A ES
ISSN: 2096-3246
Volume 58, Issue 02, May, 2026

SYNTHETIC DATA GENERATION FOR RESPONSIBLE TABULAR MACHINE LEARNING: REVIEW AND RESEARCH DIRECTIONS

1st Savita Harer 2nd Shashank Swami
1st Department of Computer Science and Engineering
Vikrant University, Gwalior, Madhya Pradesh, India
savitaharer@gmail.com
2nd Department of Computer Science and Engineering
Vikrant University, Gwalior, Madhya Pradesh, India
shashank.swami2011@gmail.com

***Abstract*—** Tabular machine learning is widely used in real-world applications such as healthcare, finance, and insurance. However, these systems face significant challenges, including data imbalance, bias, and privacy risks, which may affect model performance and reliability. Synthetic data generation has emerged as an effective solution to address these issues by creating artificial datasets that preserve statistical characteristics while protecting sensitive information.

This paper presents a comprehensive review of synthetic data generation techniques for tabular machine learning. It analyzes their effectiveness in improving data quality, mitigating bias, handling class imbalance, and ensuring privacy preservation. In addition, the study provides a structured taxonomy and comparative analysis of existing methods.

The review also identifies key research gaps, particularly the lack of unified frameworks and standardized evaluation metrics. It highlights the need for simple, scalable, and integrated solutions that balance performance, fairness, and privacy for real-world applications

***Keywords*—** *Synthetic Data Generation, Tabular Machine Learning, Data Imbalance, Bias Mitigation, Privacy Preservation, Responsible AI, Data Augmentation*

I. Introduction

Tabular data is one of the most common data formats used in machine learning. It is organized into rows and columns, where each row represents a record and each column represents a feature. This type of data is widely used in real-world domains such as banking, healthcare, education, and business systems [1], [2].

For example, in banking, customer and transaction details are stored as tabular data. In healthcare, patient records are maintained in tables, while in education, student performance data is recorded in tabular format. Because it is simple and easy to interpret, tabular data is widely used for developing machine learning models [3].

However, tabular data presents several important challenges that affect the performance of machine learning systems. One major issue is class imbalance, where some classes contain many records while others have very few. As a result, machine learning models tend to focus more on the majority class and perform poorly on the minority class [12], [13].

Another issue is bias in data. Bias occurs when some groups are not properly represented in the dataset, which can lead to unfair results. This is especially important in applications such as loan approval, hiring, and healthcare decisions. If the data is biased, the model may also produce biased outputs [10], [11].

Privacy is also a major concern in tabular data. Many datasets contain sensitive information such as personal details, financial data, or medical records. Sharing such data without proper protection can create privacy risks. Therefore, access to real-world data is often limited [8]. To address these challenges, synthetic data generation has emerged as an important approach. Synthetic data is artificial data created from patterns in real data. It resembles real data but does not contain actual personal information [1], [5].

Synthetic data offers several benefits. It can increase the number of samples in minority classes, helping to reduce imbalance. It can also improve fairness by adding more data for

underrepresented groups. At the same time, it helps protect privacy because the data is not directly linked to real individuals [6].

Numerous techniques have been developed to generate synthetic data for tabular datasets. Some methods are simple and easy to use, while others are more advanced and capable of capturing complex relationships in the data [14], [17], [19]. Although these methods are useful, they still face challenges such as maintaining data quality, fully reducing bias, and balancing privacy with model performance [8].

This paper makes the following key contributions:

1. A comprehensive and structured review of synthetic data generation techniques for tabular machine learning, focusing on imbalance, bias, and privacy challenges.
2. A clear taxonomy categorizing synthetic data methods into statistical, simulation-based, learning-based, and hybrid approaches.
3. A detailed comparative analysis highlighting the strengths and limitations of existing techniques.
4. Identification of critical research gaps, particularly the lack of unified frameworks addressing imbalance, bias, and privacy simultaneously.
5. A conceptual architecture for synthetic data generation integrating preprocessing, generation, bias handling, and privacy evaluation.

