JOURNAL OF THE ROYAL LAUREATES ACADEMY

www.rlaindia.org

SYNTHETIC DATA GENERATION AND ITS IMPACT ON ARTIFICIAL INTELLIGENCE

Harer Savita Laxman

Research Scholar, Department of Computer Science Engineering, Vikrant University, Gwalior, M.P.

Dr. Shashank Swami

Research Supervisor, Department of Computer Science Engineering, Vikrant University, Gwalior, M.P.

ABSTRACT

The promise of synthetic data to solve pressing, real-world problems in a variety of fields is becoming more apparent. To address data shortages, privacy problems, and algorithmic biases, which are prevalent in machine learning applications, they provide creative solutions. This paper provides an extensive overview of methods for creating synthetic data, such as statistical models, huge language models, diffusion models, Generative Adversarial Networks, Variational Autoencoders, and more. Improving model creation without exposing sensitive information is the main focus as the article explores the various uses of synthetic data in several industries, including healthcare, finance, cybersecurity, and intelligent systems. The paper goes on to list the main problems with synthetic data, such as bias in data distribution, less realistic results, security concerns, and problems with complying with regulations. Synthetic data, when produced and validated correctly, may greatly enhance AI innovation, facilitate privacy-preserving analytics, and allow for secure data exchange, according to the results.

Keywords: Synthetic Data, Artificial Intelligence, Privacy, Networks, Data

I. INTRODUCTION

Data of high quality must be readily available for several applications to function properly in many different domains, including healthcare, banking, cybersecurity, and most notably artificial intelligence (AI). The assumption that insights from robust datasets may change decision-making, propel innovation, and expose hitherto unseen patterns and relationships necessitates large datasets. For instance, healthcare organisations are able to use analysis of big patient records and genetic data to enable breakthroughs in personalised medicine, illness prediction, and therapy evaluation. Financial risk assessment, fraud detection, and algorithmic trading methods can all benefit from analysing massive databases of financial transactions. In addition, cybersecurity research into massive volumes of network traffic allows for the detection of anomalies and the development of preventative measures against cyberattacks. Urban planners may also use data from a wide variety of urban sensors to inform initiatives in areas such as infrastructure, traffic management, and environmental sustainability. Big data is crucial in many areas now and beyond for fostering progress, creativity, and informed decision-making, as is evident from their dependence.

A key asset and major limitation in the dynamic field of artificial intelligence, data is always changing. There is a growing problem for businesses in many sectors when it comes to collecting enough diversified, high-quality data to train AI models. According to studies, data limits are the main reason why a lot of AI initiatives never make it to production. Companies often say that data privacy is the biggest problem with AI. An in-depth analysis published in the Journal of Biomedical Informatics shows how hard it is to build reliable diagnostic tools due to a lack of data, especially for uncommon diseases for which there isn't much patient information. At the same time, data usage is subject to tight regulations, particularly those pertaining to personally identifiable information (PII), according to expanding regulatory frameworks like GDPR, HIPAA, and CCPA. For ethical AI research and development, synthetic data has emerged as a game-changer in the face of data hunger and privacy concerns.

As an alternative to more conventional ways of data collecting and processing, synthetic data provides a viable alternative by simulating the statistical features and correlations present in real-world data without actually including any observations. Recent research indicates that the synthetic data generation industry is poised for significant expansion in the next years.

This expansion is expected to be driven by regulated industries, including healthcare, financial services, and public sector applications. Financial firms are embracing synthetic data to circumvent legal limitations without sacrificing model accuracy. Investment banks are at the forefront of this trend, according to new research published in the International Journal of AI Economics. One way for organisations to tackle data scarcity, bias, and slow innovation without sacrificing ethics or compliance with regulations is to generate data that keeps the value of actual data but removes privacy issues.

