

Medical Imaging Complexity and its Effects on GAN Performance

William Cagas*, Chan Ko, Blake Hsiao, Shryuk Grandhi, Rishi Bhattacharya,
Kevin Zhu**, and Michael Lam**

Algoverse AI Research

williamgabriel.cagas@gmail.com, kevin@algoverse.us

Abstract. The proliferation of machine learning models in diverse clinical applications has led to a growing need for high-fidelity, medical image training data. Such data is often scarce due to cost constraints and privacy concerns. Alleviating this burden, medical image synthesis via generative adversarial networks (GANs) emerged as a powerful method for synthetically generating photo-realistic images based on existing sets of real medical images. However, the exact image set size required to efficiently train such a GAN is unclear. In this work, we experimentally establish benchmarks that measure the relationship between a sample dataset size and the fidelity of the generated images, given the dataset’s distribution of image complexities. We analyze statistical metrics based on delentropy, an image complexity measure rooted in Shannon’s entropy in information theory. For our pipeline, we conduct experiments with two state-of-the-art GANs, StyleGAN 3 and SPADE-GAN, trained on multiple medical imaging datasets with variable sample sizes. Across both GANs, general performance improved with increasing training set size but suffered with increasing complexity.

Keywords: GAN · Entropy · Synthetic Data Generation

1 Introduction

Machine learning in healthcare is a rapidly growing field with countless applications [28] including disease diagnosis [24], clinical treatment [26], drug development [19], and mental health [7]. The machine learning models driving these advances require the collection of high-quality, annotated medical training data, which persists as an arduous task due to privacy concerns surrounding sensitive patient data [23] and the time-intensive nature of labeling [5]. To address these issues, synthetic data—artificially generated information mimicking real-world data—has surfaced as a promising solution [8].

Currently, generative adversarial networks (GANs) remain one of the leading approaches to synthetic data generation [17]. Since its inception in 2014 [6], GANs have gained increasing attention in the medical research community due

* Lead Author.

** Senior Author.

to their ability to synthesize medical images [29]. However, achieving results with high fidelity remains a difficult task factoring the lack of medical data and prevalence of smaller datasets in the medical domain. With limited data, a GAN’s efficacy is directly affected with consequences including mode collapse, where the generator produces a limited variety of outputs [20], and overfitting, where the GAN replicates training data rather than generalizing from it [32]. Various papers such as Wang *et al.* ’s [31] transfer learning and Robb *et al.* ’s [25]

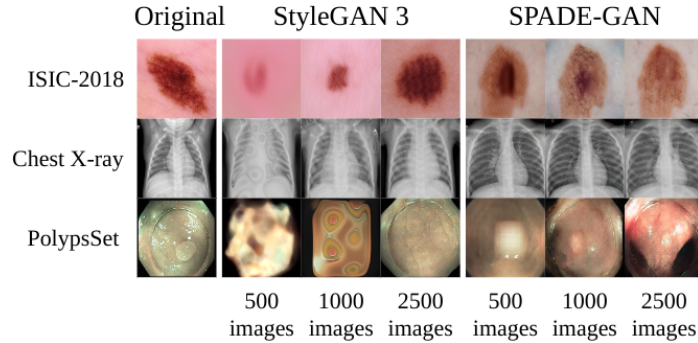


Fig. 1: Comparison between original images and synthetic images from StyleGAN 3 and SPADE-GAN based on variable image set sizes.

Few-Shot GAN (FSGAN) have tried addressing these issues as architecture-centric approaches, achieving increased training efficiency only as a result of the changes in a GAN’s structure. However, such approaches are ineffective when making alterations to a GAN’s internal structure are not feasible and when time constraints are present. As such, a data-centric approach by providing the GAN with the optimal amount of data to produce high-quality results is more appropriate. Nevertheless, the exact sample set size required to train state-of-the-art GANs is obscure.

In this study, we introduce a data-centric optimization method to create efficient GAN training for medical image synthesis. Our approach investigates how the image complexity distribution of a medical image dataset can be utilized as a measure of training difficulty for a GAN. By doing so, we can ascertain a correlation between the image complexities of the training images and the optimal training set sizes by establishing benchmarks that evaluate the relationship between a sample training set size and the fidelity of the generated images. We hypothesize that given a dataset of a specific image complexity distribution, healthcare professionals can reference the closest image fidelity curve to identify the optimal amount of experimental trials to produce superlative results. Ultimately, our approach can avoid both undertraining and wasteful overtraining by constructing a data-efficient, GAN training pipeline.

