

Evaluating Synthetic Data Generation Techniques for Improving Machine Learning Model Training

Dr. Aisling Byrne

Associate Professor, Galway Centre for Advanced Data Modeling, Ireland

Abstract

Synthetic data generation has become an important methodological approach in machine learning because predictive systems frequently encounter limitations associated with insufficient training observations, unequal class distributions, incomplete feature coverage, and restricted access to representative datasets. Conventional supervised learning methods depend on the availability of observations that adequately describe the statistical characteristics of the intended prediction problem. However, collecting additional information may require substantial resources, and existing datasets may provide limited representation of unusual operating conditions. Synthetic data generation offers computational approaches for constructing artificial observations intended to supplement genuine training records while preserving selected statistical and structural characteristics. The present paper develops a hypothetical comparative framework for evaluating synthetic data generation techniques used to improve machine learning model training. Four augmentation strategies are considered: Synthetic Minority Over-sampling Technique, class-conditional Gaussian Mixture Models, Conditional Variational Autoencoders, and Conditional Generative Adversarial Networks. A fictional industrial equipment condition dataset containing 12,000 sensor observation windows associated with 60 monitoring units is introduced to demonstrate the intended evaluation procedure. The dataset contains three classification categories representing normal, degraded, and critical equipment conditions.

The proposed methodology incorporates data preparation, class-conditional synthetic generation, quality assessment, model development, and machine-independent testing. A common Random Forest classifier is selected to examine the effects of alternative training data configurations. Two illustrative analytical tables describe the hypothetical dataset distribution and comparative classification performance. Overall accuracy and macro-averaged recall are used to demonstrate differences between aggregate recognition and sensitivity across equipment condition categories. All numerical values are manually constructed and do not represent completed synthetic data generation experiments or verified classification results. The methodological discussion examines distributional similarity, synthetic diversity, minority-class representation, information leakage, overfitting, privacy limitations, and computational efficiency.

Keywords: Synthetic Data Generation, Machine Learning, Data Augmentation, Generative Adversarial Networks, Variational Autoencoder, SMOTE, Gaussian Mixture Model, Classification Performance, Data Quality, Predictive Reliability.

1. Introduction

Machine learning has become an important component of computational decision-making because predictive algorithms can identify relationships among observations and apply those relationships to previously unseen information. The effectiveness of supervised learning depends substantially on the availability of suitable training datasets. Observations must provide sufficient information about relevant feature distributions, classification categories, and the variations encountered within the intended application. However, many practical datasets contain limitations associated with insufficient samples, unequal representation of categories, and incomplete coverage of unusual conditions. These characteristics can weaken predictive reliability, particularly when models encounter observations that differ from those available during training.

Training data limitations are particularly significant when collecting genuine observations requires specialized equipment, extensive monitoring, or rare operational events. Industrial monitoring systems, for example, may collect numerous records representing normal equipment conditions but comparatively fewer observations associated with deterioration or developing faults. Similar difficulties occur in other classification environments where uncommon events are important but inadequately represented. Expanding the dataset through additional collection may not always be practical. Consequently, researchers have investigated alternative approaches for increasing the diversity and quantity of training observations without relying exclusively on further real-world data acquisition.

Synthetic data generation involves constructing artificial observations intended to represent selected characteristics of an existing dataset or a defined data-generating process. Different methods employ distinct mathematical principles. Interpolation-based techniques create observations between existing examples, while statistical generative models estimate probability distributions from which new samples can be drawn. Deep generative models use neural architectures to learn more complex representations of the available data. These approaches offer potential advantages for training data augmentation, but synthetic observations are not automatically representative or useful. Their quality depends on the original records, modeling assumptions, generation procedure, and subsequent evaluation.

