The Effect of Hyperparameters on Quality of Synthesized Medical Images

Alexandra-Georgiana Andrei
AI Multimedia Lab
National University of Science and Technology
POLITEHNICA Bucharest
Bucharest, Romania
alexandra.andrei@upb.ro

Bogdan Ionescu
AI Multimedia Lab
National University of Science and Technology
POLITEHNICA Bucharest
Bucharest, Romania
bogdan.ionescu@upb.ro

***Abstract*—Generative Adversarial Networks have shown great promise in generating high-quality synthetic medical images, but training them remains challenging due to instabilities such as mode collapse and training collapse. In this study, we systematically investigate the impact of key hyperparameters on the performance of a Deep Convolutional GAN trained to synthesize realistic computed tomography images. Specifically, we explore variations in the optimizer parameters (β_1, β_2), learning rate, and latent dimension to understand their effects on training stability and image quality. The quality of the generated images is evaluated using the Fréchet Inception Distance (FID). Our results show that GAN performance is highly sensitive to hyperparameter settings: lower learning rates and higher β_2 values generally lead to more stable training and better image fidelity, while poor choices often result in mode collapse. The best performance is achieved with $\beta_1 = 0.5$, $\beta_2 = 0.999$, a learning rate of 0.0001, and latent dimension of 50, resulting in the lowest FID score. These findings provide practical guidance for tuning GAN hyperparameters to improve training stability and image quality in medical image synthesis tasks.**

***Index Terms*—Generative Adversarial Networks, data augmentation, medical imaging, CT imae generation**

I. INTRODUCTION

Generative Adversarial Networks (GANs) have emerged as a powerful framework for generating synthetic images, including in medical imaging applications where high-quality, realistic images can have significant value for tasks such as data augmentation, anonymization, and algorithm development. Despite their promise, GANs remain challenging to train due to fundamental issues inherent in their adversarial two-network structure, as we briefly discussed in [1]. These challenges include problems common to neural networks in general, such as training instability, vanishing gradients, and non-convergence, as well as GAN-specific difficulties like mode collapse and training collapse.

Mode collapse occurs when the generator produces a limited variety of outputs, failing to reflect the diversity present in the training dataset. As a result, the GAN may generate visually plausible but overly repetitive images that do not capture the full data distribution. Training collapse, on the other hand, refers to a situation where the GAN fails entirely, generating unrealistic, aberrant, or noisy outputs as training deteriorates or diverges. Both phenomena undermine the potential of GANs to generate clinically meaningful and diverse synthetic medical images.

Various approaches have been proposed to address these challenges, including architectural modifications, improved loss functions, and careful selection of training strategies. One critical aspect that can influence GAN performance is the choice of training hyperparameters settings that govern how the networks learn during optimization but are not themselves learned from data. These include the learning rates of the generator and discriminator, optimizer momentum terms such as β_1 and β_2, and the size of the latent input space (latent dimension), which determines the variability and expressiveness of the generated images.

In this study, we conduct a focused investigation on how key hyperparameters affect the performance and output quality of a Deep Convolutional Generative Adversarial Network (DC-GAN) trained to synthesize realistic computed tomography (CT) images. By systematically varying β_1, β_2, the latent dimension size, and learning rates for both generator and discriminator, we aim to identify hyperparameter configurations that promote stable training and improve image fidelity and diversity. Our findings provide insights into optimizing GAN training for medical image synthesis and contribute to the broader understanding of how hyperparameter tuning can mitigate common GAN failures such as mode collapse and training collapse.

II. METHODOLOGY AND IMPLEMENTATION

A. Generative Adversarial Network

Our model, depicted in Figure 1 is built upon the Deep Convolutional Genreative Adversarial Network [2] with several alterations. The generator consists of 8 deconvolutional layers with a kernel size of 4, the number of filters and strides vary from one layer to another as presented in Figure 1. Each deconvolutional layer is followed by a BatchNormalization layer and a LeakyReLU activation function with a slope of 0.2 and the last convolutional layer uses Tanh activation function. The discriminator consists of 5 convolutional layers, each followed by a LeakyReLU activation layer. The model is trained using Adam optimizer, the learning rates, β_1, β_2 and

©2025 Copyright for this paper belongs to his authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2nd Workshop on Artificial Intelligence for Multimedia,
AI Multimedia Lab, CAMPUS Research Institute, POLITEHNICA Bucharest,
Bucharest, Romania, 15 July, 2025

the latent vector dimension have been varied and combined as depicted in Table I.

