Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

Role of Synthetic Data for Improved AI Accuracy

Ketha Dhana Veera Chaitanya[1], Manas Kumar Yogi[2]

[1]B. Tech III Year CSE-AI &ML Department, Pragati Engineering College (A), Surampalem, A.P., India

[2]Assistant Professor Department, Pragati Engineering College (A), Surampalem, A.P., India

Email: [1]kethadvchaitanya@gmail.com, [2]manas.yogi@gmail.com

Abstract

Artificial Intelligence (AI) has emerged as a transformative technology across various industries, enabling advanced applications such as image recognition, natural language processing, and autonomous systems. A critical determinant of AI model performance is the quality and quantity of training data used during the model's development. However, acquiring and labeling large datasets for training can be resource-intensive, time-consuming, and privacy-sensitive. Synthetic data has emerged as a promising solution to address these challenges and enhance AI accuracy. This study explores the role of synthetic data in improving AI accuracy. Synthetic data refers to artificially generated data that mimics the distribution and characteristics of real-world data. By leveraging techniques from computer graphics, data augmentation, and generative modeling, researchers and practitioners can create diverse and representative synthetic datasets that supplement or replace traditional training data.

Keywords: AI, Synthetic Data, Privacy, Security, Bias, Fairness

1. Introduction

Synthetic data is the artificially generated data which resembles to the characteristics and properties of real world data around us. It best reciprocates the real-world data. Different algorithms and techniques can be used to find out the similarities and patterns in the real data. The primary benefit of using synthetic data is that actual data is more expensive and more

DOI: https://doi.org/10.36548/jaicn.2023.3.008
Received: 17.09.2023, received in revised form: 04.10.2023, accepted: 17.10.2023, published: 31.10.2023

difficult to gather. In these situations, the synthetic data plays a major role by helping the researchers and developers with quick synthetically generated customized data [1].

(i) Significance of Synthetic Data

The synthetic data is primarily used in training the neural networks. The generation of this data is based on algorithms. The synthetic data could also help us by protecting confidentiality and privacy of individuals [2].

(ii) Pros of using Synthetic Data

- **Enhanced Data Quality**

The real world data is very complicated and even harder to gather and sometimes it might have plenty of errors and may bias the data quality. In such cases this synthetic data will come over all these things.

- **Maintains Data Privacy**

When the real world data is being used which contains the data regarding the individuals which may cause any data misuses could be best prevented by the synthetic data as it consists a mimic of the real world data.

- **Eliminates the Redundant Data**

The synthetic data eliminates the duplicates and maintains data in unique way labeling.

(iii) Different Types of Synthetic Data

Based on the different purposes the synthetic data can be classified as of the follows

- Synthetic text data: It refers to the text which is generated artificially. Here the training models are used for the text generation [3].
- Media data: sometimes the synthetic data could also be an image, audio also where the artificial data can be made by resemblance of real world pictures in the replacement of original data.
- Tabular data: The name itself tells that this data is generated by making tabulated values upon creating rows and columns.

(iv) Methods used in Generating the Synthetic Data

The techniques which are used for creating the synthetic data and their multifaceted advantages and disadvantages using each technique in the context of improving AI accuracy are as follows [3]:

1. By known Statistical Distribution – In this type, the synthetic data is generated by drawing numbers from the distributions by observing the original distributions. If the data scientists has a clear understanding of the data, then data set with random samples could be created.

2. Agent-model technique – In this method the synthetic data will be created by observed behaviour and thus it generates the random data by the same models data. This type of fitting the data is generally called as distribution of data.

3. Using Neural Networks and Deep Learning [3]– Basically there are mainly 3 techniques namely

    - VAE – Variational Auto Encoder
    - GAN – Generative Adversarial Network
    - Diffusion Models

2. Role of Synthetic Data in the Field of AI

In developing an AI model, a large number of data sets are required for testing and training the machine learning and AI models. In such situations the real-world data might not be sufficient and not even feasible to train as well as test the model. In such cases synthetic data helps us [4].

