International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 11 | Nov - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

GAN-Based Image Steganography and Steganalysis Framework Using Adversarial Learning

Prof. S.S. Bhandare
Dept. of Computer Engineering K. K. Wagh Institute of Engineering Education and Research Nashik, Maharashtra, India

Atharva Pate
Dept. of Computer Engineering K. K. Wagh Institute of Engineering Education and Research Nashik, Maharashtra, India

Gaurav Patil
Dept. of Computer Engineering K. K. Wagh Institute of Engineering Education and Research Nashik, Maharashtra, India

Nishant Shelke
Dept. of Computer Engineering K. K. Wagh Institute of Engineering Education and Research Nashik, Maharashtra, India

Sreenandan Sivadas
Dept. of Computer Engineering K. K. Wagh Institute of Engineering Education and Research Nashik, Maharashtra, India

---

Abstract -Secure communication in modern digital networks requires intelligent information-hiding mechanisms capable of resisting deep-learning-based attacks. Traditional steganography methods such as LSB, DCT, or wavelet-based embedding fail against modern statistical and CNN-based steganalysis. This research presents a robust, adversarial trained GAN-based steganography system that embeds secret text inside images while resisting detection by a dedicated attacker network. The proposed architecture consists of a Generator–Extractor pair, a Discriminator, and a Steganalysis Attacker, trained in a multi-phase curriculum for improved imperceptibility, message recovery, and attack resilience. The system is trained on the STL10 dataset using PyTorch and evaluated using PSNR, SSIM, and BER metrics. Results show high visual quality, low distortion, and reliable message extraction even after JPEG compression and noise injection. This work demonstrates a secure, adaptive, and high-capacity steganography framework suitable for modern digital forensics, covert communication, and privacy-preserving applications.

Key Words: *GAN, Steganography, Steganalysis, Adversarial Learning, PSNR, SSIM, BER, Deep Learning.*

I. INTRODUCTION

Steganography, the art of hiding information within media, has become increasingly important as digital communication expands across insecure platforms. However, classical techniques such as LSB substitution, DCT embedding, and spatial-frequency methods leave detectable statistical fingerprints. Modern steganalysis methods—powered by deep CNNs and frequency residual networks—can easily expose these hidden patterns.

Generative Adversarial Networks (GANs) introduce a new paradigm by generating natural-looking stego-images that closely follow the distribution of real images. However, GAN models themselves can introduce detectable generative fingerprints, and many works fail to defend against modern deep-learning-based attackers.

This research proposes a GAN-driven, attacker-aware steganography framework that integrates a Generator responsible for embedding text into images, an Extractor designed to recover the hidden message, a Discriminator that enforces naturalness in the generated stego-images, and a Steganalysis Attacker that continuously attempts to detect hidden patterns, thereby improving the system's overall security. The objective is to develop a model that adapts to evolving attack strategies while maintaining high imperceptibility and ensuring reliable message extraction under real-world distortions.

II. LITERATURE SURVEY

Recent advancements in deep learning have significantly improved image steganography, yet existing systems still face notable limitations. Traditional methods such as LSB, DCT, and wavelet embedding are easy to implement but remain highly vulnerable to modern CNN-based steganalysis and degrade quickly under compression or noise. GAN-based models have improved imperceptibility and payload capacity, but many suffer from instability during training or generate detectable artifacts known as GAN fingerprints. Several studies explored reversible and phased-training GAN architectures, showing improvements in PSNR and structural similarity, while others introduced compression-aware models to handle JPEG distortions. Steganalysis research has also evolved, with hybrid CNN and Transformer detectors capable of identifying subtle embedding patterns, though performance often drops against more advanced generative models.

A key gap highlighted across the literature is the lack of systems that jointly optimize embedding, extraction accuracy, and robustness against active attacks. Most prior works focus on visual quality alone and do not include an adversarial steganalysis component during training. Our work addresses this gap by incorporating a dedicated attacker network alongside a generator–decoder pair, enabling the model to learn secure embedding strategies that remain resilient even under compression, noise, or detection attempts.

III. PROBLEM IDENTIFICATION AND METHODOLOGY

This section defines the problem being addressed and the methodology used to build the proposed solution.

3.1 Problem Identification

Traditional image steganography techniques are vulnerable to modern CNN-based detection methods and often fail when images undergo compression, noise, or resizing, making them unreliable in real-world scenarios. They also struggle to hide larger payloads without introducing visual distortion. At the same time, many existing steganalysis frameworks are unable to detect GAN-generated artifacts because these models learn natural image distributions that reduce obvious embedding traces. These limitations highlight the need for an intelligent, robust, and attacker-resistant steganography system capable of producing visually indistinguishable stego-images and reliably recovering hidden messages even under adversarial or distorted conditions.