Generative methods have expanded the methodological possibilities for constructing artificial training observations. Gaussian Mixture Models provide a statistical framework for approximating heterogeneous distributions through combinations of probability components. Variational Autoencoders learn latent representations from which new samples may be generated. Generative Adversarial Networks develop synthetic observations through a competitive learning process involving generator and discriminator networks. Conditional versions of these methods can incorporate class information to support targeted generation. Nevertheless, sophisticated models may reproduce undesirable biases, generate unrealistic feature combinations, or fail to represent minority characteristics adequately.

The present paper proposes a hypothetical comparative investigation of synthetic data generation techniques for improving machine learning model training. The study considers a fictional industrial equipment condition classification dataset containing normal, degraded, and critical operating categories. Four augmentation strategies are compared with a baseline trained using original observations only. A common Random Forest classifier provides a consistent predictive framework. Two illustrative tables demonstrate dataset composition and possible classification outcomes. No actual sensor measurements were collected, and no generative models or classifiers were executed. The paper

establishes methodological procedures for evaluating the utility and limitations of synthetic training data.

2. Objectives of the Study

The primary objective of this study is to examine the methodological effectiveness of synthetic data generation techniques for improving machine learning model training. The investigation focuses on synthetic observation quality, classification performance, minority-category representation, and evaluation reliability. Three specific objectives establish the analytical direction of the proposed comparative research framework.

2.1 Examination of Synthetic Data Quality and Statistical Representation

The first objective is to examine how different synthetic data generation techniques represent the statistical characteristics of original training observations. Synthetic datasets should preserve information relevant to the intended prediction task while avoiding unrealistic combinations of explanatory variables. The proposed investigation considers interpolation-based, probabilistic, and neural generation procedures. Particular attention is given to the extent to which synthetic observations represent feature dependencies, category-specific patterns, and variation within the original dataset. This objective emphasizes the importance of evaluating statistical quality rather than relying exclusively on the number of generated records.

The investigation further aims to distinguish synthetic diversity from simple duplication of existing observations. Increasing training sample size may provide limited benefit when additional records reproduce nearly identical examples or reinforce measurement noise. Conversely, artificial observations that differ substantially from genuine data may introduce misleading predictive patterns. Consequently, the proposed framework considers distributional similarity, feature constraints, and the coverage of relevant operating conditions. This objective seeks to establish methodological criteria for determining whether generated observations contain useful supplementary information.

2.2 Comparative Evaluation of Synthetic Data Generation Methods

The second objective focuses on establishing a consistent framework for comparing alternative synthetic data augmentation strategies. SMOTE, Gaussian Mixture Models, Conditional Variational Autoencoders, and Conditional Generative Adversarial Networks are selected to represent different generation principles. Their proposed evaluation involves training a common classifier using augmented datasets constructed under comparable conditions. Accuracy and macro-averaged recall are considered principal performance measures. The investigation seeks to identify how synthetic data strategies may influence classification outcomes without introducing differences caused by unrelated classifier architectures.

A further objective involves examining whether more complex generative approaches provide meaningful advantages over simpler interpolation-based or statistical methods. Deep generative networks can represent nonlinear feature relationships but may require more observations and extensive optimization. Simpler methods may remain competitive when the training distribution is adequately represented through local interpolation or conventional probability models. Consequently, the proposed framework emphasizes independent evaluation and computational transparency. This objective encourages selection of generation methods through reproducible evidence rather than assumptions that sophisticated neural architectures necessarily provide superior training data.

2.3 Assessment of Training Reliability and Synthetic Data Limitations

The third objective is to identify methodological conditions that may influence the reliability of machine learning models trained with synthetic observations. Information leakage, unrealistic samples, feature distortion, class imbalance, and overfitting may affect predictive outcomes. The proposed investigation considers how training-only data generation and machine-independent evaluation could reduce selected risks. Particular attention is given to ensuring that synthetic observations are not derived from validation or independent testing data.