(i) Significance of AI in Real-World Applications

What does Artificial Intelligence mean?

Artificial Intelligence is nothing but making the machines (computer systems) to think and act like humans. The AI is further divided as broad AI and narrow AI.

There a surplus number of applications of AI in real world now a days. Let's have a glance a few of the applications as follows:

(a) Healthcare System: The role of AI in healthcare is very revolutionary. Its applications are found in the wide range of examples such as:

- Cancer Diagnosis – The Path AI helps doctors to make the accurate predictions and find them in earlier stage.
- Prediction of Chronic diseases - By the help of enhanced AI microscopes now a days doctors are able to know about the harmful bacteria in patients rather than by manual testing which may require more time for test results.

(b) Agricultural Sector:[12]

- Due to shortage for the food resources and increased population. New agricultural practices need to be implemented in order to survive.
- Monitoring the crop and soil testing- Soil nutrients are very much needed for the good harvesting of a crop. As the manual testing and monitoring will take much time, instead drones like UAV can be used for capturing the images and to train the models to interpret the crops status in various aspects.
- Even while spraying the pesticides and watering the crops the smart sprayers which are embedded with the computer vision with AI ensures that each plant receives an equal distribution.

(c) AI Transportation – In order to mitigate the accidents and for the improved safety the self-driven cars would be the best examples. By implementing AI-based traffic signals, they could best predict the traffic areas and make the roadways clear in just few seconds.

• This also saves time and contributes to a pollution-free environment.

Most of the machine learning models utilises huge number of data sets for better accurate results. This is where the synthetic data will be used mostly to uplift the training size for ML models. The fig.1 shows the role of synthetic data.

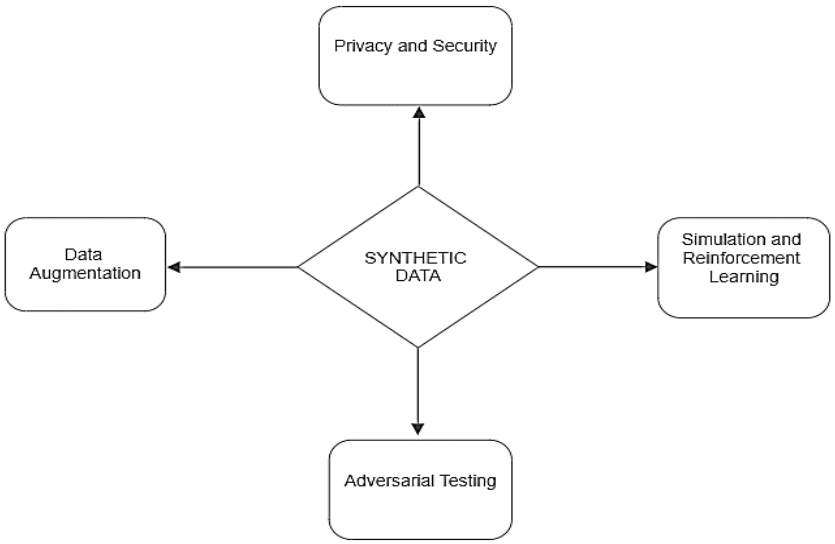

Figure 1. Role of Synthetic Data

Table 1. The Relationship Between the Synthetic Data and the AI

Sl. No.	Criteria	Synthetic Data	AI
1	Purpose	Synthetic data is primarily used for training and testing AI models and algorithms. It helps in model development when real data is limited, sensitive, or expensive to obtain.	AI is used to create systems that can perform a wide range of tasks, including data analysis, decision-making, image recognition, and natural language processing, among others.
2	Training Data	Synthetic data is often used as a supplement or substitute for real data during the training of AI models. It helps in building models, especially in cases where obtaining real data is challenging.	AI models, such as machine learning and deep learning models, require large and diverse datasets for training, which can include both real and synthetic data.
3	Data Generation	Synthetic data is generated using various algorithms and techniques, such as data generation models, to create data that resembles real data in terms of distribution and characteristics.	AI systems are designed to learn from data, and they can use both real and synthetic data for training. AI models are built using mathematical and computational approaches.