Unlike many existing studies that focus on a single issue, this paper considers data imbalance, bias, and privacy together in an integrated manner. It also presents a structured framework that combines preprocessing, data generation, bias mitigation, and privacy evaluation, making it more suitable for practical use in real-world tabular machine learning applications. Recent advancements in synthetic data generation have improved the quality, fairness, and privacy of tabular machine learning systems. Therefore, synthetic data is becoming an important solution for developing secure, reliable, and responsible AI applications. In recent years, synthetic data generation has become an important research area in machine learning because it helps improve data quality while protecting sensitive information. Modern techniques are now able to generate more realistic and useful tabular data, making machine learning systems more reliable, fair, and secure for real-world applications.

II. Review Methodology

This study follows a systematic and structured approach to review existing research on synthetic data generation for tabular machine learning. Relevant research papers were collected from well-established digital libraries, including IEEE Xplore, SpringerLink, and Scopus, using keywords such as “synthetic data,” “tabular data,” “bias,” “imbalance,” and “privacy.”

Only papers published between 2019 and 2025 were considered to ensure the inclusion of recent advancements. Priority was given to peer-reviewed journal articles and reputed conference papers. Irrelevant, duplicate, and low-quality studies were excluded based on title, abstract, and full-text screening.

The selected studies were analyzed based on methodology, application domain, performance metrics, and limitations. Special attention was given to studies addressing real-world challenges such as fairness, data imbalance, and privacy preservation.

This methodology ensures that the review is comprehensive, relevant, and focused on current research trends while maintaining quality and consistency.

III. Literature Review

A. Overview of Synthetic Data Generation

Synthetic data refers to artificially generated data that preserves the statistical properties and underlying patterns of real-world datasets without exposing sensitive information. It is widely used for data augmentation, model training, and secure data sharing.

Fonseca and Bacao [1] provide a comprehensive review of synthetic data generation techniques, highlighting that these methods range from simple statistical approaches to advanced learning-based models. Similarly, Goyal and Mahmoud [2] discuss various synthetic data generation strategies and emphasize their role in improving data availability and machine learning performance.

Papadaki et al. [3] further explain that modern synthetic data generation methods aim to preserve feature relationships and dependencies in tabular datasets. Rashidi et al. [4] propose automated platforms for generating and validating synthetic tabular data, demonstrating their effectiveness in real-world applications.

Additionally, diffusion-based approaches such as Kotelnikov et al. [5] and recent studies [22], [23], [24] highlight the growing importance of synthetic data.

Recent advancements show that diffusion models combined with privacy and large language models significantly improve data realism, diversity, and privacy preservation [22], [23], [24].

B. Synthetic Data for Handling Data Imbalance

Class imbalance is one of the most common challenges in tabular datasets. It occurs when certain classes have significantly fewer samples than others, leading to biased predictions and poor model performance. Traditional techniques such as oversampling and undersampling have been widely used to address this issue. However, He and Garcia [12] and Haixiang et al. [13] show that these methods often fail to capture complex relationships in the data and may lead to overfitting.

Synthetic data generation provides a more effective alternative by creating new samples for minority classes. Jordon et al. [8] highlight that synthetic data improves classification accuracy and enhances model generalization. Recent research also shows that synthetic data can address both class imbalance and fairness imbalance, which are often present together in real-world datasets [11], [13]. This makes synthetic data particularly useful in applications requiring fairness and balanced representation.

Most current approaches focus on improving accuracy, but they often overlook fairness across different groups. This makes them less suitable for sensitive fields like healthcare and finance, where equal and balanced representation is very important.

C. Synthetic Data for Bias Reduction

Bias in machine learning arises when datasets do not represent all groups fairly, leading to discriminatory outcomes. This is a major concern in applications such as credit scoring, hiring systems, and healthcare diagnosis.