The investigation also considers the broader reliability of synthetic data generation, including its implications for computational efficiency and information privacy. Synthetic observations may resemble source records closely enough to create disclosure concerns, while excessive augmentation can introduce unnecessary computational demands. Consequently, the proposed framework emphasizes privacy assessment, representative testing, and transparent generation procedures. This objective establishes that synthetic data usefulness depends on more than apparent similarity to genuine observations or improvements in a single classification metric.

3. Review of Literature

3.1 Statistical Sampling and Synthetic Minority Data Generation

McLachlan and Peel (2000) provided a comprehensive treatment of finite mixture models, establishing statistical foundations for representing complex probability distributions through combinations of simpler components. Chawla et al. (2002) introduced the Synthetic Minority Over-sampling Technique as a method for generating artificial minority observations through interpolation between neighboring examples. Batista et al. (2004) investigated alternative strategies for balancing classification datasets. These contributions established important foundations for constructing supplementary training observations through statistical modeling and resampling.

He and Garcia (2009) reviewed the challenges associated with learning from imbalanced datasets and examined alternative approaches for addressing unequal class representations. López et al. (2013) investigated how intrinsic data characteristics influence classification performance under imbalance. Menardi and Torelli (2014) examined methods for training and assessing classifiers using imbalanced observations. Branco et al. (2016) provided a broader review of predictive modeling involving imbalanced target distributions. Together, these studies demonstrate that synthetic sample generation must be considered alongside class overlap, label quality, feature relevance, and the characteristics of the underlying learning problem.

3.2 Development of Deep Generative Learning Approaches

Kingma and Welling (2014) developed the variational autoencoding framework through stochastic variational inference and learned latent representations. Goodfellow et al. (2014) introduced Generative Adversarial Networks, in which a generator and discriminator participate in an adversarial optimization process. Mirza and Osindero (2014) extended adversarial generation through conditioning information. These contributions established important approaches for constructing artificial observations using neural representation learning and demonstrated alternative mechanisms for approximating complex data distributions.

Sohn et al. (2015) investigated deep conditional generative models incorporating latent variables for structured prediction tasks. Salimans et al. (2016) subsequently examined techniques for improving

adversarial model training and addressing selected optimization difficulties. Patki et al. (2016) introduced the Synthetic Data Vault as a framework for generating synthetic data from structured databases. These studies expanded the methodological foundations of synthetic generation. Their contributions also demonstrate that producing realistic samples and improving downstream prediction are related but distinct evaluation objectives.

3.3 Predictive Utility, Distributional Evaluation, and Data Governance

Breiman (2001) introduced Random Forest as an ensemble learning algorithm capable of representing nonlinear relationships within structured classification datasets. Davis and Goadrich (2006) examined precision–recall evaluation, while Cawley and Talbot (2010) investigated bias associated with model selection and repeated optimization. Gretton et al. (2012) developed a kernel-based two-sample testing framework for examining differences between probability distributions. These contributions established methodological foundations for evaluating predictive performance and statistical similarity between observed and generated data.

Dwork (2006) introduced differential privacy as a formal framework for controlling information disclosure in statistical analysis. Machanavajjhala et al. (2007) examined limitations of conventional anonymization and developed additional privacy criteria. Saito and Rehmsmeier (2015) emphasized precision–recall analysis for imbalanced classification. Krawczyk (2016) reviewed continuing challenges associated with learning from imbalanced observations. Together, these publications indicate that synthetic data evaluation should consider predictive utility, statistical fidelity, privacy protection, and generalization rather than assuming that artificial records automatically address limitations in the original dataset.

4. Research Methodology

4.1 Research Design and Synthetic Data Evaluation Framework

The present paper adopts a hypothetical comparative quantitative research design for evaluating synthetic data augmentation in supervised machine learning. Five training configurations are considered: original observations only, SMOTE augmentation, class-conditional Gaussian Mixture Model generation, Conditional Variational Autoencoder generation, and Conditional Generative Adversarial Network generation. A Random Forest classifier is selected as the common downstream predictive model. The intended classification task involves recognizing three industrial equipment conditions using numerical sensor features. No generative methods or classifiers were experimentally implemented.

