IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Generation And Integration Of Synthetic Data To Develop Ai Models

Pavamana S[1], Eshwar G[2], Ullas D G[3], Yash K[4]

[1,2,3,4]Final Year B.E. Students, Department of Computer Science and Engineering
Sri Siddhartha Institute of Technology, Tumakuru, Karnataka, India

Abstract

In the era of data-driven intelligence, access to large, diverse, and high-quality datasets remains a challenge due to privacy concerns, scarcity, and labeling costs. This project presents a comprehensive approach to synthetic data generation using generative models—namely, Variational Autoencoders (VAEs), Deep Convolutional GANs (DCGANs), and Wasserstein GANs with Gradient Penalty (WGAN-GP). By leveraging these architectures, synthetic data was generated across both tabular and image domains using datasets like Iris, ADHD-200, MNIST, and Oxford 102 Flowers.

The project explores the full lifecycle: from data preprocessing and model design to training and evaluation. The VAE models were fine-tuned with different latent dimensions to balance reconstruction loss and KL divergence. Image-based GANs were trained iteratively, with evaluation metrics including Fre´chet Inception Distance (FID) applied where feasible.

The results validate that synthetic data can closely mimic real data distributions while preserving structure and variability. The findings also underscore the impact of hardware limitations in deep learning workflows, particularly during evaluation with large-scale models like InceptionV3. This work establishes a scalable and modular baseline for future exploration in privacy-preserving and data-efficient machine learning.

Index Terms— Synthetic Data, GAN, VAE, Deep Learning, FID, MNIST, ADHD, WGAN-GP

1. INTRODUCTION

In the field of Artificial Intelligence (AI), the role of data has become increasingly pivotal. As machine learning and deep learning models grow more complex, so too does their appetite for large, high-quality datasets. However, real-world datasets often present significant challenges: they can be scarce, noisy, imbalanced, or sensitive due to privacy concerns. Particularly in domains such as healthcare, finance, and security, data accessibility is restricted by ethical, legal, or infrastructural limitations. This has led to the emergence of synthetic data generation as a promising solution to address these challenges.

Synthetic data refers to data that is artificially generated rather than collected from real-world events. While traditional data augmentation techniques involve transformations of existing data (such as flipping or rotation in images), synthetic data generation can create completely new samples that are statistically representative of the original dataset. This capability is enabled by powerful generative models such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), which have shown tremendous promise in learning the underlying distribution of real data and generating novel, high-quality samples from it.

The goal of this project is to develop a comprehensive synthetic data generation system using deep learning techniques that span both tabular and image-based datasets. Our solution applies VAEs for structured/tabular data like the Iris dataset and the ADHD-200 neuropsychological dataset, while DCGAN and WGAN-GP

models are employed for unstructured image datasets such as MNIST and the Oxford 102 Flowers dataset.

These models were chosen for their effectiveness in their respective domains, and special care was taken to design and train them under hardware-constrained environments.

The motivation behind this project stems from a confluence of factors. Firstly, real-world data acquisition is expensive and often incomplete. In the case of ADHD phenotypic data, for instance, many features may be missing or labeled inconsistently, which compromises model performance. Secondly, the ability to generate privacy-preserving data is essential for public research dissemination. Thirdly, training deep learning models with synthetic data can help mitigate overfitting, especially when the real data is limited. Lastly, such models also provide interpretability into latent structures within the data and help explore potential scenarios or distributions that may not be well represented in the original dataset.

This system begins with robust preprocessing steps tailored to each dataset. For tabular data, this includes handling missing values, filtering out outliers using z-score methods, and scaling using MinMaxScaler. For images, resizing and normalization are essential to ensure that pixel distributions are suitable for model convergence. Once the datasets are prepared, the models are designed, trained, and evaluated based on their ability to replicate the structure and diversity of the input data.

For the tabular datasets, the VAE architecture includes an encoder network that compresses the input features into a lower-dimensional latent space and a decoder network that reconstructs the original data from this latent representation. The model is trained using a combination of reconstruction loss and KL divergence, with KL annealing strategies employed to improve training stability. For ADHD data, it was observed that a latent dimension of 8 increased KL divergence excessively; hence, a smaller dimension of 5 was chosen.

