



Original Research Paper

GAN-BASED AUGMENTATION TO ENHANCE DATA SET USING ADVANCED DEEP LEARNING TECHNIQUES

¹Savitha A Shenoy, ²Prathviraj, ³Panchami Pai and ⁴Shashank Poojary

^{1,2,4} Dept. of Computer Science & Engineering, Shri Madhwa Vadiraja Institute of Technology and Management, Bantakal
³ Dept. of Artificial Intelligence & Data Science, Shri Madhwa Vadiraja Institute of Technology and Management, Bantakal
¹savithashenoy2003@gmail.com, ²prathviz.111@gmail.com, ³panchamipai21@gmail.com and
⁴shashankpoojary198@gmail.com

ABSTRACT

Scientific literature has been growing exponentially, and in this regard, the need to extract and classify critical information using automated methods arises, especially concerning the identification of "problem" and "method" sentences within research papers. This paper focuses on the GAN-based data augmentation for sentence classification within scientific texts. GANs can overcome the challenge of data scarcity by generating synthetic data, thereby reducing dependence on high-frequency patterns, which usually lead to overfitting. This paper investigates the effectiveness of different architectures of GANs and data augmentation techniques for sentence classification tasks, focusing on the discrimination between problem and method sentences. Experimental results show the potential of improvements in classification accuracy, especially with augmented datasets including synthetic sentences. These findings open the door to more reliable and scalable methods for analyzing scientific literature and demonstrate the promise of GAN-driven data augmentation for NLP applications.

Keywords: GAN, data augmentation, sentence classification, scientific literature, NLP, text generation, synthetic data, imbalanced dataset, transformer model, semantic preservation, adversarial training, neural networks, model generalization, deep learning.

1. INTRODUCTION

Finding "problem" and "method" sentences in a research paper is one of the many automated information extraction tasks made possible by the vast amount of scientific literature—more than one billion published works—in the world. Applications such as automatic summarization and academic retrieval depend on these elements [1, 2]. Geometrical transformations of datasets are among the most popular methods for data augmentation. However, overfitting and decreased generalization are frequently the results of their limited capacity to provide diverse data [1]. By combining actual data and enhancing model robustness across several domains, cutting-edge techniques like Generative Adversarial Networks (GANs) offer a novel approach [2,3].

With correcting data imbalances and enhancing model performance, GANs have demonstrated special potential in medical imaging. As an illustration, Jiménez-Gaona et al. [2] utilized GAN-based augmentation in breast cancer diagnosis enhancement by adding ultrasound and mammography images for an appreciable accuracy in the classification process. For text classification, Zhang and Zhang [3] were also successful in exploiting sentence-level augmentations like back translation and techniques from BERT, thereby bettering the precision in "problem"

and "method" sentences' classification. These studies show the potential of GANs in addressing data scarcity while improving model generalization in medical and text-based tasks.

Thus, maintaining alignment between the synthetic and real data distributions helps prevent biases in high-stakes applications like healthcare. Motamed et al. [4] and Kugelman et al. [5] similarly highlighted how alignment is essential in GAN-based augmentation, especially in medical imaging tasks such as COVID-19 detection and retinal segmentation. Their results highlight the potential of GANs to produce high-quality synthetic data that improves model performance in domains where labeled datasets are sparse.

This paper addresses GAN-based data augmentation techniques to improve the classification of scientific literature sentences that denote "problem" and "method". By synthesizing diverse datasets, we aim at least to overcome the challenges facing traditional augmentation methods and eventually push the generalization of our models to higher levels. Building on such state-of-the-art advancements, this paper will significantly and directly contribute to the development of robust frameworks for the scientific analysis of literature research.

*Corresponding Author:

Received: 25th Feb, 2026; Revised: 6th June, 2026; Accepted: 8th July, 2026; Available Online: 14th Sep, 2026
(DOI): 10.70102/AEJ.2026.18.6s.33

2. RELATED WORK

The critical technique developed over the years is data augmentation for enhancing the performance of a model in diverse domains, mainly where the available data are unbalanced or in low numbers. Beyond the aforementioned applications, multiple research works investigated varied approaches and frameworks for moving this field ahead.