4	Data Augmentation	Synthetic data can be used to augment real data, providing additional examples to improve the performance and robustness of AI models.	AI models often benefit from data augmentation, which involves creating variations of real data to enhance model generalization. Synthetic data can be used in this process.
5	Data Testing	Synthetic data can be used for testing and validating AI models, helping ensure they work as expected in various scenarios.	AI systems undergo rigorous testing, including validation on real data, to assess their performance and reliability. Synthetic data can be a part of the testing process.

3. Different Applications Where the Synthetic Data is Used

Table 2. Different Aspects of Synthetic Data Applications

Reference Number	Aspect	Merits	Limitations
[6]	Data Privacy	Protects sensitive information. Complies with data regulations (e.g., GDPR).	Synthetic data might not perfectly represent real data. Risk of re-identification in some cases.
[7]	Data Availability	Overcomes data scarcity or access issues. Useful for testing and development.	Synthetic data might not capture all nuances of real data. Limited application to rare events.
[8]	Cost Efficiency	Reduces the cost of data acquisition. Saves time and resources.	Initial development cost can be high.
[8],[9]	Model Development	Helps develop and test models without real data. Accelerates innovation.	Models trained on synthetic data may not generalize well. Requires rigorous evaluation.

[9]	Scaling Data	Scales data for large-scale applications. Useful for load testing.	Synthetic data may not fully represent real-world data distributionLimited to specific use cases.
[8][9]	Bias Mitigation	Can be used to generate fair, unbiased data. Helps address fairness concerns.	Requires careful design to avoid introducing new biases. May not eliminates all biases.

Particularly synthetic data has the ability to generate the labeled data very fast and makes the research and development activities at a faster rate. The synthetic data in the various purposes of artificial intelligence indulges the following [5]

• By using synthetic data to create machine learning models, businesses can create models for which they lack the necessary data because the models are of low quality and are private.

- This synthetic data can helps in mitigation of bias as that of the real world data sets, however it can't represent the full range of information as that of the real world data.
- Synthetic data came over with three major problems namely features, aggregate, privacy of data.
- Synthetic data creates labeled data instances,which could be ready to use in training. This mitigates the time efforts for data labeling.

4. Application of Synthetic Data in Healthcare

The use of synthetic data which is generated artificially is making tremendous changes in many sectors due to its potentiality. Especially in the smart health care development synthetic data plays as a key role for research data [6].

The main reason why data about the healthcare is fewer because the data has the confidential data of real persons. As there are acts such as Health Insurance Portability and

Accountability Act -1996 which was made in United States. The 3 benefits of synthetic data can be described as follows:

- It considers the protection of privacy of individuals
- It provides faster data accessing for health care researching team
- It fills the shortage of real data sets for testing

The role of data is critical in the growth of health care, public health, and innovation, and a large amount of data is required to meet this demand. However, databases including health information are not easily accessible since they include the most personal information about people. Moreover these data sets can't be easily shared because of the regulations like Portability and Accountability Act [7].

These are some of the use cases of synthetic data in health care:

- Simulate and research prediction
- Hypothesis,algorithm testing
- Public health research

In the last two years, synthetic data has received increased attention, particularly in the case of the COVID-19 pandemic-related clinical research, which concluded that synthetic data could be used as a proxy in place of real data sets, and the analytics of both the real and synthetic data produced positive results [8].

4.1 The Research Objectives Covered in the Study

The main objective of this manuscript is to know the utilization of the synthetic data in Artificial Intelligence. This helps us to know the major terms related to the generation of the synthetic data and addresses the major issues. As well as the various applications of synthetic data have been observed.

Domain Specific Aspects includes:

- Techniques used for the generation of Synthetic Data: The manuscript explains how the synthetic data is been created by various techniques. The advanced methods includes GANs, VAEs.

