud-detection systems and model rare transaction patterns. In cybersecurity, synthetic network events can supplement limited attack examples. In computer vision, image augmentation and generative models can expand training variation. It can also support software testing, simulation, education, and controlled data sharing. In every case, generated records should be validated against domain requirements before deployment.

Challenges and Limitations
The quality of synthetic data depends on the representativeness and accuracy of the source dataset. If the original data contains bias, missing groups, or incorrect labels, a generator may reproduce or amplify these problems. Rare events may be omitted, while unrealistic combinations may be produced. GANs can experience unstable training and mode collapse; diffusion models can be computationally expensive; and statistical models may fail to capture complex dependencies.
Synthetic data does not automatically guarantee anonymity. A model may memorize rare training examples, allowing disclosure or membership inference. Privacy-preserving training can reduce some risks but may affect fidelity and downstream utility. Evaluation results can also be misleading if the real test set is not kept separate or if synthetic and real records overlap. Therefore, privacy, utility, fidelity, fairness, and reproducibility should be evaluated together.

Future Scope
Future work may compare statistical, GAN, VAE, and diffusion-based generators on common datasets using consistent evaluation protocols. Research can focus on improving generation of rare classes, mixed-type records, and time-series data. Differentially private and federated generation may support settings where data cannot be centralized. Explainable evaluation, fairness testing, and standardized

© 2026 IJSRET

benchmarks can improve trust and reproducibility. Domain-specific validation and human review will remain important before synthetic data is used in high-impact decisions.

VII. CONCLUSION

Synthetic data generation is a useful approach for addressing data scarcity, class imbalance, testing requirements, and some data-sharing limitations in machine learning. This paper described a general framework involving data collection, preprocessing, generator selection, synthetic sample creation, quality assessment, and downstream ML evaluation. Statistical methods, SMOTE, GANs, VAEs, and diffusion models provide different strengths and limitations. Reliable use requires more than realistic-looking samples: fidelity, utility, diversity, privacy, and bias must all be considered. Synthetic data should complement carefully governed real-world data and expert validation rather than be treated as an automatically safe or universally accurate replacement.

REFERENCES

1. A. Figueira and B. Vaz, “Survey on Synthetic Data Generation, Evaluation Methods and GANs,” Mathematics, vol. 10, no. 15, 2733, 2022. doi: 10.3390/math10152733.
2. J. Fonseca and F. Bacao, “Tabular and latent space synthetic data generation: a literature review,” Journal of Big Data, vol. 10, article 115, 2023. doi: 10.1186/s40537-023-00792-7.
3. M. Goyal and Q. H. Mahmoud, “A Systematic Review of Synthetic Data Generation Techniques Using Generative AI,” Electronics, vol. 13, no. 17, 3509, 2024. doi: 10.3390/electronics13173509.
4. P. Yadav et al., “Rigorous Experimental Analysis of Tabular Data Generated using TVAE and CTGAN,” International Journal of Advanced Computer Science and Applications, vol. 15, no. 4, 2024. doi: 10.14569/IJACSA.2024.01504125.
5. “Privacy Mechanisms and Evaluation Metrics for Synthetic Data Generation: A Systematic Review,” IEEE Access, vol. 12, pp. 88048–88074, 2024. doi: 10.1109/ACCESS.2024.3417608.

© 2026 IJSRET