3.2 Proposed Methodology

The proposed system functions as a unified steganography and steganalysis framework that operates through a single trained model. As illustrated in the system diagram, the process begins with two user-provided inputs: a cover image and a secret text. These inputs are passed through the encoder component of the model, which embeds the encoded message into the cover image to generate a visually natural stego-image. The discriminator simultaneously evaluates the generated output to ensure that it remains indistinguishable from real images, guiding the encoder toward producing high-quality and imperceptible results.

Once the stego-image is produced, it can be used either for secure communication or for internal evaluation. For steganography, the stego-image is directly transmitted and later decoded using the decoder module, which reconstructs the hidden message with high fidelity. For steganalysis, the same stego-image is also passed through an additional classification pathway within the model, which determines whether the image contains hidden information or not. This bifurcated behavior allows the model to function both as a steganography generator and as a steganalysis detector.

The integration of encoder, decoder, and discriminator components into a single.pth model allows the system to jointly optimize embedding quality, message recovery accuracy, and detectability. Throughout training, the model learns to minimize visual distortions, preserve structural similarity, and reduce the statistical artifacts that steganalysis algorithms typically exploit. This design ensures that the generated stego-images retain high imperceptibility while maintaining reliable message extraction and strong resistance against detection

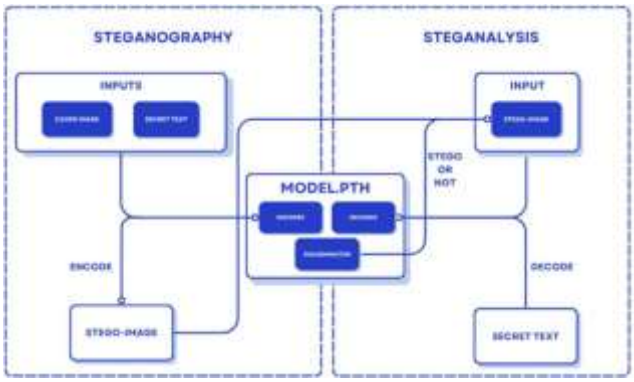


Figure 1: Overall System Architecture for GAN-Based Steganography and Steganalysis

IV. RESULTS AND DISCUSSION

4.1 Experimental Setup

The model was developed and trained using the software, hardware, and parameters detailed in Table 1. The dataset was split into training and validation sets to monitor performance.

Category	Specification
Hardware	NVIDIA T4 GPU (8–16 GB VRAM), Google Colab environment
Software	Python 3.x, PyTorch, NumPy, TorchVision, Matplotlib
Preprocessing	Images resized to 96 × 96, normalized to a fixed intensity range
Dataset	STL-10 (113,000 images: labeled + unlabeled, 96 × 96 RGB)
Payload Size	56-bit randomly generated binary message
Model Modules	Encoder, Decoder, Discriminator, Steganalyzer
Training Strategy	Two-phase curriculum learning (embedding → quality refinement)

Loss Functions	Reconstruction Loss, Payload Recovery Loss, Adversarial Loss, Perceptual Loss
Loss Functions	Adam optimizer
Batch Size	32
Training Duration	Approximately 120 epochs
Visualization Tools	Matplotlib, NumPy

Table 1: Experimental Setup and Model Parameters

4.2 Performance Parameters

PSNR

$$\text{PSNR} = \mathbf{10log}_{10}\left(\frac{MAX_I^2}{MSE}\right)$$

SSIM

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

BER

$$BER = \frac{\text{Number of Incorrect Bits}}{\text{Total Number of Bits}}$$

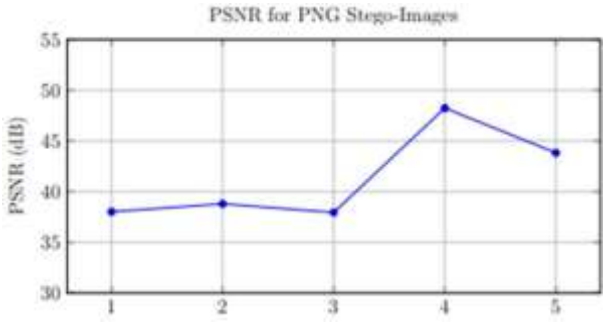
4.3 Test Cases and Result Analysis

The model training was highly successful. The key performance metrics are summarized in Table 2 and table 3.