The proposed analytical procedure consists of data quality assessment, machine-independent partitioning, generator training, synthetic sample evaluation, classifier development, and independent testing. Every generator would be fitted exclusively on the original training partition. The augmentation configurations would use equivalent target class counts to reduce differences associated solely with training dataset size. Classifier parameters would be selected using genuine validation observations. The final evaluation would use original test records that had not contributed to data preprocessing, generator training, or classifier optimization.

4.2 Hypothetical Equipment Dataset and Generation Procedures

For illustrative purposes, the study assumes a fictional dataset containing 12,000 sensor observation windows collected conceptually from 60 industrial monitoring units. Each unit contributes 200

nonoverlapping observation windows representing normal, degraded, or critical equipment conditions. The proposed dataset contains 7,200 normal observations, 3,000 degraded observations, and 1,800 critical observations. Eighteen numerical explanatory variables are assumed to describe vibration, temperature, electrical loading, pressure, and selected statistical signal characteristics. These quantities are methodological assumptions rather than actual equipment measurements.

The dataset would be divided by monitoring unit into 36 training units, 12 validation units, and 12 independent testing units. Training data would therefore contain 4,320 normal, 1,800 degraded, and 1,080 critical observations. Each augmentation method would retain the 4,320 genuine normal observations while adding 2,520 artificial degraded observations and 3,240 artificial critical observations. This would produce 12,960 training records per augmented configuration. Synthetic observations would be screened for unrealistic feature combinations and violations of predefined equipment constraints.

4.3 Synthetic Quality Assessment and Predictive Model Evaluation

SMOTE would generate minority-class observations through interpolation between neighboring examples. Class-conditional Gaussian Mixture Models would estimate probability distributions separately for the degraded and critical categories. The Conditional Variational Autoencoder would use learned latent representations and category labels to produce synthetic records. The Conditional Generative Adversarial Network would generate observations through adversarial training conditioned on the intended equipment state. Model configurations would be selected using development data, without reference to the final test outcomes.

Synthetic data quality would be assessed through marginal distribution comparisons, pairwise feature relationships, multivariate similarity, constraint violations, duplicate or nearest-neighbor analysis, and category-specific coverage. Maximum Mean Discrepancy could support comparisons of multivariate distributions. Predictive utility would be assessed using overall accuracy and macro-averaged recall on genuine test observations. A completed investigation should additionally report macro-F1, class-specific precision, probability calibration, generator training costs, and uncertainty estimates. Repeated machine-level experiments would be necessary before establishing statistically reliable differences among generation strategies.

5. Analysis

5.1 Illustrative Distribution of Original Equipment Training Data

The composition of a training dataset can influence how effectively a classifier recognizes different categories of operating conditions. Table 1 presents a hypothetical distribution of sensor observation windows across the original training, validation, and testing partitions. The dataset contains 60% normal observations, 25% degraded observations, and 15% critical observations. This unequal representation establishes a possible motivation for targeted synthetic augmentation. However, the example does not imply that all classification difficulties arise from class frequency alone.

The proposed partition separates entire monitoring units rather than randomly distributing individual observation windows across datasets. This approach is relevant because observations from the same equipment may share operational characteristics that could produce misleading results if included in both training and testing partitions. The identical class proportions across the three subsets are simplifying assumptions. A genuine investigation would require observed machine-level variation,

appropriate metadata, and verification that each partition represents the intended deployment environment.

Table 1: Hypothetical Distribution of Industrial Sensor Observations

Equipment Condition	Training Observations	Validation Observations	Test Observations	Total
Normal	4,320	1,440	1,440	7,200
Degraded	1,800	600	600	3,000
Critical	1,080	360	360	1,800
Total	7,200	2,400	2,400	12,000

The proposed augmented training configurations would contain 12,960 observations after adding 5,760 synthetic records to the original training partition. Validation and independent testing would remain restricted to the 2,400 genuine hypothetical records in each corresponding partition.

