www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 16(1), Jan-April 2025, 266-274. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2025.v16i1.1058

MASTER'S THESIS SUMMARY
Generating medical images of breast cancer using generative artificial intelligence

Student: *Yaneth Reyes-Hernández*
Polytechnical University of Pachuca (UPP)
Yane.rehyes@gmail.com

Advisor: *Jorge A. Ruiz-Vanoye*
Polytechnical University of Pachuca (UPP)
jorgeruiz@upp.edu.mx

Advisor: *Jazmin Rodríguez-Flores*
Polytechnical University of Pachuca (UPP)
jrodriguez@upp.edu.mx

Abstract. In this research work we review the fact that although there are different sets of databases with medical images of mammograms of breast cancer, these data are not really sufficient for the training of artificial intelligence systems to find signs of breast cancer in the images, so we propose the creation of new images that help to complement the data sets allowing better training in the systems that help the diagnosis of this disease.
Keywords: generative artificial intelligence, generative antagonistic networks, medical imaging, repository, data set, data set

Article Info
Received February 20, 2025
Accepted February 22, 2025

1 Introduction

Breast cancer is one of the most common health problems that have a great negative impact on public health worldwide. AI has become a very useful tool in healthcare, in the context of medical imaging aiding cancer diagnosis.

Although there are different datasets containing mammogram images, these are limited data sets because they are images that may contain sensitive information about the patients whose images were taken. This has caused that diagnostic systems do not have the necessary information to be trained to perform an efficient diagnosis.

In this research it is proposed to generate synthetic images using generative artificial intelligence models, with which good quality images are obtained and that allow to increase the repositories, allowing its use in school and medical research.

1.1 Motivations

Breast cancer is a disease with a significant impact. Arceo-Martínez et al. (2021) mention that breast cancer is recorded as the most common type of cancer worldwide, with an incidence of over 45 cases per 100,000 inhabitants, with the female population being the most affected.

With scientific and technological advancements and the rise of artificial intelligence, it is now possible to use AI systems to recognize patterns in medical images for the detection of this disease.

Currently, mammography repositories are available, and although they are highly beneficial for developing predictive systems for this disease, these systems require a vast amount of images to achieve better results.

Aguirre et al. (2021), in their article on AI applied to medical imaging, state that the main objectives of artificial intelligence range from assisting radiologists, optimizing images, recognizing structures, segmenting lesions, and even transcribing reports.

These objectives share a significant potential impact on the field of AI and, consequently, on medicine, where the role of radiology is becoming increasingly crucial.

Bastidas (2021) explains that one of the key characteristics of artificial intelligence is that systems require data to "learn" and act intelligently after being trained. He notes that, for certain applications, a few dozen data points are typically used for proof of concept, but as soon as possible, the system is scaled to use hundreds, thousands, or even tens of thousands of data points if available, to calibrate the final system.

Due to the rise of artificial intelligence, there is a growing interest in applying models that can assist in diagnosing diseases in the healthcare field. However, one of the main challenges is bias in these systems, as achieving better training requires a large amount of data.

1.2 Description of the research problem

Freire Hidalgo (2021) states that there are currently various radiological screening techniques used for the diagnosis and early detection of breast cancer. Among the most commonly used is mammography due to its low cost and minimal radiation exposure for the patient. However, these techniques require manual interpretation by radiologists, which, due to several factors such as fatigue, noise, image contrast, morphological characteristics of the breast, and an excessive number of images, can lead to diagnostic errors.

In their research, Vega et al. (2020) explain that in recent years, with the growing availability of electronically stored clinical information, the medical field has become an ideal environment for the development and application of new technologies. Machine learning has the potential to improve healthcare systems by analyzing millions of clinical data points, enabling the creation of predictive, screening, and diagnostic models.

An essential aspect of breast cancer detection is the analysis of clinical images. In the research and implementation by Díaz et al. (2021), the lack of large public databases is discussed as a significant limitation to the clinical application of AI in medical imaging, despite the vast amount of information available in current Picture Archiving and Communication Systems (PACS).

