




Review

Exploring Innovative Approaches to Synthetic Tabular Data Generation

Eugenia Papadaki ^{1,*}, Aristidis G. Vrahatis ^{1,†} and Sotiris Kotsiantis ^{2,†}¹ Bioinformatics and Human Electrophysiology Laboratory, Department of Informatics, Ionian University, 49100 Corfu, Greece; aris.vrahatis@ionio.gr² Department of Mathematics, University of Patras, 26504 Patras, Greece; sotos@math.upatras.gr

* Correspondence: epapadaki@ionio.gr

† These authors contributed equally to this work.

Abstract: The rapid advancement of data generation techniques has spurred innovation across multiple domains. This comprehensive review delves into the realm of data generation methodologies, with a keen focus on statistical and machine learning-based approaches. Notably, novel strategies like the divide-and-conquer (DC) approach and cutting-edge models such as GANBLR have emerged to tackle a spectrum of challenges, spanning from preserving intricate data relationships to enhancing interpretability. Furthermore, the integration of generative adversarial networks (GANs) has sparked a revolution in data generation across sectors like healthcare, cybersecurity, and retail. This review meticulously examines how these techniques mitigate issues such as class imbalance, data scarcity, and privacy concerns. Through a meticulous analysis of evaluation metrics and diverse applications, it underscores the efficacy and potential of synthetic data in refining predictive models and decision-making software. Concluding with insights into prospective research trajectories and the evolving role of synthetic data in propelling machine learning and data-driven solutions across disciplines, this work provides a holistic understanding of the transformative power of contemporary data generation methodologies.

Keywords: data generation; synthetic data; machine learning-based generation; statistical-based generation; healthcare; privacy preservation; evaluation metrics; fidelity metrics



Citation: Papadaki, E.; Vrahatis, A.G.; Kotsiantis, S. Exploring Innovative Approaches to Synthetic Tabular Data Generation. *Electronics* **2024**, *13*, 1965. <https://doi.org/10.3390/electronics13101965>

Academic Editor: Ping-Feng Pai

Received: 19 April 2024

Revised: 10 May 2024

Accepted: 14 May 2024

Published: 17 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In the rapidly evolving field of machine learning and neural networks, there is more demand than ever for large amounts of data [1]. The effectiveness of these advanced computational models depends crucially on large amounts of high-quality data to uncover patterns, make accurate predictions, and drive further innovative advances in a variety of domains. However, the feasibility of acquiring such huge datasets, particularly in areas such as healthcare and cybersecurity, is often constrained by diverse challenges [2].

The application of synthetic data in various machine learning tasks, with a particular focus on tabular data, has garnered significant attention in the recent literature. A previous review delivers a thorough examination of the application of synthetic data across various machine learning tasks, placing a particular emphasis on tabular data [3]. While traditional reviews may be based on keyword searches, this study distinguishes itself by offering a comprehensive examination of the use of synthetic data, presenting a detailed classification of algorithms and generation mechanisms, and discussing metrics for evaluating data quality. By addressing the existing gaps in the fragmented literature, this study aims to provide valuable information to support research on the effective use of synthetic data.

A comprehensive classification is presented that includes 70 production algorithms, as well as an explanation of six main types of production mechanisms. This study goes deeper into the discussion of metrics designed to assess the quality of synthetic data. With

the goal of bridging existing gaps in the fragmented literature, this study provides valuable insights to aid researchers and practitioners in the effective utilization of synthetic data.

Federated Learning (FL) has emerged as a decentralized approach to training statistical models by leveraging data from multiple clients without exposing their raw data, thus ensuring privacy and introducing potential security advantages [4]. Within this context, federated synthesis, employing FL for synthetic data generation, enables the amalgamation of data without compromising privacy or raw data accessibility. A scoping review of 69 articles spanning from 2018 to 2023 underscores the prevalent use of deep learning methods, notably generative adversarial networks, in federated synthesis. Although promising, further research is needed to deepen the understanding of privacy risks and develop reliable methodologies to measure them.

The pursuit of improved machine learning algorithms for personalized decision support in palliative care diagnostics underlines the need for more relevant patient data. Synthetic data generation emerges as a potential solution to address this demand, although challenges such as bias and interpretability persist. In an extensive review, the potential consequences of using synthetic data in palliative care diagnostics using machine learning are examined [5]. Furthermore, the authors provide valuable insights and practical considerations for the integration of this approach into clinical settings.

The increasing demand for large datasets is confronting the persistent obstacle of unbalanced datasets [6]. The nature of certain effects results in biased distributions, with some categories significantly outperforming others. This imbalance not only creates challenges in model training but also carries the risk of spreading biases, potentially undermining the reliability of predictive analyses. Achieving a balance in the representation of different classes becomes crucial for the robust performance of machine learning models.

Despite these challenges, the crucial issue of privacy is vital [7], especially when it comes to patients' personal data in healthcare applications in real time. Preserving sensitive information is not just a legal and ethical necessity but also a fundamental requirement to enhance trust and ensure the ethical use of data. As we navigate the complex terrain of machine learning and data-intensive applications, it becomes urgent to explore innovative solutions that reconcile the need for large, balanced datasets with the need to address privacy concerns.

This review begins a thorough exploration of contemporary research efforts that address these challenges. Taking a deep dive into the literature, we uncover a range of methodologies, with a particular focus on the innovative use of generative adversarial networks (GANs) for generating synthetic data. The necessity of large datasets, the intricacies of handling imbalanced ones, and the urgency of privacy protection are recurring issues that can be found in a wide range of domains—from cybersecurity and attack detection to healthcare and patient-oriented applications.

We employed BERT (Bidirectional Encoder Representations from Transformers) as a pivotal tool to conduct topic modeling on the collected studies. Preprocessing steps, including tokenization, lowercasing, stop-word removal, and lemmatization, were applied to ensure the cleanliness of the text data. BERTopic, a topic modeling library specifically designed to harness the power of BERT embeddings, facilitated the extraction of topics from the corpus. BERT embeddings, renowned for their ability to capture semantic nuances in text data, were generated for each document in the corpus. These embeddings served as rich representations of the textual content, enabling us to delve deeper into the underlying themes present in the literature. Clustering techniques, particularly Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), were then applied to identify cohesive clusters representing distinct topics within the corpus. After clustering, representative keywords were assigned to each topic cluster based on the most prominent terms within the documents, enhancing the interpretability of the results. The utilization of BERT embeddings in conjunction with clustering algorithms allowed for a nuanced analysis of the literature, providing insights into the key themes and implications for synthetic data generation, as presented in Figure 1.

Furthermore, our methodology leveraged the adaptability of BERTopic, which dynamically adjusts to the complexity of the data without requiring a predefined number of topics. This flexibility ensured that significant words were preserved in the topic descriptions, maintaining the semantic integrity and coherence of the extracted topics. Overall, our approach showcases the potential of deep learning-based algorithms, such as BERTopic, in the domain of text mining and processing, particularly in facilitating topic extraction and analysis from large collections of scientific articles.