B. Evaluation

We used the Fréchet Inception Distance [3] score to evaluate the generated images. FID is a metric used to evaluate GANs performance by measuring the similarity in quality and diversity between real and generated. A lower FID score indicates that the images in the training and generated sets are more similar. In our analysis, FIDs are calculated using pytorch-fid library [4]. The FID between two sets *r* and *g* is calculated as in equation (3) where μ_r and μ_g are the means and Σ_r and Σ_g are distributions of the two sets.

$$FID = \|\mu_r - \mu_g\|_2^2 + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}}) \quad (1)$$

III. Results

Table I summarizes the hyperparameter configurations tested and the results in terms of FID score or observed mode collapse. Figure 2 illustrates the FID scores for a subset of tested hyperparameter configurations. The experiments reveal several important trends regarding how these hyperparameters influence the training stability and quality of synthesized CT images.

First, the learning rate had a strong effect on training stability. Configurations with a learning rate of 0.0001 frequently resulted in successful training and reasonable FID values, whereas increasing the learning rate to 0.0002 consistently led to mode collapse regardless of other settings. For example, experiment 3 with a learning rate of 0.0002 collapsed even with otherwise moderate settings (β_1 = 0.5, β_2 = 0.9, latent dimension = 500).

The momentum term β_2 also seem to have an important impact in GAN training. Configurations with β_2 = 0.9 achieved moderate FID values (e.g., experiment 2: FID = 101.28) but became unstable at higher learning rates. When β_2 was increased to 0.99, performance decreased, with frequent mode collapse even at low learning rates (experiments 4-8). Notably, a smaller subset of configurations with β_2 = 0.99 and learning rate 0.0001 succeeded (experiments 9-11), but their FID scores varied considerably, ranging from 101.31 to 138.39.

The best performance was observed when β_2 was set to 0.999, suggesting that stronger momentum may help stabilize GAN training under certain conditions. In particular, experiment 16 (β_1 = 0.5, β_2 = 0.999, learning rate = 0.0001, latent dimension = 50) achieved the lowest FID of 98.48, indicating improved image quality. However, this result did not generalize consistently across latent dimensions: experiment 17 (same $\beta 1$, $\beta 2$, learning rate but latent dimension = 100) yielded a much higher FID of 302.82, pointing to complex interactions between β_2, latent dimension size, and learning rate.

The latent dimension itself influenced results in a nuanced way. While a latent dimension of 50 produced the best FID when combined with optimal optimizer settings, small latent dimensions did not universally prevent collapse, and larger latent spaces (e.g., 1000) showed mixed results depending on the other parameters. For instance, experiment 2 with latent dimension 1000 achieved a relatively good FID of 101.28, but experiment 19 (same latent dimension, β_2 = 0.999) resulted in a higher FID of 122.82.

TABLE I: Summary of tested hyperparameter combinations, their effects on training stability, and FID results.

#	β_1	β_2	learning rate	Latent dim.	FID / Remarks
1	0.5	0.9	0.0001	50	129.50515
2	0.5	0.9	0.0001	1000	101.27544
3	0.5	0.9	0.0002	500	Mode collapse
4	0.5	0.99	0.00001	50	Mode collapse
5	0.5	0.99	0.00001	100	Mode collapse
6	0.5	0.99	0.00001	500	Mode collapse
7	0.5	0.99	0.00001	1000	Mode collapse
8	0.5	0.99	0.0001	50	Mode collapse
9	0.5	0.99	0.0001	100	101.31832
10	0.5	0.99	0.0001	500	138.39016
11	0.5	0.99	0.0001	1000	121.73618
12	0.5	0.99	0.0002	50	Mode collapse
13	0.5	0.99	0.0002	100	Mode collapse
14	0.5	0.99	0.0002	500	Mode collapse
15	0.5	0.99	0.0002	1000	Mode collapse
16	0.5	0.999	0.0001	50	98.48011
17	0.5	0.999	0.0001	100	302.81628
18	0.5	0.999	0.0001	500	186.7802
19	0.5	0.999	0.0001	1000	122.8189
20	0.5	0.999	0.0002	50	Mode collapse
21	0.5	0.999	0.0002	100	Mode collapse
22	0.5	0.999	0.0002	500	Mode collapse
23	0.5	0.999	0.0002	1000	Mode collapse
24	0.9	0.99	0.00001	100	Mode collapse