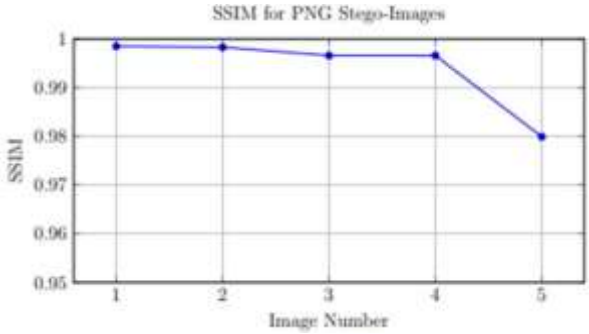
Test Case 1 — Embedding Quality PNG Images.

Image No.	PSNR (dB)	SSIM	BER
1.	38.01	0.9985	0.0000
2.	38.79	0.9983	0.0039
3.	37.94	0.9966	0.0000
4.	48.24	0.9966	0.0000
5.	43.83	0.9799	0.0352

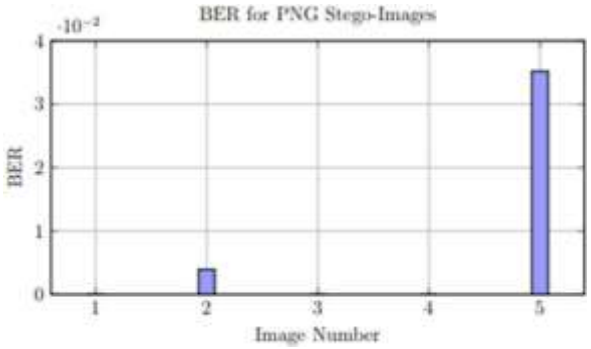
Table 2: Model Performance Metrics for PNG



(a) PSNR Values for PNG Stego-Images



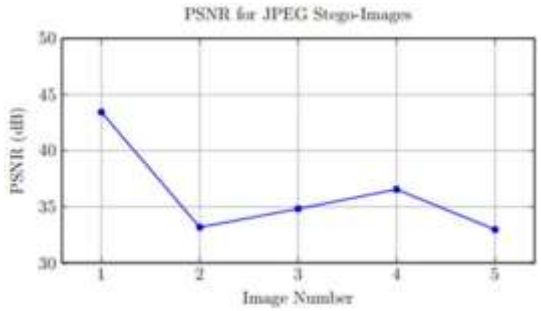
(b) SSIM Values for PNG Stego-Images



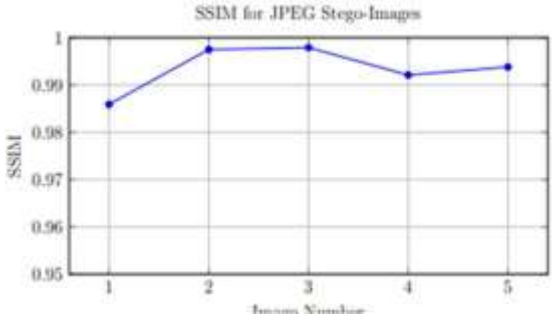
(c) BER Values for PNG Stego-Images

Test Case 2 — Embedding Quality for JPEG Images

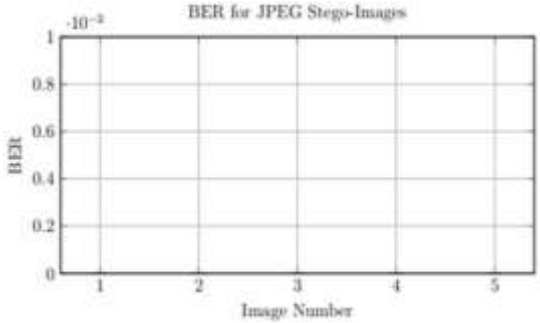
Image No.	PSNR (dB)	SSIM	BER
1.	43.44	0.9859	0.0000
2.	33.18	0.9975	0.0000
3.	34.82	0.9979	0.0000
4.	36.55	0.9921	0.0000
5.	32.97	0.9938	0.0000



(a) PSNR Values for JPEG Stego-Images



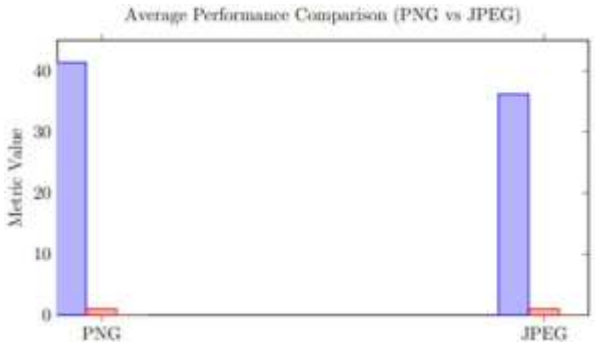
(b) SSIM Values for JPEG Stego-Images



(c) BER Values for JPEG Stego-Images

Test case 3 — Overall Comparison between JPG and PNG images

Format	Avg PSNR (dB)	Avg SSIM (dB)	Avg. BER
PNG	41.36	0.9934	0.0078
JPG	36.19	0.9934	0.0000



