UNIVERSITY OF CAPE TOWN

DEPARTMENT OF COMPUTER SCIENCE

CS/IT Honours Project Final Paper 2023

Title: Evaluating Generative Adversarial Networks on Small Medical Datasets

Author: Shaylin Chetty

Project Abbreviation: DEEPPC

Supervisor(s): Geoff Nitschke

Category	*Min*	*Max*	Chosen
Requirement Analysis and Design	*0*	*20*	
Theoretical Analysis	*0*	*25*	
Experiment Design and Execution	*0*	*20*	20
System Development and Implementation	*0*	*20*	5
Results, Findings and Conclusions	*10*	*20*	20
Aim Formulation and Background Work	*10*	*15*	15
Quality of Paper Writing and Presentation	*10*		10
Quality of Deliverables	*10*		10
Overall General Project Evaluation (*this section allowed only with motivation letter from supervisor*)	*0*	*10*	
Total marks		**80**	**80**

Evaluating Generative Adversarial Networks on Small Medical Datasets

DEEPPC: Deep Learning, Small Data Pathology Classifiers

Shaylin Chetty
CHTSHA042@myuct.ac.za
University of Cape Town
Cape Town, South Africa

ABSTRACT

The use of deep learning in medicine is constrained by the availability of data. Existing datasets are small, sparse and of poor quality. Typically, Data Augmentations are used to prevent models from overfitting when faced with limited data. We explore the use of Generative Adversarial Networks to enhance datasets by exploring the image quality generated by Vanilla GAN, WGANGP and StyleGAN2 when generating elbow X-rays in a data-constrained environment. We show that StyleGAN2 generates the highest quality imagery but is held back by inconsistent datasets. We also observe that Vanilla GAN is extremely sensitive to the choice of hyperparameters and that the image quality of Vanilla GAN and WGANGP is insufficient. Finally, we propose a domain adaption of FID scores allowing for the comparison of x-ray image generation models.

1 INTRODUCTION AND RELATED WORK

Deep learning enables computers to build representations of the world at various levels of abstraction [72]. The usage of computer vision in medicine has seen a recent surge with the promise of more efficient, accurate and less labour-intensive evaluation of medical data.

Early computer vision models used in medicine include active shape models and statistical classifiers that learn patterns in data. These models create a decision boundary in some feature space however the extraction of distinguishing features still relied on human experts [77], making it a costly and error-prone process. The transition to automated feature extraction occurred later with the introduction of Convolutional Neural Networks [40] and, later, vision transformers [16]. However, in medicine, feature extractions were generally done using principal component analysis and clustering algorithms [40].

The success of deep learning across domains can be attributed to the enhanced performance over previous state-of-the-art methods, the adoption of GPU-based training and the availability of high-quality, publically available datasets [72]. While deep learning has achieved much success in medicine [28, 81], the latter is the limiting factor to the uptake of deep learning in medicine. Typical computer vision tasks need between $10^5 \sim 10^6$ training images to learn meaningful and general representations of data [59]. In the pathology domain, unique factors inherently limit datasets to orders of magnitudes smaller than this. Considering the nature of the data captured, the domain is plagued by patient privacy and confidentiality restrictions that limit the sharing and processing of data [59]. This makes the collection of large datasets a difficult and highly regulated process. Training models on smaller datasets leads to the model overfitting the training data and achieving poor accuracy on real-world data and hold-out sets [29, 36, 78]. While this would be an undesirable result, the consequences in medicine, in particular, are severe. It has been shown that a false-positive diagnosis leads to negative short-term consequences including stress and anxiety. [56, 69]. This warrants the need for an ethical analysis of the consequences of screening and diagnosis programmes [56], particularly in automated screening processes such as those made possible by deep learning technologies. The issue of false classifications also arises in the context of data imbalance. Class imbalance in a dataset leads to a strong inclination towards sampling from and favouring majority classes engendering substantial bias in the model [55]. In medicine, there is a well-established absence of a large number of positive cases [28, 57] making these datasets inherently skewed and leading to a bias towards negative cases. This could lead to false negatives which deprive patients of an opportunity to receive treatment or medical interventions. This notion also extends to the frequency of some class labels of hard-to-detect pathologies [7]. Clearly, there is a need for balanced, complete, representative and consistent medical datasets.

Recently, there have been many large-scale research collaborations that are aimed at curating and anonymizing datasets such as *MURA* [54], *LERA* [1] and the *The Cancer Imaging Archive* [14]. Nevertheless, these target broad, general-purpose use cases and often lack consistency [26], with data for niche domains still isolated and inaccessible [26, 59]. Noticeably, these attempts provide evidence of the significant data-capturing and distributing challenge in curating such datasets [9, 29, 81]. Furthermore, creating annotated datasets is time-consuming and requires manual intervention [9]. This makes similar initiatives for niche domains or specific use cases infeasible. While annotated datasets could be circumvented by self-supervised learning, the issue of large-scale data collection is still an issue.

Therefore, the literature focuses more on using small datasets effectively as opposed to collecting more data. That is, numerous proposals have been set out to either fix data imbalance issues or to train models effectively using small datasets [29]. Data Augmentation has received much praise for its ability to apply label-preserving transformations on data. Data Augmentation involves applying geometric (flips, crops, scaling and so forth) and photometric (image recolouring, filters and so on) transformations [32, 67] on data. It has been successful in increasing the size of datasets [18, 84] and making the trained model invariant to particular transformations, preventing overfitting and improving accuracy on real-world

[1] https://aimi.stanford.edu/lera-lower-extremity-radiographs

cases [9]. However, by merely transforming images, one creates an enlarged dataset of variations of original images thereby not enhancing the diversity of the dataset [77]. That is, it increases the given sample spaces but does not explore the true sample space [28] doing little to alleviate the aforementioned data balance issues. Additionally, for our domain, data augmentation is not always effective or realistic [9]. For example, neck X-rays cannot undergo geometric transformations or recolouring as these images are standardized in colour and can only be analysed in one orientation.

As a consequence of the above, research has shifted to image generation as a means to address the aforementioned issues. The idea is to artificially inflate datasets through the generation of fake images that closely resemble real imagery but contain no personally identifiable information. This was made possible through the introduction of *Generative Adversarial Networks* (GANs) by Goodfellow et al. (2014) [20]. The initial GAN, dubbed Vanilla GAN, consists of two neural networks trained in an adversarial game to minimize a common loss function. The first network, dubbed the Generator, is responsible for studying the underlying structure of training data and uses its approximated distribution to transform a Gaussian noise input into an image. The second network, dubbed the Discriminator, receives random images from the pool of real images and generated images. The discriminator attempts to differentiate between real and fake images. A GAN trained to convergence should consist of a generator that closely approximates the real data distribution to the extent whereby the discriminator has difficulty distinguishing real and fake samples. A graphical representation of Vanilla GAN is given in figure 1.
The ability to generate unique imagery to diversify datasets [81] while ensuring that patient privacy and confidentiality issues are circumvented is a key factor driving the growth of GANs in medical imaging.