Finally, increasing β_1 to 0.9 (experiment 24) resulted in mode collapse, suggesting that deviating from the typical $\beta 1$ = 0.5 setting is not beneficial for this application.

Overall, these results highlight the strong sensitivity of GAN training to the interplay between optimizer hyperparameters, learning rate, and latent space dimensionality. The configuration yielding the lowest FID was β_1 = 0.5, β_2 = 0.999, learning rate = 0.0001, and latent dimension = 50, indicating that careful hyperparameter tuning is essential for achieving high-quality CT image synthesis and avoiding common failure modes such as mode collapse.

The Appendix provides four representative samples for each tested hyperparameter configuration to illustrate the qualitative differences in the generated CT images.

IV. Conclusions

In this work, we systematically investigated how key hyperparameters, including β_1, β_2, learning rate, and latent dimension, affect the training stability and image quality of a Deep Convolutional GAN (DCGAN) used for synthesizing CT images. Our quantitative evaluation, based on the Fréchet Inception Distance (FID), demonstrated that GAN performance is highly sensitive to these hyperparameters. The best FID score was achieved with $\beta_1 = 0.5$, $\beta_2 = 0.999$, a learning rate of 0.0001, and a latent dimension of 50 (setup #16).

However, qualitative analysis of the generated images, as illustrated in the Appendix, suggests that setups #9 and #10 produce visually superior results despite having higher FID scores. This highlights that FID alone may not fully capture

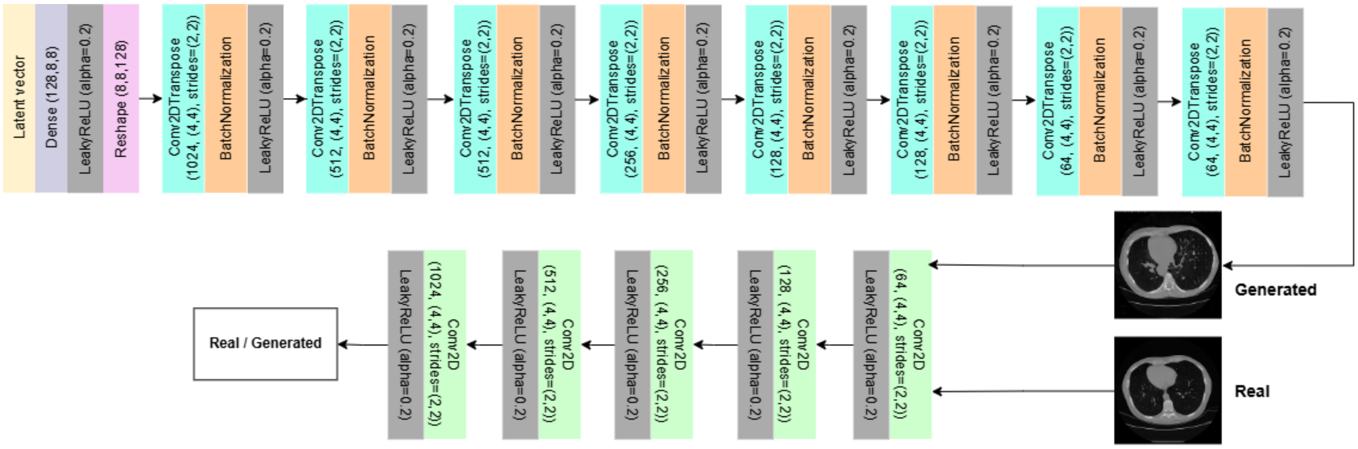

Fig. 1: Overview of the studied Generative Adversarial Network.

visual quality, especially for medical images where structural fidelity is crucial.

