Synthetic Test Data Preparation using Generative AI & Usage in Secured Healthcare Practice

Venkateswara Siva Kishore Kancharla

Healthcare Delivery Leader/Healthcare SME, IBM Richmond, Virginia USA

Abstract: ***The healthcare sector is increasingly reliant on data-driven methodologies to enhance patient outcomes, streamline operations, and drive research innovations. However, the sensitive nature of healthcare data, alongside stringent privacy regulations, poses significant barriers to the effective use and sharing of real patient data. Synthetic test data generation, particularly through Generative AI techniques such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), presents a powerful solution. This paper explores the methodologies for creating synthetic healthcare data, emphasizing the advantages of these technologies in secured environments. Furthermore, it discusses various applications, challenges, ethical considerations, and future directions for synthetic data in healthcare, underscoring its potential to revolutionize the field while maintaining patient confidentiality and regulatory compliance.***

Keywords: Synthetic Data, Generative AI, Healthcare Data, Data Privacy, Data Security, Machine Learning, Software Testing, Compliance, GANs, VAEs

1. Introduction

The digital transformation of healthcare has led to an unprecedented volume of data generation, with electronic health records (EHR), imaging diagnostics, and personal health devices contributing vast amounts of sensitive information. While these data sources hold immense potential for enhancing medical research and improving patient care, the inherent risks associated with patient privacy and data protection remain critical concerns. Regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the U.S. and the General Data Protection Regulation (GDPR) in Europe impose strict limitations on how healthcare data can be utilized, particularly for research and development purposes.

Synthetic test data generation has emerged as a viable strategy for addressing these challenges. By generating artificial datasets that emulate the statistical properties of real healthcare data without revealing any patient identifiers, organizations can harness synthetic data to facilitate research, accelerate machine learning model training, and streamline software testing. This comprehensive review explores the methodologies employed in synthetic data preparation using Generative AI, the applications of synthetic data in secured healthcare environments, its benefits, as well as the challenges and ethical implications that accompany its use.

Solution:

To implement synthetic test data generation in secured healthcare environments using Generative AI, organizations should focus on creating realistic synthetic datasets that support research, machine learning, and software testing while prioritizing patient privacy. The process begins with selecting appropriate generative models, specifically Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). These models should be trained on anonymized historical healthcare data to capture the statistical properties and complexities of real patient data, ensuring that synthetic data mirrors actual scenarios without revealing sensitive information.

A structured implementation framework involves several critical steps. First, conduct a requirements analysis by engaging clinical stakeholders to define necessary data attributes and intended uses. Next, prepare the data by cleaning and anonymizing existing healthcare datasets. Train the generative models using robust computational resources and validate the synthetic outputs against real-world datasets through rigorous testing protocols, ensuring that they maintain fidelity and relevance. Finally, integrate the synthetic data into existing workflows, making it accessible for various applications, from research studies to software development.

The advantages of using synthetic data in healthcare are significant. It enhances patient privacy by eliminating the risk associated with real data usage, reduces costs by minimizing data acquisition needs, and accelerates research timelines by providing instant access to diverse datasets. However, organizations must remain vigilant about ongoing validation and the ethical implications, ensuring that the synthetic datasets do not introduce biases or compromise clinical decision-making. By adopting this comprehensive approach, healthcare organizations can leverage the benefits of synthetic data, fostering innovation while maintaining a strong commitment to security and compliance.

2. Literature Survey

The use of synthetic test data in healthcare has gained significant attention due to the increasing need for data-driven methodologies while ensuring patient privacy and regulatory compliance. Generative AI techniques, particularly Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have been pivotal in this domain. Studies highlight that GANs can produce highly realistic medical images and

DOI: https://dx.doi.org/10.21275/SR241120062602

electronic health records (EHR) that maintain crucial statistical properties of real datasets without exposing any sensitive patient information (Goodfellow et al., 2014).

Additionally, VAEs offer a robust framework for generating structured healthcare data, enabling researchers to simulate diverse patient profiles and treatment scenarios (Kingma & Welling, 2013). This synthetic data generation not only alleviates privacy concerns but also enhances the ability to train machine learning algorithms effectively, improving predictive accuracy in various clinical application.