Al Khalil et al. [6] focused on improving medical image segmentation by using a hybrid approach that combines CNNs with attention mechanisms. Using GAN-based data augmentation, the study significantly reduced segmentation failures in cardiac MRI and improved the detection of anomalies in complex imaging data. The authors emphasized the importance of training on diverse datasets to minimize biases and enhance generalization.

Suhartono et al. [7] proposed a novel feature-based augmentation framework to detect sarcasm in NLP. Therefore, this process examined the challenges that would have been thrown into its path regarding the detection of subtle linguistic cues by considering sentiment shift patterns and contextual clues, hence bringing a considerable improvement in the performance of the model. The work showed the promise of augmentation methods for refining the comprehension of subtle expressions of linguistic meaning and enhancing performances in scarce-annotated example datasets.

Tran et al. [8] gave a comprehensive review of differentiable image data augmentation techniques that are applicable in various domains such as computer vision, NLP, and autonomous driving. The authors demonstrated the potential of unsupervised domain adaptation to reduce variations across domains and increase transfer learning among different datasets. They outlined possible key challenges or areas to develop in the future including domain shifts or adaptation in non-stationary environments, and also established potential future prospects for developing augmentation methods.

Pandey et al. [9] proposed a data augmentation method to improve the performance of ITN in ASR systems. Their approach generated synthetic spoken-written numeric pairs from out-of-domain data with minimal manual labeling, achieving a 14.44% improvement in ITN accuracy. This study underlined the effectiveness of data augmentation in reducing dependency on expensive in-domain datasets while enhancing model performance in numeric transformations.

Li et al. [10] proposed EID-GAN, a generative adversarial network designed to augment datasets with rare outlier elements, in order to address the challenge of extreme data imbalance. The model improved feature learning for anomalies by introducing a penalty function and was particularly suited for industrial anomaly detection applications. The study emphasized the adaptability of GANs in tackling imbalance issues across varied domains.

Motamed et al. [11] proposed GAN-based augmentation in medical imaging with a focus on enhancing pneumonia and COVID-19 detection from chest X-ray images. By synthesizing realistic synthetic images, the authors demonstrated how GANs could be applied to enhance generalization and robustness of the model, particularly in low data regimes. It is worth mentioning that the generation of synthetic data must be aligned with real datasets to avoid the pitfalls of potential biases in diagnostics.

Tran et al. [12] discussed the limitations of traditional data augmentation in GAN training and introduced the Data Augmentation Optimized for GAN (DAG) framework. DAG ensures alignment between augmented and original data distributions, preserving fidelity and improving synthetic image quality. This framework is particularly beneficial in applications requiring high data fidelity, such as medical imaging and anomaly detection, where traditional augmentation methods may degrade performance.

Baradad et al [13] suggested dependence on big datasets for machine learning models is fraught with difficulties like high curation expense, bias, and privacy. Learning to See by Looking at Noise discusses training vision models without actual images using procedural noise processes. They demonstrate in their results that informative visual descriptions can be learned

from generated data if it preserves the important structural properties. Similarly, in NLP, the lack of data holds back classification operations, especially for detecting problem and method sentences in scientific documents. Data augmentation through GAN mitigates this challenge by creating synthetic sentences, enhancing model generalization and lowering overfitting.

Both methods accentuate the increased importance of synthetic data in combating dataset constraints across machine learning areas.