Jordon et al. (2022) mention that many large-scale datasets are sensitive and sharing them may violate fundamental rights protected by modern privacy regulations. As a result, high-dimensional datasets, which are often scarce, are inherently vulnerable to privacy attacks, and existing anonymization techniques do not provide adequate protection. This limitation restricts the ability to share large datasets, creating a bottleneck in the development and implementation of data science and machine learning methods.

As noted by Mendizábal & Montoya (2024), with digitalization, binary codes translated some reference of reality. However, with generative AI, images are now the product of algorithms. This means that images can be self-generated; by passing the reality they once aimed to reflect, extracting their essence, and instead being created solely through algorithmic combinations, leading to self-referential productions.

Pinaya et al. (2022) highlights that the generation of synthetic data provides a promising alternative, allowing the augmentation of training datasets and facilitating large-scale medical imaging research.

Although there are various datasets with medical images related to breast cancer, they are relatively small. Significant progress has been made in detecting this disease using machine learning techniques. However, the effectiveness of these models may be limited by the quantity and quality of the data available for training, which could result in biased and less robust models.

1.3 Objectives of the thesis

The objectives of this thesis describe what is intended to be achieved in this research.

1.3.1. General Objective

Apply a generative artificial intelligence model through programming and training techniques to expand existing repositories of mammograms with breast cancer.

1.3.2. Specific Objectives

- Select generative artificial intelligence techniques through research to implement a generative model.
- Generate a synthetic dataset that closely resembles the original set using artificial intelligence techniques to augment the existing data in a repository.
- Evaluate the generated medical images using qualitative and quantitative metrics to determine their quality.

1.4 Brief description of the contribution of the thesis

This research paper proposes an alternative that addresses the lack of medical images related to breast cancer. These datasets are limited by the scarcity of images, and it is very difficult to have access to large repositories. Therefore, we propose the use of generative artificial intelligence to create new images from existing ones, in order to increase the image repositories, favoring the training and efficiency of systems that help in the detection of breast cancer.

2 Background

Table 1 shows that generative artificial intelligence began in the 1960s with the arrival of the chatbot named Eliza. According to Decide4AI (2023), there were no significant advances for several decades, leading to what was known as the "AI winter." However, major breakthroughs emerged in 2012 with deep networks and later in 2014 with the introduction of GANs.

Table 1: Background of Generative AI (Lawton, 2023)

Date	Author	Contribution
1966	MIT Professor Joseph Weizenbaum	Created the first chatbot, Eliza, which simulates conversations with a psychotherapist.
2014	Research scientist Ian Goodfellow	Developed Generative Adversarial Networks (GANs), which pit two neural networks against each other to generate increasingly realistic content.
2014	Diederik Kingma and Max Welling	Introduced Variational Autoencoders (VAEs) for generating images, videos, and text.
2015	Stanford researchers	Published a study on diffusion models in the paper Unsupervised Deep Teamwork via Nonequilibrium Thermodynamics. This technique provides a way to reverse-engineer the process of adding noise to a final image.
2017	Google researchers	Developed the concept of Transformers in the seminal paper Attention Is All You Need, inspiring further research on tools that could automatically analyze unlabeled text in large language models (LLMs).
2018	OpenAI	Launched GPT (Generative Pre-trained Transformer). Trained on approximately 40 gigabytes of data and consisting of 117 million parameters, GPT paved the way for future LLMs in content generation, chatbots, and language translation.
2021	OpenAl	Introduced DALL·E, which can generate images from text prompts. The name is a combination of WALL·E, the name of a fictional robot, and the artist Salvador Dalí.
2022	Researchers from Runway Research, Stability AI, and CompVis LMU	Released Stable Diffusion as open-source code, allowing automatic image content generation from text prompts.

2.1 Important Issue 1

Medical datasets become limited, thus contributing to low efficiency in models that aid in disease detection.