2 Background

Generative Adversarial Networks (GANs) Introduced by Goodfellow *et al.* [6], GANs are a class of generative models that consist of two neural networks: a generator G , which aims to transform its latent variable distribution $p(z)$ to closely resemble the training data distribution $p(x)$, and a discriminator D , which differentiates between the ground truth and data generated by G . Training is an adversarial process where G attempts to deceive D into classifying its outputs as real. This two-player minimax game is represented by the following loss function:

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} [\log D(\mathbf{x})] \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))] . \quad (1)$$

Many papers have tried to address data scarcity and computational costs in GAN training architecturally. One proposed approach was transfer learning [31], which consists of fine-tuning a pre-trained generator and discriminator to the desired domain. However, if the pre-trained models do not align well with the target domain, this could result in even higher data and computational demands [33]. Another approach, Few-Shot GAN (FSGAN) [25], achieved impressive adaptation even with extremely few training examples, albeit at the cost of prolonged training times. This results in the reduced quality and diversity of the synthetic data when time constraints are present.

Image Complexity Objectively, image complexity can be defined as the variety of features and details within an image. It has been shown that information entropy is a traditional, heuristic-based method of calculating the complexities of images in small-scale datasets [16].

Traditional entropy is a foundational abstraction in information theory introduced by Shannon [27]. Used as a measure of uncertainty or “surprise” in data, it is the variation in the distribution of pixel intensities of an image in grayscale format. The equation is defined as

$$H = - \sum_{i=0}^{n-1} p_i \log_b p_i, \quad (2)$$

where n denotes the number of gray levels (256 for 8-bit images), b stands for the logarithmic base (returning bits when $b = 2$), and p_i is the probability of a pixel having gray level i . However, although Shannon entropy considers compositional image information, it fails to account for spatial information, specifically the relationship between neighbouring pixels [4].

Another entropy-based metric, the Gray Level Co-Occurrence Matrix (GLCM) entropy, unlike Shannon entropy, is a measure of how often pairs of pixel values occur in a grayscale image distribution [9]. Taking into account this local spatial information, the GLCM is useful for various textural analysis tasks such as feature extraction for medical image segmentation [14]. The GLCM entropy can be represented as

$$H_g = - \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} p_{(i,j)} \log_b p_{(i,j)}, \quad (3)$$

where $p_{(i,j)}$ is the probability of two pixels having gray levels i and j at a certain angle θ and distance d away from each other. Despite GLCM better-capturing complexities within an image, it does not consider spatial patterns and global pixel relationships beyond its adjacent pairing.

3 Methodology

3.1 Image Complexity Metric

Our approach utilizes Larkin’s delentropy, a metric identical to both the Shannon entropy and the GLCM entropy, but incorporating a new density function known as the *deledensity* [15]. By analyzing the relationship between the local and global features of an image, delentropy accounts for both an image’s gradient vector field and pixel co-occurrence, encapsulating its spatial information as a whole. The deledensity, as a joint probability function, is formulated as

$$p_{(i,j)} = \frac{1}{4WH} \sum_{w=0}^{W-1} \sum_{h=0}^{H-1} \delta_{i,d_x(w,h)} \delta_{j,d_y(w,h)}, \quad (4)$$

where d_x and d_y denote the derivative kernels in the x and y direction, δ is the Kronecker delta to describe the binning operation required to generate a histogram, and H and W is the image’s dimensions (height and width) [13]. By obtaining this, we can then calculate delentropy as

$$\text{DE} = -\frac{1}{2} \sum_{i=0}^{I-1} \sum_{j=0}^{J-1} p_{(i,j)} \log_b p_{(i,j)}, \quad (5)$$

such that I and J represent the number of bins (discrete cells) in the 2D distribution, and the $\frac{1}{2}$ is derived from Papoulis’ generalized sampling expansion [21].

To interpret this measure, yielding a high delentropy suggests an image has a high range of variation in pixel intensities and more sophisticated details. A low delentropy can be interpreted as a result of having a uniform distribution of pixel intensities, indicating simple structure and a less-detailed image.

Prior to any calculations, each image was preprocessed into an 8-bit, grayscale image. This ensured delentropy was calculated in a consistent, single-channel format throughout each dataset.