These findings underline the importance of careful hyperparameter tuning and suggest that a combined quantitative and qualitative evaluation is essential when assessing GAN-based medical image synthesis. In future work, we will further investigate these observations by exploring alternative evaluation methods that better reflect perceptual and clinical relevance, as well as by extending our analysis to other generative architectures and medical imaging modalities.

REFERENCES

[1] A.-G. Andrei and B. Ionescu, “Generative adversarial networks (gans) in medical imagining,” 2024.

[2] M. S. C. A. Radford, L, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv*, 2015.

[3] T. U. B. N. S. H. M. Heusel, H. Ramsauer, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” in *Advances in Neural Information Processing Systems*, 2017.

[4] M. Seitzer, “pytorch-fid: FID Score for PyTorch,” https://github.com/mseitzer/pytorch-fid, August 2020, version 0.2.1.

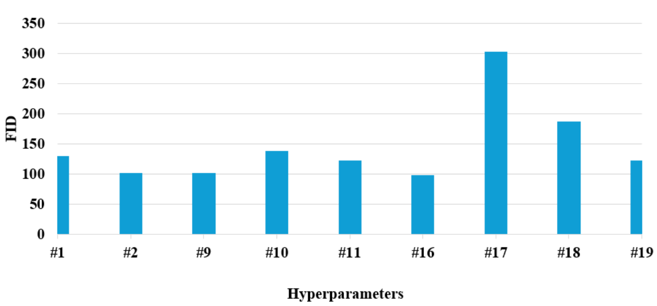

Fig. 2: FID scores across tested hyperparameter configurations

V. Appendix

Fig. 3: Generated images for hyperparameter configurations #1 (first four) and #2 (last four).

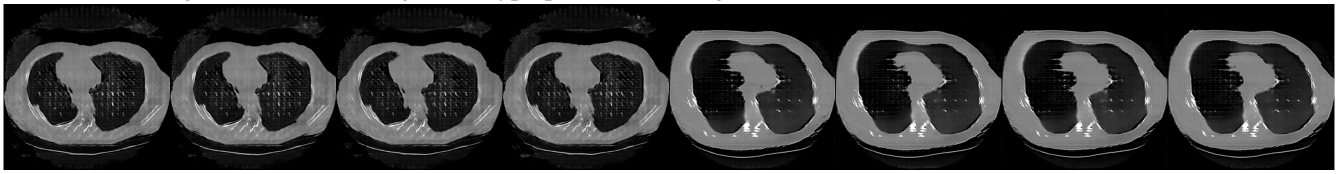

Fig. 4: Generated images for hyperparameter configurations #3 (first four) and #4 (last four).

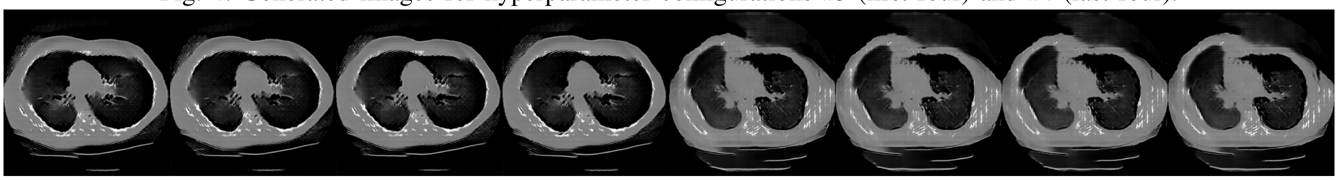

Fig. 5: Generated images for hyperparameter configurations #5 (first four) and #6 (last four).

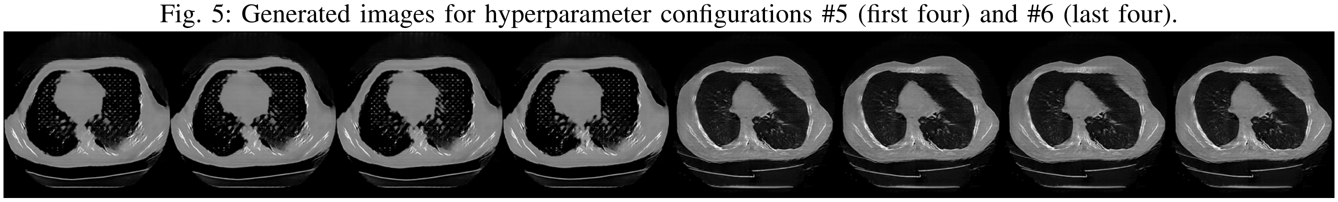

Fig. 6: Generated images for hyperparameter configurations #7 (first four) and #8 (last four).