II. THE GENERATION OF SYNTHETIC DATA

In most contexts, "real data" is information gathered from the actual world, and this might include everything from text and photographs to video and audio. However, problems like data imbalance and data discrimination emerge in real-world applications as a result of its incompleteness and intrinsic constraints. Researchers are using a variety of techniques to generate synthetic data from existing actual data as it is challenging to meet the demand using just real data. From more conventional statistical models to cutting-edge deep learning-based approaches, these methods cover the gamut.

Statistical Models

In order to create synthetic data with comparable attributes, statistical models for creating datasets generally try to mimic the distribution, correlations, and features of actual data. The key approaches are outlined below:

- **Distribution-based methods**

The data's distribution properties are what this approach is trying to replicate. Probability density functions (PDFs) are utilised to characterise the distribution of continuous data, whilst probability mass functions (PMFs) are employed for discrete data. One way to add more information to a synthesis is to take a random sample from the distribution of the current data.

- **Interpolation and Extrapolation**

Creating additional data points between or beyond current data points is what interpolation and extrapolation are all about. Data pertaining to time series, locations, etc., benefit greatly

from this. The linear connection between two neighbouring known points is used to derive the value of a new point in linear interpolation, a popular interpolation method.

- **Monte Carlo Simulation**

To mimic the unpredictability of real-world systems, Monte Carlo simulation makes use of random sampling. Data synthesis makes use of this technique to create new samples by selecting at random from preexisting distributions. It is widely used in modelling physics, engineering, and finance.

- **Model-based Sampling**

This approach forecasts future data points by applying statistical models to previously collected data. By fitting a linear regression model to the available data, for instance, and then randomly picking the model's parameters, one might generate additional data points. Data showing linear connections benefit greatly from this method.

- **Kernel Density Estimation**

To estimate the probability density function, kernel density estimation requires enclosing each data point with a kernel (usually a Gaussian kernel) and then computing the contribution at each point. It is possible to use the estimated probability density function to conduct random sampling when creating fresh samples. When trying to capture the multimodality and complexity of data distributions, this is helpful.

Deep learning based generation

Due to the significant progress in deep learning in the past several years, researchers are focusing more and more on using deep learning techniques to create synthetic datasets. Deep learning techniques, in contrast to more conventional statistical methods, may learn more complex feature representations of data automatically, without requiring human designers to create feature extractors. Since they are inherently nonlinear, they can easily adjust to the complicated nonlinear relationships in data, allowing for a better capture of the data's important features. The next section provides an overview of many common methods for generating deep learning-based data.

- **Variational Auto-Encoder(VAE)**

VAE is a form of probabilistic generative model, which leverages the encoder-decoder architecture. Parameters of the variational distribution are used by the encoder to map the input data to the underlying latent space, and features from the latent space are projected back into the input space by the decoder. The VAE is able to collect the distribution of latent space characteristics and use that information to create several samples that all adhere to the same distribution.The produced data is more reflective of the complexity in real datasets since the VAE's intrinsic randomness provides a degree of variability.

- **Generative Adversarial Networks(GAN)**

In 2014, Good fellow et al. initially suggested GAN, which includes both a generator and a discriminator. To create samples that are similar to the training set, the generator randomly selects values from the latent space. At the same time, the discriminator determines if the data sample is real or artificial. In a never-ending cycle of adversarial learning, the two components switch off updating parameters until the generator can provide high-quality samples that mislead the discriminator.

- **Diffusion models**

By using a methodical approach to mimic complex temporal correlations in real data, the diffusion model provides a reliable framework for building synthetic datasets. This approach incorporates these models into the data generation process, drawing on the principles of diffusion processes, which involve the gradual spread of information through a network. This allows for the recreation of complex relationships and patterns found in real data. The key is to catch the ever-changing information as it changes over time. This method, which is based on diffusion models, is able to accurately reflect the complicated temporal dynamics found in real-world datasets while simultaneously accurately reproducing statistical features. An advanced tool for modeling actual data situations, this algorithmic technique generates synthetic datasets that are crucial for training and evaluating machine learning models in many domains. Image classification, object recognition, and semantic segmentation are just a few of the vision tasks that have benefited greatly from the synthetic data produced by diffusion models.