For image datasets, DCGAN was used for MNIST due to its efficiency and suitability for grayscale images. The model was trained iteratively up to 500 epochs, where we observed significant qualitative improvements in digit formation and consistency. For more complex RGB image generation tasks like Oxford Flowers, WGAN-GP was implemented due to its stable training dynamics and ability to model high-resolution image details. It also introduced a gradient penalty to enforce the Lipschitz constraint, replacing the unstable loss metrics of standard GANs.

Evaluation plays a crucial role in validating the quality of generated data. For tabular data, kernel density estimation (KDE) plots were used to compare real and synthetic data distributions. For images, visual inspection over epochs provided early signs of learning, while FID (Fréchet Inception Distance) was considered as a quantitative metric. However, due to hardware limitations, computing FID scores at scale, especially for the Oxford Flowers dataset, was challenging.

Despite these constraints, the models produced high-fidelity synthetic data that could feasibly augment real-world datasets in downstream applications. Furthermore, this project highlights the practical trade-offs between model performance and computational resources. In environments with limited hardware, strategies like reducing batch size, model simplification, and early stopping become vital.

In summary, this project successfully implements a suite of generative models to tackle the problem of data scarcity using synthetic generation. Through structured methodology, diverse dataset support, and modular evaluation pipelines, it serves as a practical blueprint for integrating synthetic data in machine learning workflows. As the field moves toward responsible AI and data-centric development, such tools are not just optional but necessary.

2. LITERATURE REVIEW

1. Comprehensive Exploration of Synthetic Data Generation This paper outlines how synthetic data can augment real-world datasets, improve model generalization, and support data privacy. It categorizes generation techniques into GAN-based, VAE-based, and rule-based, emphasizing the importance of domain knowledge in selecting appropriate methods. It also highlights evaluation challenges such as distribution similarity and task-specific metrics. It provides a clear taxonomy of synthetic data types and use cases in AI, healthcare, finance, and computer vision .

2. Machine Learning for Synthetic Data Generation: A Review This review focuses on the role of ML in synthesizing tabular, image, and time-series data. It discusses GAN variants (CGAN, WGAN, BigGAN), data utility vs. privacy trade-offs, and highlights key limitations such as bias replication and training instability.

The paper calls for better performance benchmarks and unified frameworks for evaluating synthetic data utility and fidelity .

3. Survey on Synthetic Data Generation, Evaluation Methods and GANs A technical overview is presented on state-of-the-art GAN architectures used for synthetic data generation. It discusses FID and IS metrics in depth and categorizes GANs into unsupervised, semi-supervised, and conditional. Special focus is given to domain-specific applications like medical imaging and autonomous driving. It also discusses the issue of mode collapse and proposes hybrid evaluation frameworks .

4. Synthetic Data: What, Why, and How This foundational paper distinguishes synthetic data into fully synthetic, partially synthetic, and hybrid datasets. It provides an analytical viewpoint on when to prefer synthetic data and outlines legal and ethical considerations in data anonymization. The work emphasizes reproducibility and utility as the primary goals and recommends simulation-based approaches when real data is limited or unavailable .

5. Wasserstein GAN (WGAN) WGAN introduced the Earth Mover's distance to address convergence issues and instability in classical GANs. It removed the sigmoid activation in the discriminator (critic) and replaced it with a continuous scalar output. The Lipschitz constraint was initially enforced by weight clipping, and later improved using gradient penalty (WGAN-GP), enhancing training robustness and sample diversity .

6. Generation of Synthetic Data with Generative Adversarial Networks This paper explores how GANs can be applied across different data modalities—images, text, and tabular data—for synthetic data generation. It explains the core structure of GANs (generator and discriminator), and their adversarial training process. The authors examine various GAN variants like DCGAN, CGAN, and CycleGAN, focusing on their suitability for different types of data.

3. METHODOLOGY

1. **Problem Analysis**

- Identify the type of data required for AI model training.
- Analyze the challenges with existing datasets such bias, privacy concerns, or data scarcity.

2. **Collection of Referential Data**

- Select real-world datasets for benchmarking synthetic data performance.
- **Tabular Data:** Iris, ADHD dataset.
- **Image Data:** OXford Flower 102,MNIST dataset.