There are three types of image generation techniques which GANs can perform: *unconditional image synthesis*, which is evaluated in this study, *conditional image synthesis* and *cross-modality image synthesis*, a type of conditional image synthesis involving image *translation* between treatments (domains) [75] usually through the use of Conditional GANs [43, 44], which feeds explicit instructions to the generator, and CycleGAN [71].

Our focus is on unconditional image synthesis. Under this paradigm, the generator receives no guidance on how the final image should look but rather uses the discriminator's classification to guide the generation process into producing results that *confuse* the discriminator. We will explore two GANs used to achieve this, both of which has been covered in literature, namely StyleGAN2 [76] and WGAN [48]. These are preferred over Vanilla GAN due to the improved training stability and their ability to capture finer images in data. With recent advancements, the quality of synthetic imagery produced by GANs has drastically improved to the degree where radiologists have difficulty discerning real images from fake images [58, 59, 76].

Although extensive research has been done on GANs in medicine, the use cases generally focus on smaller pathologies such as histology [2] [12, 25, 55, 71], skin lesions [1], retina [26] and liver scans

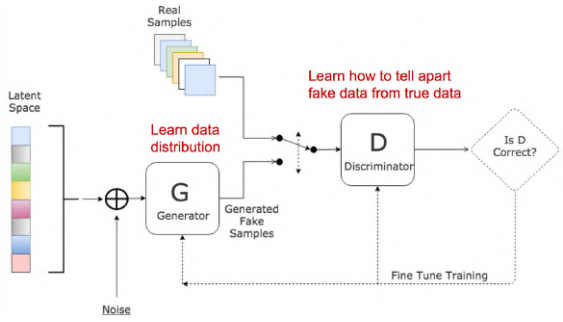


Figure 1: Vanilla GAN architecture. The latent space is a Gaussian distribution from which a noise sample is drawn [19]

[62, 76] or on chest and brain x-ray synthesis [27, 48, 57, 75]. However, little work has been done on the extremities such as the elbow. As per our knowledge, only one such paper has investigated this [39]. The authors applied AC-GAN to a limb dataset for use in a hip fracture detection under a transfer learning approach. Hence, this study aims to determine whether unconditional image synthesis with GANs is feasible for enhancing small medical datasets, specifically focusing on the elbow. Numerous GANs have been proposed exclusively to deal with the shortcomings of medical data including MI-GAN for retinal vessel images [26], PathologyGAN for generating cancer tissue phenotypes in tumours [52], MedWGAN for electronic health record generations [7, 63] and healthGAN for image generation factoring time-series [15]. While effective, these were constructed for niche use cases making them less general. Therefore, we focus on evaluating Vanilla GAN, WGAN-GP and StyleGAN2. Due to data constraints, we ignore class conditioning in this study. However, this study lays the foundation for future class-conditioned image generation once data collection and data consistency measures improve.

We start off by describing the methods used, including a brief overview of our proposed MedFID metric which is detailed in appendix A. We then discuss the experimental setup in section 3. Our results are presented in section 4 and discussed in section 5. Finally, we conclude our work, provide recommendations for future research and summarize our key findings in section 6. All models run are available on our GitHub repo [3] and the latex version of this report is available on Overleaf [4].

2 METHODS

In this section, we first present an overview of the dataset used in this study, followed by introducing Vanilla GAN. In section 2.3 we discuss some of the common pitfalls of Vanilla GAN which led to the development of WGAN and StyleGAN, each of which are discussed below. Finally, we discuss the metrics used to quantify image quality. Note that the exact implementation details are discussed in section 3.

[2]Histology is the study of animal and plant tissues through staining and sectioning under a microscope

[3]https://github.com/Shaylin-UCT/DEEPPC
[4]https://www.overleaf.com/9972285551xpcvkbrfgmgj

2.1 Data-sets

We use a collection of X-ray and CT scans of the elbow and neck provided by Dr. Kruger of Groote Schuur Hospital, Cape Town. The dataset consists of 5159 images of the elbow and 1107 images of the neck. The elbow consists of a roughly equal split between AP radiographs (elbow kept straight but oriented 45 degrees in internal rotation) and lateral radiographs (elbow flexed at 90 degrees). All images have been anonymized and have been ethically cleared as part of a larger study by Dr. Kruger. Running a generative model using the entire elbow dataset will result in the distribution of each subset being merged together, as discussed in Appendix D. Therefore, we focus primarily on generating lateral elbow scans using the 2618 LAT elbow scans in the dataset. Note that the results are expected to be similar for AP elbow. Due to the small size of the training dataset, we ignore class conditioning the GAN.

2.2 Vanilla GAN

Goodfellow et al. (2014) [20] introduced a novel image generation technique to generate synthetic images that follow the underlying distribution of original data called Generative Adversarial Networks (GANs). Their original work is now referred to as Vanilla GAN. These consist of two multi-layer perceptrons (MLP), which are simultaneously trained via an adversarial game, namely the Generator and the Discriminator. Both perceptrons are trained to optimize the value function in Equation 1. First, we take the real data distribution to be P_r. The Generator is responsible for generating synthetic images by approximating the underlying statistical structure behind the training data. The Generator receives a random noise input sampled from a uniform or Gaussian distribution, $z \sim p(z)$, which is mapped to data space via the MLP, $G(z, \theta_g)$, where θ_g are model hyperparameters. This mapping is done under the Generator's approximation of the real data distribution, P_g. The Discriminator will attempt to discern synthetic data from real data by predicting the probability that the data came from the original training set, $D(x)$. The aim of the training is for the Generator to minimize the loss function, that is to create images close to the training data, while the Discriminator tries to maximize the loss function. A well-trained generator should produce samples that cause the Discriminator to produce a probability of 50% - the Discriminator should experience difficulty discerning real and generated samples. See figure 1 for a graphical overview.

$$V(D,G) = \mathbb{E}_x[logD(x)] + \mathbb{E}_z[log(1 - D(G(z)))] \tag{1}$$

2.3 Vanilla GAN Training Issues

2.3.1 Simulataneous Training Issues. GAN training aims to get $P_g \approx P_r$, that is to produce synthetic samples that are close to real samples. To quantify the distance between the distributions, Goodfellow et al. used the *Jensen-Shannon divergence* as a loss function [20]. The *Jensen-Shannon divergence* consists of a sigmoid-style curve resulting in gradients close to 0 for very large or very small loss values. The implication is that as the distance between generated samples increases, the gradient of the curve approaches zero stopping the learning process. Furthermore, as the distance between distributions increases, the Discriminators ability to discriminate data improves, approaching a perfect discriminator, that is $D(x) = 1 \forall x \in P_r$ and $D(x) = 0 \forall x \in P_g$. This leads to the *vanishing gradient problem* whereby the loss function approaches zero and the gradient of the loss function decreases, slowing learning to the point where the real and approximated distributions fail to converge.

2.3.2 Mode Dropping and Mode Collapse. GANs are notoriously susceptible to *mode collapse*, where the Generator continuously maps noise input to the same output, reducing the diversity of synthetic images [81] even though images may, visually, be of high quality. *Mode dropping* occurs when some hard-to-represent models are disregarded by the Generator [6], resulting in synthetic images that are of a high quality yet less general.