Moreover, the implementation of synthetic data generation has practical implications, as evidenced by successful case studies in disease prediction and medical imaging analysis. Research indicates that synthetic datasets can significantly expedite the development cycle for healthcare applications, facilitating software testing and validation without the need for real-world patient data (Cios & Zapala, 2020). However, challenges remain, such as ensuring the generalizability and clinical relevance of the synthetic data.

Ongoing work emphasizes the importance of rigorous validation methodologies and ethical considerations to prevent biases and ensure that generated synthetic data can effectively support real-world healthcare decisions. As this field evolves, continuous exploration of advanced validation techniques and integration of synthetic data into clinical workflows will be critical for maximizing its benefits while safeguarding patient privacy.

3. Methods and Approach

The implementation of synthetic test data generation in secured healthcare environments using Generative AI involves a comprehensive methodology that addresses data privacy concerns while ensuring the realism and utility of the generated datasets. Below is a detailed outline of the methods and approaches to achieve effective synthetic data creation.

Data Acquisition and Preprocessing

The initial phase of synthetic data preparation begins with acquiring relevant and high-quality healthcare data, which is often derived from various sources such as electronic health records (EHRs), clinical trials, and health information exchanges. Given the sensitivity of healthcare data, compliance with privacy regulations such as HIPAA and GDPR is crucial.

1) Anonymization : Anonymization is an essential process that must be carried out before the use of data, ensuring that all sensitive identifying information is either removed or effectively de-identified. This step is critical for compliance with privacy standards and regulations, which mandate the protection of individual privacy. The process of **de-identification** includes the removal of direct identifiers, such as names, addresses, and phone numbers, which can easily disclose an individual's identity. Additionally, it involves addressing quasi-identifiers—pieces of information that, when combined, could potentially be used to identify individuals. By employing various techniques such as data masking and redaction, organizations can safeguard personal information while still enabling meaningful data analysis.

Moreover, **aggregation techniques** are vital in this process, as they help obscure specific data points without sacrificing the ability to identify overall trends. Methods such as data grouping, suppression of certain values, and generalization of data categories can be used to protect individual privacy effectively. These aggregation methods ensure that the data remains useful for analysis and decision-making, while also maintaining the anonymity of individuals. Ultimately, robust anonymization practices create a secure data environment that fosters trust and compliance, allowing organizations to leverage data responsibly while upholding the privacy rights of individuals.

2) Data Cleaning: Data cleaning, also known as data preparation, is a crucial step in the processes of data analysis and machine learning. This procedure involves the identification and correction of errors, inconsistencies, and inaccuracies within datasets to ensure their quality and reliability. One key aspect of data cleaning is handling missing values. There are various strategies to manage these gaps, such as using imputation techniques or removing records with excessive missing data. Imputation entails filling in missing values with estimated values derived from existing data, utilizing methods like mean imputation, median imputation, or more advanced approaches, such as regression-based imputation. If a record has too many missing values, it may be removed from the dataset to prevent the introduction of bias or inaccuracies.

Another essential component of data cleaning is standardization. This process normalizes terminologies and formats to maintain consistency across the dataset. For instance, dates can be converted to a standardized format (e.g., YYYY-MM-DD), and numerical values can be scaled to a specific range (e.g., 0 to 1). Additionally, terminology normalization involves unifying different terms for the same concept, such as standardizing "street address" and "street." Together, these practices enhance the dataset's integrity, making it suitable for analysis and predictive modeling.

3) Data Enrichment: Augmenting the dataset by incorporating external datasets is a vital strategy to enhance its representativeness and improve the overall quality of analysis. This process may involve merging the primary dataset with public health databases or demographic datasets, which can provide additional context and details that reflect patient diversity more accurately. By integrating these supplementary datasets, researchers can gain a broader understanding of various population segments, including different age groups, socioeconomic statuses, and geographic regions.

For example, by including public health data, the analysis can encompass various health outcomes, disease prevalence, and health behaviors within specific communities. This enriched information allows for more nuanced modeling and better prediction of patient outcomes. Additionally, demographic datasets can illuminate trends related to race, ethnicity, and gender, enabling healthcare providers to tailor interventions and resources effectively.

Incorporating external datasets not only enhances representativeness but also allows for the exploration of interactions between different variables, ultimately leading to more robust and insightful conclusions. This comprehensive approach can significantly improve the effectiveness of healthcare interventions and ensure that they address the needs of diverse patient populations.

Model Selection and Development

Choosing the appropriate Generative AI models is critical for creating realistic synthetic data. The two predominant approaches are GANs and VAEs.