- **Large language models**

Recent years have seen the rise of large language models (LLMs) as a game-changing method for creating artificial datasets. One example is the ability of large language models (LLMs) to generate synthetic datasets; models like GPT-3.5 showcase this capability with their remarkable in-context learning skills and vast pre-trained linguistic knowledge. To overcome the problem of data scarcity in some fields, this feature allows models to be trained on smaller domains. A collection of OpenAI-created Natural Language Processing (NLP) models called Generative Pre-trained Transformers (GPTs) use the Transformer architecture to detect input sequences with long-distance dependencies. Predicting the next word in a given context is one use of GPT, which is unsupervised trained on large-scale Internet text corpora. Using this pre-training technique, the model can learn the ins and outs of language statistics and contextual connections, which allows it to do a wide range of NLP tasks in a pre-training-fine-tuning fashion.

III. SHORTCOMINGS OF SYNTHETIC DATA

Data Distribution Bias

Synthetic datasets differ noticeably from their real-life equivalents in a number of important statistical respects, including feature distribution, class distribution, and others. This bias imparts a predisposition for models to create false prognostications within practical applications, consequently weakening their faithfulness to properly represent real-world occurrences.

Incomplete Data

The existence of incomplete or missing data in synthetic datasets, purportedly caused by mistakes, defects, or a failure to capture the many changes that occur in real datasets as they are generated. The model's robustness and practical applicability might be affected by its inability to correctly predict or handle situations with partial data due to this lack of complete information.

Inaccurate Data

The emergence of mistakes, disturbances, or discrepancies in artificial datasets, markedly differing from the truth of actual datasets. This inconsistency may stem from computational flaws, noise interference, or additional variables that lead to errors. As a result, the model could assimilate flawed patterns, leading to skewed predictions and compromising its overall efficacy and dependability when faced with authentic data.

Insufficient Noise Level

Synthetic datasets may exhibit an excessive sterility, devoid of the diverse noise and complexities found in authentic data. In authentic circumstances, data inevitably comprises varied interferences, mistakes, and uncertainties.

Over-Smoothing

During synthetic data generation, some models might excessively diminish or simplify the data, leading to a reduced representation lacking the intricate details and variety found in genuine datasets. This tendency may engender difficulties for the model in integrating intricate variances within authentic data.

Neglecting Temporal and Dynamic Aspects

Certain approaches for synthetic data production may insufficiently capture temporal and dynamic subtleties, characteristics that are essentially important inside real datasets. The significant inability to accurately replicate these temporal complexities may result in the ineffectiveness of models in practical scenarios.

Inconsistency

The scarcity of variability in synthetic datasets compared to the intricate diversity found in real datasets. The former generally represents variances derived from various origins, historical periods, and environmental contexts, elements that synthetic datasets typically do not capture. This deficiency may provide obstacles for models in adjusting to the complex fluctuations arising from various origins, time periods, and environmental factors, thereby leading to a reduction in generalisation efficacy across different datasets.

IV. RISKS IN SYNTHETIC DATA APPLICATION

General Model Performance

The extensive utilisation of synthetic data might limit the generalisation capabilities of AI models. For example, datasets such as PersonX, produced by a game data engine, could differ considerably from actual data. In natural language processing, depending on fine-tuning with extensive language models could confine subsequent tasks to the capabilities and biases of the chosen model. In the medical field, a surplus of fictitious instances during model training might diminish trust in diagnostic outcomes among healthcare practitioners and patients.

Ethical and Social Implications

The utilisation of synthetic data may engender ethical and societal issues. The generation of fictitious characters or scenarios using synthetic data prompts enquiries over AI accountability in producing imaginative material, which may result in inaccuracy, misconceptions, or the spread of inaccurate information with harmful social consequences.

Security and Adversarial Risks

The emergence of synthetic data introduces problems related to security and adversarial assaults. The nefarious application of synthetic data might destabilise AI models in the face of adversarial assaults, since these models would fail to grasp the intricacies and variability of the actual world during their training. This vulnerability heightens the probability of deceit or manipulation, endangering the integrity and safety of the system.