3.2 GAN Selection

Core to the experimental approach was the selection of two state-of-the-art GANs, SPADE-GAN [22] and StyleGAN 3 [11] on which to run the experimental pipeline. These networks have been widely adopted by the medical image synthesis community and empirically observed to produce superior-quality medical images when compared to predecessor GANs [29]. StyleGAN 3’s large community support and wide availability of its code repository along with its numerous configurations for different training settings were taken into account as well.

3.3 GAN Pipeline

For experiments, given our data-centric approach, StyleGAN 3 and SPADE-GAN were run with the official, publicly available implementations with default hyperparameters and no augmentations to each network’s architecture.

Preprocessing We first set all images to a consistent 512x512 resolution. As such, training parameters were based on the size of the preprocessed images, as documented in the official implementations. SPADE-GAN additionally relies on segmentation masks to produce synthetic data. We used pre-existing annotations for ISIC-2018 and the Polyps Set. Because the Chest X-ray dataset did not have such annotations, masks were generated using TorchXRyVision [3]. All experiments were performed on one NVIDIA A100 and three NVIDIA A40 GPUs.

Training and Generation The experimental pipeline was designed to identify the role of image dataset size in the image generation fidelity of selected GANs, for which to be compared to the image complexity distribution of each dataset. To that end, for each GAN training run, all parameters were held constant with the exception of the image set size, which was subsequently set to 500, 1000, and 2500 images, randomly sampled from the same dataset for each experimental run, respectively. The pipeline was designed to incorporate 500 images as a baseline for GANs to train with limited data. We facilitated experiments with 2500 images for a more comprehensive training run to better capture the underlying image distribution and use an intermediary set of 1000 images serving as a middle ground. For StyleGAN 3, all experimental runs were trained for 100 epochs; for SPADE-GAN, training iterated 50 epochs. The trained adversarial network was then used to generate synthetic images, the fidelity of which was then evaluated for each training set size.

Evaluation The Fréchet Inception Distance (FID) [10] is a common metric used to evaluate the fidelity of the synthetically generated images for GANs [1]. Defined as the distance between the distributions of the ground truth and the generated images respectively, in our paper, we use the FID to assess the performance of the GAN (i.e. image fidelity) for each experimental run across both GANs. **A lower FID score signifies that a GAN is more proficient at generating synthetic data close to its target distribution.** From these data, we obtained fidelity curves for each dataset that describe how FID scores trend with increasing training set size.

4 Experimental Results

Datasets We employed three medical image datasets: International Skin Imaging Collaboration 2018 Challenge (ISIC-2018) [2], Chest X-Ray Images (Chest X-ray) [12], and Colonoscopy Polyp Detection and Classification (Polyps Set) [30]. These datasets were chosen for their diversity in both perceptual complexity,

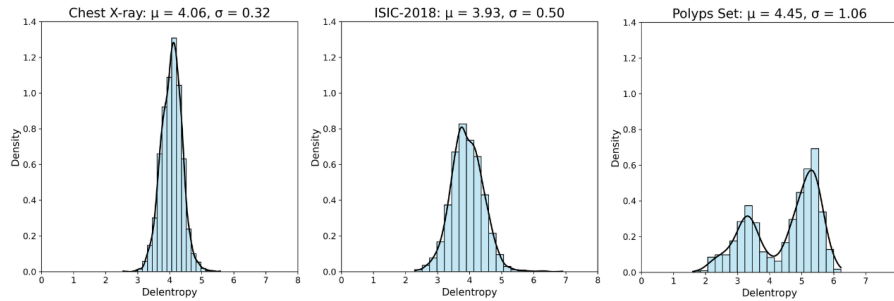


Fig. 2: Delentropy distributions across each medical image dataset. A higher mean delentropy μ indicates a dataset with more complex images.

ranging from relatively skin lesions to complex colon polyps, and imaging modality (dermoscopy vs. x-ray vs. colonoscopy).

We carried out delentropy calculations as described in Section 3 by using a publicly available implementation from Marchesoni [18]. To effectively capture the overall complexity of each image dataset, we captured each dataset’s delentropy distribution as displayed in Fig. 2.

Across the experimental runs, FID scores consistently decreased with increasing dataset size. On StyleGAN 3, synthesized images that had been generated by a GAN trained on 2500 images exhibited an average FID score reduction of 48% when compared to those generated by a StyleGAN 3 that had been trained on a mere 500 images (Fig. 3). SPADE-GAN experienced an analogous 31% FID score reduction on average, though it is worth noting that FID reduction plateaued after only 1000 training images.