Mehrabi et al. [11] provide a detailed survey on bias and fairness in machine learning, emphasizing the need for balanced datasets. Synthetic data can help reduce bias by generating additional samples for underrepresented groups, thereby improving fairness. However, Barocas et al. [10] highlight a critical limitation: if the original dataset is biased, synthetic data may also inherit this bias. Therefore, bias mitigation requires careful design of data generation techniques.

Recent approaches focus on combining synthetic data generation with data balancing strategies to improve fairness across demographic groups [5], [11], [25]. These methods show promising results in reducing bias while maintaining model performance. Recent work also emphasizes generating synthetic data that simultaneously preserves fairness and privacy, addressing both ethical and regulatory concerns [25].

A key limitation of current bias mitigation techniques is that they rely heavily on the quality of the original dataset. If the source data contains inherent bias, synthetic data generation may amplify these issues rather than resolve them.

D. Synthetic Data for Privacy Preservation

Privacy preservation is a major concern in tabular data, especially in sensitive domains such as healthcare and finance. Sharing real data can lead to privacy breaches and legal issues.
Synthetic data provides a solution by generating artificial records that do not correspond to real individuals. Goncalves et al. demonstrate that synthetic patient data can be used for analysis while reducing the risk of data leakage [6].
Similarly, Murtaza et al. highlight that synthetic data enables privacy-preserving data sharing without compromising data utility [7]. Jordon et al. [8] proposed PATE-GAN to improve privacy preservation in synthetic data generation. This highlights the ongoing challenge of balancing data utility and privacy protection.
Despite advancements in privacy-preserving techniques, a significant trade-off exists between data utility and privacy. Strong privacy guarantees often reduce the usefulness of synthetic data for downstream machine learning tasks.

E. Types of Synthetic Data Generation Techniques

Synthetic data generation techniques can be broadly classified into three categories:

Statistical Methods: These methods rely on probability distributions to generate new data. They are simple, interpretable, and suitable for small datasets.

Simulation-Based Methods: These approaches use domain knowledge and predefined rules to generate data. They are commonly used in healthcare and engineering applications.

Learning-Based Methods: These methods learn patterns from real data and generate new samples accordingly. They are capable of capturing complex relationships between features.

A large-scale review by Kotelnikov et al. [5] analyzes multiple generation techniques and highlights their advantages and limitations. Similarly, Papadaki et al. [3] emphasize that modern methods focus on preserving feature relationships and correlations in tabular data.

F. Evaluation of Synthetic Data Quality

Evaluating synthetic data is a critical aspect of research in this field. Researchers use multiple metrics to assess the quality and usefulness of synthetic datasets.

Common evaluation criteria include:

- Statistical similarity between real and synthetic data
- Machine learning model performance
- Privacy risk

Mehrabi et al. [11] highlight the importance of using multiple evaluation metrics to ensure both data utility and privacy. Benchmarking studies also emphasize the need for standardized evaluation frameworks. However, there is currently no universally accepted evaluation standard [1], [2], [3], which remains a significant challenge.

IV. Taxonomy of Synthetic Data Generation

Synthetic data generation techniques for tabular machine learning can be systematically categorized based on their underlying principles and data modeling approaches. A clear

taxonomy helps in understanding the strengths, limitations, and applicability of each method, as shown in Table I.

A. Statistical-Based Methods

Statistical methods generate synthetic data using probability distributions derived from real datasets. These approaches assume that the underlying data follows a known distribution such as Gaussian, Poisson, or Multinomial. These methods are simple, computationally efficient, and easy to interpret. However, they often fail to capture complex relationships [1] between features, especially in high-dimensional datasets. As a result, their performance is limited in real-world scenarios where data dependencies are nonlinear. For example, in a banking dataset, if synthetic income data is generated using a normal distribution, it may not reflect the actual relationship between income, loan amount, and repayment behavior.