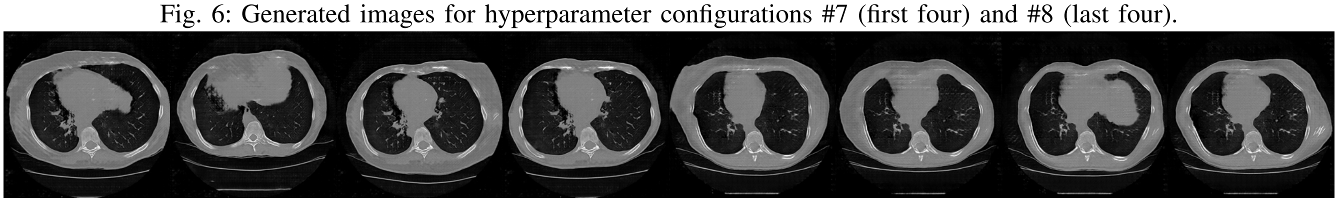

Fig. 7: Generated images for hyperparameter configurations #9 (first four) and #10 (last four).

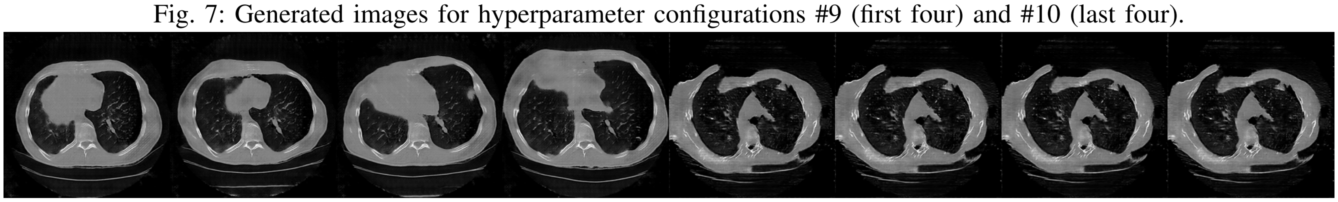

Fig. 8: Generated images for hyperparameter configurations #11 (first four) and #12 (last four).

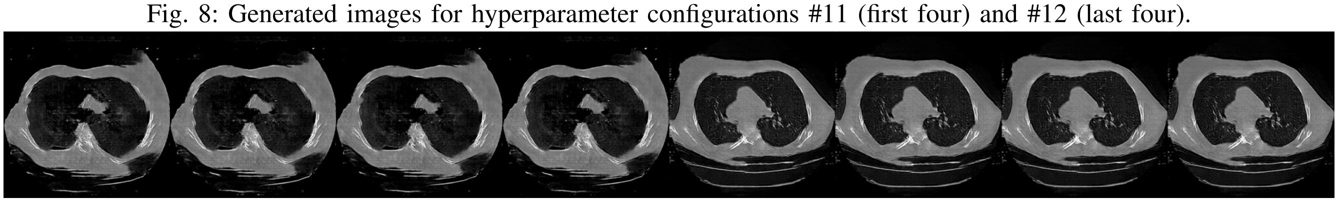

Fig. 9: Generated images for hyperparameter configurations #13 (first four) and #14 (last four).

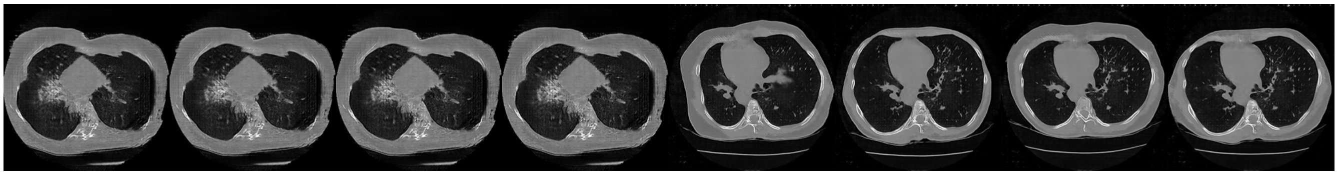

Fig. 10: Generated images for hyperparameter configurations #15 (first four) and #16 (last four).