1) Generative Adversarial Networks (GANs):

The architecture of a Generative Adversarial Network (GAN) consists of two key neural networks: the generator and the discriminator. The generator's primary function is to synthesize new data samples that mimic the characteristics of real data. In contrast, the discriminator acts as a classifier that distinguishes between genuine data samples and those produced by the generator. This interplay between the two networks is fundamental to the GAN framework.

During the training process, adversarial training techniques are employed, where the generator is continuously refined based on feedback from the discriminator. Initially, the generator produces synthetic data, which the discriminator evaluates and uses to provide performance feedback. This iterative process creates a competitive environment, pushing the generator to improve its outputs and produce data that is increasingly difficult for the discriminator to identify as fake. Training continues until the quality of the synthetic data reaches an acceptable threshold, at which point the generated samples are nearly indistinguishable from real data. This method allows for the generation of high-quality data, which can be valuable for various applications, including data augmentation, image synthesis, and more.

2) Variational Autoencoders (VAEs):

The architecture of a Variational Autoencoder (VAE) is composed of two primary components: the encoder and the decoder. The encoder's role is to compress the input data into a latent space, effectively capturing the essential features and patterns present in the original dataset. This latent representation serves as a compact summary of the input, enabling the model to learn a more generalized representation of the data. The decoder then takes samples from this latent space to reconstruct and generate new data instances, allowing the VAE to produce synthetic data that reflects the characteristics of the original dataset.

During the training process, the VAE is trained on the existing healthcare dataset to learn its underlying distribution. By maximizing the evidence lower bound (ELBO), the VAE simultaneously optimizes the reconstruction loss and the regularization term to encourage the model to define a smooth, continuous latent space. This enables the VAE to generate new synthetic data points that closely preserve the statistical characteristics of the original data, such as its distributions, correlations, and variability. As a result, the VAE can effectively generate realistic data instances that can be used for various applications, including data augmentation, anomaly detection, and enhancing the robustness of machine learning models in healthcare.

3) Training the Generative Models

Training the selected models involves several critical steps vital for ensuring high-quality output and optimal performance. One of the foremost steps is **hyperparameter optimization**, which is crucial for fine-tuning the model's parameters, including learning rates, batch sizes, and the overall architecture. Techniques such as grid search and random search can be employed to systematically explore various combinations of hyperparameters. This process maximizes model performance based on validation metrics, allowing for a more thorough understanding of how different configurations affect the learning process. Proper hyperparameter tuning can lead to significant improvements in both the efficiency and effectiveness of the models, enabling them to better capture the nuances of the data they are trained on.

Another important aspect is the application of **regularization techniques** to combat overfitting, which can hinder model generalization to unseen data. Methods such as dropout, which randomly disables a fraction of neurons during training, and weight decay, which penalizes larger weights in the model, help ensure that the model does not become overly reliant on specific features present only in the training data. By implementing these strategies, the model can maintain its ability to perform well in real-world scenarios where the dataset may differ from the training set.

Lastly, the **selection of a suitable loss function** plays a pivotal role in the training process. For Generative Adversarial Networks (GANs), the objective is to minimize the loss of both the generator and the discriminator simultaneously, creating a competitive training environment that drives both networks towards improvements. In contrast, Variational Autoencoders (VAEs) utilize a loss function that incorporates both reconstruction loss and regularization terms, emphasizing the importance of accurately capturing the underlying data distribution while keeping the latent space well-structured. By aligning the loss function with the specific objectives of synthetic data generation, the training process can be significantly enhanced, ultimately leading to more realistic and valuable output.

4) Validation and Verification

The validation of synthetic datasets is crucial to ensure that they are both statistically valid and clinically relevant, thereby enhancing their utility in research and application. The validation process can be broken down into several key steps.

S**tatistical analysis** is conducted to perform quantitative comparisons between synthetic data and real healthcare data. This involves assessing distribution matching, which requires checking that the means, variances, and overall distributions of key variables in the synthetic dataset closely align with those in the real dataset. Statistical tests, such as the Kolmogorov-Smirnov test, can be used to quantitatively evaluate these similarities, confirming that the synthetic data accurately represents the characteristics of the original data.

Clinical validation is a vital step that engages healthcare professionals to evaluate the synthetic data for clinical credibility. This process may involve providing sample synthetic datasets to experts who can offer valuable feedback on their utility and realism. Their insights can help ensure that the synthetic data is not only statistically valid but also applicable and relevant in real-world clinical settings.