B. Simulation-Based Methods

Simulation-based methods rely on domain knowledge and predefined rules to generate synthetic data. These approaches are commonly used in fields such as healthcare, finance, and engineering. They provide high control over data generation and ensure realistic behavior based on domain constraints. However, they lack generalization capability and require expert knowledge for designing simulation rules.

For example, in an insurance system, a simulation model may generate claims data based on age, policy type, and risk factors. However, it may fail to capture unexpected real-world variations.

C. Learning-Based Methods

Learning-based methods generate synthetic data by learning patterns directly from real datasets. These approaches are capable of capturing complex feature interactions and high-dimensional relationships [17], [19], [26], [27]. They provide high-quality synthetic data and are widely used in modern applications [17], [19]. However, they are computationally expensive and often require large datasets for training.

For example, in a healthcare dataset, a learning-based method can generate realistic patient records by learning relationships among age, symptoms, diagnosis, and treatment. This makes the synthetic data more useful for training machine learning models.

Recent hybrid approaches combining GANs with transformers have further improved data quality and representation learning in tabular datasets [26], [27].

D. Hybrid Methods

Hybrid approaches combine statistical and learning-based techniques to improve performance [6]. These methods aim to balance interpretability and accuracy. They often provide better results than individual methods but increase system complexity.

For example, in a loan prediction dataset, a hybrid method may use statistical techniques to generate basic features such as age and income, while a learning-based model captures complex relationships such as credit behavior and repayment patterns.

V. Architecture Diagram

This paper proposes a Responsible Synthetic Data Generation Framework (RSDGF) designed to address data imbalance, bias, and privacy in a unified manner.

The framework consists of the following components:

- Data Preprocessing Module: Handles missing values, encoding, and normalization
- Synthetic Data Generator: Uses statistical or learning-based models

- Bias Mitigation Module: Ensures fair representation across groups
- Privacy Preservation Module: Applies privacy-aware mechanisms
- Evaluation Module: Assesses utility, fairness, and privacy
- This framework provides a structured approach for generating high-quality synthetic data suitable for real-world machine learning applications.

Figure 1 illustrates Architecture of Synthetic Data Generation illustrates the overall architecture of the synthetic data generation process for tabular machine learning. The workflow begins with real tabular data collected from domains such as banking, healthcare, or education.

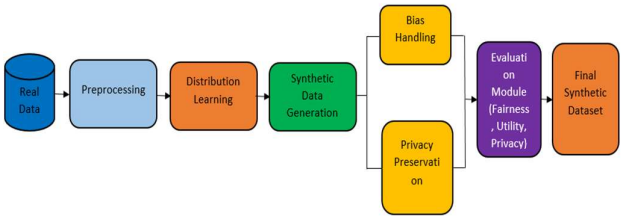

Figure 1. Architecture of Synthetic Data Generation

Initially, the data undergoes a preprocessing stage, where it is cleaned by handling missing values and removing inconsistencies. In addition, categorical features are encoded into suitable formats, and numerical features are normalized to ensure uniformity. This step prepares the dataset for effective learning.

Following preprocessing, the system performs distribution learning. At this stage, the model captures the underlying patterns and relationships present in the data using statistical or learning-based approaches. This enables the system to understand how the original data is structured.

Next, the trained model generates synthetic tabular records that preserve the key statistical properties of the original dataset. These generated records are then evaluated using multiple quality measures such as data similarity, model utility, fairness, and privacy risk.

Finally, the validated synthetic dataset can be used for machine learning tasks, data sharing, testing, or research purposes without directly exposing sensitive real-world information.

Based on the learned distribution, synthetic data is then generated. The generated data mimics the statistical properties of the real dataset while ensuring that it does not contain any sensitive or identifiable information.

Subsequently, bias handling and data balancing techniques are applied to ensure fair representation of different classes or groups. This step helps reduce class imbalance and minimize bias in the dataset[14].