Antoniou et al [15] talked about Data augmentation is a key method for enhancing the robustness and generalization of machine learning models, especially in low-data scenarios. Data Augmentation Generative Adversarial Networks (Antoniou et al., 2018) propose DAGAN, a framework that employs GANs to create diverse augmented samples and enhances performance in low-data settings. Their research shows that GAN-based augmentation enables models to learn improved representations by augmenting sample diversity without compromising critical data features. In scientific text classification, analogous methods have been used to create synthetic problem and method sentences to counteract data sparsity and overfitting. The success of GAN-driven augmentation in vision and NLP tasks highlights its promise in improving model

reliability and scalability. Biswas et al. [16] investigated GAN-based data augmentation for improved model performance through generating varied synthetic samples. The research showed enhancement in classification tasks by reducing data scarcity and overfitting. Vision and NLP tasks benefit equally from GAN-driven augmentation with improved dataset balance and accuracy.

3. METHODOLOGY

This section illustrates the hybrid approach, consolidating computer vision technique and deep learning technique for implementing image augmentation. Our framework comprises dataset limitation in the form of image manipulation and image generation technique.

1. Image Data Collection, Preprocessing and Augmentation

Preparation and augmentation of the image dataset to inject variability in the enhancement of models ability to learn accurate data distributions at the very first stage.

1.1 Input Data Collection

A set of diverse images is collected systematically from specified sources and stored for subsequent processing. These images have similar structure and represent the data which is getting augmented.

1.2 Input Data Preprocessing

The processing implements standardization protocols across all the original input images. The images undergo 256×256 pixels dimension normalisation for traditional augmentation using opencv, while the images undergo 64×64 pixels dimension normalisation for augmentation using deep learning.

1.3 Image Augmentation Techniques

Augmentation techniques are performed on the dataset to ensure enriching the content while preserving the core features of the dataset. These augmentation techniques include:

Blending: Combines two random images using weighted pixel-wise operations introducing smooth transitions

Rotation: Random rotations between a range such as -30° to +30° simulating different views.

Flipping: Vertical and horizontal mirror image generation.

Noise Addition: Adds Gaussian noise simulating real-world noise.

Color Space Alignment: Randomly adjusting brightness and contrast to compensate for lighting changes.

1.4 Output

The transformed images are stored in an output directory organized by transformation type, allowing for easy dataset organization. This increases the diversity of datasets.

2. Generating Images with GANs

This step is based on training a GAN to produce a realistic synthetic image that captures the characteristics of the original dataset.

2.1 Overview

A GAN uses a Generator, which creates images from random noise vectors, and a Discriminator that observes the images as real or fake. This dual approach better fine-tunes the quality of the images output from the generator.

2.2 GAN Architecture

Generator: maps a random noise (latent vector) to an image of high dimensionality using transposed convolution layers by gradually adding detail and increasing resolution.

Discriminator: It provides adaptive feedback through complementary convolution layers that enable the computer to learn complex image distributions .

2.3 Dataset Preparation

The augmented images are preprocessed (resized 64×64 pixels and normalized[-1,1]) and then divided into training and validation sets to normalize input for GAN training. These preprocessing methods ensure stabilized input in most cases.

2.4 GAN Training

The training alternates between the Generator and Discriminator:

Discriminator Update: Trains on real and fake images, learns to differentiate between the two.

Generator Update: Trains to fool the Discriminator by producing increasingly realistic images.

Binary Cross-Entropy Loss is used in both networks for optimizing performance. The training runs for multiple epochs, improving GAN's ability to generate images progressively.

Training parameters:-

min-batch size : 64 samples

epochs : 500

learning rate : 1e-4

momentum parameters : $\beta_1=0.5$, $\beta_2=0.999$

2.5 Model Saving and Image Generation

Model Persistence: The trained Generator and Discriminator models are saved (in pth format) for later use.

Image Synthesis: The Generator produces new images by inputting random noise vectors simulating the statistical characteristics of the original dataset [7-19].

Suhartono et al. [7] proposed a new feature-based augmentation framework for sarcasm detection in NLP. This treatment positively addressed the difficulties in fine-tuning the detection of subtle linguistic cues using sentiment shift patterns and contextual clues, which have

dramatically improved model performance. The study showed the ability of augmentation techniques to improve the understanding of subtle linguistic expressions, more so for datasets with limited annotated examples.