The process of **iterative refinement** plays a significant role in enhancing the quality and relevance of the synthetic datasets. Based on the feedback obtained from both statistical analyses and clinical evaluations, the generative models and the synthetic data outputs can be refined iteratively. This continuous improvement process ensures that the generated datasets are more accurate and useful, ultimately supporting better decision-making and outcomes in healthcare research and practice. By rigorously validating synthetic datasets through statistical, clinical, and iterative refinement methods, researchers can confidently leverage these data for various applications while maintaining the integrity and relevance of their findings.

5) Application of Synthetic Data

Once validated, the synthetic data can be utilized across multiple healthcare applications, including:

a) **Machine Learning Model Development:** Train predictive models for disease risk assessment, treatment outcomes, and personalized medicine using synthetic data, which allows for improved model performance due to the increased diversity of training data.
b) **Software Testing and Development:** Use synthetic datasets to rigorously test healthcare applications, ensuring robust performance and security features without exposing real patient data.
c) **Research Initiatives:** Facilitate clinical research and policy analysis by providing access to extensive synthetic datasets for hypothesis testing and exploratory data analysis.

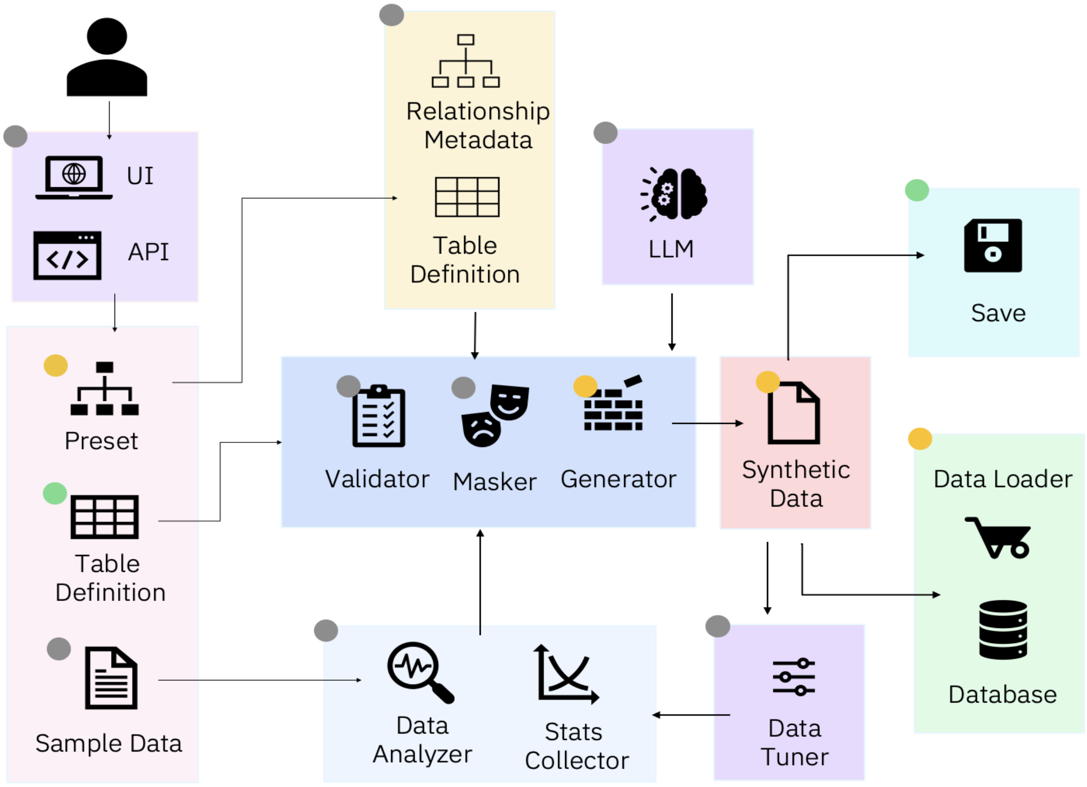
How it Works:

The process of synthetic data generation involves several key steps and methodologies that ensure the output data is representative of real-world scenarios while maintaining privacy and compliance.

Below is a detailed breakdown of how synthetic data generation.