The generated data is then evaluated for privacy to ensure that it does not leak sensitive information from the original dataset. Finally, the validated synthetic data is used for training and evaluating machine learning models, where performance is assessed using metrics such as accuracy, fairness, and privacy.

VI. Challenges in Existing Research

Despite the advantages of synthetic data, several challenges remain:

- High computational complexity in advanced methods
- Difficulty in maintaining data quality

- Trade-off between privacy and accuracy
- Lack of standard evaluation metrics
- Limited applicability to real-world datasets[5]

Kotelnikov et al. [5] highlight that diffusion-based models require large datasets and substantial computational resources, which limits their practical use.

VII. Research Gaps

The analysis reveals that existing research lacks a unified framework capable of simultaneously addressing data imbalance, bias, and privacy preservation. Most existing approaches focus on individual challenges, resulting in fragmented solutions.

Furthermore, there is a lack of standardized evaluation metrics that jointly assess data utility, fairness, and privacy. This limitation makes it difficult to compare different methods effectively and leads to inconsistent conclusions across studies. In addition, the absence of benchmark datasets and unified evaluation protocols further complicates the validation of proposed techniques.

These gaps highlight the need for a simple, scalable, and integrated synthetic data generation framework that balances performance, fairness, and privacy for real-world applications. Such a framework should not only improve data quality but also ensure ethical and reliable decision-making in sensitive domains.

A. Single-Problem Focus

Most studies address only one issue (either imbalance, bias, or privacy), while very few works consider all three challenges together within a unified framework. Current evaluation methods primarily measure accuracy or statistical similarity but often ignore fairness and privacy aspects. There is no standard metric that simultaneously evaluates data utility, bias reduction, and privacy preservation. Additionally, many approaches lack cross-domain validation, limiting their general applicability.

B. Trade-off Problem

Improving privacy often reduces model performance, while reducing bias may distort the original data distribution. This creates a fundamental trade-off among accuracy, fairness, and privacy. Existing methods fail to provide an optimal balance among these factors. There is a need for adaptive techniques that dynamically manage these trade-offs based on application requirements and domain constraints.

C. Limited Real-World Applicability

Many approaches are tested only on small, clean, or benchmark datasets, which do not reflect real-world complexities. Real-world tabular data, such as loan records, healthcare data, and insurance datasets, are often noisy, incomplete, biased, and highly sensitive. Existing models struggle to generalize effectively in such environments. Moreover, issues such as scalability, computational cost, and deployment feasibility are not adequately addressed.

Based on the literature, the following research gaps are identified:

- ***Lack of unified frameworks addressing bias, imbalance, and privacy together [1], [2].***

- ***Over-dependence on complex and computationally expensive models .***
- ***Limited focus on simple, interpretable, and scalable methods.***
- ***Absence of standardized and comprehensive evaluation metrics.***
- ***Insufficient validation on large-scale and real-world datasets.***
- ***Lack of explainability and transparency in synthetic data generation models.***

Addressing these gaps requires the development of a robust, interpretable, and scalable synthetic data generation framework that can effectively balance data utility, fairness, and privacy while being applicable to real-world scenarios.

VIII. Comparative Analysis

The comparison of different synthetic data generation methods, as presented in Table I, indicates that each approach demonstrates strengths in specific areas while also exhibiting certain limitations.

Statistical methods are simple, interpretable, and computationally efficient. However, their applicability to real-world problems is limited, as they struggle to capture complex and non-linear relationships within data. Although they can partially address class imbalance, they are less effective in reducing bias and generating high-quality synthetic data for advanced applications [1].

In contrast, learning-based methods, including deep learning and generative adversarial network (GAN)-based approaches such as CTAB-GAN and Conditional GAN, show strong performance in handling data imbalance and generating realistic datasets. These methods are capable of capturing complex feature interactions and high-dimensional relationships, thereby providing high-quality synthetic data. However, they require substantial computational resources and large training datasets. Furthermore, they may introduce privacy risks if sensitive information from the original dataset is unintentionally learned [3], [17], [19]. GAN methods outperform statistical models in nonlinear data representation but suffer from instability and privacy leakage, limiting their real-world deployment.