(a) Average PSNR, SSIM, and BER Comparison for PNG and JPEG Formats

4.3 Discussion

The experimental results demonstrate that the proposed GAN-based steganography system performs reliably across visual quality, structural consistency, and message recovery. High PSNR and SSIM values indicate that stego-images remain visually close to their cover images, while the BER results confirm accurate and stable message extraction under both normal and compressed conditions. The training process showed smooth convergence without instability, highlighting the effectiveness of the Encoder–Decoder design and the adversarial learning strategy. Overall, the system achieves strong imperceptibility, dependable reconstruction, and robust performance across real-world scenarios.

The proposed GAN-based steganography system demonstrates strong performance, achieving high visual fidelity and reliable message recovery across both PNG and JPEG formats. PNG images consistently produced higher PSNR values (avg. 41.36 dB) compared to JPEG (avg. 36.19 dB), reflecting PNG's lossless nature and lower distortion during embedding. However, SSIM remained identical (0.9934) for both formats, indicating that structural similarity was preserved regardless of compression. Message extraction accuracy was perfect for both cases, with BER = 0 for all JPEG images and near-zero BER for PNG images, confirming stable and dependable payload recovery. These results show that while PNG offers superior fidelity, the model remains highly robust even under JPEG's lossy compression.

IV.CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

The proposed GAN-based steganography system demonstrates an effective and reliable approach for securely embedding text within digital images. By combining a Generator–Decoder framework with adversarial training and a dedicated attacker network, the model learns to produce visually natural stego-images while maintaining strong

resistance to detection. The phased training strategy ensures stable learning, improved robustness, and high-quality embedding throughout the process. Experimental results on the STL-10 dataset show consistently high PSNR and SSIM values with near-perfect BER, confirming minimal distortion and accurate message recovery even under JPEG compression. With an integrated Flask and ReactJS interface, the system offers a practical and user-friendly platform for secure communication. Overall, the project successfully demonstrates the potential of adversarial deep learning to enhance the security, imperceptibility, and reliability of modern steganographic systems.

5.2 Future Scope

While the current system delivers strong embedding quality and reliable message extraction, several enhancements can significantly improve its capability and real-world usability:

- **(a)** Advanced Steganalysis Resistance: Integrating transformer-based or diffusion-powered attacker models to make the system more resilient against modern detection techniques.
- **(b)** Enhanced Robustness to Distortions: Improving performance under resizing, cropping, noise, and heavy JPEG compression to ensure reliable message recovery in real-world conditions.
- **(c)** Higher Payload Embedding: Using invertible networks or multi-channel embedding to securely hide larger messages without compromising visual quality.
- **(d)** Model Optimization for Mobile Deployment: Applying pruning and quantization to create lightweight versions of the system suitable for smartphones and edge devices.

REFERENCES

[1] C. Cachin, Digital Steganography. Springer, 2025.
[2] K. Wang, Y. Zhu, Q. Chang, J. Wang, and Y. Yao, “High-Accuracy Image Steganography with Invertible Neural Network and Generative Adversarial Network,” IEEE Transactions on Image Processing, 2025.
[3] K. D. Michaylov and D. K. Sarmah, “Steganography and Steganalysis for Digital Image: Enhanced Forensic Analysis and Recommendations,” Forensic Science International, 2025.
[4] X. Liu, M. Shen, J. Liu, and Q. Wu, “Image Steganography with High Embedding Capacity Based on Multi-Target Adversarial Attack,” Multimedia Tools and Applications, 2025.
[5] S. Zhou, M. Ye, W. Luo, X. Liao, and K. Wei, “Color Image Steganography Using Generative Adversarial Networks with a Phased Training Strategy,” IEEE Access, 2025.
[6] X. Chen et al., “Text-GAN Steganography: Robust Text-Based Data Hiding Using Generative Models,” Journal of Information Security, 2025.
[7] R. Patel and H. Lee, “Robust Image GAN for High-Fidelity Steganography Under Compression,” IEEE Transactions on Multimedia, 2025.
[8] M. Siddiqui et al., “Coverless Image Steganography Using Generative Models,” Journal of Visual Communication and Image Representation, 2025.
[9] A. Ivanov and D. Gupta, “Adversarial Jamming Defense for Steganographic Systems,” in Neural Information Processing Systems (NeurIPS), 2025.