5.2 Illustrative Comparison of Synthetic Data Training Strategies

The usefulness of synthetic training data can be examined through its influence on classifier performance when evaluated using independent genuine observations. Overall accuracy describes the proportion of correct predictions, while macro-averaged recall gives equal weight to recognition across all classification categories. Table 2 presents hypothetical classification counts for five training configurations. The constructed test dataset contains 1,440 normal, 600 degraded, and 360 critical observations. All numerical outcomes were manually selected and do not represent executed machine learning experiments.

The illustrative comparison assigns the highest overall accuracy to the Conditional Variational Autoencoder configuration, whereas Conditional GAN augmentation receives the highest macro-averaged recall. This pattern demonstrates a possible trade-off between aggregate classification performance and the recognition of less frequently represented conditions. However, the numerical differences cannot establish the superiority of either generation approach.

Table 2: Hypothetical Classification Performance Under Synthetic Data Augmentation

Training Configuration	Correct Normal Predictions	Correct Degraded Predictions	Correct Critical Predictions	Accuracy (%)	Macro Recall (%)
Original Data Only	1,330	492	216	84.92	78.12
SMOTE Augmentation	1,308	509	264	86.71	83.00

Gaussian Mixture Generation	1,320	500	250	86.25	81.48
Conditional VAE Generation	1,316	514	280	87.92	84.94
Conditional GAN Generation	1,298	516	292	87.75	85.75

The illustrative results demonstrate that a model may achieve slightly higher overall accuracy while another identifies a greater proportion of observations across less frequently represented categories. The appropriate comparison therefore depends on the evaluation objective and the operational importance of different classification errors.

6. Discussion

6.1 Synthetic Representation and Training Data Diversity

Synthetic data generation provides an approach for expanding the statistical representation available during machine learning development. Interpolation-based methods construct observations within local neighborhoods of existing data, while probability-based and neural methods attempt to represent broader distributional characteristics. These strategies may provide useful training information when the original dataset contains insufficient examples of relevant conditions. However, their effectiveness depends on the extent to which generated observations preserve meaningful relationships among features and categories.

The proposed framework emphasizes the distinction between increasing dataset size and increasing useful information. Artificial observations that closely duplicate existing records may reinforce patterns already represented during training without improving generalization. Conversely, synthetic examples located outside plausible feature relationships may introduce misleading information. The hypothetical comparison demonstrates one possible scenario in which augmentation changes classification performance. Nevertheless, the assigned values do not establish that the generated records would be realistic or beneficial in an actual industrial monitoring environment.

6.2 Statistical Generation and Deep Generative Learning

Gaussian Mixture Models provide a statistical approach for representing heterogeneous distributions through multiple probability components. Their usefulness depends on the suitability of the assumed distributional structure and the quality of parameter estimation. Variational Autoencoders provide an alternative through learned latent representations, while Conditional GANs employ adversarial learning to generate category-specific observations. These methods differ in computational requirements and their capacity to represent complicated feature relationships. Consequently, their comparative usefulness should be examined empirically rather than inferred from architectural complexity.

The illustrative performance table assigns slightly different advantages to Conditional VAE and Conditional GAN augmentation. The VAE configuration receives favorable aggregate accuracy, while the GAN configuration receives higher macro-averaged recall. These constructed differences

demonstrate why synthetic data methods should be evaluated using more than one performance measure. A genuine comparison would also examine generator stability, category coverage, constraint violations, and computational costs. The most complex generation method would not necessarily represent the most efficient or dependable solution.

6.3 Synthetic Data Utility and Model Generalization

Synthetic data utility should be assessed according to how generated observations influence predictive performance on data that remain independent of the generation process. High similarity between synthetic and training distributions may indicate statistical fidelity, but it does not automatically establish improved generalization. Similarly, a classifier may perform favorably on synthetic observations while producing limited improvements on genuine test records. Consequently, independent real-data evaluation is necessary for establishing whether artificial training examples contribute information that benefits downstream classification.