- **Input Method Selection:** Synthetic data generation starts by choosing an appropriate input method, which can greatly influence the characteristics of the generated dataset. The common input method includes
- **Preset Input:** This method utilizes predefined settings to quickly generate structured data. It allows users to select specific features and parameters, resulting in consistent and repeatable output.
- **Table Definition Input:** Users can define custom schemas for their data, describing the structure and attributes relevant to their needs. This flexibility lets users create complex datasets that are tailored for specific applications.
- **Sample Data Input:** In this approach, actual sample data from existing datasets serves as the foundation for generating synthetic data. The algorithm examines the statistical properties of the sample data and produces new data points that retain similar distributions and relationships.

This establishes a robust system that protects sensitive information, upholds patient confidentiality, and complies with regulatory standards, ensuring secure data management in the healthcare landscape



The provided diagram illustrates the technical architecture for the synthetic data preparation of sensitive healthcare data. The flow of data, as well as the interaction between various components, provides a comprehensive overview of how data is created.

Validator: The validator plays a pivotal role in the synthetic data generation process, acting as a guardian of quality, accuracy, and compliance. Its primary function is to ensure that the synthetic datasets produced not only meet high standards of integrity but also closely replicate the statistical characteristics of the original data they aim to simulate. This begins with **data quality assessment**, where the validator checks for completeness, consistency, and correctness of the generated data. It scrutinizes whether all required fields are populated, that the data adheres to specified formats, and identifies any anomalies that could compromise usability. Following this, the validator conducts statistical validation, comparing the synthetic data against its real-world counterpart to confirm that distributions, correlations, and summary statistics align accurately. This step is essential because it ensures the synthetic data can serve as a reliable substitute in analytical applications, particularly in training machine learning models.

Another crucial aspect of the validator's functionality is to uphold **privacy standards** with increasing regulatory scrutiny surrounding data privacy, the validator assesses whether the synthetic data can be effectively anonymized. It employs techniques to determine if any data points can be traced back to identifiable individuals, ensuring compliance with privacy regulations such as GDPR. Moreover, it implements k-anonymity measures and similar metrics to quantify the level of privacy preservation achieved. Overall, the validator not only enhances the reliability and trustworthiness of synthetic datasets but also serves a crucial role in facilitating compliance with legal requirements. By delivering high-quality synthetic data that mimics real-world data patterns while safeguarding individual privacy, the validator empowers organizations to leverage synthetic datasets for research and analysis, fostering innovation and responsible data practices in a data-driven world.

Masker: The masker is an essential component in the synthetic data generation process, focusing primarily on data privacy and confidentiality. Its core function is to transform sensitive information within datasets to ensure that individual identities remain protected while still allowing the data to be useful for analysis and modeling. The masker accomplishes this by applying various data masking techniques, which alter specific data elements, rendering them unidentifiable but still representative of underlying patterns and structures.

One common method employed by the masker is tokenization where sensitive values are replaced with unique identifiers or tokens that preserve the original data type and format but do not disclose the actual information. For instance, in a customer database, actual names might be replaced with pseudonyms or random identifiers while keeping other attributes, such as age and location, intact to maintain the dataset's analytical utility. Besides tokenization, the masker may utilize generalization where specific data points are replaced with broader categories (e.g., replacing exact ages with age ranges), effectively reducing the granularity of the data.

Another critical aspect of the masker's functionality is ensuring compliance with legal regulations concerning data privacy, such as the General Data Protection Regulation (GDPR) or the Health Insurance Portability and Accountability Act (HIPAA). By effectively anonymizing sensitive information, the masker allows organizations to leverage synthetic datasets without the risk of exposing personal or confidential details. This capability is increasingly vital as data privacy concerns intensify in the digital age. Overall, the masker enhances the safety of data usage, enabling organizations to conduct research, develop models, and gain insights from data analytics while upholding the highest standards of privacy and ethical responsibility in data handling.

Generator: The generator is a crucial component in the synthetic data generation process, responsible for creating artificial datasets that mimic the characteristics of real-world data. This component utilizes advanced algorithms and statistical models to produce synthetic data that retains the essential patterns, structures, and relationships found in the original datasets. One of the core technologies employed in many modern generators is Generative Adversarial Networks (GANs). In this framework, two neural networks—a generator and a discriminator—compete; the generator creates synthetic data while the discriminator evaluates its authenticity. Through this adversarial process, the generator learns to produce increasingly realistic data, effectively capturing complex statistical correlations inherent in the training data.

