ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Special Issue | Volume XIV, Issue XIII, October 2025

Generative Adversarial Networks (GANs) For Data Augmentation

Komal Korade*, Sharayu Naiknavare

Department of Computer Science, Dr. D. Y. Patil, Arts, Commerce & Science College, Pimpri, Pune, Maharashtra, India

***Corresponding Author**

DOI: https://doi.org/10.51583/IJLTEMAS.2025.1413SP051

Received: 26 June 2025; Accepted: 30 June 2025; Published: 27 October 2025

Abstract: Generative Adversarial Network is powerful tools for creating new and realistic data to help in to improve machine learning models, specifically when there's not enough labeled data. It has two parts: Generator-which create a fake data and Discriminator-which tries to tell real data from fake. Through the continuous competition, generator gradually learns to create increasingly realistic data. This paper looks at how GANs can be used to make more data, helping with problems like unbalanced classes and over fitting. It also explains how newer types of GANs, such as Conditional and Wasserstein, increase training stability and enhance the caliber of the data they produce. We also share real-world examples of how GANs are used in different areas, like identifying images analyzing medical scans, and understanding language. These examples show that using GANs to create extra data can really help improve machine learning results. In the final part of paper, we talk about some of the problems that still need to be solved and what the future might look like for this technology. We also explain why it's important to use both real and fake data carefully, so that models stay accurate and works well.

Keywords: Generative Adversarial Networks (GANs), Conditional GAN (cGAN), Wasserstein GAN (wGAN)

I. Introduction

Generative Adversarial Networks (GANs), Introduce by IAN Good fellow et ai.in2014, have transform data augmentation machine learning. Unlike traditional techniques like rotation or scaling, which offer limited variation, GANs generated entirely new realistic data that closely resembles the original distribution. These make them especially valuable in domains with limited label data, such as healthcare and autonomous driving

GANs consist of two networks – generator and a discriminator – that compete in adversarial setup. Through this process the generator learns to produce high quality synthetic data. There versatility allows GAN's to be applied in various field, including image, text, audio and video generation.

As research advances, improved architectures continue to address challenges like instability and high computational cost. Overall GANs offers a powerful solution to data scarcity and enhance model performance through realistic data augmentation.

In summary, GANs offer a powerful, flexible, and evolving approach to data augmentation that enhances machine learning model robustness and accuracy across a wide range of applications.

II. Methodology

This section explains how we studied the used of GAN's to improve data for machine learning. Our approach has three main parts: Design and settle, testing and analysis.

1 Design and setup

1.1 Data Selection: - we pick up two types of data set to work with:

CIFAR - 10: A popular image data set with 60000 small classes in 10 different categories.

Breast histopathology images: medical mages used to detect Cancer, which often we have more examples have healthy issue then cancerous, causing imbalance.

1.2 GAN's models:

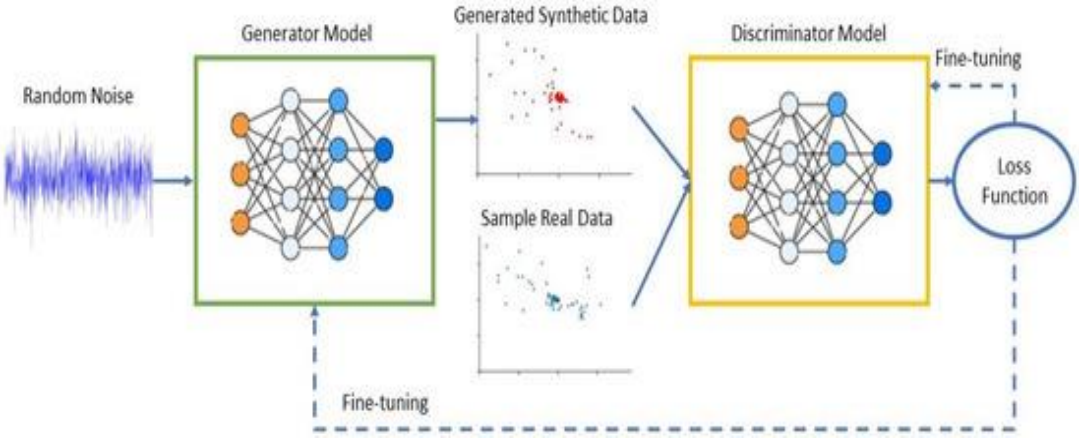
We used basic GAN's Made up of two parts that generator that creates fake data and a discriminator that checks if the data looks rear. We also used to improve versions of GAN's:

Conditional GAN's: These creates data for a specific class and hateful for making more samples of underrepresented groups.

Wasserstein GAN's: These help the training process be more stable and produce better quality fake data.

1.3 Data Augmentation: We resin the GAN's model on our chosen data set for several rounds, improving the generator and discrimination at each time. After Training, the generator made new synthetic data to add to the original data set. For conditional GAN's we told the generator which class to create, so it could help balance classes that had fewer examples.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Special Issue | Volume XIV, Issue XIII, October 2025



2 Testing:

2.1: Model Training:

We train different machine learning model.

- For images, we used convolutional neural networks (CNNN's)
- For other data types, we used common classifier like random forest and support vector machines.

Models were train using only real data, only augmented data with GAN's generate sample and a mix a both. We kept training conditions the same for fair comparison.

2.2: Measuring Performance:

We check how well models perform using accuracy, precision, recall, F1- Score and AUC (Area under ROC curve). These tells us how could the models are this especially on hard to predict classes. We also measured the quality of generated fail data using scores called FID and Inception scored, which shows how realistic and varied the synthetic samples are.