Hybrid methods attempt to combine the advantages of statistical and learning-based approaches by balancing performance, bias reduction, and privacy preservation. These methods are particularly effective in domain-specific applications, such as healthcare. However, their effectiveness may not generalize across diverse datasets, and they often increase system complexity [6].

Frameworks such as Synthetic Data Vault provide flexible and practical solutions for synthetic data generation by supporting multiple modeling techniques. While they are useful in real-world applications, their performance heavily depends on proper parameter tuning and configuration [14]. Similarly, benchmarking studies play an important role in evaluating and comparing different approaches, but they do not provide a standardized solution or unified evaluation framework [8].

Overall, the analysis clearly shows that no single method can effectively address data imbalance, bias, and privacy simultaneously. Most existing approaches focus on one or two aspects while overlooking others, which highlights the need for a unified and balanced synthetic data generation framework.Table I. Comparative Analysis Of Synthetic Data Methods

Ref	Method Type	Type	Imbalance	Bias	Privacy	Strength	Limitations
1	Statistical Models	Statistical	Medium	Low	Medium	Simple, interpretable	Limited complexity
3	Learning-Based Models	Deep Learning	High	Medium	Low	High accuracy	High computation

6	Healthcare Synthetic	Hybrid	Medium	Medium	High	Domain-specific	Limited generalization
8	Privacy-Preserving GAN	Evaluation	High	Medium	Medium	Good comparison	No standard metric
14	SDV Framework	Learning	High	Low	Medium	Flexible	Needs tuning
17	CTAB-GAN	Learning	High	Medium	Medium	Captures relationships	Complex training
19	Conditional GAN	Learning	High	Medium	Low	Good performance	Privacy risk

IX. Mathematical Formulation

TO EVALUATE THE QUALITY OF SYNTHETIC DATA, STANDARD MACHINE LEARNING PERFORMANCE METRICS AND FAIRNESS MEASURES ARE USED. THESE METRICS ASSESS CLASSIFICATION PERFORMANCE, PREDICTIVE CAPABILITY, AND POTENTIAL BIAS IN THE GENERATED DATA.

A. Accuracy:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

B. Precision:

$$\text{Precision} = TP / (TP + FP) \quad (2)$$

C. Recall:

$$\text{Recall} = TP / (TP + FN) \quad (3)$$

D. F1-Score:

$$\text{F1 score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

E. Bias Measurement:

$$\text{Bias (SPD)} = P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1) \quad (5)$$

Where:

$\hat{Y}$ = Predicted output

A = Sensitive attribute (e.g., gender, age)

These metrics collectively provide a comprehensive evaluation of synthetic data quality in terms of predictive performance, fairness, and reliability [11].

F. Disparate Impact (Fairness):

$$DI = \hat{P}(Y=1|A=0) / \hat{P}(Y=1|A=1) \quad (6)$$

G. Differential Privacy:

$$P(M(D)) \approx P(M(D')) \quad (7)$$

The differential privacy equation shows that the output of the model remains almost the same even if a single record in the dataset is changed or removed. This helps protect sensitive personal information while allowing the data to be used for analysis and machine learning tasks.

H. GAN objective function

$$\min_G \max_D V(D, G) = \log D(x) + \log(1 - D(G(z))) \quad (8)$$

The GAN objective function describes the working process between two models called the generator and the discriminator. The generator creates synthetic data samples, while the discriminator checks whether the data is real or generated. The term log D(x) represents the ability of the discriminator to correctly identify real data, whereas log (1 − D(G(z))) shows how well it can detect synthetic data produced by the generator. During training, both models improve continuously, which helps generate more realistic and high-quality synthetic tabular data.