2.3.3 Leaky Augmentations. As discussed, Data Augmentations is a good way to increase the size of a dataset however training a GAN on data that underwent traditional data augmentation would cause the Generator to model the distribution of training data as if the augmentation is part of the original image. Therefore, generated images will contain the augmentation [29]. This is referred to as leaky augmentation.

2.3.4 Instability. The aforementioned issues contribute to potential instabilities that could be observed during the GAN training process. Various solutions have been proposed to address these issues including using smaller learning rates, using gradient clipping [5], using a more robust loss function [5, 22], using better activation functions such as ReLU and LeakyReLU [53, 79], using batch normalization on all layers excluding the Generator's output and the Discriminator's input layers or using stridden convolutions for downsampling instead of pooling layers [53].

2.4 WGAN

WGAN improves on the Vanilla GAN loss function [22] by using the Wasserstein distance to measure the distance between P_r and P_g. The Wasserstein distance effectively measures the minimum cost of transporting data to convert one distribution to another. The new loss function is given by Equation 2 - D is a continuous 1-Lipschitz function which is a learnt by the discriminator [22].

$$L = \mathbb{E}_{x \sim P_r}[D(x)] - \mathbb{E}_{G(x) \sim P_g}[D(G(x)] \tag{2}$$

The structure of the MLPs used in the Generator and Discriminator remains unchanged except for replacing the sigmoid activation function in the output layer of the Discriminator with a linear activation function. The sigmoid activation function limited output values to the range $[0, 1]$ whereas the WGAN implements a linear activation function capable of generating any real number, preventing vanishing gradients. Additionally, this can be considered a measurement of how *real* synthetic data is [74]. To prevent an *exploding gradient* problem, gradient clipping is applied to limit the Discriminator's values to $[-c, c]$, where c is a hyperparameter [5]. However, the choice of hyperparameter is crucial: a value too small slows learning whilst a value too large leads to an exploding gradient and a divergent model. Therefore, we select the approach of Gulrajani et al. [22] that introduces a *gradient penalty* to penalize the model if the gradient norm moves too far from the

target value of 1. The new loss function is given by equation 3 ($\hat{x}$ is obtained from a uniform distribution between points in P_r and P_g as discussed in [22]). The benefits of this include a more stable training process, decreased risk of vanishing gradient issues and less hyperparameter tuning. We use the default gradient penalty (λ) of 10 as proposed by [22].

$$L = \mathbb{E}_{x\sim P_r}[D(x)] - \mathbb{E}_{G(x)\sim P_g}[D(G(x))] + \lambda \mathbb{E}_{\hat{x}\sim \mathbb{P}(x)}[(||\nabla_{\hat{x}} D(\hat{x})||_2 - 1)^2] \tag{3}$$

Overall, WGANGP drastically improves the performance of Vanilla GAN through more stabilised training, a loss function that improves the behaviour of the gradient during backpropagation.

2.5 StyleGAN2

StyleGAN is currently the state-of-the-art GAN method for high-resolution image synthesis by leveraging a style-based approach. Most GAN models, including Vanilla GAN and WGAN, focus on the discriminator, the optimization of the loss function to ensure stability and hyperparameter tuning. For example, many advanced GAN models improve the discriminator [74] while neglecting the generator leaving them to "operate as black boxes" [30] such that much of the image generation process is poorly understood [30]. While conditional GANs help guide the generator in producing samples of the target domain, there is no way to control the generated images themselves, that is, the *style* of images. This includes elements such as the background and foreground of images as well as finer image details. Karras et al. (2019) [30] introduced a style-based generator to achieve this. A conventional GAN uses a random noise sample as input to the Generator, StyleGAN performs a non-linear transformation, through an 8-layer MLP, on the random noise input. This new latent variable undergoes an affine transformation into *styles* that controls Adaptive Instance Normalization (AdaIN) at each layer allowing each convolutional layer to learn a different style (input *A* in figure 2). Additionally, Gaussian noise input at each stage in the generation process allows for stochastic variation in images (input *B* in figure 2). StyleGAN uses the WGANGP loss (equation 3) as a loss function due to its convergence properties over the original GAN loss (equation 1) [30].

StyleGAN also introduces *style mixing* - a process whereby two latent codes are propagated through the Generator network with the generator switching between styles at random points. Since the transformed noise introduces style at each layer, style mixing allows for diverse and unique images. That is, the generator creates an image based on samples drawn from the distribution of each style. StyleGAN2 improved upon StyleGAN through the use of perceptual path length regularization to ensure consistency amongst generated shapes and remove any StyleGAN-specific characteristic artefacts. Bias and noise terms are now applied on normalized data instead of having bias and noise terms undergoing normalization as shown in figure 2. While state-of-the-art, the major drawback for its uptake in medical imaging is the computation power needed to train the model. Often, costs are reduced at the expense of image resolution [47, 49] however we show that, on our small medical dataset, StyleGAN2 training should be cut short to achieve optimal results reducing the impact of this issue.

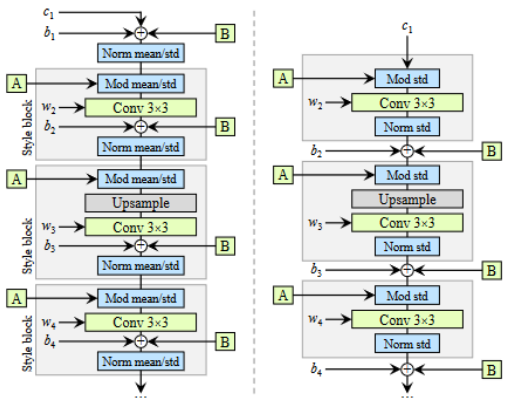


Figure 2: The architecture of StyleGAN [31]. On the left is the detailed architecture of StyleGAN. The AdaIN layers consist of the normalization and modulation layers. *A* is the *style* passed to each layer while *B* is a gaussian noise input for stochastic variation. On the right is StyleGAN2. For both models, the weights and bias terms are shown

2.6 Metrics

Generative models are typically evaluated through manual evaluation [17, 41, 68] however this process is cumbersome [8], biased towards the visual quality of images [8], subjective and produces variant results. Additionally, visual inspection may fail to detect nuances in the data meaning mode dropping will go undetected [8]. Furthermore, most GANs contain no objective loss function, making it difficult to compare models [58]. Quantitatively evaluating GANs is widely recognized to be challenging [80]. However, the *Frétchet Inception Distance* (FID), as proposed in [24], has become the standard for GAN evaluation due to its robustness and discriminability characteristics as discussed in [80]. FID scores use a pre-trained Inception Network [65], trained on the ImageNet dataset, to extract feature vectors for a synthetic and real image by removing the final classification layer of the network. The distance between the feature vectors is calculated using equation 4. The d^2 is the FID score, with a lower score indicating similar feature vectors and, thus, similar images. One should note that the applicability of FID to the medical discipline requires further investigation [70] as the ImageNet dataset it is trained on contains no medical data [11, 70, 76]. It has been experimentally shown by Woodland et al. (2022) [76] that there is a negative correlation between FID scores and human perceptual vision however the validity of the assumption requires further research. As such, we propose a domain-specific adaption of FID scores, which we dub MedFID, that captures the features unique to medical X-rays. We use the same pre-trained Inception V3 network as FID scores but add four additional fully connected layers, of which the final layer generates feature representations of an image for use in the FID calculation. We apply dropout at a rate of 0.2 to prevent overfitting. Lastly, a softmax classifier is used. The additional layers were first trained, keeping the base model frozen, on a dataset of X-rays of the upper and lower extremities. The entire model was then fine-tuned to improve classification accuracy. As supplementary work to this

study, a complete overview of the problem, motivations for the given approach and an empirical evaluation of the proposed metric are laid out in Appendix A. This provides a metric based on domain knowledge allowing for a more appropriate comparison between models in the field, as suggested by [62]. Like FID, MedFID is a comparative metric.