In Shoshan et al. (2023) an attempt is made to show that when data is sparse, synthetic data can be used to increase the available repository, allowing AI models to be trained with a larger dataset. A new method for calibrating the synthetic error estimate to match that of the real domain is also presented.

Similarly, Yang et al. (2023) explores the innovative concept of leveraging AI-generated images as new data sources, reshaping traditional modeling paradigms in visual intelligence. It is also mentioned that unlike real data, AI-generated data show remarkable advantages, including unmatched abundance and scalability, rapid generation of vast datasets, and effortless simulation. Thus, the potential of its generated data is explored in a range of applications, from training machine learning models to simulating scenarios for computational modeling, testing and validation.

2.2 Major Theme 2

It is difficult to have access to large datasets due to the sensitivity of these datasets as they contain important patient information. With the advent of GANs, the generation of high quality images has been opened up, making it possible to increase repositories.

In McNulty et al. (2024), a reusable and open-source synthetic image generation process, the GAN image synthesis tool (GIST), was developed that is easy to use and implement. The process helps to improve and standardize AI algorithms in digital health by generating high-quality synthetic image data that are not tied to specific patients.

2.3 Related work

Table 2 details some of the related works found that are relevant to the proposed project, contributing to a better understanding of the problem addressed.

Table 2: Related Works

Project Name	Author	Description
Synthesis of 2D medical images using a probabilistic diffusion model with transformer-based noise removal	(Pan et al., 2023)	A medical image synthesis framework is presented to address the challenge of limited training datasets for AI models.
Alignment of synthetic medical images with clinical knowledge using human feedback	(Sun et al., 2023)	Introduces a framework with a pathologist in the loop to generate clinically plausible synthetic medical images.
Medical Diffusion: Probabilistic diffusion models with noise removal for 3D medical image generation	(Khader et al., 2022)	Demonstrates that probabilistic diffusion models can synthesize high-quality medical image data, validated for MRI and CT images.
On the usability of synthetic data to enhance the robustness of deep learning-based segmentation of cardiac MRI images	(Al Khalil et al., 2023)	Investigates the effectiveness and usability of a diverse synthesized database of realistic CMR images for cardiac MRI segmentation.
Generation of synthetic ground-glass nodules using generative adversarial networks (GANs)	(Wang et al., 2022)	Develops a generative adversarial network (GAN) model to generate synthetic pulmonary lesions that mimic ground-glass nodules (GGNs).
Reproduction of brain tumor images with multi-tasking using diffusion models	(Rouzrokh et al., 2022)	Describes a DDPM for performing multiple repainting tasks on 2D axial slices of brain MRIs with various sequences, presenting proof-of-concept examples of its performance across different evaluation scenarios.

Brain image generation with latent diffusion models	(Pinaya et al., 2022)	Explores the use of latent diffusion models to generate synthetic images from high-resolution 3D brain scans.
Deep learning-based patient re-identification can leverage the biometric nature of medical chest X-ray data	(Packhäuser et al., 2022)	Demonstrates that a well-trained deep learning system can retrieve patient identity from chest X-ray data.
Adapting pre-trained vision and language foundation models to medical imaging domains	(Chambon et al., 2022)	Investigates and expands the representation capabilities of large pre-trained foundation models for medical concepts, specifically leveraging stable diffusion models to generate domain-specific images found in medical imaging.
Study of generative adversarial networks for synthetic data generation and their application to optoacoustic tomography	(Lopina et al., 2023)	Proposes the use of a generative adversarial network (GAN) to perform data augmentation aimed at improving image reconstruction in optoacoustic tomography (OAT) systems.

3 Proposed Solution Approach

The GAN model described in Figure 1, uses two neural networks, one to generate images and one to discriminate the generated images. The discriminator aims to differentiate when an image is false or real and the generator seeks to generate images that cannot be recognized as false by the discriminator. The GAN model was chosen because it allows the generation of realistic and quality images, which will contribute significantly to the expansion of the repository.