- Data security and privacy: Privacy is the foremost important of anything .Particularly when talking about the banking sector the synthetic data can be helpful for reducing the risks by generating the statistical copies without disclosing sensitive information.

- The same can be applicable to the health sector where the privacy details such as the patient's diseases and medical history can be hidden by the synthetic data.

4.2 Design Specific Limitations of Synthetic Data Addressed and Overcoming the synthetic Data Limitations

- The manuscript point outs the critical issues when generating the synthetic data in a realistic manner so that it won't be biased in any corner.

- The synthetic data could be used as outliers and thus helps the models to recognize and handle them in real world scenarios.

4.3 Synthetic Data in Medical Health Industry

One of the applications of the synthetic data is that there is a mobile app named M-Sense. The app is created to help migraine patients to control and reduce the risk of migraine. This app collects data from the patients and this data along with the synthetic data can be used for future studies.

5. Synthetic Data Vs. AI Technology in Smart Health Care

Artificial intelligence has its applications in various domains of medicine and health care, from the basic things such as recognising patterns and diagnosing the treatment to speed recovery. But the main problem arises when the sufficient data is not provided for the development of AI model [9].

In such cases the synthetic data which is fake but almost familiar with real world data. So, by this the importance of data for AI in health systems is understood.

Sometimes the synthetic data also helps by filling the missing data in the AI datasets so as to make them a real thing and ensuring the patients good health. The different aspects of synthetic data in healthcare are depicted in the fig.2 below.

Figure 2. Different Aspects of Synthetic Data in Healthcare

Nowadays there is a drastic development in Artificial intelligence (AI) and Machine Learning (ML). As the AI and ML models requires large data sets for building a best model. But the main barrier is that the training model is costly as well as time taking. Here comes the best use case of synthetic data which offers a solution that is the data that resembles the real datasets. This artificial data produced by the computer could be utilised in place of manually compiling the data. The final result is that the AI&ML models might be trained faster with less cost and the algorithms can produce new outcomes and help [10]

1. Faster discovery of newly spreading diseases,
2. Detailed study of the diseases and primary preventions to be taken.
3. Enhance the process of drug or vaccine discovery at faster rate.

In this it can be justified that the synthetic data can be used for acquiring improved results when AI technologies are used in smart healthcare. How the synthetic data could be useful in the healthcare system and its applications

6. Challenges and Problems Faced when using Synthetic Data with AI in Healthcare Field.

Besides the benefits of the synthetic data there are also some of the flaws in it particularly with AI in healthcare. The major primary problems faced when using synthetic data with AI in healthcare is that the data quality problem.

Even though synthetic data is produced using algorithms and models that mimic real-world data, there is always a possibility of erroneous data and the possibility of potential errors, which are particularly significant in the healthcare industry.

The healthcare field needs accurate and reliable data which is going to impact directly on the patient's health. So, the data created should be very clear and error free. Synthetic data is generated based on existing real-world data, which means that sometimes it may not accurately show appropriate results [11]. The challenges existing in the application of synthetic data in AI is shown in figure. 3 below.

The problem that are existing in the application of synthetic data are as follows:

1. In machine learning there might be a problem of data leakage might occur when the information of the test data is used by the training process thus leading to a problem.
2. The synthetic data can also be expensive, particularly in the case when used for large scale projects and research activities which will be a drawback for small scale researchers.
3. The neural network systems which generate synthetic data are prone to cyber-attacks. So, there is a possibility that hackers can hack the synthetic data generators and miss use them.