$$d^2((m, C), (m_w, C_w)) = ||m - m_w||_2^2 + Tr(C + C_w - 2(CC_w)^{1/2}) \quad (4)$$

3 EXPERIMENTAL SETUP

We evaluated each of the aforementioned models for a range of hyperparameter choices to determine the sensitivity of the quality of synthetic images to model configurations. The progress made by a GAN is heavily dependent on the choice of hyperparameters [41] hence the hyperparameter configurations used in this study come from optimal configurations in recent literature as recommended by [37]. Some tests were done using variations of suggested approaches to uncover deeper insights into how certain parameters influence generated images to aid in the hyperparameter tuning process in further studies. The table of hyperparameters for Vanilla GAN and WGANGP are presented in table 1 and table 2 respectively. All models were trained to generate images at a resolution of 256x256 to capture fine details present in our x-rays. A resolution larger than this is expected to generate unrealistic samples [59]. Lastly, the Adam optimizer was used in all experiments as it converges faster than Stochastic Gradient Descent [9].
All models are available on our GitHub repo [5].

3.1 Vanilla GAN

The experiments, and their respective hyperparameters, presented in table 1 were selected to validate recent literature as well as to evaluate the effect of the learning rate and optimizer hyperparameter configurations on GAN training. The base code is taken from GitHub [6].

Vanilla GAN				
Experiment	Lr	β_1	β_2	Reference
Vanilla1	0.002	0.5	0.999	Source implementation
Vanilla2	0.0002	0.5	0.9	Hu et al. [25]
Vanilla3	0.001	0.5	0.9	-
Vanilla4	0.001	0.5	0.999	-
Vanilla5	0.0001	0.5	0.9	Zhu et al. [86]

Table 1: Table of experiments for Vanilla GAN. All models use a batch size of 64.

3.2 WGANGP

The baseline implementation is that of Gulrajani et al. (2017) [22]. Our models, and related literature, keep the parameters of the Adam optimizer fixed at $b1 = 0.5$ and $b2 = 0.9$ suggesting that the model is stable at these values. We keep our models consistent with this notion and focus on varying the learning rate. The models are summarized in table 2. The base code is taken from GitHub [7].

WGAN-GP				
Experiment	Lr	β_1	β_2	Reference
WGANGP1	0.0001	0.5	0.9	Xiao et al. [78]
WGANGP2	0.00005	0.5	0.9	Gulrajani et al. [22]
WGANGP3	0.001	0.5	0.9	Kim et al. [34]
WGANGP4	0.0002	0.5	0.9	Hu et al. [25]
WGANGP5	0.002	0.5	0.9	-

Table 2: Table of experiments: WGAN-GP. All models use a batch size of 64 and a gradient clip value of 0.01

3.3 StyleGAN2

StyleGAN has proved to be invariant to the choice of hyperparameters as the latent noise fed in at different steps in the generation process allows the model to capture finer details without needing any hyperparameter searches or training [45, 76]. As such, we kept the baseline configurations with $b1 = 0$ and $b2 = 0.99$ for the Adam optimizer and a learning rate of 0.0025. The official PyTorch implementation of StyleGAN2-ADA [8] is used without ADA for these experiments.

3.4 Equipment

All models were run in PyTorch on the Center for High-Performance Computing cluster [9], using an Nvidia V100 GPU equipped with Cuda 11.6. We use Python 3.11 in a conda environment.

4 RESULTS

To evaluate the effectiveness of each GAN, the images generated by each GAN were evaluated using the MedFID score along with a visual analysis of generated images. In figures 3 and 7, we present the MedFID scores of each experiment over 8000 iterations. We then perform a side-by-side comparison of the MedFID score for images generated at 200 epochs for each Vanilla GAN and WGAN experiment in figure 6. This result for StyleGAN2 is covered in section 4.3. The statistical significance of the differences within experiments are presented in table 3. In figures 8 and 9 we present a series of images generated by Vanilla GAN and WGAN at 800, 5600 and 10800 iterations to monitor the training progress made by each GAN in terms of image quality. Image generation progress made by StyleGAN is shown in figure 10.

4.1 Vanilla GAN

In Figure 3 (left), we present the MedFID scores for each Vanilla GAN model over 8000 iterations. The model proposed by Zhu et al [86] performed the best while, overall, we see significant variations in MedFID scores highlighting the sensitivity to model parameters [58]. For example, a small change to the learning rate as observed when moving from Vanilla1 to Vanilla5 decreases the MedFID score

[5] https://github.com/Shaylin-UCT/DEEPPC
[6] https://github.com/eriklindernoren/PyTorch-GAN/blob/master/implementations/gan/gan.py
[7] https://github.com/eriklindernoren/PyTorch-GAN/blob/master/implementations/wgan_gp/wgan_gp.py
[8] https://github.com/NVlabs/stylegan2-ada-pytorch
[9] https://www.chpc.ac.za/

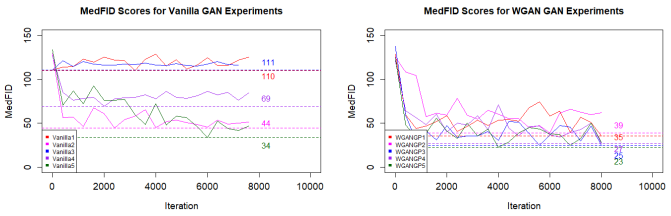


Figure 3: MedFID scores for the Vanilla GAN and WGAN experiments defined in table 1 and 2.

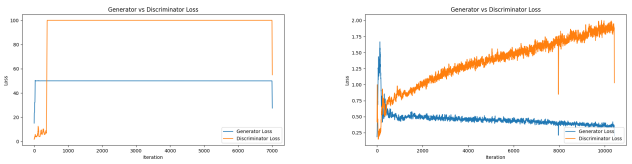


Figure 4: The generator and discriminator losses showing the effect of changing the learning rate from 0.001 (Vanilla3 - left) to 0.0001 (Vanilla5 - right). The increased learning rate, while holding Adam hyperparameters constant, enabled learning even though the model failed to converge.

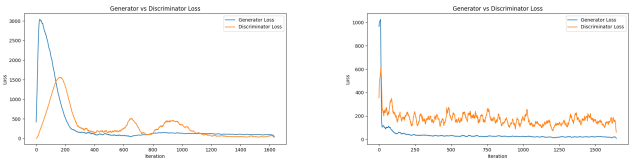


Figure 5: The generator and discriminator losses for WGANGP2 (left) and WGANGP3 (right).

by roughly 70 points. Each of the models are compared via pairwise t-tests in the table 3 (left). In figure 8, we present samples of images generated at set training intervals. Consistently, Vanilla GAN fails to produce recognisable imagery however the shape of the lateral elbow is captured indicating mode collapse: we capture and repeat broader details without adding depth and finer details.