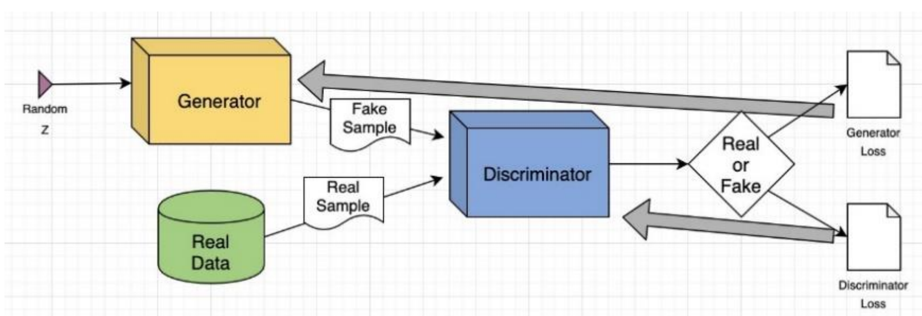

Figure 1 Methodology of the GAN training process Amazon Web Services, Inc. (s/f-b)

4 Experimental results

The GAN model was applied in Python and implemented in different environments to determine which requirements are best for effective training.

1. Computer without GPU: A CPU-based system with the characteristics described in Table 3 was used.

Reyes-Hernández et al. / International Journal of Combinatorial Optimization Problems and Informatics, 16(1) 2025, 266-274.

Table 3: PC Specifications

Brand	DELL
Computer Type	CPU
Processor	Intel Core I5
RAM	32GB
Storage	500GB

During the implementation of the model, significant issues were observed:

- The process was very slow, requiring small 128x128 images and small batch sizes to minimize execution time.
- Difficulties completing the training process.
- The model failed to generate images due to excessive resource usage, causing the process to be interrupted.

Image 2 displays some images generated using the following parameters:

- Input Images: 1000 (128x128)
- Epochs: 40
- Batch Size: 8
- Execution Time: 7 hours 55 minutes

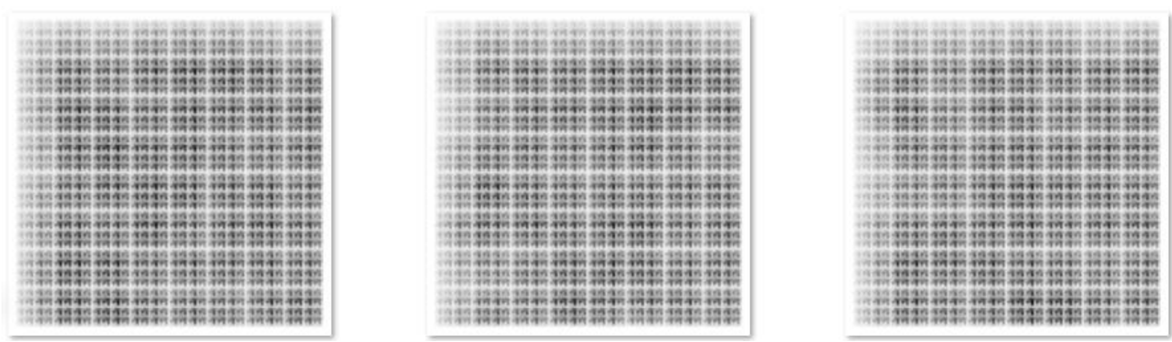

Image 2: Images Generated on a CPU-Based System Without GPU

2. Google Colab with GPU: Google Colab was chosen (Table 4) as it provides free resources, such as RAM and graphics cards, to improve execution times.

Table 4: Google Colab Specifications

Brand	Google Colab
Computer Type	Virtual
GPU	15GB
RAM	12.7GB
Storage	112.6GB

Observations:

- Memory filled up, causing the training process to be interrupted.
- After approximately 15 minutes, the session would automatically close, requiring re-execution. However, with the saved training progress, it was easier to resume and complete the process.