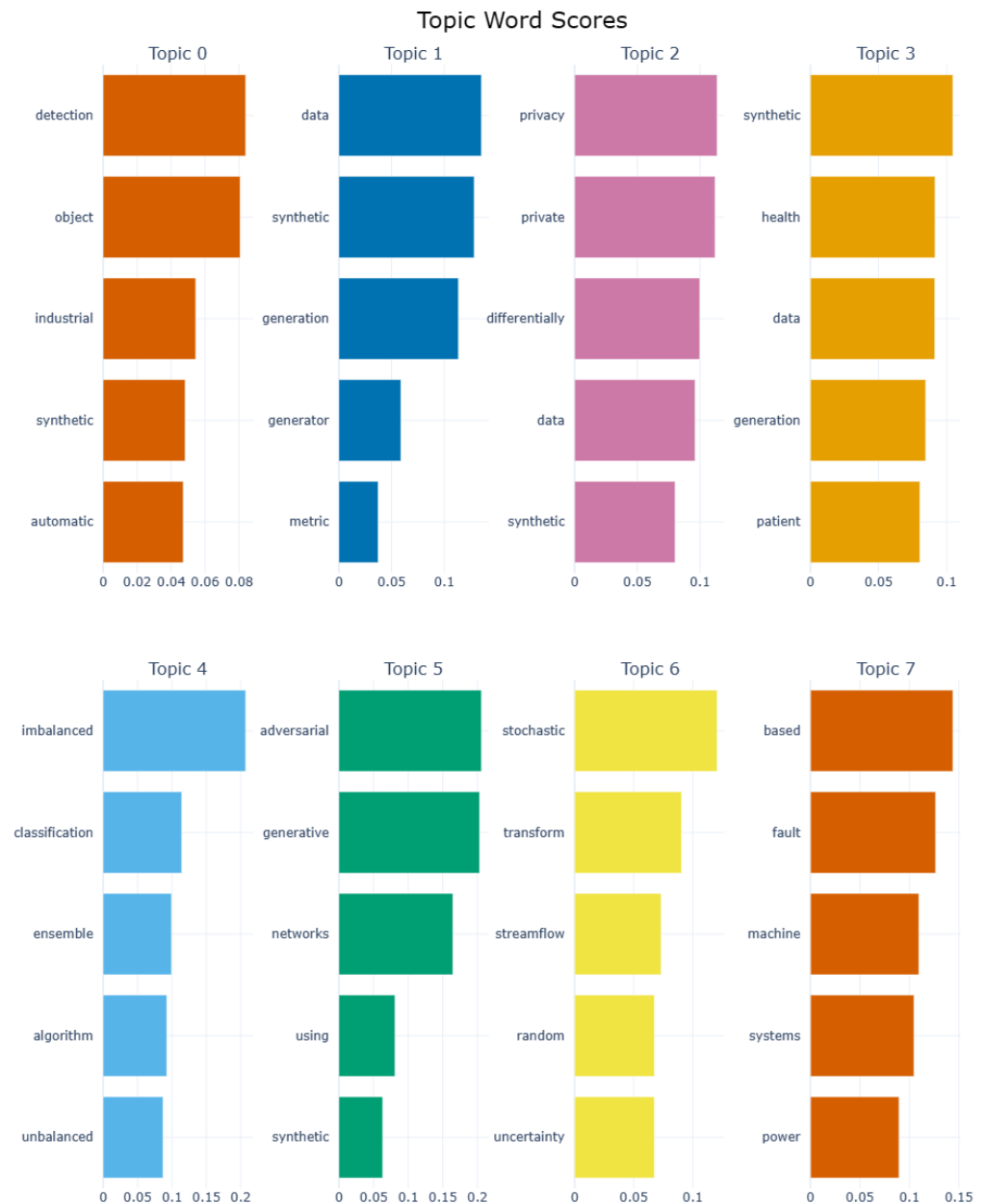


Figure 1. Results from BERTopic illustrating four distinct topics generated via BERT. Each bar chart displays the top five words or terms associated with a topic. The x-axis represents the c-TF-IDF scores, which quantify term relevance by assessing both frequency within the topic and uniqueness across the document corpus. The top word in each chart, indicated by the highest score, highlights the most defining and unique term for that specific topic, serving as the central theme around which the other terms are contextually aligned.

Figure 1 displays the results obtained using BERTopic. Specifically, the bar charts represent four distinct topics generated by BERT, showcasing the top five associated words

or terms for each topic. The x-axis of these bar charts represents the c-TF-IDF score, which quantifies the relevance of terms by evaluating their frequency within a particular document and their distinctiveness across the entire document corpus.

As we navigate the varied terrain of synthetic data generation, we aim to distill knowledge, draw connections, and provide a comprehensive overview of the evolving methodologies that address these sophisticated challenges. Through a rigorous inspection of the reviewed papers, we seek to unravel the potential confluences between the need for extensive, balanced datasets, the intricacies of handling imbalances, and privacy requirements—paving the way for a more informed and ethical approach to data-intensive applications in the field of machine learning and neural networks.

2. Related Work

In this section, we will explore existing research and methodologies in the field of synthetic data generation across various domains, such as healthcare and retail. This exploration will encompass various aspects of data generation, such as numerical and temporal data. In addition, we will dig further into metrics for evaluating synthetic data, examining how datasets are used to assess their fidelity and usefulness. We will highlight key studies and approaches that have contributed to advancing the understanding and application of generative models, such as generative adversarial networks (GANs) and variational autoencoders (VAEs), in addressing challenges related to data scarcity, imbalance, and privacy concerns. By reviewing the latest studies, we aim to contribute to the ongoing debate on the use of synthetic data in a variety of applications.

2.1. Data Generation Approaches

As data science continues to evolve, the struggle to create meaningful datasets from limited or sensitive sources remains a pressing challenge. In response to this, recent developments have fostered the development of innovative approaches to data generation. In the sections below, we try to explore these strategies in more depth.

2.1.1. Statistical-Based Generation

In the field of data science, the challenge of generating meaningful datasets from limited or sensitive sources remains a critical task. Addressing this challenge requires innovative approaches. In this context, recent developments in statistics-based generation techniques have generated considerable interest, highlighting promising avenues for overcoming the limitations imposed by small datasets. Two such notable contributions, GenerativeMTD and the divide-and-conquer (DC) strategy for a tabular data dictionary (STD), have emerged as potential solutions in this area. Through the accurate implementation of contemporary methodologies, these approaches offer new perspectives on data generation, demonstrating superior performance and efficiency compared to conventional methods.

GenerativeMTD [8] stands out as an innovative solution to the complex issues associated with synthetic data generation from small datasets. By using pseudo-real data through mega-trend diffusion and K-Nearest Neighbor techniques, GenerativeMTD demonstrates a unique approach that outperforms other methods in terms of effectiveness and performance. The study navigates the complexities associated with small datasets and highlights the crucial aspects of maintaining data fidelity and privacy. GenerativeMTD's ability to outperform established methods for tabular data generation is particularly remarkable, pointing out its benefits and its immediate application and setting it as a benchmark for future efforts in this field. Overall, the study makes a significant contribution, not only by confronting a pressing challenge in the field but also by advancing the latest technology in synthetic data generation through the careful integration of innovative techniques.

In the field of data generation, remarkable progress is being made with the innovation of a new approach—the divide-and-conquer (DC) strategy—for a tabular data dictionary (STD) [9]. This strategy aims not only to create artificial datasets but also to preserve the complex logical relationships within the data, ensuring a more accurate representation of

real healthcare scenarios. The potential of the method lies in its practical validation, where the DC method outperforms a technique based on conditional sampling on three different disease datasets. A notable feature is the method's commitment to data balance during the STD generation process, resulting in a more robust and unbiased synthetic dataset. This research not only contributes to the developing field of synthetic data in healthcare but also highlights the application and efficiency of the DC strategy in capturing the complexity of healthcare data.

2.1.2. Machine Learning-Based Generation

In the field of data generation methodologies, machine learning-based techniques have emerged as powerful tools for the creation of datasets with diverse applications in numerous domains. This section explores several innovative studies that leverage machine learning frameworks to address specific challenges in data generation and augmentation. The use of advanced techniques, such as conditional generative adversarial networks (cGANs), tabulated generative adversarial networks (TGANs), and variational autoencoders, highlights the extent of ongoing work in this area. These methodologies seek to not only improve prediction accuracy in critical healthcare scenarios, such as fluid overload prediction and cancer diagnosis, but also extend their reach into diverse areas, such as agriculture and software engineering. Taking advantage of the potential of machine learning, these studies not only advance the boundaries of data generation but also reflect a future where tabular data will play a vital role in enhancing the reliability and effectiveness of machine learning models in various data contexts.

In order to predict fluid overload in critical care patients, the development of an ensemble machine learning model has been proposed [10]. This novel approach integrates both original and synthetically generated datasets. The distinctive feature of employing conditional generative adversarial networks (cGANs) for synthetic data generation adds a further complexity factor to the proposed approach. The primary finding, where the meta-learner trained on the combined dataset outperforms models based solely on the original data, highlights the potential of synthetic data to enhance prediction models for critical care scenarios. This study not only contributes to the expanding field of research on the applications of synthetic data in healthcare but also highlights their utility in improving the accuracy and reliability of prediction models, ultimately impacting patient care in intensive care units.

In an attempt to improve crop classification accuracy, the contribution of conditional tabular generative adversarial networks (CTGANs) to generate synthetic training data is innovative [11]. The application of deep learning-driven synthetic data generation addresses a critical challenge related to the scarcity of training data for minor crops in SAR–optical data-based classification. The application of CTGANs highlights not only the adaptive ability of creative adversarial networks but also their effectiveness in generating high-quality synthetic data, resulting in a significant improvement in the accuracy of small crop classification. The study's findings provide valuable insights into the potential of leveraging advanced deep learning techniques for enhancing classification performance in scenarios where obtaining sufficient real-world training data is a challenge. This research not only provides a major advance in the field of agricultural remote monitoring but also highlights the wide range of applications of synthetic data generation to improve the reliability of machine learning models in data-limited scenarios.

In an attempt to address the complexity of utilizing production data in the manufacturing industry for data-driven analytics, a potential solution involving the application of variational autoencoders, a form of synthetic data generation method, was proposed [12]. This research demonstrates the effective results of this approach in enhancing prediction models, which is particularly beneficial for manufacturing companies faced with limited and unbalanced data. Variational autoencoders are highlighted as the most suitable among various methods, including generative adversarial networks and synthetic minority over-sampling techniques. This is a result of the ability of variational autoencoders to efficiently

generate crucial features from small sample datasets. This research offers valuable insights for construction companies seeking to enhance the accuracy and certainty of predictions by maximizing synthetic data generation power in the field of industrial data analytics.