4.2 WGANGP

In figure 3 (right), we present the MedFID scores of each WGANGP model over 8000 iterations. Visually, we see how each GAN has reached some form of convergence in figure 9, with the final MedFID scores being nearly identical for most models as seen by the table of statistical comparisons in table 3 (right). The added gradient penalty proves to effectively regularize the model while the new loss function improves stability and rids the model of the mode collapse experienced in Vanilla GAN.

In figure 6, we display the MedFID scores at 200 epochs for the different WGAN and Vanilla GAN models. The statistical tests shown in table 3.

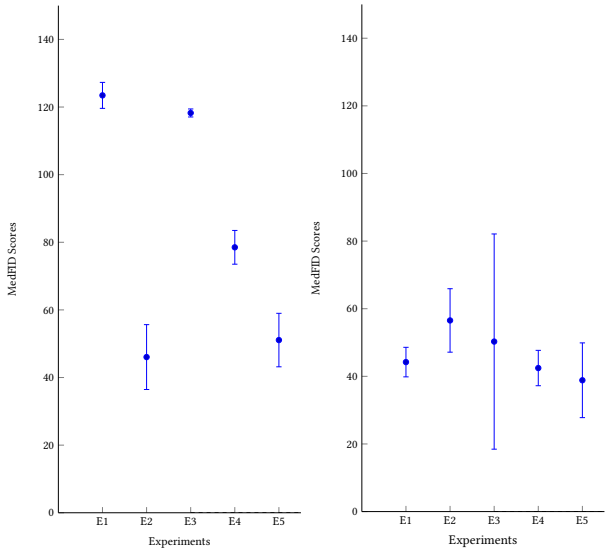


Figure 6: The MedFID scores at 200 epochs for Vanilla GAN (left) and WGAN (right) experiments. Standard deviations are added. As seen, WGAN is more stable than Vanilla GAN and is less sensitive to hyperparameters with few statistical differences between models.

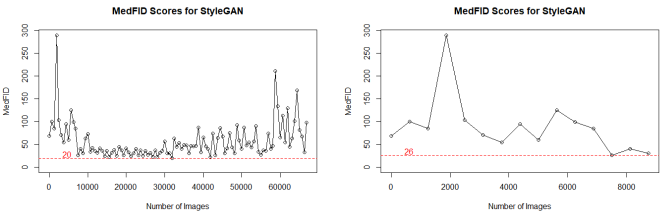


Figure 7: MedFID Scores for StyleGAN2 (left). With noticeable instabilities after ∼ 8400 iterations, we stop training and report the MedFID scores up to that point on the right.

4.3 StyleGAN2

In figure 7 (left), we present the MedFID scores during the learning process of StyleGAN2. As motivated in section 3.3, only one model was run. The model becomes progressively unstable once the distribution has been learnt, possibly indicating an attempt to overfit, as seen in appendix C. Over 60000 iterations, we reach a minimum MedFID of 20. We select to stop training at ∼ 8400 iterations as this is the point where training has stabilized. Beyond this, training oscillates slowly and becomes progressively worse. The minimum MedFID score is now 26 and on par with those of the best WGANGP models. This is surprising as the visual quality of StyleGAN2, as shown in figure 11, is superior to WGANGP. We discuss and hypothesise the reasons for this observation in the discussion. StyleGAN2 is successful in capturing finer image details, including the various orientations of images in the training dataset and image labels while producing diverse output. This, however, increases the MedFID score as we shall discuss.

	Vanilla2	Vanilla3	Vanilla4	Vanilla5
Vanilla1	✓	✓	✓	✓
Vanilla2	.	✓	✓	✗
Vanilla3		.	✓	✓
Vanilla4			.	✓
Vanilla5				.

	WGANGP2	WGANGP3	WGANGP4	WGANGP5
WGANGP1	✓	✗	✗	✗
WGANGP2	.	✗	✓	✓
WGANGP3		.	✗	✗
WGANGP4			.	✗
WGANGP5				.

	WGANGP1	WGANGP2	WGANGP3	WGANGP4	WGANGP5
StyleGAN2	✗	✓	✗	✗	✗

Table 3: Statistical differences of the Vanilla GAN and WGANGP experiments, as well as a StyleGAN vs WGANGP comparison, using pairwise t-tests with $H_0 : \mu_{MedFID_1} = \mu_{MedFID_2}$ evaluated at the 5% significance level.

5 DISCUSSION

In this section, we analyse the results presented above. We focus on each GAN individually and expose its pitfalls and strong points. In section 5.4, we present a recommendation on the most suitable GAN for medical image synthesization and provide a few recommendations for improvements.

5.1 Vanilla GAN

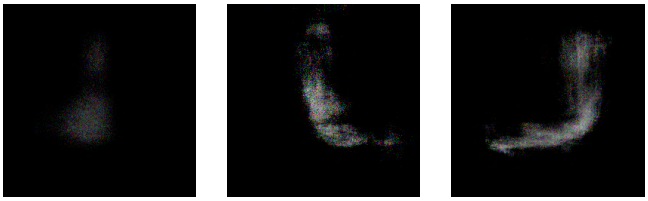
Through our experiments, we have shown that all of the Vanilla GAN training problems discussed in 2.3 are present when applied to medical datasets. In figure 3, the behaviour of the models differ substantially. VanillaE1 and VanillaE3 never converge but rather oscillate around its starting MedFID value indicating that the model has failed to learn sufficient patterns in the data. The other experiments, however, see rapidly decreasing MedFID scores as the generator's approximation of the real data distribution improves. However, at around 1000 iterations, the MedFID scores plateau and begin to oscillate indicating that learning has stopped and that the training has now become unstable. This can also be seen in figure 8, where the generated images remain relatively unchanged between 800 and 10800 iterations. The general trend is that a lower learning rate decreases the MedFID scores as the model captures more intricacies in the data. For example, when changing the learning rate from 0.001 (Vanilla3) to 0.0001 (Vanilla5), holding the Adam parameters constant, the MedFID score decreases by ~ 70 points (p-value<0.05) and the model no longer collapses as seen in figure 4. While training doesn't converge with the updated learning rate, the model now begins to learn image features. Changing to a slightly higher learning rate of 0.0002 (Vanilla2) achieves the same result (p-value < 0.05) providing clear evidence for a smaller learning rate. The choice of the β_2 parameter of the Adam optimizer is also crucial. We fix the learning rate at 0.001 but change the β_2 value, as in the case of Vanilla3 ($\beta_2 = 0.9$) and Vanilla4 ($\beta_2 = 0.999$). Under this change, the higher β_2 value leads to a ~ 40 point improvement in the MedFID score (p-value < 0.05). However, the visual quality of both experiments (figure 8 (c) and (d)) are poor. At $\beta_2 = 0.999$, the model fails to capture any semantics of the training images and generates blank output indicating that no learning has been done. $\beta_2 = 0.9$ captures some pixels however generated objects are spread around the image with no clear elbow pattern present. These results are summarized in figure 6 where the range of MedFID values for Vanilla GAN clearly indicate the variations in model performance as a result of slightly different hyperparameter choices. This supports the well-known notion that Vanilla GAN is extremely sensitive to the choice of hyperparameters [41]. It is this training instabilities and hyperparameter complexity that led to the development of more mature GAN variants in most application domains. Lastly, Vanilla GAN fails to produce any recognisable imagery, often resorting to blocks of pixels towards the centre of the screen in an elbow pattern. The capturing of the general pattern suggests mode dropping, where details beyond the shape are ignored by the Generator.