- The GPU memory is only available for a limited time.

Image 3 displays some images generated using the following parameters:
- Input Images: 1000 (512x512)
- Epochs: 40
- Batch Size: 8
- Execution Time: 2 hours 45 minutes

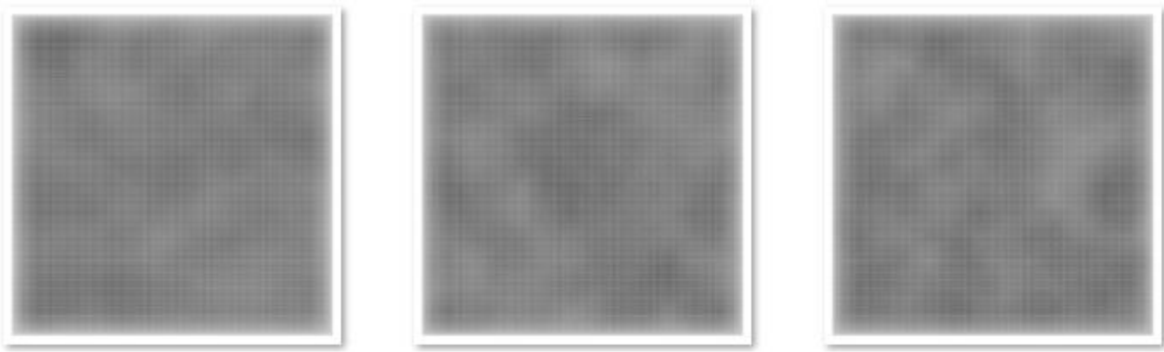

Image 3: Images Generated in Google Colab with GPU

3. Computer with GPU: Due to the issues encountered in the previous execution environment, a laptop was acquired (Table 5) with different specifications that improved the training process.

Table 5: Laptop Specifications

Brand	MSI
Tipo de equipo	Laptop
RAM	16GB
GPU	8GB
Storage	500GB

Issues Encountered:
- The entire dataset was used, including images of both the right and left breast, causing the generated images to be inaccurate.
- Tests were conducted using images from only one side, but after accumulating multiple steps, the images stopped improving and began to deteriorate. This issue was identified as being caused by an imbalance in the discriminator.

Image 4 displays some generated images based on the following parameters:
- Input Images: 1281 (128x128)
- Epochs: 500
- Batch Size: 64
- Execution Time: 2 hours 25 minutes

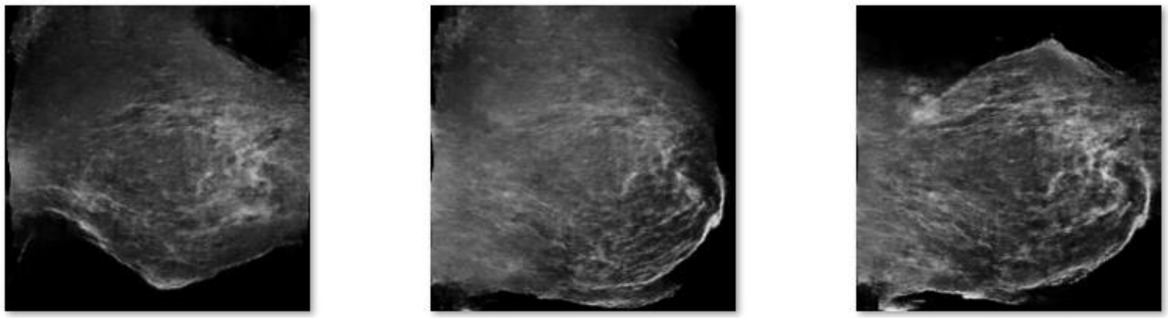

Image 4: Images Generated on Laptop with GPU

5 Conclusions

Although images identical to those in the original repository have not yet been obtained, a close approximation has been achieved, demonstrating significant progress in applying the GAN model.