Recognizing the success of GANs in generating tabular data but also trying to address the performance and interpretability limitations associated with these traditional models, GANBLR was introduced [13]. This innovative model, inspired by Naive Bayes and Logistic Regression, surpasses its existing counterparts in terms of performance and also introduces explicit feature interactions, enhancing its interpretability. GANBLR's outperformance is demonstrated through its ability to generate tabular data with improved precision. Importantly, the research provides valuable insights into feature importance throughout the data generation process, highlighting the factors influencing the model's decisions. This work contributes significantly to the evolving field of data generation, offering a promising solution for applications that require insights into both performance and interpretability within the domain of tabular data.

In order to deal with class imbalance in a cancer intracellular signaling dataset, a new machine learning-based strategy was presented [14]. Employing a generative adversarial network (GAN), this study introduced synthetic data generation for the minority class. The outcomes of this approach underline a notable increase in classification accuracy, indicating the effectiveness of this innovative methodology in handling imbalanced data scenarios. This research contributes to advancing the application of generative adversarial networks in addressing current challenges, providing valuable insights for improving classification performance in datasets characterized by class imbalance, particularly in the domain of intracellular signaling in cancer.

In an effort to predict software efforts in dynamic environments, a recent study explored the potential of improving the Conditional Variational Autoencoder (CVAE) [15]. This approach proved to be promising, demonstrating superior performance in comparison to existing methods, as indicated by accuracy metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and R-squared (R²). Additionally, the model demonstrates efficiency in generating synthetic data that closely mimic real-world data. Leveraging these synthetic data for software effort estimation, as opposed to real data, results in enhanced performance on baseline machine learning models. This research both advances the application of CVAE in regression synthetic project generation and highlights the potential of synthetic data for improving software effort estimation in dynamic environments.

In the ongoing attempt to advance medical image classification, a recent study presented an innovative approach called the Multi-Task Process (MTP) for sacred gap (SH) classification [16]. This new method uses modern deep learning techniques to address the inherent challenges posed by limited medical datasets. Impressively, the MTP achieves a success rate ranging from 90% to 93% in SH classification. The core of this approach is the use of a synthetic dataset created through the innovative implementation of a generative adversarial network (GAN). Remarkably, the features of the matrices are converted into images, facilitated by a two-dimensional embedding algorithm. Subsequently, these transformed images serve as inputs to Convolutional Neural Networks (CNNs). This pioneering strategy highlights the inherent link between synthetic data fed by a GAN and CNN, presenting a promising solution to the common problem of data scarcity in medical image classification tasks.

In the constantly evolving field of machine learning (ML) services, a recent study introduced a new data synthesis technique that utilizes generative deep learning models [17]. The proposed approach encompasses three distinct variants: standard VAEs, β -VAEs, and Introspective VAEs, strategically designed to confront the challenge of hesitant consumers in the ML service market. Addressing the critical issue of data privacy, this technique enables consumers to preview the performance of an ML service without revealing their actual data. The overarching goal is to cultivate a more fluid and accessible market for ML services. Through robust experiments, the study documented the effectiveness of these

variants in meeting challenging data quality requirements, opening the way for a dynamic and flexible market for ML services. This research pushes the boundaries of data synthesis methodologies and will help build a future where transparency and accessibility define the ML service landscape.

In a significant stride toward advancing breast cancer diagnosis and prognosis, a recent study presented a pioneering deep learning-based approach [18]. This innovative methodology focuses on the generation of synthetic data, aiming to facilitate the early diagnosis and prediction of breast cancer. In particular, the approach goes beyond the potential of previous deep learning methods, demonstrating its potential to reshape the landscape of breast cancer patient care. By harnessing the power of synthetic data, this research not only elevates the performance standards of existing techniques but also envisions a future where more accurate and timely diagnostics contribute to improved patient outcomes. This work stands at the forefront of integrating deep learning and generative frameworks, offering promising possibilities for enhancing clinical practices and ultimately advancing the well-being of breast cancer patients.

In the area of healthcare technology, a recent study introduced an innovative approach to the automated detection of shock-inducing rhythms in automated external defibrillators (AEDs) [19]. Focused on addressing the challenges posed by imbalanced datasets of electrocardiograms, this research employed a hybrid generative adversarial network (GAN)-based deep learning methodology. Notably, this innovative approach utilizes GANs to create synthetic data, achieving remarkable performance levels in the detection of shockable rhythms. It is impressive that the proposed solution not only meets but also exceeds the criteria set by the American Heart Association for AEDs, marking a significant advance in automated life-saving technologies. This research promises to revolutionize the healthcare technology scene by demonstrating the potential of advanced deep learning techniques in improving the reliability and effectiveness of crucial medical interventions.

In the dynamic terrain of predicting the demand for electric kickboards, a recent study introduced an innovative approach that leverages generative adversarial networks (GANs) [20]. The research proposed a model designed to generate synthetic time-series data, aiming to enhance mobility demand prediction. This model incorporates modifications to the Wasserstein GAN with a gradient penalty and integrates a regression-based blending ensemble technique, forming a comprehensive strategy for improved prediction performance. Through the results of several experiments, this study shows the increased prediction accuracy as well as the increased performance of the model compared to previous models. This research contributes to the specific area of mobility demand forecasting and also highlights the broader potential of GANs in generating synthetic data to refine predictive analyses in evolving and dynamic contexts.

In an effort to improve the prediction of mortality in patients with acute pancreatitis (AP), a recent study used advanced machine learning and data augmentation techniques [21]. The main goal of this research was to overcome challenges such as the lack of high-quality datasets and class imbalance, which often prevent accurate predictions. To address these issues, the authors strategically combined three datasets, creating a more comprehensive and robust dataset. While using advanced techniques to handle missing values, the study examined various sampling techniques, revealing that Random Forest and deep neural networks (DNNs) emerge as the most effective. This research provides insights into the critical importance of early detection for the management of AP, as well as demonstrates the effectiveness of advanced models for identifying high-risk patients. The development of a novel integrative approach has an important technical impact on the field of medical prognosis, highlighting the potential ingenuity of using both machine learning and data augmentation for predicting mortality in AP patients.

In a preliminary study, a new deep learning tool, the tabular generative adversarial network (TGAN), was presented [22], with the aim of generating synthetic full-scale burst test data specifically tailored to corroded pipelines. This tool integrates the joint probability distribution of five random variables, effectively identifies outliers, and enhances the

authenticity of synthetic data. The study further validated the tool by training machine learning models using a combination of real and synthetic data, revealing that the synthetic data accurately mirror the distribution observed in real-world scenarios. Beyond its immediate applications, this innovative TGAN tool holds promising implications for advancing pipeline integrity management models, offering a robust solution for assessing and maintaining the integrity of corroded pipelines through the generation of realistic synthetic data.

In the continuously evolving computer science field landscape, a recent study stands out as a significant step forward in addressing the complexities of incomplete data and synthetic data generation. This research introduced three novel data imputation methods harnessing the power of generative adversarial networks (GANs) [23]. What differentiates this work is the ability to integrate these computational techniques into a computational method designed for generating synthetic tabular data. Notably, the methods exhibit comparable or superior performance to existing state-of-the-art techniques. Beyond their technical capacities, the study recognizes and addresses the broader legal, ethical, and privacy concerns associated with real-world datasets. This comprehensive approach lays the foundation for the responsible production of privacy-sensitive synthetic data, with far-reaching implications in various fields.

2.2. Data Characteristics

In the field of data science, the generation of diverse datasets plays a vital role in addressing many challenges. From addressing data scarcity to privacy concerns, the generation of synthetic data has emerged as a powerful tool for researchers in various domains. This section explores various aspects of data generation, which is the basis of data analysis.

2.2.1. Numerical Data Generation

Numerical data generation is a crucial aspect of modern data science, allowing researchers to address a variety of challenges. This process involves generating numerical datasets that mimic real-world scenarios, often leveraging advanced computational techniques such as machine learning and statistical modeling. By generating synthetic datasets, researchers can overcome the limitations imposed by data scarcity, privacy concerns, and complex relationships within the data. Moreover, the continuous improvement of numerical data generation methodologies promises to unlock new opportunities for innovation and discovery across a wide range of fields, further establishing their importance in the continuously evolving terrain of data science.

In an effort to broaden the data in the era of cognitive psychology research, an innovative study introduced a deep learning approach specifically designed to generate synthetic datasets [24]. Focusing on the renowned Stroop task, the authors compared their method with traditional random generation techniques. The results reveal the remarkable capability of the proposed deep learning approach, highlighting its ability to preserve the statistical characteristics of the original dataset more efficiently than conventional methods. This work not only marks a significant advance in the generation of synthetic data but also highlights the method's exceptional ability to accurately reflect and retain the statistical detail that is vital for cognitive psychology research.