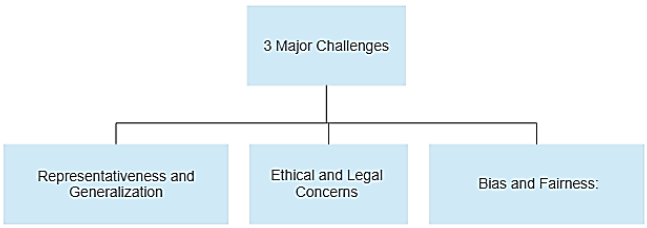

Figure 3. Main Challenges for Application of Synthetic Data in AI

7. Comparison of The AI Accuracy Achieved by the Synthetic Data and the Real-World Data in the Health Care

Table 3. Comparison of AI Accuracy using Synthetic Data vs. Real Data

Sl. No	Application	Dataset Used	Source	F1 Score (Real Data)	F1 Score (Synthetic Data)
1	Image Recognition	Weather Image Recognition This dataset contains labeled 6862 images of different types of weather.	https://www.kaggle.com/datasets/jehanbhathena/weather-dataset	93.245	94.677
2	Natural Language Processing	IMDB Dataset of 50K Movie Reviews	https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews	96.221	96.330
3	Healthcare Diagnosis	Lung Cancer patient detection dataset	https://data.world/cancerdatahp/lung-cancer-data/workspace/file?filename=cancer+patient+data+sets.xlsx	97.211	97.219
4	Anomaly Detection	Smartphone Dataset for anomaly detection in Crowds. This dataset was	https://archive.ics.uci.edu/dataset/613/smartphone+dataset+for+anomaly+detection+in+crowds	95.887	95.914
6	Fraud Detection	Credit Card Transactions Fraud Detection Dataset. Simulated Credit Card Transactions generated using Sparkov.	https://www.kaggle.com/code/badmangamingsv/credit-card-fraud-detection	97.656	98.011

8. Results of Efficiency Comparison between Real time data and Synthetic Data

The figures 4 – 7 illustrates the efficiency and data utility aspects comparison between real time data and synthetic data on diverse applications.

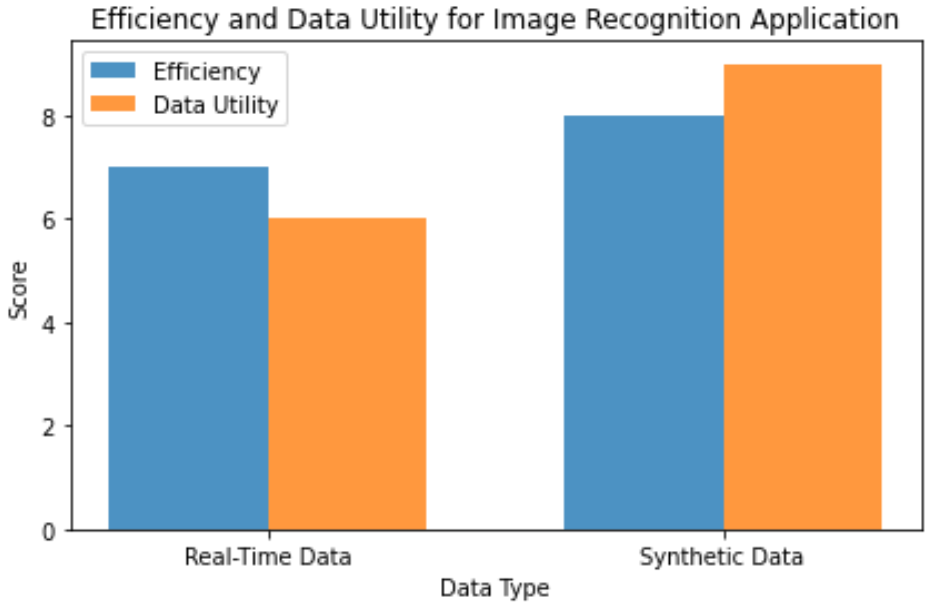

Figure 4. Efficiency and Data Utility Aspects Comparison Between Real Time Data and Synthetic Data for Image Recognition Application [12]