3. **Data Preprocessing**

- Normalize, scale, and clean referential data.
- Handle missing values, remove duplicates, and balance class distributions if required.

4. **Generative Model Selection**

- Choose an appropriate generative models based on data type:

– **For Image Data:** Generative Adversarial Networks (GANs) (e.g., DCGAN, StyleGAN,WGAN).

– **Tabular Data:** Variational Autoencoders (VAEs) or CTGAN (Conditional Tabular GAN).

5. **Synthetic Data Generation**

- Implement and fine-tune generative models to create synthetic data.
- Ensure data variability and realism while avoiding mode collapse in GANs.

6. **Evaluation Metrics:** Define the metrics to access the quality of synthetic data for training AI models.

7. **Integration with AI models**

- Train AI models on real, synthetic, and mixed datasets.
- Evaluate performance differences using metrics like accuracy, F1-score, precision-recall curves.
- Test synthetic data impact on models like CNNs (for images) and Decision Trees, Random Forest, KNNs or Neural Networks .

The overall architecture of the synthetic data generation system is illustrated in Figure 1. It outlines the flow from dataset collection and preprocessing to model training, evaluation, and integration with AI models.

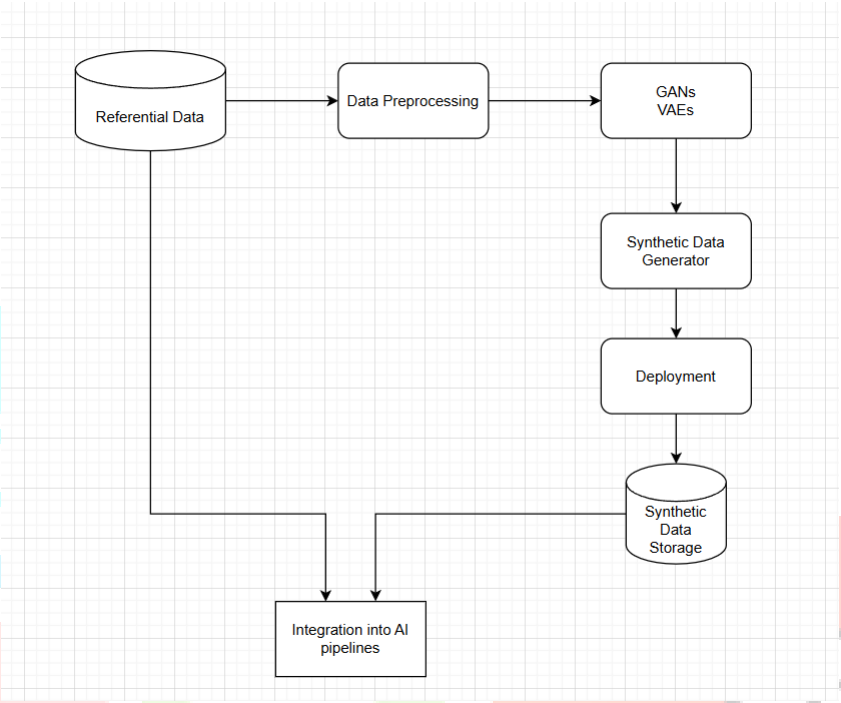

Figure 1: Architecture of Synthetic Data Generation Engine

3.1 Hyperparameter Settings

The following table lists the key hyperparameters used for training the VAE, DCGAN, and WGAN-GP models:

Table 1: Hyperparameters Used Across Models

Model / Component	Hyperparameter	Value
VAE (Iris/ADHD-200)	Latent Dimension Batch Size Learning Rate Epochs	5 (reduced from 8) 32 0.001 300
DCGAN (MNIST)	Latent Dimension Batch Size Learning Rate Epochs	100 128 0.0002 250

WGAN-GP (Oxford Flowers)	Latent Dimension	128
	Batch Size	64
	Learning Rate (Generator)	0.0001
	Learning Rate (Critic)	0.0001
	Epochs	250
	Gradient Penalty Weight	10
	Critic Iterations	5

3.2 Dataset Description

The following datasets were used for training and evaluating synthetic data generation models:

Table 2: Summary of Datasets Used

Dataset	Type	Samples	Features / Dimensions	Algorithm Used
Iris	Tabular	150	4	VAE Training
ADHD-200	Tabular	∼150	5	VAE Training
MNIST	Image (Gray)	70,000	28x28	DCGAN
Oxford Flowers	Image (RGB)	8,189	64x64	WGAN-GP

4. RESULTS AND DISCUSSION

This section outlines the experimental results obtained from synthetic data generation across various datasets using VAE, DCGAN, and WGAN-GP architectures. Both qualitative and quantitative methods were employed to evaluate the performance and realism of the generated data.

0.1. Tabular Data Generation with VAE

Iris Dataset: The VAE model effectively learned the low-dimensional structure of the Iris dataset. Using a latent dimension of 5, the model achieved stable training and reliable reconstructions. Kernel Density Estimation (KDE) plots confirmed close alignment between the real and synthetic data distributions. KL divergence was managed through annealing to improve convergence.

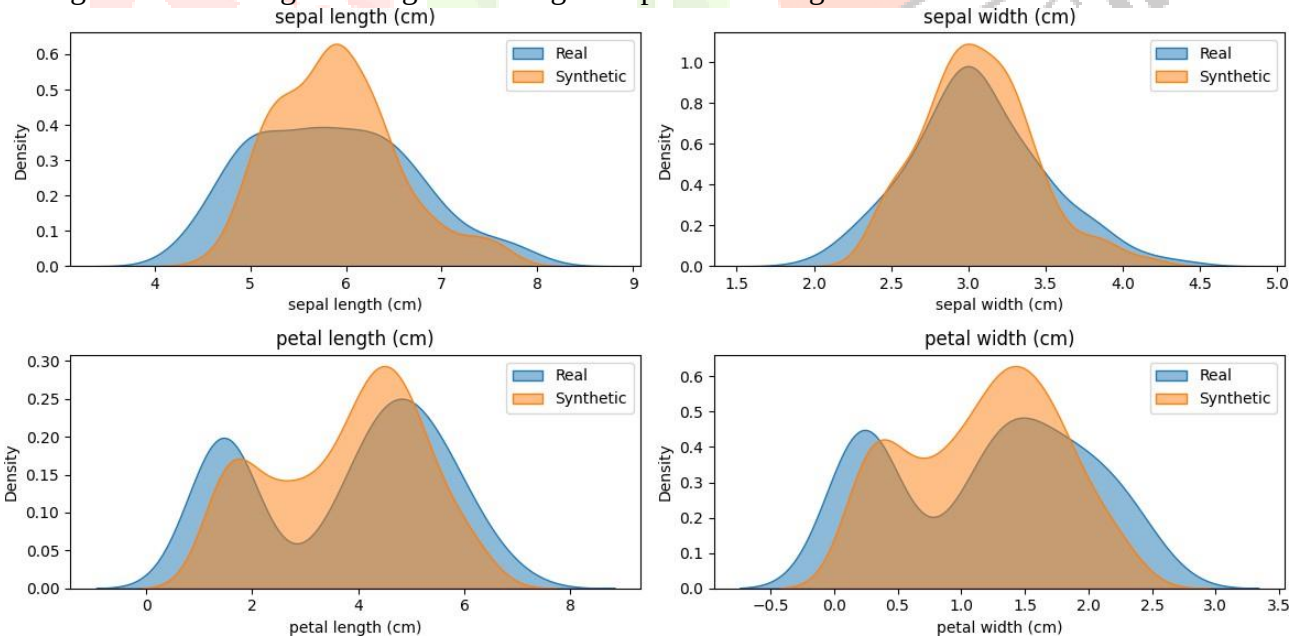

Figure 2: Synthetic vs Real Distribution (IRIS Dataset)

ADHD-200 Dataset: The ADHD dataset, being more complex, required careful preprocessing and train- ing. The VAE again used a latent dimension of 5 to avoid instability. Generated data preserved feature distributions like age and IQ scores, validated using KDE and statistical checks. Results showed potential for privacy-preserving neuropsychological data modeling.