2.2.2. Categorical Data Generation

The generation of categorical data is a fundamental part of data science, facilitating the research and analysis of various aspects in different fields. This process involves the generation of synthetic datasets that include categorical variables that imitate characteristics or features of the real world. By composing categorical data, researchers can overcome challenges related to data sparsity, privacy concerns, and the complexity of categorical relationships within the dataset. Whether applied to marketing research, social sciences, or customer segmentation, categorical data generation plays a vital role in identifying patterns,

trends, and insights that are vital for making informed decisions. As these methodologies continue to evolve and improve, the opportunities for innovation and discovery in various fields are expanding, underscoring the importance of categorical data generation in shaping the future of data science.

In an attempt to address the persistent challenge of imbalanced data, a proposed solution involves a hybrid generative adversarial network (GAN) approach, specifically utilizing a conditional Wasserstein GAN with a gradient penalty to generate tabular data [25]. To enhance focus on minority classes, an auxiliary classifier loss is incorporated into this approach. The proposed method also introduces the concept of PacGAN in the discriminator architecture to mitigate the mode collapse problem. The results of comprehensive experiments underscore the effectiveness of this approach in significantly improving the performance of recommendation systems when confronted with imbalanced data. This work introduces an innovative strategy for addressing data imbalances, potentially paving the way for more robust and equitable system recommendations.

2.2.3. Temporal Data Generation

The generation of temporal data allows researchers and analysts to investigate patterns and trends over time in various fields. This process involves the generation of datasets that capture temporal data dynamics, such as time series or sequences of events, that reflect real-world processes. By generating synthetic time datasets, professionals can address the challenges associated with data scarcity, temporal dependencies, and the complexity of temporal relationships within the data. These synthetic datasets provide valuable insights into temporal behaviors, facilitating decision-making processes in areas such as finance, healthcare, and climate science. Moreover, refining temporal data generation methodologies promises to unlock new opportunities for understanding and predicting dynamic systems, thus advancing knowledge and innovation in various fields.

In an attempt to address the need for large datasets, a recent study introduced an approach with MTS-TGAN, a novel generative adversarial network (GAN) architecture tailored to generate multivariate time-series data closely resembling real-world datasets [26]. The results showed that MTS-TGAN was effective in capturing the distribution and characteristics of real data and potentially reduced errors in predictive and discriminative scores. This work also represents a valuable contribution to the field of synthetic time-series data generation and also serves as a promising starting point for further exploration and progress in this area.

In the process of improving energy consumption prediction, a new approach involves the integration of time-series energy consumption data with various data augmentation techniques, including generative adversarial networks (GANs), to generate synthetic data [27]. The proposed machine learning model, when combined with these synthetic data, demonstrates a remarkable capability to enhance the accuracy of energy consumption predictions. Notably, the model effectively reduces prediction errors and improves accuracy when integrated with the original data. Furthermore, it successfully addresses the challenge of mode collapse and exhibits faster convergence compared to existing GAN models for synthetic data generation. This paper contributes to the field of energy consumption forecasting, as well as introduces an innovative framework that utilizes synthetic data for increased accuracy and efficiency in forecasting models.

2.3. Privacy Preservation

In the area of the privacy preservation of data, innovative methods are constantly being developed to address the escalating challenges posed by cybersecurity threats and privacy concerns. A notable area of focus is the generation of synthetic data, which serves as a fundamental tool to strengthen cybersecurity defenses and facilitate the release of private research data to the public. Recent developments in this area include the use of generative adversarial networks (GANs) to create realistic data that closely resemble the original datasets, thus enabling the development of powerful machine learning and deep

learning solutions for detecting and countering cyber threats. Furthermore, in areas where user privacy is of ultimate importance, innovative frameworks such as Duo-GAN offer promising solutions by creating synthetic data that preserve the complexity of the original dataset while addressing privacy concerns.

As a contribution to the field of cybersecurity, an innovative method for generating botnet data in a tabular format through two distinct generative adversarial network (GAN) models was introduced [28]. The study systematically explored the performance of these models across different epoch sizes, demonstrating that, after 1000 epochs, they achieved approximately 80% similarity to real data. This method holds promise for data augmentation in cybersecurity applications, particularly for the development of robust machine learning (ML) and deep learning (DL) solutions geared toward detecting and countering botnet attacks.

The escalating challenges in cybersecurity stemming from the digitization of everyday services have led to a methodology for generating realistic zero-day attack data employing generative adversarial networks (GANs) [29]. In this context, this study's methodology started from the generation of zero-day-type, yet realistic, data in a tabular format and concluded with the evaluation of a neural network as a zero-day attack detector, which was trained with and without synthetic data. The results show that the generation of zero-day attack data in a tabular format reaches an equilibrium after about 5000 iterations and produces data that are almost identical to the original data samples. Notably, the neural network model trained with the dataset containing ZDGAN-generated samples outperforms the same model trained solely on the original dataset, showcasing high validation accuracy and minimal validation loss. In this way, this approach presents an effective strategy for fortifying cybersecurity defenses against the landscape of zero-day attacks.

An alternative approach presents two innovative methods designed to create private tabular research data products that are ready for release to the public [30]. Employing a pseudo-posterior mechanism, these methods strategically mitigate the risk of identification disclosure while upholding a robust level of privacy. Notably, the approach ensures global probabilistic differential privacy, considering the distribution of both survey outcomes and weights. This framework guarantees precise estimates of tabular cells and their standard errors, effectively addressing the bias in survey sampling. The performance of these methods surpasses that of the commonly employed Laplace Mechanism, as evidenced by real data applications and simulations. Remarkably, these techniques enable the release of microdata to the public, facilitating further analysis without compromising privacy.

In the dynamic landscape of intrusion detection, the application of deep learning models is promising but faces challenges such as increased false positive rates and detection difficulties, especially when datasets are unbalanced and have small sample sizes. In response to these challenges, a modified iteration of EC-GAN was exploited, leveraging synthetic data generated by WCGAN-GP, to address the intricacies of network flow classification, specifically on the CIC-IDS-2017 dataset [31]. This innovative methodology yields superior results compared to conventional approaches, even when trained on a mere 25% of the dataset. As shown in this study, this method not only contributes to advancing the effectiveness of intrusion detection but also underscores the potential of EC-GAN in improving classification performance in scenarios characterized by imbalanced and limited-sample datasets.

In the field of Intelligent Systems, where user privacy stands paramount, a recent study presented a possible solution focused on the intricate challenges posed by heavily unbalanced data, particularly in domains like Fraud Detection. This research introduced Duo-GAN, a cutting-edge framework leveraging generative adversarial networks (GANs) [32]. The strength of Duo-GAN lies in the generation of synthetic data that reflect the complexity of the original dataset and preserve the user's privacy. Through experiments, the study demonstrated only a 5% difference in F1 scores between classifiers trained with genuine data and those trained with synthetic data and then tested with real data. This study show-

cases the efficacy of Duo-GAN in augmenting unbalanced datasets for improved model training and performance and addresses the privacy concerns inherent in sensitive data.

2.4. Evaluation Metrics

Before we go deeper into evaluating the quality of synthetic data, it is important to understand the metrics and standards used for evaluation. This section explores several evaluation metrics that play a critical role in assessing the fidelity and usefulness of synthetic datasets. These metrics provide researchers with standardized tools for measuring the effectiveness of various data generation techniques in a variety of domains.

2.4.1. Fidelity Metrics

In the field of synthetic data evaluation, significant steps have been taken toward the establishment of standardized metrics for assessing the fidelity and utility of synthetic datasets. A notable effort in the healthcare domain highlights the importance of an inclusive evaluation pipeline that captures the dimensions of similarity, utility, and privacy. This standardized approach aims to provide a robust framework for comparing different methods for generating synthetic data, thereby enabling better performance evaluation and helping to select the most appropriate technique for specific datasets and healthcare applications. In addition, a prior contribution comes in the form of the TabSynDex metric, which serves as a universal measure for evaluating the strength of tabular synthetic data. This innovative metric enables a comprehensive evaluation by measuring the similarity between real and synthetic datasets through a set of component scores. With the introduction of such standardized evaluation approaches and universal metrics, researchers and practitioners are provided with powerful tools for making informed decisions about the effectiveness of different generative models in the field of synthetic data generation.

In the terrain of synthetic data in the health domain, a standardized evaluation approach is described [33]. This study emphasizes the necessity of assessing synthetic health data through a comprehensive pipeline, inspecting three pivotal dimensions: resemblance, utility, and privacy. Recognizing the complex nature of data quality assessment across these dimensions, the study aimed to provide a robust framework for distinguishing the effectiveness of different methods for generating synthetic data. By adopting this standardized pipeline, the research achieved a better performance of existing methods and aided in the discernment of the most suitable synthetic data generation technique tailored to specific datasets and applications within the health domain.