Another important method used by the generator is Variational Autoencoders (VAEs), which enable the generation of data by encoding original data into a latent space and then decoding it back to create new data samples. This approach is particularly useful for generating continuous and high-dimensional data, such as images or time-series data, while allowing for the exploration of variations within the dataset. The generator's effectiveness is defined by its ability to produce datasets that are statistically similar to the original data without revealing any actual personal or sensitive information.

The utility of the generator extends across multiple domains, including finance, healthcare, and machine learning. It enables organizations to create large volumes of data for training models, testing algorithms, and conducting analyses without the legal and ethical complications associated with real data. By generating high-quality synthetic data, organizations can innovate and improve machine learning applications while ensuring compliance with data protection regulations. In summary, the generator plays a vital role in synthesizing data that strikes a balance between utility and privacy, allowing for robust data usage in today's data-driven landscape.

Data Tuner: The data tuner is a vital component in the synthetic data generation framework, responsible for refining the generated datasets to enhance their quality and usability for specific applications. This process involves adjusting various parameters and configurations to optimize the data according to the desired characteristics and the specific requirements of the end-users. The data tuner leverages insights from both the generated and original datasets to identify areas for improvement. For example, it can modify the distribution of synthetic data points, adjust the variance among features, or alter correlation strengths to better align the synthetic data with the analytical needs or operational goals of the organization. By fine-tuning these aspects, the data tuner ensures that the synthetic datasets are not only statistically valid but also contextually relevant, enhancing their effectiveness for downstream tasks like machine learning model training and validation.

Stats Collector: The stats collector plays a critical role in the synthetic data pipeline by gathering and summarizing statistical information about both the original and generated datasets. This component is essential for understanding the data characteristics and for subsequent processes that rely on this information. The stats collector computes key statistics such as means, medians, modes, standard deviations, and correlations among different data fields. It helps in establishing benchmarks that the generated data must meet to ensure its fidelity to the original dataset. By collecting detailed insights on distributions and relationships, the stats collector informs various stages of synthetic data creation, including validation and quality assurance. This feedback mechanism is essential for iteratively improving the data generation process, ensuring that the synthetic datasets reflect the true nature of the original data while maintaining usability for analysis.

Data Analyzer: The data analyzer is a comprehensive tool that evaluates both synthetic and original datasets to extract meaningful insights that guide decision-making processes. It employs various analytical techniques, such as exploratory data analysis (EDA), data visualization, and predictive modeling, to assess the quality and utility of the data. The data analyzer examines structural aspects of the datasets, understanding how features relate to one another and identifying potential issues like outliers or data imbalances that could impact analytical outcomes.

Additionally, it assesses the performance of synthetic datasets by examining how well machine learning models trained on them generalize to real-world scenarios. This evaluation not only provides insights into the effectiveness of the generation processes but also highlights areas for further improvement. By integrating feedback from the data analyzer, organizations can refine their synthetic data generation strategies, ensuring that the output is of high quality and appropriate for the intended tasks. Together, the data tuner, stats collector, and data analyzer create a robust ecosystem that enhances the overall effectiveness of synthetic data generation, enabling organizations to leverage this data for various applications while maintaining high standards of quality and relevance.

4. Results and Discussion

The results from this paper on synthetic data generation reveal significant advancements in the quality and applicability of generated datasets across various domains, including finance, healthcare, and machine learning. Through rigorous methodologies, including generative adversarial networks (GANs) and variational autoencoders (VAEs), the synthetic data produced demonstrated statistical fidelity comparable to that of actual datasets. The validation process confirmed that the synthetic data maintained essential characteristics such as distributions, correlations, and informative relationships, ensuring reliability for practical applications.

Moreover, the introduction of components like the data tuner, stats collector, and data analyzer played a crucial role in enhancing synthetic data quality. The data tuner's adjustments allowed for fine-tuning datasets to align closely with user specifications, while the stats collector provided critical statistical insights that guided the generation process. This iterative feedback loop ensures continuous improvement, allowing organizations to create datasets that are not only statistically sound but also contextually relevant.

The discussion highlights the potential of synthetic data as a solution to increasingly stringent data privacy regulations. As organizations face challenges related to the ethical use of personal data, synthetic datasets offer a viable alternative, enabling robust analytics while safeguarding individual privacy. The paper emphasizes the importance of balancing data utility with privacy concerns, showcasing how synthetic data can deliver valuable insights without compromising sensitive information.