Legal Compliance Challenges

In some fields, the utilisation of synthetic data may provide difficulties in adhering to legal regulations. For example, using synthetic data for risk evaluation in the banking industry may face regulatory obstacles. Regulatory bodies frequently want models that are transparent and comprehensible, and the synthetic data creation process may struggle to fulfil these criteria.

V. CONCLUSION

The development of synthetic data has emerged as a crucial technique for the progression of artificial intelligence, while safeguarding data privacy and adhering to legal standards.

Contemporary methodologies like GANs, VAEs, diffusion models, and extensive language models have markedly enhanced the quality and applicability of synthetic datasets in several fields. Despite ongoing issues about data accuracy, bias, security, and ethical implications, persistent research is improving the dependability and efficacy of synthetic data production techniques. As technological progress continues, synthetic data is anticipated to play a crucial role in facilitating safe, privacy-conscious, and reliable AI systems, fostering innovation in sectors such as healthcare, finance, cybersecurity, education, and other data-heavy domains.

REFERENCES: -

Alabdulwahab, S., Kim, Y.-T., & Son, Y. (2024). Privacy-preserving synthetic data generation method for IoT-sensor network IDS using CTGAN. *Sensors, 24(3),* 7389. https://doi.org/10.3390/s24227389

Berg, A. M., Mol, S. T., Kismihok, G., & Sclater, N. (2016). The role of a reference synthetic data generator within the field of learning analytics. *Journal of Learning Analytics, 3*(1), 107–128. https://doi.org/10.18608/jla.2016.31.7

Chim, J., Ive, J., &Liakata, M. (2025). Evaluating synthetic data generation from user-generated text. *Computational Linguistics, 51*(1), 191–233. https://doi.org/10.1162/coli_a_00540

Giomi, M., Boenisch, F., Wehmeyer, C., & Tasnádi, B. (2023). A unified framework for quantifying privacy risk in synthetic data. *Proceedings on Privacy Enhancing Technologies, 2023*(2), 312–328. https://doi.org/10.56553/popets-2023-0055

Govindankutty, S. (2025). Synthetic data generation and privacy-preserving AI. *International Research Journal on Advanced Engineering and Management (IRJAEM), 3*(5), 1681–1688. https://doi.org/10.47392/IRJAEM.2025.0270

Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., & Rankin, D. (2022). Synthetic data generation for tabular health records: A systematic review. *Neurocomputing, 493(3),* 28–45.

Ji, E., Ohn, J. H., Jo, H., Park, M.-J., Kim, H., Shin, C., & Ahn, S. (2025). Evaluating the utility of data integration with synthetic data and statistical matching. *Scientific Reports, 15*(1), 1–25. https://doi.org/10.1038/s41598-025-01514-0

Kanagarla, K. (2024). The role of synthetic data in ensuring data privacy and enabling secure analytics. *European Journal of Engineering and Technology Research, 10*(11), 75–79. https://doi.org/10.5281/zenodo.14012320

Kanagarla, K. (2024). The role of synthetic data in ensuring data privacy and enabling secure analytics. *European Journal of Engineering and Technology Research, 10*(11), 75–79. https://doi.org/10.5281/zenodo.14012320

Liu, Q., Shakya, R., Jovanovic, J., Khalil, M., & de la Hoz-Ruiz, J. (2025). Ensuring privacy through synthetic data generation in education. *British Journal of Educational Technology, 56*(3), 1053–1073. https://doi.org/10.1111/bjet.13576

Miletic, M., &Sariyar, M. (2024). Challenges of using synthetic data generation methods for tabular microdata. *Applied Sciences, 14*(14), 5975. https://doi.org/10.3390/app14145975

Velocia, C., & Kumar, K. (2025). Synthetic data in AI healthcare research: A review of use cases, benefits, and risks. *AI-Driven Synthetic Health Data, 1*(2), 312–317.