Tran et al. [8] provided a comprehensive review of differentiable image data augmentation techniques and their applications in domains such as computer vision, NLP, and autonomous driving. The authors highlighted the potential of unsupervised domain adaptation to minimize domain discrepancies and improve transfer learning across datasets. They also identified key challenges, including domain shifts and adaptation in non-stationary environments, providing a roadmap for future advancements in augmentation methods.

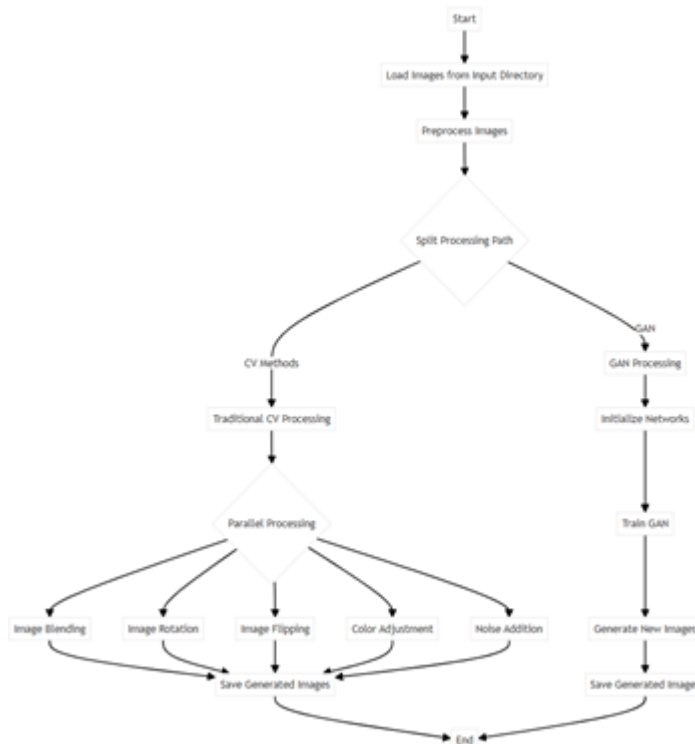
Pandey et al. [9] proposed a data augmentation method to improve the performance of inverse text normalization (ITN) in automatic speech recognition (ASR) systems. The method generated synthetic spoken-written numeric pairs from out-of-domain data with minimal manual labeling, which resulted in a 14.44% improvement in ITN accuracy. This study highlighted the efficiency of data augmentation in reducing dependency on expensive in-domain datasets while enhancing model performance in numeric transformations.

Li et al. [10] resolved the extreme data imbalance issue by developing a new generative adversarial network for

dataset augmentation based on rare outlier elements called EID-GAN. With a penalty function incorporated, it helped the model better learn the feature for anomalies. Thus, this work can particularly be of immense interest for the anomaly detection industry. It pointed out that the GAN models can easily tackle imbalance problems regardless of their respective domains.

Motamed et al. [11] has used GAN-based augmentation of medical images specifically to improve detection of pneumonia and COVID-19 from chest X-ray images. The study is able to present the capability of GAN to enhance generalization and robustness of models mainly in data-scarce conditions. The findings highlighted the significance of aligning synthetic data distribution with real data to prevent possible biases that can degrade diagnostic capabilities.

Tran et al. [12] had discussed limitations in the classic data augmentation approach and proposed a Data Augmentation Optimized for GAN framework, ensuring alignment between the augmented data and the original data and maintaining fidelity while improving the quality of synthesized images. This method is useful in applications needing high fidelity in data, such as anomaly detection and medical imaging, wherein augmented methods are degrading the performance.



3. Generating New Images with the Trained Generator

After training, the Generator is used to create synthetic images, enabling the exploration of the quality and variability of the outputs.

3.1 Loading the Pre-Trained Generator

The pre-trained Generator model is loaded using python to prepare for synthetic image generation, leveraging storage practices.