Comparing both Fig. 2 and Fig. 3, one can see a general relationship between the delentropy distribution and the training performance of both GANs. As the spread of image complexities increases from a slender, peaked distribution to a broader, bimodal one, we see a corresponding increase in FID scores for each dataset sample size. The Chest X-ray dataset with the most homogeneous image complexities shown by a tall and narrow distribution, yields the lowest FID score after being trained for 2500 images, indicating that both GANs had easier training runs with this dataset. On the contrast, the Polyps Set—the dataset with the widest distribution and multiple complexity peaks—correlates with the highest FID scores for each dataset sample size, which suggests that the GAN was faced with a more challenging and unstable training run. Ultimately, this pattern shows a general inverse relationship—GAN performance decreases with an increasing spread of image complexities within a dataset.

5 Discussion

SPADE-GAN outperformed StyleGAN 3 across *all datasets and training sizes*, with FID scores averaging 33% lower, likely due to its architecture that incorporates segmentation masks for structural information, whereas StyleGAN 3

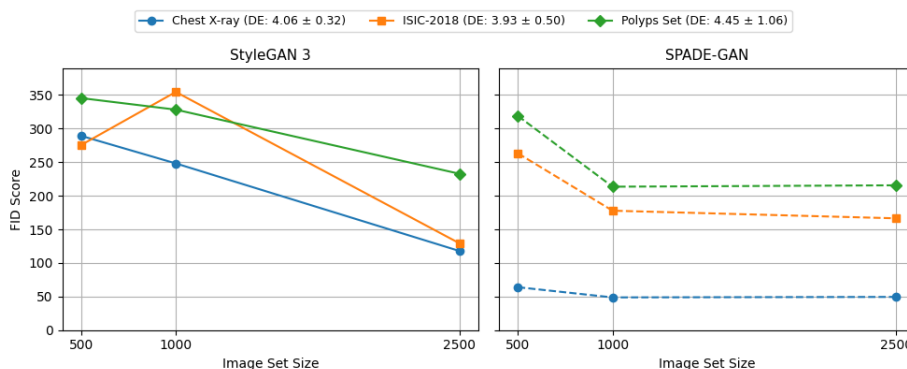


Fig. 3: Fréchet Inception Distance (FID) curves comparing StyleGAN 3 and SPADE-GAN across each medical image dataset with varying sample sizes. Lower FID scores correspond to higher fidelity synthetic images.

trained on raw image data alone, making it more difficult to generalize to high-entropy datasets. Moreover, ISIC-2018 being an outlier can be attributed to its fluctuations in image complexity, reflected by the standard deviation in delentropy (Fig. 2). Despite having a lower mean delentropy, its spread likely resulted in difficulties in GAN training and learning the images’ distribution, contrasting with the Chest X-ray dataset.

While the experimental results generally reflected an intuitive understanding of how image complexity and training data influence GAN training, the FID curves provide insightful details, offering a deeper perspective on these effects. SPADE-GAN exhibits both better quality results than StyleGAN 3 in the form of lower FID scores and more consistent training as evidenced by the smooth, non-overlapping FID curves (Fig. 3). As aforementioned, performance plateaued after 1000 training images, suggesting that additional training data past that point may not help increase GAN performance as measured by FID score. This is also apparent in the generated images themselves, which exhibit little perceptual difference between those generated after 1000 training images and those generated after 2500 (Fig. 1). Contrast this with the StyleGAN 3 curves, which do not reach any noticeable plateau between 500 and 2500 training images. In fact, the increasingly negative slope values of the StyleGAN 3 graphs imply that StyleGAN 3 begins to better capture the images’ features at a point past 1000 images, the exact whereabouts of which would need to be determined by a separate study.

The FID curves generated by this set of experiments set up a useful benchmark to which other potential training image data sets can be compared. For training sets that are of similar delentropy distributions and used to train StyleGAN 3 or SPADE-GAN, it is not unreasonable to predict that their training curves will be similar to those represented in Fig. 3, though many more training

set sizes and image sets are required before a truly comprehensive representation can be reached.