The proposed methodology uses machine-independent partitions to reduce the risk of information leakage through related sensor observations. However, machine-level separation does not remove every possible source of dependency or distributional bias. Synthetic generation may still reinforce systematic limitations in the original training equipment. Future implementation should therefore examine performance across multiple machine populations and operating conditions. Statistical utility, representational diversity, and predictive generalization should be evaluated as complementary dimensions rather than interchangeable indicators of synthetic data quality.

7. Challenges

7.1 Synthetic Data Fidelity and Unrealistic Feature Relationships

Synthetic data generation can produce observations that resemble original records in selected statistical characteristics while violating important relationships among explanatory variables. Industrial sensor measurements may contain physical dependencies involving temperature, mechanical load, vibration, and pressure. Generative models that fail to represent these dependencies can produce combinations that are mathematically possible but operationally unrealistic. Consequently, statistical similarity alone may provide an incomplete assessment of synthetic data quality.

Generation methods may also inadequately represent rare operational conditions when the original dataset contains limited examples. A model can reproduce frequent patterns while overlooking less common but important variations. Conditional generation may increase the number of minority-category observations without ensuring that their internal diversity is appropriately represented. The proposed framework therefore emphasizes category-specific coverage, physical constraint assessment, and the examination of unusual feature combinations. These safeguards would require expert validation before synthetic records could be considered suitable for training operational systems.

7.2 Overfitting, Mode Collapse, and Artificial Sampling Bias

Deep generative models can encounter difficulties associated with unstable optimization and inadequate representation of the original data distribution. Generative Adversarial Networks may exhibit mode collapse, in which the generator produces a restricted collection of patterns rather than representing the full diversity of observations. Variational Autoencoders can produce overly simplified representations

under unsuitable model configurations. Such limitations may reduce the usefulness of generated records even when their individual values appear plausible.

Sampling bias can also arise when synthetic observations amplify systematic errors or underrepresented structures within the original dataset. A generator trained on unrepresentative equipment conditions may reproduce those limitations rather than correct them. Similarly, excessive augmentation can increase the influence of generated observations relative to genuine measurements. The proposed methodology therefore emphasizes controlled augmentation quantities, repeated evaluation, and comparison with an original-data baseline. These procedures are essential for determining whether synthetic generation improves or weakens classification reliability.

7.3 Data Privacy, Information Leakage, and Computational Requirements

Synthetic data generation is sometimes considered an alternative to sharing original datasets, but artificial observations do not automatically provide privacy protection. A poorly trained generator may reproduce individual source records or preserve sensitive relationships that permit unwanted information disclosure. Formal privacy guarantees require appropriate mechanisms and evaluation rather than reliance on the synthetic appearance of generated records. Although the proposed industrial dataset does not contain personal records, these considerations remain relevant when synthetic generation is applied to confidential or sensitive information.

Information leakage presents another challenge when synthetic generators are trained using records later designated for independent evaluation. Such practices can introduce knowledge of test distributions into the training process and produce optimistic classification outcomes. Deep generative architectures also introduce substantial computational requirements during training and parameter optimization. Consequently, a method producing favorable predictive results may still impose operational costs that limit its usefulness. The proposed framework emphasizes training-only generation, independent testing, computational transparency, and appropriate privacy assessment before practical adoption.

8. Conclusion

Synthetic data generation provides a methodological approach for supplementing machine learning training datasets when genuine observations are limited or unevenly distributed. The present paper developed a hypothetical comparative framework involving SMOTE, Gaussian Mixture Models, Conditional Variational Autoencoders, and Conditional Generative Adversarial Networks. A common Random Forest classifier was selected to support consistent evaluation of alternative training configurations. The proposed methodology demonstrates how artificial observations could be assessed through statistical quality measures and their contribution to independent classification performance.