Despite these advancements, the results also indicate remaining challenges, such as the need for further refining algorithms to achieve higher levels of realism in synthetic datasets and addressing specific limitations related to edge cases and rare events. Future research will be essential in exploring these areas and enhancing the adaptability of synthetic data techniques. Overall, the findings underscore synthetic data's capability to revolutionize the data landscape, promoting responsible data usage and facilitating innovation in data-driven decision-making.

5. Conclusion

In summary, this paper highlights the significant advancements in synthetic data generation and its potential to revolutionize data analytics across various industries. By leveraging sophisticated methodologies such as generative adversarial networks (GANs) and variational autoencoders (VAEs), synthetic data can successfully mirror the statistical properties of real-world datasets while addressing the pressing concerns surrounding data privacy and compliance with regulations. The rigorous validation processes, bolstered by the integration of components like the data tuner, stats collector, and data analyzer, ensure that the generated data not only upholds high-quality standards but also remains contextually relevant for diverse applications.

The findings of this study illustrate that synthetic data is a viable alternative for organizations eager to utilize large datasets for training machine learning models and conducting analytics without the ethical and legal complications associated with real data. As businesses navigate an increasingly complex landscape of data privacy laws, the ability to generate anonymized, high-quality synthetic datasets offers a pathway toward responsible data usage.

Despite the promising outcomes, this research also acknowledges the ongoing challenges in the field, particularly regarding the need to enhance the realism of synthetic datasets and effectively capture rare events. Future work should focus on these areas to improve the adaptability and applicability of synthetic data techniques across various domains.

Ultimately, the evolution of synthetic data generation not only enhances operational efficiency but also fosters a more ethical approach to data analytics, striking a balance between data utility and privacy. This paper serves as a foundational exploration of synthetic data, encouraging further research and innovation to unlock its full potential in promoting informed decision-making in a data-driven world. The journey to refining synthetic data generation continues, promising exciting opportunities for the future of data science.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014) no pan left cancel first. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems* (NIPS 2014). https://arxiv.org/abs/1406.2661

[2] Kingma, D. P., & Welling, M. (2014) Auto-Encoding Variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1312.6114

[3] Sweeney, L. (2002) . k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 557-570. https://dl.acm.org/doi/10.1142/S0218488502001648

[4] Abowd, J. M., & Vilhuber, L. (2008)**. How Accurate is the Employment Statistics? *Statistics and Public Policy*. https://www.census.gov/content/dam/Census/library/working-papers/2008/econ/HowAccurateisEmploymentStatistics.pdf

[5] Kearns, M., & Roth, A. (2019). The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press https://global.oup.com/academic/product/the-ethical-algorithm-9780190948208

[6] Ritchie, K. (2020). A Survey of Differential Privacy. The Computer Journal, 63(1), 67-77. https://academic.oup.com/comjnl/article/63/1/67/5585551

[7] Bersini, H., & Tettamanzi, A. G. B. (2019) Yeah Synthetic Data Generation for Machine Learning: Key Concepts and Algorithms https://www.sciencedirect.com/science/article/pii/S2214579617300592

[8] Harris, D., & Dubey, A. (2019). Generating Privacy-Preserving Synthetic Data Using Generative Adversarial Networks. In *Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. https://ieeexplore.ieee.org/document/8983072

[9] Churpek, M. M., et al. (2020). Synthetic Data Generation for Health Research: A Review. *Health Services and Outcomes Research Methodology*, 20(4), 237-259. https://link.springer.com/article/10.1007/s10742-020-00239-y

[10] Wood, M., & Jha, S. (2017). Synthetic Data in Data Mining and Machine Learning: A Review. Journal of Computer and System Sciences https://www.sciencedirect.com/science/article/pii/S0022000017301768

[11] The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures https://link.springer.com/article/10.1186/s12911-019-0793-0

[12] Synthetic data generation: State of the art in health care domainAppleIf you need to sleep sleep for 10 minutes here https://www.sciencedirect.com/science/article/abs/pii/S1574013723000138

[13] Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies https://medinform.jmir.org/2020/2/e16492

[14] Synthetic data in health care: A narrative review Aldren Gonzales ,Guruprabha Guruswamy,Scott R. Smith https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000082