The laptop equipped with a GPU has shown better performance in terms of response times and resource management, facilitating the tests conducted on the selected GAN model. This has been particularly useful for continuing the necessary experimental tests to achieve the expected results.

It is important to highlight that training can be time-consuming due to various challenges that may arise. Among the most common issues are overfitting and imbalance between the generator and discriminator networks, which negatively impact the model's ability to effectively generate high-quality images.

6 References

Aguirre, D. F., Carballo, D. L., González, D. X., & Gigirey, D. V. (2021). INTELIGENCIA ARTIFICIAL APLICADA A LA IMAGEN MÉDICA. *Revista de Imagenología*, *24*(2), Article 2.

Al Khalil, Y., Amirrajab, S., Lorenz, C., Weese, J., Pluim, J., & Breeuwer, M. (2023). On the usability of synthetic data for improving the robustness of deep learning-based segmentation of cardiac magnetic resonance images. *Medical Image Analysis*, *84*, 102688. https://doi.org/10.1016/j.media.2022.102688

Amazon Web Services, Inc. (s. f.). *¿Qué es una GAN? - Explicación sobre las redes generativas antagónicas - AWS*. Amazon Web Services, Inc. Recuperado 4 de diciembre de 2024, de https://aws.amazon.com/es/what-is/gan/

Arceo-Martínez, M. T., López-Meza, J. E., Ochoa-Zarzosa, A., Palomera-Sanchez, Z., Arceo-Martínez, M. T., López-Meza, J. E., Ochoa-Zarzosa, A., & Palomera-Sanchez, Z. (2021). Estado actual del cáncer de mama en México: Principales tipos y factores de riesgo. *Gaceta mexicana de oncología*, *20*(3), 101-110. https://doi.org/10.24875/j.gamo.21000134

Bastidas, O. (2021, febrero 23). ¿Qué cantidad de datos necesita un sistema de inteligencia artificial? *Medium*. https://oscarbastidas.medium.com/qu%C3%A9-cantidad-de-datos-necesita-un-sistema-de-inteligencia-artificial-ea952fdc0d9a

Chambon, P., Bluethgen, C., Langlotz, C. P., & Chaudhari, A. (2022). *Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains* (No. arXiv:2210.04133). arXiv. https://doi.org/10.48550/arXiv.2210.04133

decide4AI. (2023, julio 11). IA Generativa: Qué es, historia, tipos y casos de uso. *Decide*. https://decidesoluciones.es/ia-generativa-que-es-historia-tipos-y-casos-de-uso/

Díaz, O., Rodríguez-Ruiz, A., Gubern-Mérida, A., Martí, R., & Chevalier, M. (2021). ¿Son los sistemas de inteligencia artificial una herramienta útil para los programas de cribado de cáncer de mama? *Radiología*, *63*(3), 236-244. https://doi.org/10.1016/j.rx.2020.11.006

Freire Hidalgo, J. M. (2021). *Machine Learning y clustering para detección de cáncer de mama a través de imágenes de mamografía*. http://dspace.utpl.edu.ec/jspui/handle/20.500.11962/29121

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., & Weller, A. (2022). *Synthetic Data—What, why and how?* (No. arXiv:2205.03257). arXiv. http://arxiv.org/abs/2205.03257

Khader, F., Mueller-Franzes, G., Arasteh, S. T., Han, T., Haarburger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baessler, B., Foersch, S., Stegmaier, J., Kuhl, C., Nebelung, S., Kather, J. N., & Truhn, D. (2022, noviembre 7). *Medical Diffusion: Denoising Diffusion Probabilistic Models for 3D Medical Image Generation*. arXiv.org. https://arxiv.org/abs/2211.03364v7

Lawton, G. (2023, septiembre). *¿Qué es la IA generativa? Todo lo que necesita saber*. https://www.computerweekly.com/es/definicion/Que-es-la-IA-generativa-Todo-lo-que-necesita-saber