As expected, vanilla GAN performs poorly on medical data which aligns with recent literature that supports style-based or convolutional-based models. However, the optimal Vanilla GAN model appears to be Vanilla2 and Vanilla5, noting the lack of statistical difference between the two.

5.2 WGANGP

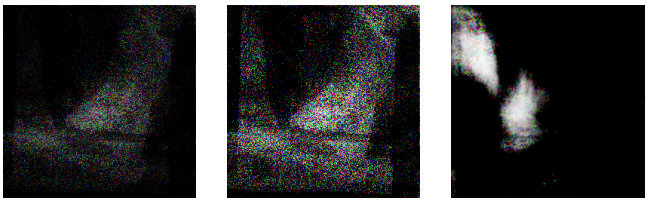
Unlike Vanilla GAN, WGANGP models consistently pick up broad features such as shape and depth however, visually, the images are still of low quality. The progression of WGANGP learning, experimentally, seems to move from smudges to more distinct elements as training evolves. This can be attributed to the improved loss function with gradient penalty that carefully controls the gradient, preventing the gradient from misbehaving, that is prevents vanishing and exploding gradients. This stability is seen in figure 9, where all WGANGP models learn features. However, as in the case of Vanilla GAN, training oscillates after ~ 1000 iterations.

The effect of learning rates is discussed next. As mentioned, we keep the parameters of the Adam optimizer fixed at $\beta_1 = 0.5, \beta_2 = 0.9$. In general, a lower learning rate improves the MedFID score. Large learning rates (0.001 and 0.002) achieve similar MedFID scores (p-value<0.05) to lower learning rates however are much more variant in their output. The learning rate also prevents the model from entirely converging as seen in figure 5 (right). This results in the model being unable to capture significant details of the image. This is visually demonstrated in part (c) of figure 9: these learning rates often produce the correct shape but with extra emphasis concentrated in small parts of the image as seen by the white blob towards the centre of the image. Slowing the learning rate down too much, as in the case of WGANGP2 (lr = 0.00005), results in training taking too long to produce images of the quality experienced by higher learning rates (part (b) of figure 9). Doubling the learning rate decreases

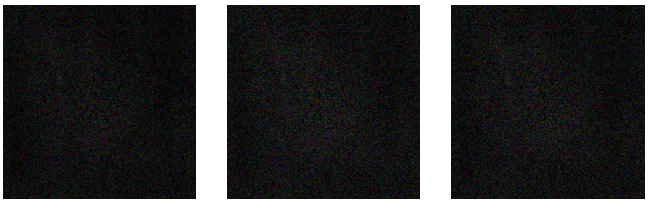


(a) Vanilla1: $lr = 0.002, \beta_1 = 0.5, \beta_2 = 0.999$

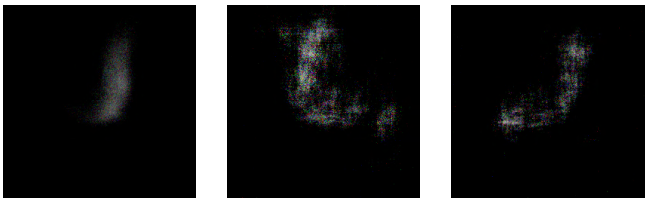
(b) Vanilla2: $lr = 0.0002, \beta_1 = 0.5, \beta_2 = 0.9$



(c) Vanilla3: $lr = 0.001, \beta_1 = 0.5, \beta_2 = 0.9$



(d) Vanilla4: $lr = 0.001, \beta_1 = 0.5, \beta_2 = 0.999$

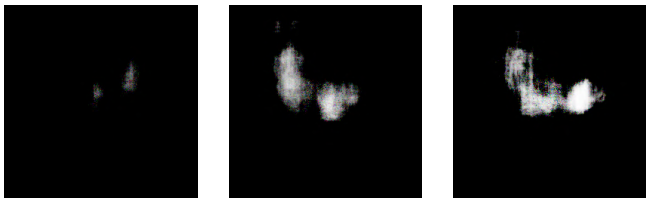


(e) Vanilla5: $lr = 0.0001, \beta_1 = 0.5, \beta_2 = 0.9$

Figure 8: Sample output from Vanilla GAN training for 800, 5600 and 10800 iterations (left to right).

(a) WGANGP1: $lr = 0.0001, \beta_1 = 0.5, \beta_2 = 0.9$

(b) WGANGP2: $lr = 0.00005, \beta_1 = 0.5, \beta_2 = 0.9$



(c) WGANGP3: $lr = 0.001, \beta_1 = 0.5, \beta_2 = 0.9$

(d) WGANGP4: $lr = 0.0002, \beta_1 = 0.5, \beta_2 = 0.9$

(e) WGANGP5: $lr = 0.002, \beta_1 = 0.5, \beta_2 = 0.9$

Figure 9: Sample output from WGANGP training for 800, 5600 and 10800 iterations (left to right).

the MedFID score by $\sim$ 12 points to $\sim$ 44.2, a clear improvement in image quality (p-value<0.05). It is worth noting that this model (lr = 0.00005) is statistically different to all models besides WGANGP3. We predict that this is due to the higher learning rate of 0.001 in WGANGP3 capturing fewer details of the image as seen in part (c) of figure 9.

Additionally, WGANGP1 and WGANGP2 move away from an optimal MedFID value and converge again suggesting that an optimal minimum was achieved but the lower learning rate captured finer details over a period of time, throwing the MedFID scores off a minimum but allowing it to slowly return as more fine details are captured. This is seen by the fluctuating discriminator loss in figure 5 (left).

Notibly, learning rates $lr \in [0.0002, 0.002]$ show no statistical differences indicating that WGANGP is invariant to larger learning rates however a learning rate of 0.001, although statistically indifferent, produces much more variant results. This is further motivated by the lack of statistically significant differences when compared to a model with a learning rate 20 times smaller (WGANGP2) Hence, optimal learning rates for WGANGP on our dataset seem to be

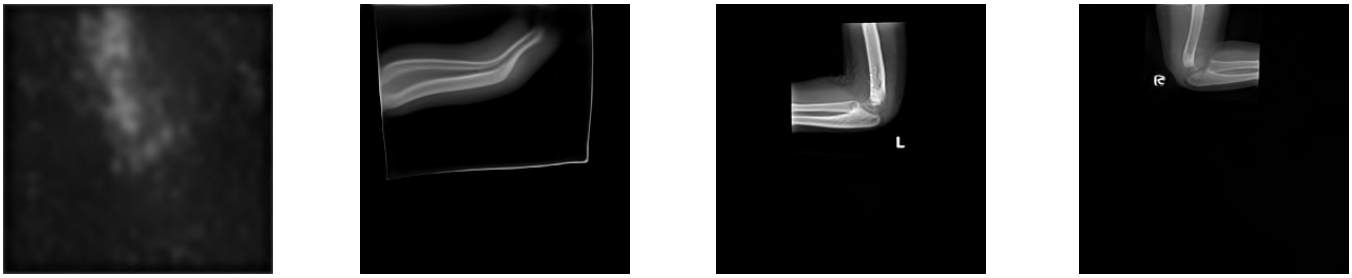


Figure 10: Image demonstrating the progress of StyleGAN2 (random seed values used for each generation)

Figure 11: A selection of generated images showing some of the features generated due to dataset inconsistency. From left to right observe cropped images with white borders, cropped images, shrunken images and rotated images. These inflate the MedFID score as these rare occurrences in the database filter into most generated images, deteriorating the similarity between real and generated images.