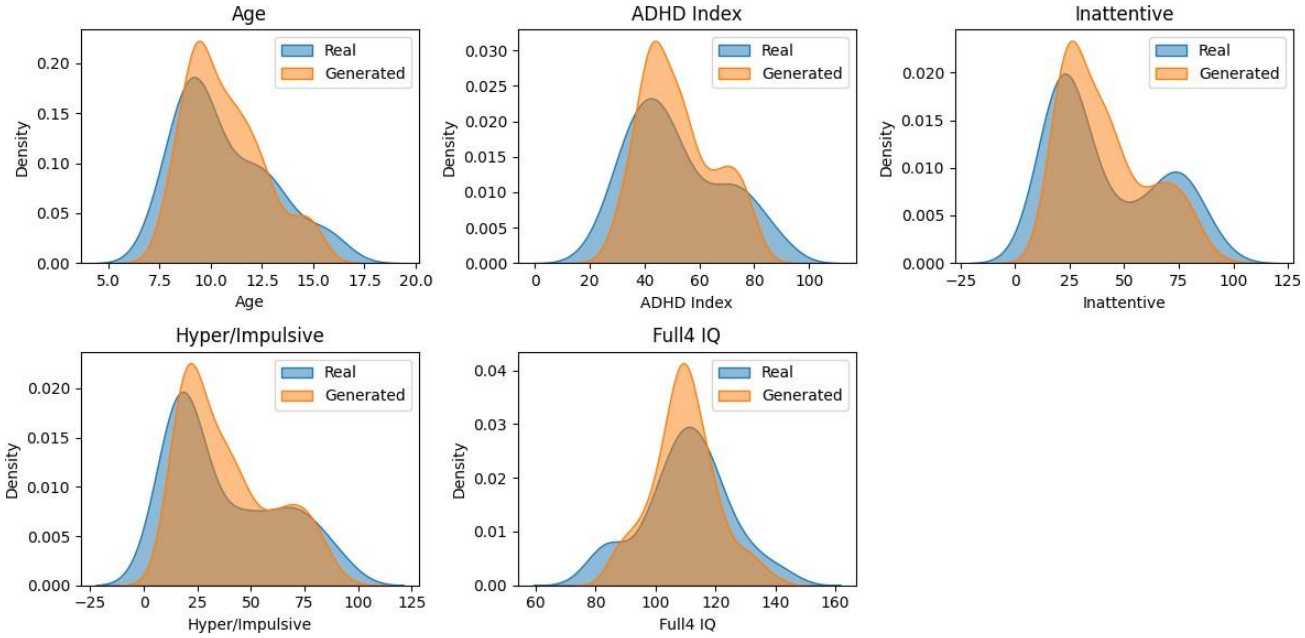

Figure 3: Synthetic vs Real Distribution (ADHD Dataset)

0.2. Image Generation with GANs

MNIST Dataset: DCGAN was trained for 500 epochs. The generator successfully produced clear and consistent handwritten digits. Visual inspection confirmed diversity and sharpness in outputs. FID scores improved over training, reducing from 127 to 70, validating the model's convergence.

Oxford Flowers Dataset:

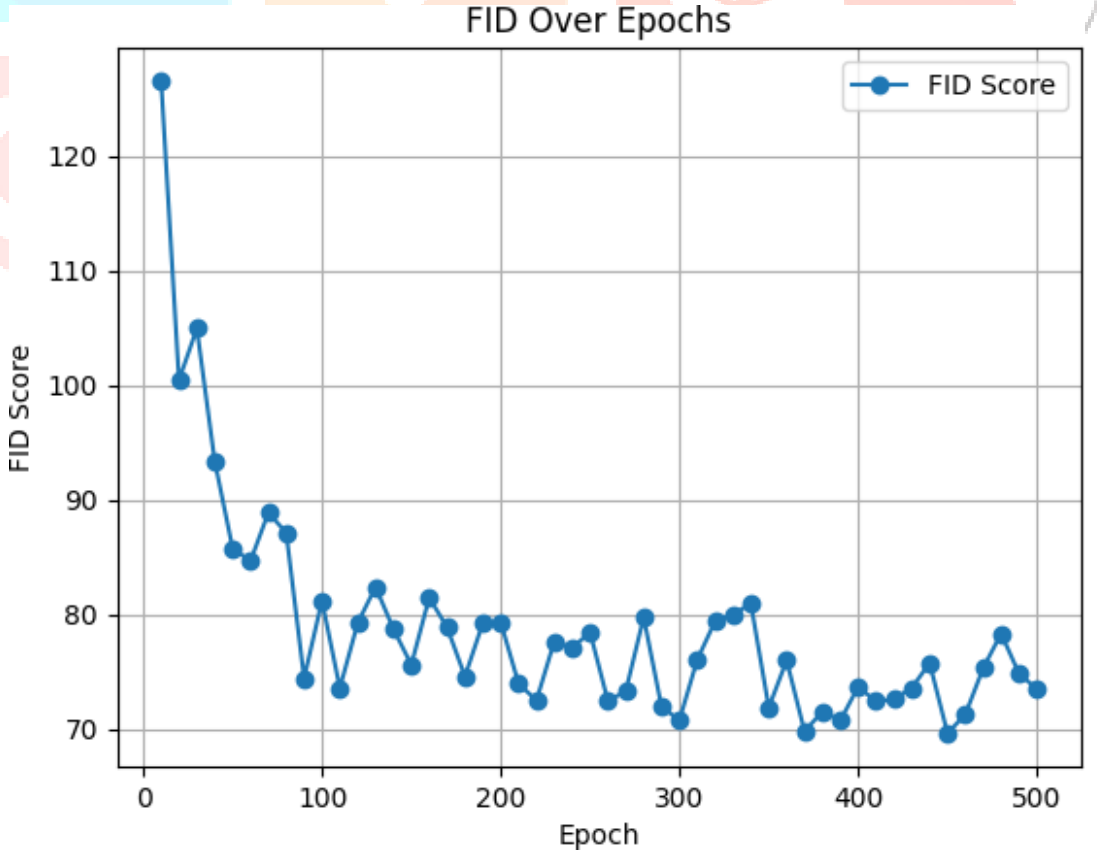

Figure 4: FID Over Epochs

WGAN-GP was employed to generate high-resolution (64×64) RGB flower images. Training remained stable, and visual inspection showed increasingly detailed and diverse outputs. FID computation for this dataset was infeasible due to resource constraints, which highlights the importance of sufficient hardware for

deep generative models.

FID Score Evaluation – Oxford 102 Flowers

Fréchet Inception Distance (FID) was planned as a key metric to evaluate the image quality of samples generated by the WGAN-GP model on the Oxford 102 Flowers dataset. FID is widely used in generative modeling to assess how close the generated images are to the real data distribution by comparing feature vectors extracted using the InceptionV3 model.

However, in our case, FID calculation proved to be highly resource-intensive. The evaluation could not be completed successfully due to repeated **memory exhaustion errors** on both Kaggle (with access to 16GB RAM) and our local hardware setups. The failure was primarily due to:

- The **large size of the Oxford Flowers dataset** (8,189 RGB images).
- The need to **resize all images to 299×299** to match InceptionV3 input requirements.
- The **high VRAM usage** during batch processing of real and generated images to extract feature statistics.

This experience highlights a critical insight: **computational resources are a bottleneck in deep learning evaluation**. While our WGAN-GP model showed good qualitative performance and stable training, the inability to compute FID reliably emphasizes the importance of high-end GPUs or memory-optimized solutions in production-grade synthetic data evaluation.

We plan to reattempt FID evaluation with reduced batch sizes or by leveraging cloud infrastructure in future iterations.

Evaluation Summary

- VAE models yielded reliable reconstructions for both tabular datasets.
- DCGAN outperformed expectations on MNIST, supported by FID trends.
- WGAN-GP delivered promising qualitative results, though quantitative evaluation was constrained by hardware.
- All models showed adaptability to their respective data types, proving the feasibility of a multi-modal synthetic data generator.

Overall, the generated datasets retained key statistical properties and visual characteristics of the original data. These results support the use of synthetic data in scenarios where privacy, availability, or cost are limiting factors.