Broader Impacts Our research on GANs for medical image synthesis may have positive and negative societal implications. On the positive side, it can enhance healthcare outcomes by improving the training of machine learning models with realistic synthetic data, therefore protecting patient privacy. Contrarily, potential negative impacts include the risk of maliciously generating fraudulent synthetic data and the possibility of reinforcing biases due to a lack of diversity of representing patient populations. These considerations demonstrate the importance of addressing both the benefits and potential risks associated with the use of GANs in the medical domain.

6 Conclusion

In this work, we highlight the impact of image complexity on GAN performance in medical image synthesis. We empirically demonstrate a general inverse relationship: higher image complexity leads to poorer image fidelity results and lesser performance in GANs. Furthermore, we demonstrate FID curves showing healthcare professionals the possibility for the use of our benchmarks to gauge an estimate of data training requirements to achieve desirable results based on the image complexity distribution of a medical image dataset.

7 Limitations

Due to limited resources, experiments were only run on 500, 1000, and 2500 training images, leading to coarse-grained results. An extended study with a larger range and finer-grained increments would better elucidate exactly how FID scores respond to changes in training image dataset size. The use of FID scores as a sole evaluation metric also has its limitations, not necessarily correlating with human perceptual interpretations, something that is extremely important in the medical field where human doctors are still largely the source of truth. Skandarani *et al.* [29] shows that a lower FID score may not be a good measure of how well synthetic images can perform on a downstream task as well. Although this study specifically focused on GANs as the primary generative model, similar experiments extrapolated across other generative models such as stable diffusion may prove more relevant in the context of recent advancements in generative AI. More research with larger resources involving multiple evaluations of a similar experimental setup is required.

Acknowledgements

We are grateful to Michael Lam and Kevin Zhu for their excellent mentorship, constructive feedback, and unwavering support throughout our research.

References

1. Borji, A.: Pros and cons of gan evaluation measures: New developments. *Computer Vision and Image Understanding* **215**, 103329 (2022). <https://doi.org/https://doi.org/10.1016/j.cviu.2021.103329>, <https://www.sciencedirect.com/science/article/pii/S1077314221001685> 5
2. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic) (2018), <https://arxiv.org/abs/1902.03368>, licensed under the Creative Commons Attribution-NonCommercial (CC-BY-NC) license. 5
3. Cohen, J.P., Viviano, J.D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M.P., Chaudhari, A., Brooks, R., Hashir, M., Bertrand, H.: TorchXRrayVision: A library of chest X-ray datasets and models. In: *Medical Imaging with Deep Learning* (2022), <https://github.com/mlmed/torchxrayvision> 5
4. Gao, P., Li, Z., Zhang, H.: Thermodynamics-based evaluation of various improved shannon entropies for configurational information of gray-level images. *Entropy* **20**(1) (2018). <https://doi.org/10.3390/e20010019>, <https://www.mdpi.com/1099-4300/20/1/19> 3
5. Gilbert, A., Marciniak, M., Rodero, C., Lamata, P., Samset, E., Mcleod, K.: Generating synthetic labeled data from existing anatomical models: An example with echocardiography segmentation. *IEEE Transactions on Medical Imaging* **40**(10), 2783–2794 (2021). <https://doi.org/10.1109/TMI.2021.3051806> 1
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. *Advances in neural information processing systems* **27** (2014) 1, 3
7. Graham, S., Depp, C., Lee, E.E., Nebeker, C., Tu, X., Kim, H.C., Jeste, D.V.: Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports* **21**, 1–18 (2019) 1
8. Guo, X., Chen, Y.: Generative ai for synthetic data generation: Methods, challenges and the future. *arXiv preprint arXiv:2403.04190* (2024) 1
9. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-3**(6), 610–621 (1973). <https://doi.org/10.1109/TSMC.1973.4309314> 3
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 30 (2017), <https://arxiv.org/abs/1706.08500> 5
11. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan (2020), <https://arxiv.org/abs/1912.04958>, licensed under the Nvidia Source Code License. 4
12. Kermany, D., Zhang, K., Goldbaum, M.: Labeled optical coherence tomography (oct) and chest x-ray images for classification (2018). <https://doi.org/10.17632/rscbjbr9sj.2>, <https://doi.org/10.17632/rscbjbr9sj.2>, licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. 5
13. Khan, T.M., Naqvi, S.S., Meijering, E.: Leveraging image complexity in macro-level neural network design for medical image segmentation. *Scientific Reports* **12** (2022). <https://doi.org/10.1038/s41598-022-26482-7>, <https://doi.org/10.1038/s41598-022-26482-7> 4