0.002 and 0.0002.

While an optimal learning rate has been found, a visual inspection of images shows that the generated samples are not of sufficient quality to be used in practice. The mode collapse experienced in Vanilla GAN is no longer present and the images contain more depth, owing to the improved loss function. That is, textures and shapes become more apparent however the results are still not comparable to real imagery.

5.3 StyleGAN2

By visual inspection, the quality of images generated by StyleGAN2 is far superior to those generated by WGANGP and Vanilla GAN, as shown in figure 11. Items such as bone structures and joints are now clearly present in each image. We predict that each noise sample inserted at each step in the generation process allows the model to capture finer details of images such as the "L" and "R" markings and the blank space around the images. However, in light of this, the MedFID scores are similar to that of the best-performing WGANGP model identified in section 5.2, with no statistical difference between StyleGAN2 and WGANGP4 ($lr = 0.0002$) and WGANGP5 ($lr = 0.002$) at the 5% significance level. A visual inspection of images reveals why this is the case: StyleGAN2 captures each feature in images across the dataset and replicates this in generated images. This results in features only present in a few images being replicated, and sometimes merged, into other images, indicating some degree of overfitting. This is evidence of StyleGAN2's ability to capture finer image details, making it an attractive model as we fully learn the training distribution. However, our dataset is not consistent, that is, the elements are heterogeneous as seen in Appendix B. The high variability of the data leads to training instabilities regardless of image resolution [85]. Therefore, given the constraints of a non-consistent dataset, StyleGAN2 is visually better, and on par in terms of MedFID scores, when compared to WGANGP. However, if the dataset were consistent, StyleGAN2 is expected to out-perform WGAN [29, 35, 42].

The training of StyleGAN2 is much slower than that of WGANGP and Vanilla GAN due to increased regularization and additional layers in the mapping and image synthesis process [3] presenting a tradeoff between computational complexity and image quality.

StyleGAN2 capturing each feature is advantageous however it poses a risk to its application in pathology as it would require standardized images to be useful. Given the small-size, sparse-nature limitations of medical datasets, including the one used in this study, this is not always the case. These shadows doubt on the usage of StyleGAN2 in current medical practices. Future and ongoing work at Groote Schuur Hospital includes the derivation of larger standardized datasets being manually devised and annotated however this is a slow and laborious process.

5.4 Summary

Our experiments support the notorious claim that training a GAN is difficult [50] and would require extensive hyperparameter tuning and other *tricks* to achieve success [37]. Vanilla GAN consistently fails to produce good imagery and performs much worse than the WGANGP and StyleGAN2 providing clear evidence that more complexity is needed in the model. WGANGP, at a learning rate in the

range of 0.002 and 0.0002, performs well - these produce low MedFID scores and capture the shape of the elbow with depth but the quality of images is subpar and not suitable for downstream tasks. StyleGAN2 proves to be the best-performing model. Contrasting StyleGAN2 and WGANGP, it is clear that the image quality for StyleGAN2 is superior, and on par with that of medical imagery, even if the MedFID scores are similar, as discussed above. Therefore, StyleGAN2 is the better choice of model [30] for our use case. However, there are two major drawbacks to StyleGAN2, as exposed in this study: Firstly, StyleGAN2 is computationally more expensive than WGANGP but offers better quality images [3, 42]. Secondly, StyleGAN2's ability to recall fine image details which manifests itself as overfitting on small datasets. The lack of consistency in the dataset forces StyleGAN2 to recall all image details with variation only existing in the placement of artefacts. As discussed in section 5.3, given a consistent dataset StyleGAN is expected to perform better than WGANGP. At present, little is known about the image generation process [3] but StyleGAN2 enables researchers to implicitly control image quality and the nature of outputs through the injection of *styles* throughout the generation process. Given the small nature and lack of consistency in small datasets, the performance of StyleGAN2 provides evidence that generating medical imagery using GANs is possible however auxiliary work needs to be completed.

6 CONCLUSIONS AND FUTURE WORK

The recent excitement surrounding generative AI, particularly GANs, have propagated into the field of pathology. In this study, we have provided empirical evidence supporting style-based models in medical image generation over models that optimize a loss function or improve the discriminator. These models capture finer image details however do tend to overfit on small, inconsistent datasets. However, our StyleGAN2 experiments demonstrate that high-quality image generation is possible on small datasets (<3000 images). However, to improve the quality of images, there needs to be data consistency measures put in place. Overall, there is promise for GANs in augmenting medical datasets. However, at present, neither of the tested models is suitable for deployment without auxiliary work being done as the risks involved with inaccuracies in medical-focused deep learning models require GANs to produce images as realistic as possible before they can be implemented [27]. We have, however, proved that there is great potential for StyleGAN2 but there needs to be initiatives that promote the curation of large, consistent datasets.

It is important to note that while we applied the selected models for the entire dataset set, to overcome a class imbalance, models should be conditioned on minority classes with the generated images added to the minority class. This was not done in this study due to the small size of the dataset available.

The main contributions of this paper are:

- We have shown that StyleGAN2 performs well on small medical datasets but is limited by inconsistent data. Should data collection and consistency improve, StyleGAN2 proves to be a good foundation for future research.
- We show that the training time for StyleGAN2 on small medical datasets should be cut short for optimal results as the model becomes unstable at higher epochs.
- We present a novel X-ray-orientated metric for objective synthetic x-ray evaluation.

As a foundation for future work, we recommend that the above experiments be repeated with consistent datasets to confirm the hypothesis of inconsistency causing StyleGAN2 issues. Further tests would have to be done on MedFID to ensure its applicability to the field. Generating images of higher resolution would be ideal to pick up finer details. In our study, this was prevented by GPU capacity. Future work should exploit computational power of better GPUs or look into course-to-fine-grained [46] or progressively growing GANs [60]. The use of advanced regularization methods that increase the probability of model convergence [73, 82, 83] should also be explored. We also suggest that applying transfer-learning to GANs in the medical domain be explored further due to the expected performance increase when faced with limited data [2, 76].

7 ACKNOWLEDGEMENTS

This work is based on the research supported in part by the National Research Foundation of South Africa (Reference number: PMDS22071841724).