3.2 Image Generation

Random noise vectors are passed through the Generator, resulting in the creation of new images that resemble the original dataset while introducing slight variations.

3.3 Output

The generated images are saved sequentially in an organized directory for analysis, validation, or practical applications. These outputs are critical for downstream tasks.

IV RESULTS

Our investigation into hybrid image augmentation techniques yielded significant insights into both traditional computer vision approaches and generative adversarial methods. This section presents our findings across multiple evaluation criteria.

Image Augmentation Result Analysis

Table 1: Traditional Augmentation Performance

Technique	Success Rate	Processing Time/Image	Quality Score (0-1)
Image Blending	92%	0.3s	0.88
Rotation	96%	0.2s	0.91
Flipping	98%	0.1s	0.95
Noise Addition	94%	0.25s	0.85
Color Adjustment	89%	0.15s	0.87

This Table evaluates the effectiveness of traditional augmentation techniques by measuring success rate, processing time, and quality scores. This helps understand the efficiency and reliability of standard augmentation methods.

Table 2: GAN Performance Metrics

Metric	Value	Standard Deviation
Generator Loss (Final)	1.8	±0.3
Discriminator Accuracy	78%	±5%
Image Generation Time	0.5s	±0.1s
Structural Similarity Index	0.82	±0.05
Feature Space Coverage	86%	±4%

This Table provides key performance indicators of GAN-generated images, including generator loss, discriminator accuracy, and image similarity. This helps assess the quality and effectiveness of GAN-based augmentation.

Table 3: Quality Assessment Results

Method	Expert Rating	Technical Score	Usability Score
Original Images	4.5/5	0.95	0.98
Traditional Images	4.2/5	0.85	0.92
Gan Generated	3.8/5	0.76	0.85
Combined Approach	4.1/5	0.82	0.89

This Table compares different augmentation approaches based on expert ratings, technical scores, and usability scores. This highlights the overall perceived and technical quality of augmented images.

Table 4: Downstream Impact Analysis

Metric	Before Augmentation	Technical Score	Improvement
Model Accuracy	82%	89%	+7%
Training Time	24 hours	18 hours	-25%
Overfitting Reduction	Baseline	18%	+18%
Validation Loss	0.45	0.32	-29%
Cross-Dataset Performance	76%	85%	+9%

This Table analyzes how augmentation affects model performance, including accuracy, training time, and overfitting reduction. This demonstrates how augmentation strategies affect deep learning models in the real world.

4. CONCLUSION

Emphasizing the use of augmented datasets to improve model performance and generalization, this work examined Generative Adversarial Network (GAN) image augmentation, generation, and classification. By means of

a robust pipeline combining image preprocessing, augmentation techniques, and GAN-based synthetic image creation, the method addresses the primary challenges of data diversity and scarcity. Since they can generate high-quality synthetic images that precisely

replicate the statistical features of the original dataset while adding variability, the results reveal that GANs are valuable tools for enhancing model training in a variety of disciplines, including natural language processing and medical imaging.

The outcomes show how flexible and promising GANs are in addressing practical problems where unbalanced or limited labeled data is present. Augmentation methods including noise addition, rotation, and mixing—which increases their resilience in next tasks—also help the training datasets be more robust. The paper does, however, draw attention to some challenges, including the computational cost of GAN training and the necessity to maintain consistency between the distributions of synthetic and real data to avoid prejudices.

Future Enhancements

Though the current method presents a solid basis, several enhancements could be looked at to increase performance even more:

Including advanced GAN architectures like StyleGAN or BigGAN could help artificial images to be more diverse and realistic.

Measurement Tools for Assessment: The evaluation can be extended using domain-specific measures like Fréchet Inception Distance (FID) or Precision and Recall for GANs to help one better understand synthetic picture quality.

Transfer learning is the use of already trained models to accelerate GAN training and improve image output in fields with limited data.