X. Discussion

The literature clearly demonstrates that synthetic data generation is a powerful approach for improving tabular machine learning systems. It plays a significant role in addressing key challenges such as data imbalance, bias, and privacy risks. One major observation is that most existing approaches focus on solving individual problems rather than providing a unified solution. For instance, several methods effectively address class imbalance but fail to ensure fairness across different demographic groups. Similarly, privacy-focused approaches often reduce data utility, which negatively affects model performance.

Another important finding is the inherent trade-off between data utility and privacy. Increasing privacy protection often results in reduced model accuracy, while improving fairness may require altering data distributions, which can impact predictive performance. This trade-off remains one of the most critical challenges in synthetic data research.

The literature also reveals a strong dependence on complex learning-based models. Recent benchmarking studies indicate that diffusion-based models often outperform GAN and VAE-based approaches in terms of data quality and robustness, although they require higher computational resources [28]. While these methods generate high-quality synthetic data, they require large datasets, substantial computational resources, and expert-level implementation. This limits their applicability in real-world scenarios, particularly in small-scale industries and organizations with limited resources.

Furthermore, the absence of standardized evaluation frameworks makes it difficult to compare different synthetic data generation techniques. Different studies use different performance metrics, leading to inconsistent conclusions. From a practical perspective, real-world datasets such as loan approval systems, healthcare records, and insurance data are highly complex, noisy, and biased. Existing methods often fail to fully address these real-world challenges.

The comparative analysis indicates that no single method is universally optimal. Statistical models are suitable for low-complexity datasets, while learning-based approaches perform better

in high-dimensional environments. However, their computational cost and privacy risks limit their real-world deployment. Hybrid approaches offer a promising direction but require further research to improve scalability and generalization.

XI. Conclusion and Future Work

This paper presented a comprehensive literature review of synthetic data generation techniques for addressing key challenges such as data imbalance, bias, and privacy in tabular machine learning. The study examined various approaches, including statistical, simulation-based, learning-based, and hybrid methods, and analyzed their strengths and limitations.

The analysis indicates that learning-based methods generally achieve better performance in capturing complex data patterns; however, they are computationally expensive and may introduce privacy risks. In contrast, statistical methods are simple and interpretable but are limited in handling complex relationships. Hybrid approaches provide a balance between performance and interpretability, although they increase overall system complexity. A key observation from this study is that no existing method effectively addresses imbalance, bias, and privacy simultaneously, highlighting the need for a unified and practical framework.

Future research should focus on developing simple, scalable, and efficient solutions that can handle multiple challenges while maintaining data quality, fairness, and privacy. In particular, there is a need to design unified frameworks that integrate imbalance handling, bias mitigation, and privacy preservation within a single system.

Additionally, future work should explore lightweight and computationally efficient models suitable for real-world deployment. The development of standardized evaluation metrics that jointly assess data utility, bias reduction, and privacy preservation is also essential for the consistent comparison of different methods. Furthermore, validating proposed approaches on real-world datasets, such as healthcare records, financial transactions, and insurance data, will improve their practical applicability.

Finally, there is a growing need for explainable and interpretable synthetic data generation models that provide transparency and build trust in machine learning systems. Addressing these directions will contribute to the development of reliable and responsible synthetic data solutions for real-world applications. The proposed research direction can support the development of trustworthy and responsible AI systems in sensitive real-world domains.

References

[1] J. Fonseca and F. Bacao, “Tabular and latent space synthetic data generation: A literature review,” Journal of Big Data, vol. 10, no. 115, 2023.

[2] M. Goyal and Q. H. Mahmoud, “A systematic review of synthetic data generation techniques,” Electronics, vol. 13, no. 17, 2024.

[3] E. Papadaki, A. G. Vrahatis, and S. Kotsiantis, “Exploring innovative approaches to synthetic tabular data generation,” Electronics, vol. 13, no. 10, 2024.

[4] H. H. Rashidi et al., “A novel and fully automated platform for synthetic tabular data generation and validation,” Scientific Reports, vol. 14, 2024.