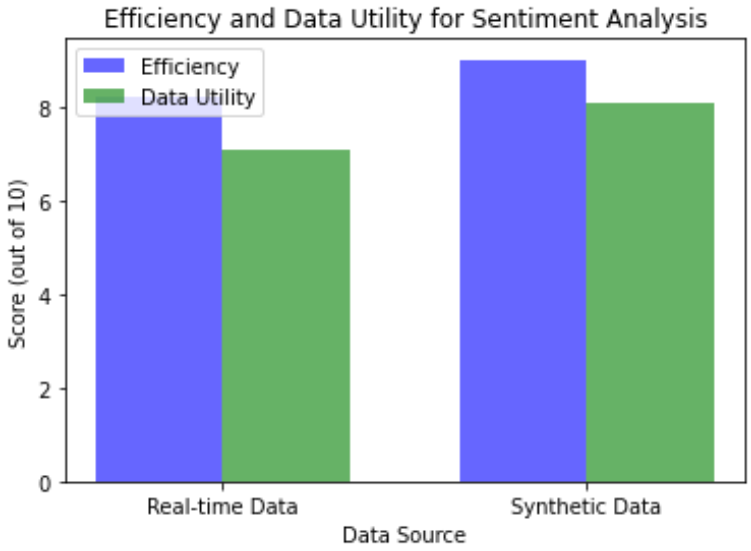

Figure 5. Efficiency and Data Utility Aspects Comparison Between Real Time Data and Synthetic Data for Sentiment Analysis Application [13]

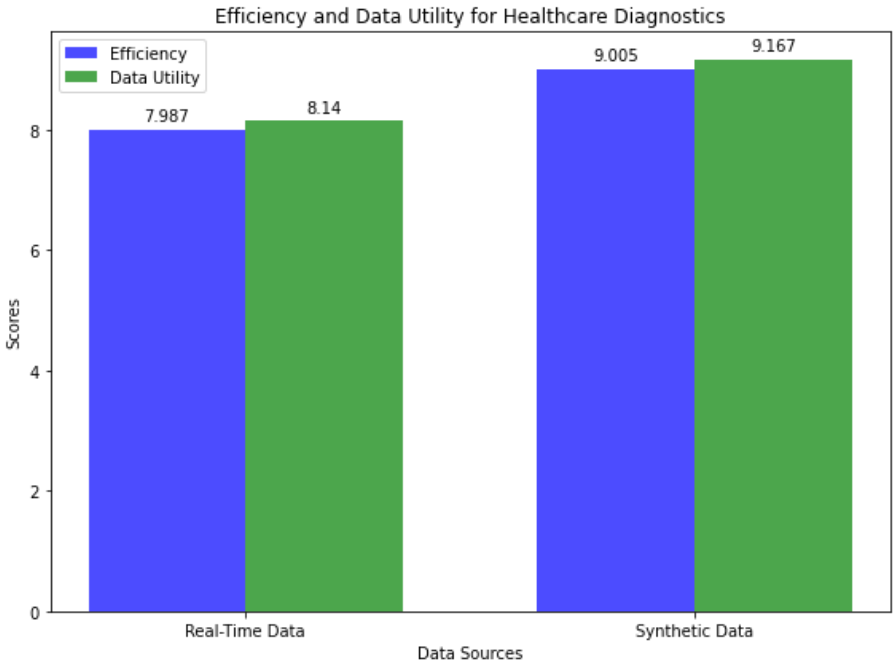

Figure 6. Efficiency and Data Utility Aspects Comparison Between Real Time Data and Synthetic Data for Healthcare Diagnostics Application [14]