5. CONCLUSION

The process of synthetic data generation plays a pivotal role in modern AI development, enabling scalable data augmentation, privacy preservation, and robust model training. This report presented a comprehensive pipeline using both Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) including DCGAN and WGAN-GP, applied across diverse datasets—tabular (Iris and ADHD-200) and visual (MNIST and Oxford 102 Flowers).

The project involved detailed preprocessing, model implementation, hyperparameter tuning, and qualitative evaluation of synthetic outputs. VAE-based approaches proved effective for structured tabular data, while DCGAN and WGAN-GP offered high-quality image generation, with WGAN-GP yielding sharper, more stable outputs.

Key challenges included managing memory-intensive training, selecting optimal latent dimensions, and model instability. Despite these, the pipeline successfully produced synthetic data closely resembling the original distributions.

This synthetic data framework contributes toward solving problems of limited real-world data, promot- ing privacy-aware learning, and enhancing AI model generalization—forming a foundational asset for future machine learning applications.

Future and Scope As the demand for data privacy, fairness, and scalability grows in the field of Artificial Intelligence, synthetic data is poised to play a vital role in real-world deployments. While the models implemented in this project demonstrate the ability to generate high-quality synthetic data, there remains significant scope for future exploration and improvement.

One promising direction is the enhancement of generative models with larger and deeper architectures such as StyleGAN or Diffusion Models, which have been shown to produce state-of-the-art results in image synthesis. These could be explored especially for complex datasets like Oxford Flowers, where finer details and texture quality are critical.

Additionally, performance metrics can be improved by integrating more robust and real-time evaluation techniques. For example, deploying Fréchet Inception Distance (FID) tracking during training in a resource-efficient manner, or incorporating Precision-Recall for Distributions (PRD) can provide a more nuanced understanding of model quality.

On the tabular data side, there is a need to move beyond purely numerical data and explore categorical or mixed-type data generation using models such as CTGAN or TVAE. Integrating such models can broaden the application of synthetic data generation to domains like healthcare, banking, and census data.

Another area of expansion lies in real-time deployment and synthetic dataset APIs. The integration of our models into a web-based interface (e.g., using FastAPI or Flask) could allow researchers and developers to generate synthetic samples on-demand for training, testing, or educational purposes.

Finally, future work can explore the ethical and regulatory aspects of synthetic data. By embedding privacy-preserving mechanisms such as differential privacy or membership inference resistance, synthetic data can be made not just realistic but also safe and compliant with data protection regulations like GDPR.

With ongoing research and rapidly advancing hardware, synthetic data generation will continue to evolve and become a cornerstone of responsible and scalable AI development.

ACKNOWLEDGEMENT

We express our heartfelt gratitude to our beloved Principal, **Dr. M.S Raviprakasha**, for providing us with an excellent academic environment and constant encouragement to pursue research and innovation.

We are sincerely thankful to the Head of the Department, **Dr. Raviram V**, Department of Computer Science and Engineering, Sri Siddhartha Institute of Technology, Tumakuru, for his guidance, support, and providing the necessary infrastructure to carry out this project.

We are deeply grateful to our project guide, **Mrs. Sindhu TN**, for her invaluable guidance, technical insights, timely feedback, and constant support throughout the duration of this work. Her mentorship played a key role in shaping the quality and direction of this project.

Future work may explore integrating differential privacy or privacy-preserving training techniques into the generative models to further ensure regulatory compliance and robustness against adversarial privacy breaches.

We also thank our teaching faculty, lab instructors, and peers who provided constructive suggestions and motivation during the course of the project.

References

[1] Alkahtani, H., et al. *Survey on Synthetic Data Generation, Evaluation Methods and GANs*, 2020.

[2] Zhao, Q., & Saligrama, V. *Synthetic Data: What, Why and How*, 2023.

[3] Chen, J., & Huang, Z. *Machine Learning for Synthetic Data Generation: A Review*, 2022.

[4] Cai, Q., et al. *Comprehensive Exploration of Synthetic Data Generation*, 2021.

[5] Frid-Adar, M., et al. *Generation of Synthetic Data with Generative Adversarial Networks*, 2018.

[6] Arjovsky, M., Chintala, S., & Bottou, L. *Wasserstein GAN*, 2017.