A pioneering contribution to the field of synthetic tabular data evaluation is the TabSynDex metric [34]. This innovative evaluation metric stands as a universal measure designed to assess the robustness of synthetic tabular data. TabSynDex distinguishes itself by enabling a comprehensive evaluation measuring the similarity between real and synthetic data through a set of component scores. This study specifically focused on providing valuable insights into the performance of neural network-based methods. In addition, baseline models for comparative analysis were presented, highlighting the efficacy of TabSynDex over existing metrics in evaluating the quality of synthetic data. With this universal metric, researchers and practitioners gain a powerful tool to make informed decisions about the effectiveness of various generative models in the realm of synthetic data generation.

2.4.2. Utility Metrics

In the arena of utility metrics for evaluating synthetic data, recent studies have focused on comparing and evaluating different techniques on different datasets. While one study emphasized the evaluation of different models using preference and cluster-log metrics, challenging the prevailing view that favors generative adversarial networks (GANs), another study thoroughly evaluated four different techniques for generating anonymous network traffic data. These research efforts provide valuable insights into utility metrics for

synthetic data evaluation, emphasizing the need for careful consideration when choosing data generation techniques across various domains.

A comprehensive study comparing various techniques for tabular synthetic data generation across diverse datasets was conducted [35]. The study's focus was on evaluating the efficacy of different models using two key metrics: propensity and cluster-log. Surprisingly, the results challenged the prevailing notion that generative adversarial networks (GANs) are the preferred models for synthetic data generation. Instead, the Classification And Regression Tree (CART) model consistently outperformed other techniques, consistently delivering superior results. This research broadened the horizons of the field of synthetic data generation, prompting a reconsideration of the dominant role often attributed to GANs in this domain.

A comprehensive evaluation of four distinct techniques employed for the generation of anonymous network traffic data within the domain of traffic classification took place in a recent study [36]. The selected methodologies encompassed oversampling techniques, conditional tabular generative adversarial networks, the variational autoencoder, and the copula generative adversarial network. The effectiveness of each approach was meticulously assessed through the application of both traditional and modern machine learning models, with particular emphasis on the impact of these techniques in comparison to the original dataset. The findings underscored the efficacy of oversampling techniques and the copula generative adversarial network in generating high-quality and diverse anonymous network traffic data. Meanwhile, conditional tabular generative adversarial networks and the variational autoencoder showed promising results, but with some limitations. This study highlights the importance of using suitable techniques to generate anonymous network traffic data for valid applications in machine learning.

2.5. Application Domains

As we dig deeper into different application areas, this section gives insights into the role of synthetic data in addressing crucial challenges and revealing new opportunities. In various sectors, such as healthcare and retail, synthetic data generated through advanced computational techniques are emerging as a promising solution to alleviate data scarcity problems and enhance innovation.

2.5.1. Healthcare

In recent medical data analysis studies, there has been a growing interest in using algorithms to generate synthetic data on various medical datasets in tabular form. One study emphasized ensuring the fidelity of synthetic data using valid statistical metrics, which were then used to train machine learning classifiers, while another study focused on improving prognostic models for cardiovascular disease by evaluating techniques for generating synthetic categorical data, particularly by addressing class imbalance. Furthermore, a study targeting the diagnosis of chronic liver disease demonstrated significant improvements by incorporating Genetic Adversarial Networks (GANs) and synthetic minority oversampling (SMOTE) techniques, enhancing the performance of classifiers and potentially leading to more efficient and accessible diagnoses. Collectively, these findings highlight the promising role of synthetic data in advancing medical research and diagnostic processes.

In the domain of medical data analysis, this study provided a framework by utilizing algorithms for generating synthetic data on eight different tabular medical datasets [37]. With an eye on ensuring the fidelity of the synthetic data, the research used valid statistical metrics to measure the integrity of the data. Subsequently, these synthetically generated datasets became pivotal in training machine learning classifiers. The findings underscore the potential of synthetic data as a valuable solution to counter challenges related to the scarcity of high-quality datasets and concerns about patient privacy within medical data. However, the varied nature of classification performance across different datasets and algorithms for generating synthetic data introduces a complex perspective on the effectiveness of synthetic data in the medical field.

With the aim of improving predictive decision support models for cardiovascular disease, this study was concerned with the evaluation of synthetic categorical data generation techniques [38]. Recognizing the challenge posed by class imbalance, this research systematically tested a number of oversampling techniques on synthetic data to determine their impact on model performance. Notably, the study revealed the superiority of GAN-based models, particularly when coupled with linear classifiers, surpassing other oversampling methods in both predictive accuracy and the identification of crucial risk factors. These findings underscore the promising role of synthetic data in elevating the precision of machine learning models dedicated to predicting cardiovascular diseases and elucidating associated risk factors.

By addressing the challenges of diagnosing chronic liver disease, this study sought to improve the affordability and accuracy of the diagnostic process [39]. Employing five distinct machine learning algorithms—Logistic Regression, K-Nearest Neighbor, Decision Tree, Support Vector Machine, and Artificial Neural Network—the research investigated the impact of generative adversarial networks (GANs) and synthetic minority oversampling (SMOTE) on refining prediction accuracy. The findings demonstrate that the integration of these techniques significantly improves the overall performance of the classifier, showcasing their potential to contribute to more effective and accessible diagnoses in the field of chronic liver disease.

2.5.2. Retail

In retail, a fundamental domain in the modern world, the application of synthetic data represents a vital frontier poised to address critical challenges and unlock new opportunities. Synthetic data generated through advanced computational techniques offer a promising solution to mitigate the data scarcity issues commonly encountered in retail analytics. By simulating realistic yet artificial datasets, synthetic data facilitate the development and testing of innovative strategies without relying on limited or sensitive real-world data. This approach holds immense potential across various aspects of retail operations. As the retail landscape continues to evolve in the digital age, the adoption of synthetic data methodologies emerges as a transformative force driving efficiency, competitiveness, and customer-centricity in the retail sector.

A recent study addressed the challenge of data scarcity in estimating the Water Quality Index (WQI) by leveraging artificial intelligence (AI) to generate synthetic datasets [40]. Over a three-year period, synthetic data were systematically generated with varying synthetic multiplier values (10%, 25%, and 50%). While this prior research provides valuable information, the study acknowledges some limitations. The absence of estimates of changes in hydrological or land use patterns is recognized as a weakness. This research proposes a further study that will improve the accuracy of the WQI by incorporating a more extensive dataset for GP model training. Additionally, the paper goes deeper into the discussion of metrics designed to assess the quality of synthetic data.

To summarize, in this review on synthetic data generation, as shown in Table 1, data generation approaches are dissected, encompassing statistical-based methods and the burgeoning field of machine learning-based techniques. Noteworthy studies are cited to underscore the breadth and depth of research in each approach. Moreover, the review sheds light on different data characteristics, emphasizing numerical, categorical, and temporal data generation, while also addressing the paramount issue of privacy preservation in synthetic data generation. Evaluation metrics play a pivotal role, with the paper elucidating the distinction between fidelity metrics and utility metrics, which are crucial for gauging the quality and applicability of the generated data. Furthermore, the review delineates the diverse application domains of synthetic data, particularly focusing on healthcare and retail sectors, offering insights into practical implementations and real-world implications. This synthesis of reviewed studies serves as a resource, furnishing readers with a comprehensive understanding of synthetic data generation's multifaceted landscape and its profound impact across various domains.

Table 1. Table of reviewed studies.

Categories	Subcategories	Journals Reviewed
Data generation approaches	Statistical-based generation	[8,9]
	Machine learning-based generation	[10–23]
Data characteristics	Numerical data generation	[24]
	Categorical data generation	[25]
	Temporal data generation	[26,27]
Privacy preservation		[28–32]
Evaluation metrics	Fidelity metrics	[34,41]
	Utility metrics	[35,36]
Application domains	Healthcare	[37–39]
	Retail	[40]

3. Discussion

The review presented above highlights the significant advancements and innovative approaches in data generation methodologies, particularly focusing on statistical-based, machine learning-based, and privacy-preserving techniques across various application domains. These methodologies have demonstrated their potential to overcome challenges such as data scarcity, privacy concerns, and class imbalance, thereby opening new avenues for research and applications in diverse fields. Unlike typical reviews that are based solely on an exploration of the literature, this study incorporates a synthesis of existing methodologies with a critical analysis and broadens the discussion by offering new insights and perspectives.

A significant aspect of these studies is the growing importance of synthetic data generation techniques in addressing the limitations of real-world datasets. Statistical-based approaches, such as GenerativeMTD and the divide-and-conquer (DC) strategy, have shown promise in accurately representing complex data relationships while ensuring data balance. These methods offer practical solutions for generating synthetic datasets that closely resemble real-world scenarios, particularly in domains like healthcare and software engineering.