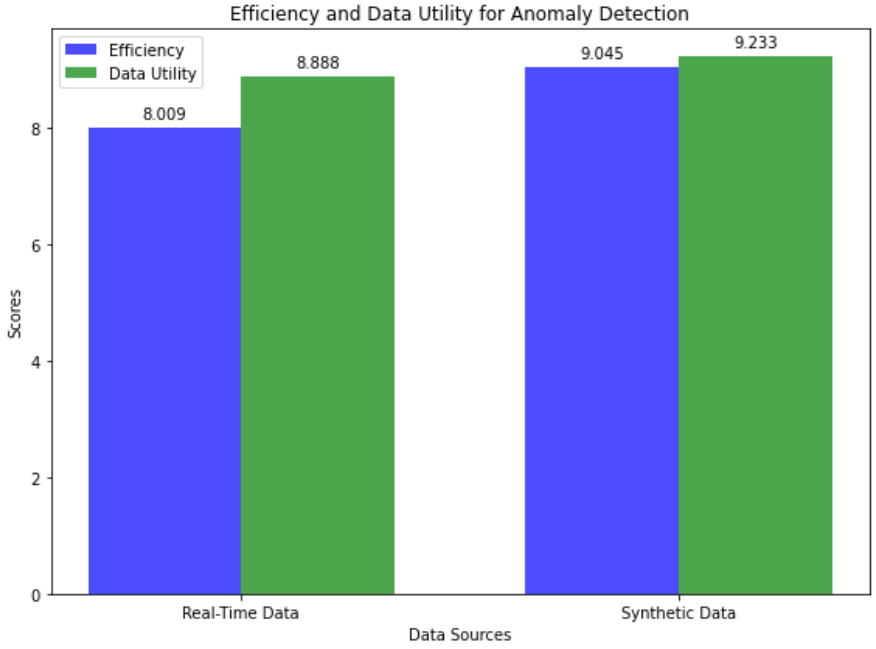

Figure 7. Efficiency and Data Utility Aspects Comparison Between Real Time Data and Synthetic Data for Anomaly Detection Application [15]

9. Conclusion

In conclusion, the use of synthetic data presents a promising avenue for boosting AI accuracy. By augmenting or replacing traditional training data, synthetic data mitigates challenges related to data scarcity, privacy concerns, and resource constraints. As AI continues to advance, the integration of synthetic data into the training pipeline, it holds the potential to catalyse breakthroughs in various domains, enabling AI systems to achieve higher levels of accuracy and generalization. However, careful consideration of synthetic data's quality, biases, and limitations remains essential for responsible and effective implementation.

References

[1] Nikolenko, Sergey I. Synthetic data for deep learning. Vol. 174. Springer Nature, 2021.

[2] Abowd, John M., and Lars Vilhuber. "How protective are synthetic data?." International Conference on Privacy in Statistical Databases. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[3] Patki, Neha, Roy Wedge, and Kalyan Veeramachaneni. "The synthetic data vault." 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2016.

[4] Jordon, James, et al. "Synthetic Data--what, why and how?." arXiv preprint arXiv:2205.03257 (2022).

[5] Hu, Qixin, Alan Yuille, and Zongwei Zhou. "Synthetic Data as Validation." arXiv preprint arXiv:2310.16052 (2023).

[6] Assefa, Samuel A., et al. "Generating synthetic data in finance: opportunities, challenges and pitfalls." Proceedings of the First ACM International Conference on AI in Finance. 2020.

[7] Hyun, Jayun, et al. "Synthetic Data Generation System for AI-Based Diabetic Foot Diagnosis." SN Computer Science 2.5 (2021): 345.

[8] Kurapati, Shalini, and Luca Gilli. "Synthetic data: A convergence between Innovation and GDPR." Journal of Open Access to Law 11.2 (2023): 12-12.

[9] Gonzales, Aldren, Guruprabha Guruswamy, and Scott R. Smith. "Synthetic data in health care: a narrative review." PLOS Digital Health 2.1 (2023): e0000082

[10] Dahmen, Jessamyn, and Diane Cook. "SynSys: A synthetic data generation system for healthcare applications." Sensors 19.5 (2019): 1181.

[11] Giuffrè, Mauro, and Dennis L. Shung. "Harnessing the power of synthetic data in healthcare: innovation, application, and privacy." NPJ Digital Medicine 6.1 (2023): 186.

[12] https://www.kaggle.com/datasets/jehanbhathena/weather-dataset

[13] https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

[14] https://data.world/cancerdatahp/lung-cancer data/workspace/file?filename=cancer+patient+data+sets.xlsx

[15] https://archive.ics.uci.edu/dataset/613/smartphone+dataset+for+anomaly+detection+in+crowds

[16] https://www.kaggle.com/code/badmangamingsv/credit-card-fraud-detection