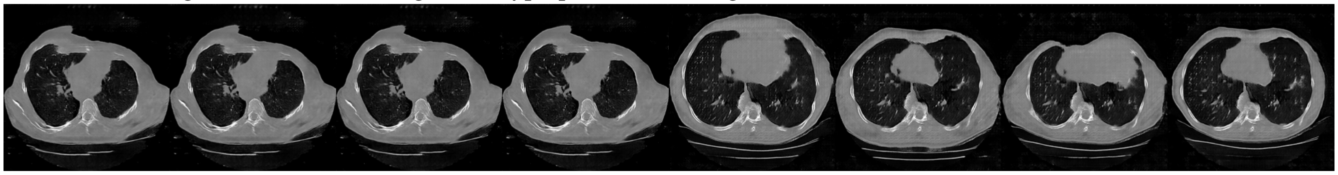

Fig. 11: Generated images for hyperparameter configurations #17 (first four) and #18 (last four).

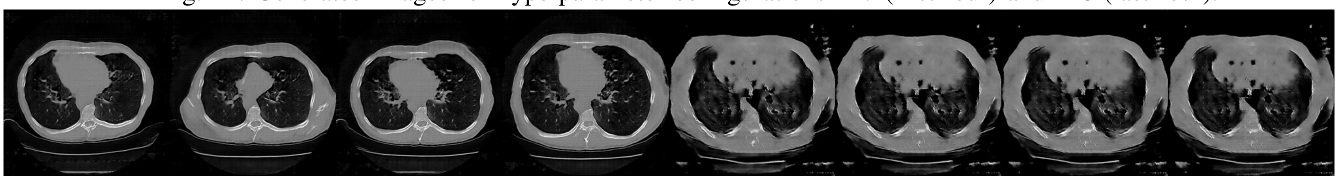

Fig. 12: Generated images for hyperparameter configurations #19 (first four) and #20 (last four).

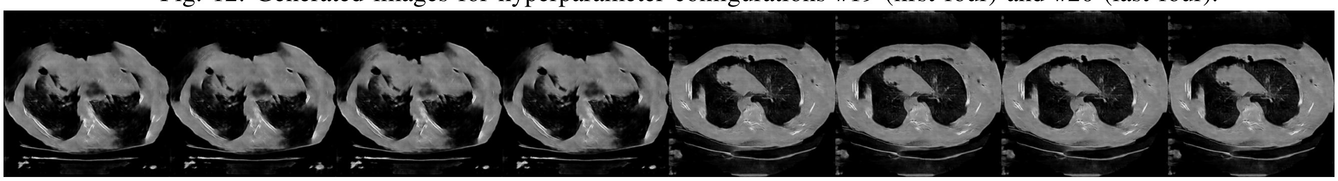

Fig. 13: Generated images for hyperparameter configurations #21 (first four) and #22 (last four).

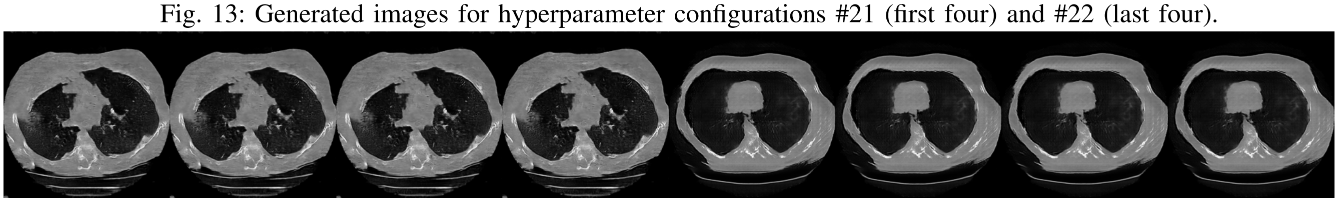

Fig. 14: Generated images for hyperparameter configurations #23 (first four) and #24 (last four).