Another aspect worth discussing is the transferability of synthetic data generation techniques from one domain to another. While certain methodologies may demonstrate effectiveness in specific application domains, their adaptability to diverse datasets and contexts remains a topic of interest. For example, although these statistical-based approaches have shown promise in generating synthetic datasets that closely resemble real-world scenarios, further research is warranted to explore their applicability beyond the domains of healthcare and software engineering into sectors like finance, manufacturing, and telecommunications.

A systematic review addresses the challenges and potential of synthetic data generation for tabular health records [33]. The promising outcomes observed in previous studies are related to the synthesis of tabular data, particularly in fields such as healthcare and energy consumption. Despite these advancements, the study highlights the need for a reliable and privacy-preserving solution in handling valuable healthcare data. Focusing on methodologies developed within the past five years, the review elaborates on the role of generative adversarial networks (GANs) in healthcare applications. The evaluation of GAN-based approaches reveals progress, yet it underscores the necessity for further research to pinpoint the most effective model for generating tabular health data.

Similarly, machine learning-based generation techniques, including conditional generative adversarial networks (cGANs), variational autoencoders, and deep learning models, have demonstrated remarkable capabilities in generating synthetic data with high fidelity. Yet, their generalizability across different datasets and domains requires careful consideration. Understanding the nuances of dataset characteristics and domain-specific challenges is crucial for assessing the transferability of these models and methodologies.

These approaches not only improve prediction accuracy in critical healthcare scenarios but also extend their applicability to diverse domains such as agriculture and energy consumption forecasting.

Furthermore, privacy preservation emerges as a critical concern in data generation, particularly in domains where sensitive information is involved. Innovative frameworks like Duo-GAN and hybrid GAN-based methodologies offer promising solutions for generating synthetic data while preserving user privacy. These techniques enable the release of private research data to the public while diminishing the risk of identification disclosure. Examining the robustness of these privacy-preserving techniques across diverse datasets and application domains can provide valuable insights into their generalizability and practical utility.

A thorough review delved into the realm of creating synthetic data for Intrusion Detection Systems (IDSs) using generative adversarial networks (GANs) [42]. The study examined several GAN architectures, notably VanillaGAN, WGAN, and WGAN-GP, alongside specific models such as CTGAN, CopulaGAN, and TableGAN. The evaluation focused on their performance using the NSL-KDD dataset. The findings of the study demonstrated the effectiveness of GANs in generating realistic network data for IDS applications. However, the study underscored a crucial point: the choice of GAN architecture and model significantly influenced the quality of the synthetic data. This review provides valuable insights for researchers and practitioners in the field of cybersecurity, emphasizing the necessary considerations when employing GANs for synthetic data generation in the context of IDS datasets.

Evaluation metrics play a crucial role in assessing the fidelity and utility of synthetic datasets. Standardized metrics such as TabSynDex enable comprehensive evaluations, measuring the similarity between real and synthetic data across various dimensions. By establishing universal measures for evaluating synthetic data quality, researchers and practitioners can make informed decisions about the effectiveness of different generative models.

While existing evaluation metrics provide valuable insights into the quality of synthetic tabular data, it is essential to acknowledge their limitations and strengths. One limitation of existing evaluation metrics is their reliance on specific statistical measures, which may not comprehensively capture the complexity of real-world data distributions. Metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) focus primarily on quantifying the discrepancy between synthetic and real data distributions but may overlook nuances in data relationships and patterns. Additionally, these metrics may not adequately account for the diversity and variability present in real-world datasets, leading to potential inaccuracies in assessing synthetic data quality.

Moreover, existing evaluation metrics may exhibit limited suitability for evaluating different types of synthetic data, particularly across diverse application domains. While certain metrics like the Kolmogorov–Smirnov (KS) test and Wasserstein distance (WD) are commonly used to assess the similarity between distributions, their effectiveness may vary depending on the characteristics of the data and the underlying generation methodologies. For instance, metrics tailored to assess the fidelity of continuous data may not be directly applicable to categorical or mixed-type data, necessitating the development of domain-specific evaluation metrics.

Despite these limitations, existing evaluation metrics offer valuable insights into the quality and performance of synthetic tabular data generation techniques. Metrics such as TabSynDex enable comprehensive evaluations, measuring the similarity between real and synthetic data across various dimensions. By establishing universal measures for evaluating synthetic data quality, researchers and practitioners can make informed decisions about the effectiveness of different generative models.

The application domains discussed in this study underscore the broad impact of synthetic data generation techniques. In healthcare, synthetic data facilitate medical research and diagnostic processes by addressing challenges related to data scarcity and privacy.

Similarly, in retail, synthetic data offer a solution to data scarcity issues, enabling the development and testing of innovative strategies without relying on limited real-world data.

In a comprehensive survey, the complex domain of synthetic data generation was examined, with a specific emphasis on the innovative domain of generative adversarial networks (GANs) [43]. Synthetic data become invaluable in scenarios where original data are scarce or of degraded quality, helping to improve the performance of machine learning models. This study covers various aspects, including GAN architectures, challenges and breakthroughs in their training, algorithms for synthesizing data, diverse applications, and methodologies for evaluating the synthetic data's quality. A special feature of this research is the unique combination of synthetic data generation and GANs, offering a perspective to researchers entering this field. As is shown, this review explores the main techniques for evaluating the quality of synthetic data, with a particular focus on tabular data, offering readers a comprehensive and insightful resource for digging deeper into the complex field of synthetic data creation and GANs.

The reviewed methodologies and models offer diverse approaches to handling complex data relationships and interactions within tabular datasets, each with its strengths and limitations. Understanding how these techniques operate in various scenarios is crucial for assessing their effectiveness and applicability.

Statistical-based approaches, such as GenerativeMTD and the divide-and-conquer (DC) strategy, excel in capturing complex data relationships by leveraging mathematical principles and statistical inference. These methodologies analyze the underlying structures of tabular datasets and generate synthetic data that closely resemble real-world distributions. For example, GenerativeMTD effectively models temporal dependencies in time-series data, making it suitable for scenarios where sequential patterns are prevalent, such as financial forecasting and sensor data analysis. Similarly, the DC strategy partitions the dataset into smaller subsets, allowing for the localized modeling of complex relationships and interactions. However, these methodologies may struggle with high-dimensional datasets or nonlinear relationships, requiring careful parameter tuning and preprocessing steps to achieve satisfactory results.

Machine learning-based generation techniques, including conditional generative adversarial networks (cGANs), variational autoencoders (VAEs), and deep learning models, offer powerful tools for capturing intricate data relationships and generating synthetic data with high fidelity. cGANs, for instance, excel in generating data samples conditioned on specific attributes or features, enabling fine-grained control over the synthesized output. VAEs, on the other hand, learn latent representations of the data distribution, allowing for continuous interpolation between data points and the exploration of the latent space. Deep learning models leverage hierarchical representations to capture spatial and temporal dependencies in tabular datasets. These techniques demonstrate remarkable capabilities in scenarios where data relationships are nonlinear or high-dimensional, such as image recognition, natural language processing, and time-series forecasting. However, they may require large amounts of training data and computational resources, and their interpretability can be limited compared to statistical-based approaches.

Despite their strengths, both statistical-based and machine learning-based techniques may struggle with certain challenges, such as data sparsity, imbalanced class distributions, and outliers. For instance, generating synthetic data that accurately represent rare events or minority classes can be challenging, leading to biased or unrealistic outcomes. Moreover, ensuring the diversity and generalizability of synthetic datasets across different application domains remains an ongoing research area.

In addition, considering the growing concerns around data privacy and fairness, it is imperative to thoroughly explore the ethical implications of synthetic data generation techniques. The development of algorithms based on machine learning techniques must take into account concepts such as data bias and fairness [44]. While the scientific literature proposes numerous techniques to detect and evaluate these problems in real datasets, less attention has been dedicated to methods generating intentionally biased datasets, which

could be used by data scientists to develop and validate unbiased and fair decision-making algorithms [45]. Synthetic data, emerging as a rich source of exposure to variability for algorithms, present unique ethical challenges. For instance, the deliberate modeling of bias in synthetic datasets using probabilistic networks raises questions about fairness and transparency in algorithmic decision-making processes. Moreover, the incorporation of synthetic data into machine learning algorithms reconfigures the conditions of possibility for learning and decision-making, warranting careful consideration of the ethicopolitical implications of synthetic training data. In light of these considerations, it is essential to assess the ethical implications of synthetic data generation techniques and develop potential mitigation strategies to ensure the responsible and equitable use of synthetic data in algorithmic decision-making processes.