[5] K. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko, “TabDDPM: Modelling tabular data with diffusion models,” in Proc. 40th Int. Conf. Machine Learning (ICML), 2023.

[6] A. Goncalves et al., “Generation and evaluation of synthetic patient data,” Neurocomputing, vol. 493, pp. 28–45, 2022.

[7] H. Murtaza et al., “Synthetic data generation in healthcare: A review,” Computer Science Review, vol. 48, 2023.

[8] J. Jordon, J. Yoon, and M. van der Schaar, “PATE-GAN: Generating synthetic data with differential privacy,” in Proc. Int. Conf. Learning Representations (ICLR), 2020.

[9] K. El Emam, L. Mosquera, and R. Hoptroff, *Practical synthetic data generation: Balancing privacy and the broad availability of data*. Sebastopol, CA, USA: O'Reilly Media, 2020.
[10] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. Cambridge, MA, USA: MIT Press, 2019.
[11] N. Mehrabi et al., "A survey on bias and fairness in machine learning," ACM Computing Surveys, vol. 54, no. 6, 2021.
[12] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.
[13] G. Haixiang et al., "Learning from class-imbalanced data: A review," Expert Systems with Applications, vol. 73, pp. 220–239, 2017.
[14] N. Patki, R. Wedge, and K. Veeramachaneni, "The synthetic data vault," in Proc. IEEE Int. Conf. Data Science and Advanced Analytics (DSAA), 2016, pp. 399–410.
[15] J. Yoon, D. Jarrett, and M. van der Schaar, "Time-series generative adversarial networks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2019.
[16] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in Proc. Int. Conf. Learning Representations (ICLR), 2014.
[17] Z. Zhao, A. Kunar, R. Birke, and L. Y. Chen, "CTAB-GAN+: Enhancing tabular data synthesis," IEEE Access, vol. 11, pp. 56789–56802, 2023.
[18] Z. Zhao et al., "VT-GAN: Cooperative tabular data synthesis," in Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW), 2023.
[19] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," in Advances in Neural Information Processing Systems (NeurIPS), 2019.
[20] A. Assefa et al., "Generating synthetic data in finance," in Proc. ACM Int. Conf. AI in Finance (ICAIF), 2020.
[21] T. DeVries and G. W. Taylor, "Dataset augmentation in feature space," in Proc. Int. Conf. Learning Representations (ICLR), 2017.
[22] C. Zhu, Y. Wang, and X. Liu, "DP-TLDM: Differentially Private Tabular Latent Diffusion Model for Synthetic Data Generation," 2024.
[23] T. Sattarov, A. Filippone, and M. Petrov, "Differentially Private Federated Learning of Diffusion Models for Synthetic Tabular Data Generation," in Proc. IEEE Int. Conf. Data Mining Workshops (ICDMW), 2024, pp. 1–8.
[24] Y. Chen, H. Zhao, and J. Li, "Generating Realistic Synthetic Tabular Data with Integrated LLM and Diffusion Models," Neurocomputing, vol. 612, pp. 128945, 2025.
[25] F. Sarmin, M. Hasan, and S. Rahman, "Privacy-Preserving Fair Synthetic Tabular Data Generation," Expert Systems with Applications, vol. 258, pp. 125041, 2025.
[26] H. Lee, J. Kim, and S. Park, "TAEGAN: Generating Synthetic Tabular Data for Data Augmentation," in Proc. Int. Joint Conf. Neural Networks (IJCNN), 2024, pp. 1–8.
[27] R. Sharma and P. Gupta, "TabTransGAN: A Hybrid GAN–Transformer Approach for Tabular Data Synthesis," IEEE Access, vol. 13, pp. 45678–45691, 2025.
[28] A. Kumar, S. Verma, and R. Singh, "Tabular Data Generation Models: Benchmarking GAN, VAE, and Diffusion Methods," Pattern Recognition Letters, vol. 189, pp. 45–58, 2025.