The illustrative tables demonstrate possible procedures for reporting original data distributions and classification performance following synthetic augmentation. However, all numerical values were manually constructed and do not represent actual industrial observations, generated synthetic records, or completed machine learning experiments. The hypothetical comparison assigns different performance advantages to Conditional VAE and Conditional GAN configurations, but these differences cannot establish experimental superiority. Reliable conclusions would require implemented generators, authentic or reproducibly simulated source data, independent testing, and appropriate statistical uncertainty analysis.

Overall, the study identifies statistical fidelity, synthetic diversity, minority-class representation, overfitting, information leakage, privacy, and computational complexity as important considerations in synthetic data research. Future empirical investigations should evaluate these factors across different datasets, classification architectures, and augmentation levels. Synthetic observations may improve machine learning development when they preserve meaningful characteristics and provide information not adequately represented in the original training dataset. Nevertheless, their benefits must be established through independently verified predictive evidence rather than assumptions that larger or more sophisticated synthetic datasets necessarily produce better models.

References

1. Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017). Generating multi-label discrete patient records using generative adversarial networks. In Machine Learning for Healthcare Conference (pp. 286–305). PMLR. <https://proceedings.mlr.press/v68/choi17a.html>
2. Fonseca, J., & Bacao, F. (2023). Tabular and latent space synthetic data generation: A literature review. *Journal of Big Data*, 10, Article 115. <https://doi.org/10.1186/s40537-023-00792-7>
3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 27. Curran Associates.
4. Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, 20, Article 108. <https://doi.org/10.1186/s12874-020-00977-1>
5. Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493, 28–45. <https://doi.org/10.1016/j.neucom.2022.04.053>
6. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 33, 6840–6851.
7. Jordon, J., Yoon, J., & van der Schaar, M. (2019). PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations (ICLR)*. <https://openreview.net/forum?id=S1zk9iRqF7>
8. Kotelnikov, A., Baranchuk, D., Rubachev, I., & Babenko, A. (2023). TabDDPM: Modelling tabular data with diffusion models. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 17564–17579). PMLR. <https://proceedings.mlr.press/v202/kotelnikov23a.html>
9. Liu, T., Tao, D., & Xu, C. (2017). A survey on data augmentation for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*.
10. Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 399–410). IEEE. <https://doi.org/10.1109/DSAA.2016.49>
11. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6, Article 60. <https://doi.org/10.1186/s40537-019-0197-0>
12. Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems*, 32. Curran Associates. <https://arxiv.org/abs/1907.00503>

13. Yoon, J., Jarrett, D., & van der Schaar, M. (2019). Time-series generative adversarial networks. In Advances in Neural Information Processing Systems, 32. Curran Associates. <https://papers.nips.cc/paper/8789-time-series-generative-adversarial-networks>
14. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. <https://doi.org/10.1613/jair.953>
15. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE International Joint Conference on Neural Networks (pp. 1322–1328). IEEE. <https://doi.org/10.1109/IJCNN.2008.4633969>
16. Xu, L., Veeramachaneni, K., et al. (2021). Synthesizing tabular data using generative models: A review of methods and applications. Proceedings of the IEEE International Conference on Data Science and Advanced Analytics.
17. Dankar, F. K., Ibrahim, M., & Ismail, L. (2022). A multi-dimensional evaluation of synthetic data generators. IEEE Access, 10, 11147–11158.
18. Esteban, C., Hyland, S. L., & Rätsch, G. (2017). Real-valued (medical) time series generation with recurrent conditional GANs. arXiv. <https://arxiv.org/abs/1706.02633>
19. Nikolenko, S. I. (2021). Synthetic data for deep learning. Springer. <https://doi.org/10.1007/978-3-030-75178-4>
20. Shi, R., Wang, Y., Du, M., Shen, X., & Wang, X. (2025). A comprehensive survey of synthetic tabular data generation. arXiv. <https://arxiv.org/>