In addition to privacy concerns and biases, addressing the ethicopolitical implications of synthetic data is crucial for fostering transparency and accountability in algorithmic decision-making. Synthetic data promise to place algorithms beyond the realm of risk by providing a controlled environment for training and testing, yet their usage raises questions about the societal impact of algorithmic decision-making. As machine learning algorithms become deeply embedded in contemporary society, understanding the role played by synthetic data in shaping algorithmic models and decision-making processes is paramount [46]. Moreover, developing guidelines and best practices for the ethical use of synthetic data can help mitigate potential risks and ensure that algorithmic decision-making processes uphold principles of fairness, transparency, and accountability in diverse societal domains.

Taking into account all of these factors, deploying synthetic data generation techniques in different sectors or industries presents various practical challenges and implementation barriers that warrant careful consideration. Understanding these challenges is essential for effectively leveraging synthetic data generation methods in real-world applications.

One practical challenge is the availability of high-quality training data representative of the target domain. While synthetic data generation techniques offer a means to augment limited or unavailable real-world datasets, ensuring the fidelity and diversity of synthetic data remains a key concern. In many sectors, obtaining labeled training data that accurately reflect the underlying data distributions and capture domain-specific nuances can be challenging. Moreover, maintaining the balance between data diversity and privacy preservation introduces additional complexities, especially in highly regulated industries such as healthcare and finance.

Implementation barriers also arise from the computational and resource-intensive nature of certain synthetic data generation techniques. For instance, machine learning-based approaches, such as generative adversarial networks (GANs) and deep learning models, often require significant computational resources and expertise to train and deploy effectively. In sectors with limited access to computational infrastructure or data science expertise, deploying and maintaining such techniques can be prohibitively challenging. Additionally, ensuring the scalability and efficiency of synthetic data generation pipelines to accommodate large-scale datasets and real-time data generation further complicates implementation efforts.

Furthermore, the interpretability and explainability of synthetic data generation techniques pose challenges in sectors where transparency and accountability are paramount. Understanding how synthetic data are generated and their implications for downstream decision-making processes is crucial for building trust and confidence among end-users and stakeholders. Providing transparent documentation of the synthetic data generation process and validation methodologies is essential for fostering trust and facilitating adoption in diverse sectors and industries.

Despite all of these challenges, the future evolution of synthetic tabular data generation techniques holds great promise, driven by advancements in machine learning, artificial intelligence, and data generation methodologies. Envisioning the trajectory of this field

involves anticipating key trends and developments that are likely to shape the landscape of synthetic data generation in the coming years.

One key direction for future research is the development of more sophisticated generative models capable of capturing complex data relationships and distributions with greater fidelity. Machine learning techniques such as deep generative models, including deep neural networks and variational autoencoders, are poised to play a central role in this evolution. By leveraging hierarchical representations and advanced optimization algorithms, these models offer the potential to generate synthetic data that closely resemble real-world datasets across diverse domains and application scenarios.

Moreover, the integration of domain knowledge and expert insights into the synthetic data generation process is expected to enhance the realism and utility of generated datasets. Hybrid approaches that combine statistical modeling techniques with machine learning algorithms enable the incorporation of domain-specific constraints and priors into the generative process. For instance, incorporating structural equation modeling or Bayesian networks to encode domain knowledge can improve the interpretability and fidelity of synthetic data, making them more suitable for downstream applications such as predictive modeling and decision support systems.

Another important direction for future research is the development of robust evaluation methodologies and benchmarking frameworks for assessing the quality and utility of synthetic tabular data. As synthetic data generation techniques continue to evolve, it becomes increasingly important to establish standardized evaluation metrics and datasets that enable fair comparisons across different methodologies. By promoting transparency and reproducibility in the evaluation process, researchers can facilitate the adoption and validation of novel techniques and accelerate innovation in the field.

Furthermore, addressing ethical and societal implications remains a critical aspect of the future evolution of synthetic data generation techniques. As synthetic data become more prevalent in various sectors and industries, ensuring fairness, transparency, and accountability in their generation and usage is paramount. Interdisciplinary collaborations between researchers from data science, ethics, law, and social sciences can help navigate the complex ethical landscape of synthetic data generation and develop guidelines for responsible data usage.

Overall, we tried to highlight the growing potential of synthetic data generation methodologies in overcoming data-related challenges across various domains. Constant research and development in this field are essential for advancing data science and unlocking new opportunities for innovation and breakthroughs.

4. Conclusions

In summary, the reviewed literature underscores the critical role of synthetic data generation methodologies in addressing data scarcity, privacy concerns, and complex relationships within datasets across diverse domains. From statistical-based approaches to machine learning-based techniques, the advancements highlighted offer promising avenues for generating high-fidelity synthetic data. Moreover, the development of innovative privacy-preserving frameworks emphasizes the importance of maintaining user privacy while generating realistic synthetic datasets. Standardized evaluation metrics further enhance the reliability and comparability of synthetic data quality assessments. Across healthcare, retail, and other application domains, synthetic data emerge as an evolutionary tool, enabling improved medical research, diagnostic processes, and retail analytics. Looking ahead, continuous research and development in synthetic data generation methodologies will be essential for driving innovation and supporting responsible data use in various real-world contexts.

Author Contributions: Conceptualization, S.K.; investigation, A.G.V. and E.P.; writing—original draft preparation, E.P.; writing—review and editing, A.G.V. and E.P.; visualization, S.K.; supervision, A.G.V.; project administration, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data sharing is not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ghasemaghahi, M. Understanding the impact of big data on firm performance: The necessity of conceptually differentiating among big data characteristics. *Int. J. Inf. Manag.* **2021**, *57*, 102055. [\[CrossRef\]](#)
2. Choi, H.R.; Lee, S.W.; Kim, Y.; Lee, J.H.; Koh, H.; Kim, H.C. The necessity and case analysis of bigdata quality control in medical institution. *J. Bigdata* **2017**, *2*, 67–74.
3. Fonseca, J.; Bacao, F. Tabular and latent space synthetic data generation: A literature review. *J. Big Data* **2023**, *10*, 115. [\[CrossRef\]](#)
4. Little, C.; Elliot, M.; Allmendinger, R. Federated learning for generating synthetic data: A scoping review. *Int. J. Popul. Data Sci.* **2023**, *8*, 2158. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Hahn, W.; Schütte, K.; Schultz, K.; Wolkenhauer, O.; Sedlmayr, M.; Schuler, U.; Eichler, M.; Bej, S.; Wolfien, M. Contribution of Synthetic Data Generation towards an Improved Patient Stratification in Palliative Care. *J. Pers. Med.* **2022**, *12*, 1278. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Krawczyk, B. Learning from imbalanced data: Open challenges and future directions. *Prog. Artif. Intell.* **2016**, *5*, 221–232. [\[CrossRef\]](#)
7. Mehmood, A.; Natgunanathan, I.; Xiang, Y.; Hua, G.; Guo, S. Protection of big data privacy. *IEEE Access* **2016**, *4*, 1821–1834. [\[CrossRef\]](#)
8. Sivakumar, J.; Ramamurthy, K.; Radhakrishnan, M.; Won, D. GenerativeMTD: A deep synthetic data generation framework for small datasets. *Knowl.-Based Syst.* **2023**, *280*, 110956. [\[CrossRef\]](#)
9. Kang, H.Y.J.; Batbaatar, E.; Choi, D.W.; Choi, K.S.; Ko, M.; Ryu, K.S. Synthetic Tabular Data Based on Generative Adversarial Networks in Health Care: Generation and Validation Using the Divide-and-Conquer Strategy. *JMIR Med. Inform.* **2023**, *11*, e47859. [\[CrossRef\]](#)
10. Rafiei, A.; Ghiasi Rad, M.; Sikora, A.; Kamaleswaran, R. Improving mixed-integer temporal modeling by generating synthetic data using conditional generative adversarial networks: A case study of fluid overload prediction in the intensive care unit. *Comput. Biol. Med.* **2024**, *168*, 107749. [\[CrossRef\]](#)
11. Mirzaei, A.; Bagheri, H.; Khosravi, I. Enhancing Crop Classification Accuracy through Synthetic SAR-Optical Data Generation Using Deep Learning. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 450. [\[CrossRef\]](#)
12. Neunzig, C.; Möllensiepe, D.; Hartmann, M.; Kuhlentötter, B.; Möller, M.; Schulz, J. Enhanced classification of hydraulic testing of directional control valves with synthetic data generation. *Prod. Eng.* **2023**, *17*, 669–678. [\[CrossRef\]](#)
13. Zhang, Y.; Zaidi, N.; Zhou, J.; Li, G. Interpretable tabular data generation. *Knowl. Inf. Syst.* **2023**, *65*, 2935–2963. [\[CrossRef\]](#)
14. Sánchez-Gutiérrez, M.E.; González-Pérez, P.P. Addressing the class imbalance in tabular datasets from a generative adversarial network approach in supervised machine learning. *J. Algorithms Comput. Technol.* **2023**, *17*, 17483026231215186. [\[CrossRef\]](#)
15. Marco, R.; Ahmad, S.S.S.; Ahmad, S. Improving Conditional Variational Autoencoder with Resampling Strategies for Regression Synthetic Project Generation. *Int. J. Intell. Eng. Syst.* **2023**, *16*, 372.
16. Kilic, F.; Korkmaz, M.; Er, O.; Altin, C. A CNN-Based Novel Approach for Classification of Sacral Hiatus with GAN-Powered Tabular Data Set. *Elektronika Ir Elektrotechnika* **2023**, *29*, 44–53. [\[CrossRef\]](#)
17. Panfilo, D.; Boudewijn, A.; Saccani, S.; Coser, A.; Svara, B.; Chauvenet, C.R.; Mami, C.A.; Medvet, E. A Deep Learning-Based Pipeline for the Generation of Synthetic Tabular Data. *IEEE Access* **2023**, *11*, 63306–63323. [\[CrossRef\]](#)
18. Inan, M.S.K.; Hossain, S.; Uddin, M.N. Data augmentation guided breast cancer diagnosis and prognosis using an integrated deep-generative framework based on breast tumor’s morphological information. *Inform. Med. Unlocked* **2023**, *37*, 101171. [\[CrossRef\]](#)
19. Dahal, K.; Ali, M.H. A Hybrid GAN-Based DL Approach for the Automatic Detection of Shockable Rhythms in AED for Solving Imbalanced Data Problems. *Electronics* **2022**, *12*, 13. [\[CrossRef\]](#)
20. Chatterjee, S.; Byun, Y.C. Generating Time-Series Data Using Generative Adversarial Networks for Mobility Demand Prediction. *Comput. Mater. Contin.* **2023**, *74*, 5507–5525. [\[CrossRef\]](#)
21. Hameed, M.A.B.; Alamgir, Z. Improving mortality prediction in Acute Pancreatitis by machine learning and data augmentation. *Comput. Biol. Med.* **2022**, *150*, 106077. [\[CrossRef\]](#)
22. He, Z.; Zhou, W. Generation of synthetic full-scale burst test data for corroded pipelines using the tabular generative adversarial network. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105308. [\[CrossRef\]](#)
23. Neves, D.T.; Alves, J.; Naik, M.G.; Proença, A.J.; Prasser, F. From missing data imputation to data generation. *J. Comput. Sci.* **2022**, *61*, 101640. [\[CrossRef\]](#)
24. Choi, J.G.; Nah, Y.; Ko, I.; Han, S. Deep Learning Approach to Generate a Synthetic Cognitive Psychology Behavioral Dataset. *IEEE Access* **2021**, *9*, 142489–142505. [\[CrossRef\]](#)
25. Shafqat, W.; Byun, Y.C. A Hybrid GAN-Based Approach to Solve Imbalanced Data Problem in Recommendation Systems. *IEEE Access* **2022**, *10*, 11036–11047. [\[CrossRef\]](#)