Lopina, A. S., González, M. G., & Vera, M. (2023). Estudio de redes generativas de confrontación para generación de datos sintéticos y su aplicación a tomografía optoacústica. *Elektron*, *7*(2), Article 2. https://doi.org/10.37537/rev.elektron.7.2.185.2023

McNulty, J. R., Kho, L., Case, A. L., Fornaca, C., Johnston, D., Slater, D., Abzug, J. M., & Russell, S. A. (2024). Synthetic Medical Imaging Generation with Generative Adversarial Networks For Plain Radiographs. *arXiv.org*. https://doi.org/10.48550/arXiv.2403.19107

Mendizábal, I. R., & Montoya, M. I. V. (2024). La imagen en la era de la inteligencia artificial. *Uru: Revista de Comunicación y Cultura*, *10*, Article 10. https://doi.org/10.32719/26312514.2024.10.12

Packhäuser, K., Gündel, S., Münster, N., Syben, C., Christlein, V., & Maier, A. (2022). Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest X-ray data. *Scientific Reports*, *12*(1), 14851. https://doi.org/10.1038/s41598-022-19045-3

Pan, S., Wang, T., Qiu, R. L. J., Axente, M., Chang, C.-W., Peng, J., Patel, A. B., Shelton, J., Patel, S. A., Roper, J., & Yang, X. (2023). 2D medical image synthesis using transformer-based denoising diffusion probabilistic model. *Physics in Medicine & Biology*, *68*(10), 105004. https://doi.org/10.1088/1361-6560/acca5c

Parakatta, A. (2023, junio 15). VAE v/s GAN — A case study. *Medium*. https://medium.com/@parakatta/vae-v-s-gan-a-case-study-b09c7169ac02

Pinaya, W. H. L., Tudosiu, P.-D., Dafflon, J., Da Costa, P. F., Fernandez, V., Nachev, P., Ourselin, S., & Cardoso, M. J. (2022). Brain Imaging Generation with Latent Diffusion Models. En A. Mukhopadhyay, I. Oksuz, S. Engelhardt, D. Zhu, & Y. Yuan (Eds.), *Deep Generative Models* (pp. 117-126). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-18576-2_12

Rouzrokh, P., Khosravi, B., Faghani, S., Moassefi, M., Vahdati, S., & Erickson, B. J. (2022, octubre 21). *Multitask Brain Tumor Inpainting with Diffusion Models: A Methodological Report*. arXiv.org. https://arxiv.org/abs/2210.12113v2

Shoshan, A., Bhonker, N., Kviatkovsky, I., Fintz, M., & Medioni, G. (2023). *Synthetic Data for Model Selection* (No. arXiv:2105.00717). arXiv. http://arxiv.org/abs/2105.00717

Sun, S., Goldgof, G. M., Butte, A., & Alaa, A. M. (2023, junio 16). *Aligning Synthetic Medical Images with Clinical Knowledge using Human Feedback*. arXiv.org. https://arxiv.org/abs/2306.12438v1

Vega, M. Á., Mora, L. M. Q., & Badilla, M. V. C. (2020). Inteligencia artificial y aprendizaje automático en medicina. *Revista Medica Sinergia*, *5*(8), Article 8. https://doi.org/10.31434/rms.v5i8.557

Wang, Z., Zhang, Z., Feng, Y., Hendriks, L. E. L., Miclea, R. L., Gietema, H., Schoenmaekers, J., Dekker, A., Wee, L., & Traverso, A. (2022). Generation of synthetic ground glass nodules using generative adversarial networks (GANs). *European Radiology Experimental*, *6*, 59. https://doi.org/10.1186/s41747-022-00311-y

Yang, Z., Zhan, F., Liu, K., Xu, M., & Lu, S. (2023). *AI-Generated Images as Data Source: The Dawn of Synthetic Era* (No. arXiv:2310.01830). arXiv. http://arxiv.org/abs/2310.01830