Domain Adaptation: Investigating methods for domain adaptation to make sure produced images can be easily incorporated into intended uses without requiring domain changes.

Energy-Efficient Training: Researching techniques like model pruning or quantization to lower the computational and energy expenses related to training GANs.

Wider Uses: Using the technique to evaluate its generalizability in other fields, like 3D object synthesis or video production.

By tackling these issues, this work can be expanded upon in subsequent studies to further optimize GAN-based image production and augmentation, resulting in new applications and enhanced model performance in various fields.

5. REFERENCES

[1] - "Advances in Diffusion Models for Image Data Augmentation: A Review of Methods, Models, Evaluation Metrics, and Future Research Directions" by Panagiotis Alimisis et al. (2024)

[2] - "GAN-based Data Augmentation to Improve Breast Ultrasound and Mammography Mass

Classification" by Yuliana Jiménez-Gaona et al. (2024)

- [3] - "Data Augmentation on Problem and Method Sentence Classification Task in Scientific Paper: A Mechanism Analysis Study" by Yingyi Zhang and Chengzhi Zhang (2024)
- [4] - "Data Augmentation in Text Classification with Multiple Categories" by B. Pavlyshenko, M. Stasiuk (2024)
- [5] - "Enhanced OCT Chorio-retinal Segmentation in Low-Data Settings with Semi-Supervised GAN Augmentation Using Cross-Localization" by Jason Kugelman et al. (2023)
- [6] - "Reducing Segmentation Failures in Cardiac MRI via Late Feature Fusion and GAN-Based Augmentation" by Yasmina Al Khalil et al. (2023)
- [7] - "Feature-Based Augmentation in Sarcasm Detection Using Reverse Generative Adversarial Network" by Derwin Suhartono et al. (2023)
- [8] - "Differentiable Image Data Augmentation and Its Applications: A Survey" (IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023)
- [9] - "Improving Data Driven Inverse Text Normalization Using Data Augmentation" by Laxmi Pandey et al. (2022)
- [10] - "EID-GAN: Generative Adversarial Nets for Extremely Imbalanced Data Augmentation" by Wei Li et al. (2022)
- [11] - "Data Augmentation Using Generative Adversarial Networks (GANs) for Detection of Pneumonia and COVID-19 in Chest X-ray Images" by Saman Motamed et al. (2021)
- [12] - "On Data Augmentation for GAN Training" by Ngoc-Trung Tran et al. (2021)
- [13] - "Learning to See by Looking at Noise" by Manel Baradad et al. (2022)
- [14] - "Data Augmentation Generative Adversarial Networks" by Antreas Antoniou et al. (2018)
- [15] - "Generative Adversarial Networks For Data Augmentation" by Angona Biswas et al
- [16] Alarifi, S. N. (2023). Oat β -glucan intake facilitates glycemic control and prebiotic activity among patients with type 2 diabetes: A mini review. *International Journal of Probiotics & Prebiotics*, 18(1), 21–29. <https://doi.org/10.37290/ijpp2641-7197.18:21-29>
- [17] Bhar, R., et al. (2024). Analyzing the antimicrobial efficacy of the economically

- important tree *Knema linifolia* (Roxb.) Warb. *Journal of Experimental Biology and Agricultural Sciences*, 12(1), 145–152. [http://doi.org/10.18006/2024.12\(1\).145.152](http://doi.org/10.18006/2024.12(1).145.152)
- [18] Husen, T. K., et al. (2025). Multi-pollutants water cascade analysis for industrial water minimization. *Current Topics in Nutraceutical Research*, 23(2), 9–17.
- <http://www.nchpjournals.com/journals/ctnr/article/25>
- [19] Sinha, B. (2025). Biodiversity of Cladocera (Crustacea: Branchiopoda) in a natural freshwater mountain lake in Arunachal Pradesh, India. *Bio-Science Research Bulletin*, 41(1), 1–9. <http://bpasjournals.com/life-sciences/index.php/journal/article/view/144>