26. Yadav, P.; Gaur, M.; Fatima, N.; Sarwar, S. Qualitative and Quantitative Evaluation of Multivariate Time-Series Synthetic Data Generated Using MTS-TGAN: A Novel Approach. *Appl. Sci.* **2023**, *13*, 4136. [\[CrossRef\]](#)
27. Hazra, D.; Shafqat, W.; Byun, Y.C. Generating Synthetic Data to Reduce Prediction Error of Energy Consumption. *Comput. Mater. Contin.* **2022**, *70*, 3151–3167. [\[CrossRef\]](#)
28. Peppes, N.; Alexakis, T.; Demestichas, K.; Adamopoulou, E. A Comparison Study of Generative Adversarial Network Architectures for Malicious Cyber-Attack Data Generation. *Appl. Sci.* **2023**, *13*, 7106. [\[CrossRef\]](#)
29. Peppes, N.; Alexakis, T.; Adamopoulou, E.; Demestichas, K. The Effectiveness of Zero-Day Attacks Data Samples Generated via GANs on Deep Learning Classifiers. *Sensors* **2023**, *23*, 900. [\[CrossRef\]](#)
30. Hu, J.; Savitsky, T.D.; Williams, M.R. Private Tabular Survey Data Products through Synthetic Microdata Generation. *J. Surv. Stat. Methodol.* **2022**, *10*, 720–752. [\[CrossRef\]](#)
31. Zekan, M.; Tomičić, I.; Schatten, M. Low-sample classification in NIDS using the EC-GAN method. *JUCS J. Univers. Comput. Sci.* **2022**, *28*, 1330. [\[CrossRef\]](#)
32. Ferreira, F.; Lourenco, N.; Cabral, B.; Fernandes, J.P. When Two are Better Than One: Synthesizing Heavily Unbalanced Data. *IEEE Access* **2021**, *9*, 150459–150469. [\[CrossRef\]](#)
33. Hernandez, M.; Epelde, G.; Alberdi, A.; Cilla, R.; Rankin, D. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing* **2022**, *493*, 28–45. [\[CrossRef\]](#)
34. Chundawat, V.S.; Tarun, A.K.; Mandal, M.; Lahoti, M.; Narang, P. A Universal Metric for Robust Evaluation of Synthetic Tabular Data. *IEEE Trans. Artif. Intell.* **2024**, *5*, 300–309. [\[CrossRef\]](#)
35. Pathare, A.; Mangrulkar, R.; Suvarna, K.; Parekh, A.; Thakur, G.; Gawade, A. Comparison of tabular synthetic data generation techniques using propensity and cluster log metric. *Int. J. Inf. Manag. Data Insights* **2023**, *3*, 100177. [\[CrossRef\]](#)
36. Cullen, D.; Halladay, J.; Briner, N.; Basnet, R.; Bergen, J.; Doleck, T. Evaluation of Synthetic Data Generation Techniques in the Domain of Anonymous Traffic Classification. *IEEE Access* **2022**, *10*, 129612–129625. [\[CrossRef\]](#)
37. Rodriguez-Almeida, A.J.; Fabelo, H.; Ortega, S.; Deniz, A.; Balea-Fernandez, F.J.; Quevedo, E.; Soguero-Ruiz, C.; Wägner, A.M.; Callico, G.M. Synthetic Patient Data Generation and Evaluation in Disease Prediction Using Small and Imbalanced Datasets. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 2670–2680. [\[CrossRef\]](#)
38. García-Vicente, C.; Chushig-Muzo, D.; Mora-Jiménez, I.; Fabelo, H.; Gram, I.T.; Løchen, M.L.; Conceição, G.; Soguero-Ruiz, C. Evaluation of Synthetic Categorical Data Generation Techniques for Predicting Cardiovascular Diseases and Post-Hoc Interpretability of the Risk Factors. *Appl. Sci.* **2023**, *13*, 4119. [\[CrossRef\]](#)
39. Alauthman, M.; Aldweesh, A.; Al-qerem, A.; Aburub, F.; Al-Smadi, Y.; Abaker, A.M.; Alzubi, O.R.; Alzubi, B. Tabular Data Generation to Improve Classification of Liver Disease Diagnosis. *Appl. Sci.* **2023**, *13*, 2678. [\[CrossRef\]](#)
40. Chia, M.Y.; Koo, C.H.; Huang, Y.F.; Di Chan, W.; Pang, J.Y. Artificial Intelligence Generated Synthetic Datasets as the Remedy for Data Scarcity in Water Quality Index Estimation. *Water Resour. Manag.* **2023**, *37*, 6183–6198. [\[CrossRef\]](#)
41. Hernandez, M.; Epelde, G.; Alberdi, A.; Cilla, R.; Rankin, D. Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions. *Methods Inf. Med.* **2023**, *62*, e19–e38. [\[CrossRef\]](#) [\[PubMed\]](#)
42. Bourou, S.; El Saer, A.; Velivassaki, T.H.; Voulkidis, A.; Zahariadis, T. A Review of Tabular Data Synthesis Using GANs on an IDS Dataset. *Information* **2021**, *12*, 375. [\[CrossRef\]](#)
43. Figueira, A.; Vaz, B. Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics* **2022**, *10*, 2733. [\[CrossRef\]](#)
44. Barbierato, E.; Vedova, M.L.D.; Tessera, D.; Toti, D.; Vanoli, N. A Methodology for Controlling Bias and Fairness in Synthetic Data Generation. *Appl. Sci.* **2022**, *12*, 4619. [\[CrossRef\]](#)
45. Jacobsen, B.N. Machine learning and the politics of synthetic data. *Big Data Soc.* **2023**, *10*, 20539517221145372. [\[CrossRef\]](#)
46. Dahal, K.; Ali, M.H. Imposing Fairness Constraints in Synthetic Data Generation. In Proceedings of the 27th International Conference on Artificial Intelligence and Statistics, Valencia, Spain, 2–4 May 2024; pp. 2269–2277.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.