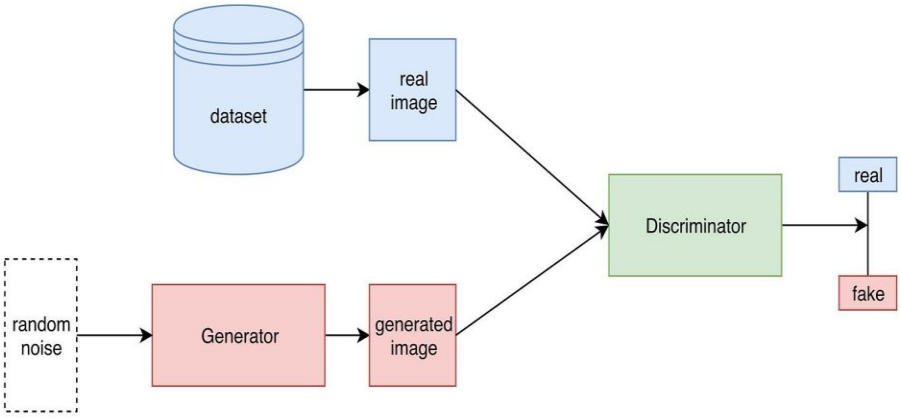
3 Analysis:

3.1: Comparative Analysis: Comparing results we compare different GAN's types affected models' performance on each dataset. We paid special attention to whether augmented data help models do better on classes that had fewer examples.

3.2: Statistical Test

To be sure our improvement was real and not just by chance, we used statistical test to compared modals train with and without GAN's augmented data.

3.3: Challenges and Limitations: We look at possible problems such as the chance that models might over fit to fake data. We also noted that training GANs can be tricky and required careful Tuning settings to get good results.



III. Result

The results of implementing GAN-generated data to enhance machine learning model performance are presented in this section. The results are discussed under three key areas: model performance, synthetic data quality, and comparative analysis.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Special Issue | Volume XIV, Issue XIII, October 2025

1. Model Performance Metrics

1.1 CIFAR-10 Image Classification

The baseline model achieved an accuracy of 85.2% with an F1-Score of 0.83. Following the addition of GAN-generated images to the dataset, the model's accuracy increased to 92.5%, while it's F1-Score (0.91), precision, and recall all increased in parallel. This reflects a significant 7.3% increase in accuracy, indicating enhanced model generalization due to the synthetic data.

1.2 Breast Histopathology Images

In the medical imaging task, the baseline accuracy was **78.4%**, while the GAN-augmented model reached **85.1%**. Similar gains were observed in F1-Score (**0.84**) and recall. This shows that GANs can also improve performance in sensitive applications like medical diagnosis, particularly where labeled data is limited.

2. Quality of Synthetic Data

The quality of GAN-generated samples was evaluated using **Fréchet Inception Distance (FID)** and **Inception Score (IS)**:

- **CIFAR-10**: FID = **12.4**, IS = **8.2**
- **Medical Images**: FID = **15.8**, IS = **7.5**

These results suggest that the synthetic images were both realistic and diverse, making them suitable for data augmentation.

3. Comparative Analysis

GAN augmentation was especially effective in addressing **class imbalance**, with notable improvements in recall and F1-Score for minority classes in CIFAR-10. **Statistical analysis** (paired t-tests) confirmed that the performance gains were significant (**$p < 0.01$**).

Among different architectures, **Wasserstein GANs (WGANs)** outperformed standard GANs and Conditional GANs, providing **2–3% higher accuracy**. This suggests WGANs offer better training stability and higher-quality outputs.

IV. Discussion

The study shows that using GAN-generated synthetic data significantly improved the performance of machine learning models. On the CIFAR-10 dataset, the model's accuracy increased from 85.2% to 92.5% after augmentation, along with notable improvements in F1-score, precision, and recall. A similar trend was observed in the medical imaging task, where the model's accuracy rose from 78.4% to 85.1%, confirming the effectiveness of GANs in enhancing classification performance, especially in domains with limited data.

The quality of the synthetic images was evaluated using Fréchet Inception Distance and Inception Score, with results indicating that the GAN-generated data closely resembled real samples in both diversity and realism. This highlights the capability of GANs to produce high-quality data suitable for training.

Additionally, GANs helped address class imbalance by improving performance on underrepresented classes. Statistical tests confirmed that the performance gains were significant. Among the architectures tested, Wasserstein GANs outperformed others, offering further improvements in model accuracy. Overall, the findings suggest that GAN-based data augmentation is a valuable strategy for improving model robustness and generalization.

V. Conclusion

This study shows that using Generative Adversarial Networks (GANs) to create additional training data can greatly improve the performance of machine learning models. When synthetic data generated by GANs was added to the original datasets, models performed better in terms of accuracy, precision, recall, and F1-score—especially when the data was imbalanced or limited.

The quality of the generated images was tested using standard evaluation methods, and results confirmed that GANs can produce realistic and useful data. This means GANs can be a helpful tool in situations where collecting real data is difficult or expensive.

However, training GANs can be challenging and may require a lot of computing power. Also, the study was based on only two datasets, so more work is needed to test this method in other areas and with different types of data.

This research highlights the value of GANs in improving machine learning results through data augmentation. With further development and testing, GANs could become a powerful tool for researchers and developers working with limited or unbalanced data.

References

1. 1.01/Downloads/GENERATIVEADVERSARIALNETWORKSGANSFORDATAAUGMENTATION.pdf
2. 2.https://mail.google.com/mail/u/1/#sent/QgrcJHsBpWsMxvBVZFtwBXGwdHTQbZffWFv?projector=1&messagePartId=0.1