14. Khan, Z.F., Alotaibi, S.R.: Computerised segmentation of medical images using neural networks and glm. In: 2019 International Conference on Advances in the Emerging Computing Technologies (AECT). pp. 1–5 (2020). <https://doi.org/10.1109/AECT47998.2020.9194196> 3
15. Larkin, K.G.: Reflections on shannon information: In search of a natural information-entropy for images (2016), <https://arxiv.org/abs/1609.01117> 4
16. Liu, S., Zhao, L., Chen, D., Song, Z.: Contrastive learning for image complexity representation (2024), <https://arxiv.org/abs/2408.03230> 3
17. Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., Fu, T., Wei, W.: Machine learning for synthetic data generation: A review (2024), <https://arxiv.org/abs/2302.04062> 1
18. Marchesoni, F.: How do i know if an image after image enhancement is better than before? (March 2023), <https://ai.stackexchange.com/questions/39483/how-do-i-know-if-image-after-image-enhancement-is-better-than-before-image-pre> 6
19. Nordon, G., Koren, G., Shalev, V., Horvitz, E., Radinsky, K.: Separating wheat from chaff: joining biomedical knowledge and patient data for repurposing medications. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9565–9572 (2019) 1
20. Pan, Z., Yu, W., Yi, X., Khan, A., Yuan, F., Zheng, Y.: Recent progress on generative adversarial networks (gans): A survey. *IEEE access* 7, 36322–36333 (2019) 2
21. Papoulis, A.: Generalized sampling expansion. *IEEE Transactions on Circuits and Systems* 24(11), 652–654 (1977). <https://doi.org/10.1109/TCS.1977.1084284> 4
22. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization (2019), <https://arxiv.org/abs/1903.07291>, licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license. 4
23. Pezoulas, V.C., Zaridis, D.I., Mylona, E., Androutsos, C., Apostolidis, K., Tachos, N.S., Fotiadis, D.I.: Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and Structural Biotechnology Journal* 23, 2892–2910 (2024). <https://doi.org/https://doi.org/10.1016/j.csbj.2024.07.005>, <https://www.sciencedirect.com/science/article/pii/S2001037024002393> 1
24. Rahman, S., Ibtisum, S., Bazgir, E., Barai, T.: The significance of machine learning in clinical disease diagnosis: A review. *arXiv preprint arXiv:2310.16978* (2023) 1
25. Robb, E., Chu, W.S., Kumar, A., Huang, J.B.: Few-shot adaptation of generative adversarial networks. *arXiv preprint arXiv:2010.11943* (2020) 2, 3
26. Shang, J., Ma, T., Xiao, C., Sun, J.: Pre-training of graph augmented transformers for medication recommendation. *arXiv preprint arXiv:1906.00346* (2019) 1
27. Shannon, C.E.: A mathematical theory of communication. *The Bell System Technical Journal* 27(3), 379–423 (1948). <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x> 3
28. Shi, Z.R., Wang, C., Fang, F.: Artificial intelligence for social good: A survey. *arXiv preprint arXiv:2001.01818* (2020) 1
29. Skandarani, Y., Jodoin, P.M., Lalande, A.: GANs for Medical Image Synthesis: An Empirical Study. *Journal of Imaging* 9(3), 69 (2023). <https://doi.org/10.3390/jimaging9030069>, <https://www.mdpi.com/2313-433X/9/3/69> 2, 4, 8

30. Wang, G.: Replication Data for: Colonoscopy Polyp Detection and Classification: Dataset Creation and Comparative Evaluations (2021). <https://doi.org/10.7910/DVN/FCBUOR>, <https://doi.org/10.7910/DVN/FCBUOR>, licensed under the CC0 1.0 Universal (CC0 1.0) Public Domain Dedication. 5
31. Wang, Y., Wu, C., Herranz, L., Van de Weijer, J., Gonzalez-Garcia, A., Raducanu, B.: Transferring gans: generating images from limited data. In: Proceedings of the European conference on computer vision (ECCV). pp. 218–234 (2018) 2, 3
32. Webster, R., Rabin, J., Simon, L., Jurie, F.: Detecting overfitting of deep generative networks via latent recovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11273–11282 (2019) 2
33. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., He, Q.: A comprehensive survey on transfer learning. Proceedings of the IEEE 109(1), 43–76 (2020) 3