REFERENCES

[1] Ibrahim Saad Aly Abdelhalim, Mamdouh Farouk Mohamed, and Yousef Bassyouni Mahdy. 2021. Data augmentation for skin lesion using selfattention based progressive generative adversarial network. *Expert Systems with Applications* 165 (March 2021), 113922. https://doi.org/10.1016/j.eswa.2020.113922

[2] Harold Achicanoy, Deisy Chaves, and Maria Trujillo. 2021. StyleGANs and Transfer Learning for Generating Synthetic Images in Industrial Applications. *Symmetry* 13, 8 (Aug. 2021), 1497. https://doi.org/10.3390/sym13081497

[3] Hamed Amiri, Ivan Vasconcelos, Yang Jiao, Pei-En Chen, and Oliver Plümper. 2023. Quantifying microstructures of earth materials using higher-order spatial correlations and deep generative adversarial networks. *Scientific Reports* 13, 1 (Jan. 2023), 1805. https://doi.org/10.1038/s41598-023-28970-w

[4] I. Andreou and N. Mouelle. 2023. Evaluating Generative Adversarial Networks for particle hit generation in a cylindrical drift chamber using Fréchet Inception Distance. *Journal of Instrumentation* 18, 06 (June 2023), P06007. https://doi.org/10.1088/17480221/18/06/P06007

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. http://arxiv.org/abs/1701.07875 arXiv:1701.07875 [cs, stat].

[6] Sanjeev Arora and Yi Zhang. 2017. Do GANs actually learn the distribution? An empirical study. http://arxiv.org/abs/1706.08224 arXiv:1706.08224 [cs].

[7] Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. 2019. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association* 26, 3 (March 2019), 228–241. https://doi.org/10.1093/jamia/ocy142

[8] Ali Borji. 2018. Pros and Cons of GAN Evaluation Measures. http://arxiv.org/abs/1802.03446 arXiv:1802.03446 [cs].

[9] Lorenzo Brigato and Luca Iocchi. 2021. A Close Look at Deep Learning with Small Data. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 24902497. https://doi.org/10.1109/ICPR48806.2021.9412492 ISSN: 10514651.

[10] John E. Burkhardt, Karamjeet Pandher, Phillip F. Solter, Sean P. Troth, Rogely Waite Boyce, Tanja S. Zabka, and Daniela Ennulat. 2011. Recommendations for the Evaluation of Pathology Data in Nonclinical Safety Biomarker Qualification Studies. *Toxicologic Pathology* 39, 7 (Dec. 2011), 11291137. https://doi.org/10.1177/0192623311422082

[11] Junxiao Chen, Jia Wei, and Rui Li. 2021. TarGAN: TargetAware Generative Adversarial Networks for Multimodality Medical Image Translation. http://arxiv.org/abs/2105.08993 arXiv:2105.08993 [cs, eess].

[12] Mingyu Chen, Bin Zhang, Win Topatana, Jiasheng Cao, Hepan Zhu, Sarun Juengpanich, Qijiang Mao, Hong Yu, and Xiujun Cai. 2020. Classification and mutation prediction based on histopathology H&E images in liver cancer using deep learning. *npj Precision Oncology* 4, 1 (June 2020), 14. https://doi.org/10.1038/s4169802001203

[13] Maria J. M. Chuquicusma, Sarfaraz Hussein, Jeremy Burt, and Ulas Bagci. 2018. How to Fool Radiologists with Generative Adversarial Networks? A Visual Turing Test for Lung Cancer Diagnosis. http://arxiv.org/abs/1710.09762 arXiv:1710.09762 [cs, qbio].

[14] Kenneth W. Clark, Bruce A. Vendt, Kirk E. Smith, John B. Freymann, Justin S. Kirby, Paul Koppel, Stephen M. Moore, Stanley R. Phillips, David R. Maffitt, Michael Pringle, Lawrence Tarbox, and Fred W. Prior. 2013. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *J. Digit. Imaging* 26, 6 (2013), 1045–1057. https://doi.org/10.1007/s10278-013-9622-7

[15] Saloni Dash, Andrew Yale, Isabelle Guyon, and Kristin P. Bennett. 2020. Medical Time-Series Data Generation Using Generative Adversarial Networks. In *Artificial Intelligence in Medicine*, Martin Michalowski and Robert Moskovitch (Eds.). Vol. 12299. Springer International Publishing, Cham, 382–391. https://doi.org/10.1007/978-3-030-59137-3_34 Series Title: Lecture Notes in Computer Science.

[16] Luca Deininger, Bernhard Stimpel, Anil Yuce, Samaneh AbbasiSureshjani, Simon Schönenberger, Paolo Ocampo, Konstanty Korski, and Fabien Gaire. 2022. A comparative study between vision transformers and CNNs in digital pathology. http://arxiv.org/abs/2206.00389 arXiv:2206.00389 [cs, eess].

[17] Emily Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. 2015. Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks. http://arxiv.org/abs/1506.05751 arXiv:1506.05751 [cs].

[18] Serge Dolgikh. 2021. *Analysis and Augmentation of Small Datasets with Unsupervised Machine Learning*. preprint. Health Informatics. https://doi.org/10.1101/2021.04.21.21254796

[19] Al gharakhanian. [n. d.]. Generative adversarial networks - hot topic in machine learning. https://www.kdnuggets.com/2017/01/generative-adversarial-networks-hot-topic-machine-learning.html

[20] Ian J. Goodfellow, Jean PougetAbadie, Mehdi Mirza, Bing Xu, David WardeFarley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. http://arxiv.org/abs/1406.2661 arXiv:1406.2661 [cs, stat].

[21] Qing Guan, Xiaochun Wan, Hongtao Lu, Bo Ping, Duanshu Li, Li Wang, Yongxue Zhu, Yunjun Wang, and Jun Xiang. 2019. Deep convolutional neural network Inceptionv3 model for differential diagnosing of lymph node in cytological images: a pilot study. *Annals of Translational Medicine* 7, 14 (July 2019), 307307. https://doi.org/10.21037/atm.2019.06.29

[22] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved Training of Wasserstein GANs. http://arxiv.org/abs/1704.00028 arXiv:1704.00028 [cs, stat].

[23] Tianyu Han, Sven Nebelung, Christoph Haarburger, Nicolas Horst, Sebastian Reinartz, Dorit Merhof, Fabian Kiessling, Volkmar Schulz, and Daniel Truhn. 2020. Breaking medical data sharing boundaries by using synthesized radiographs. *Science Advances* 6, 49 (Dec. 2020), eabb7973. https://doi.org/10.1126/sciadv.abb7973

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2018. GANs Trained by a Two TimeScale Update Rule Converge to a Local Nash Equilibrium. http://arxiv.org/abs/1706.08500 arXiv:1706.08500 [cs, stat].

[25] Bo Hu, Ye Tang, Eric I.Chao Chang, Yubo Fan, Maode Lai, and Yan Xu. 2019. Unsupervised Learning for Celllevel Visual Representation in Histopathology Images with Generative Adversarial Networks. *IEEE Journal of Biomedical and Health Informatics* 23, 3 (May 2019), 13161328. https://doi.org/10.1109/JBHI.2018.2852639 arXiv:1711.11317 [cs].

[26] Talha Iqbal and Hazrat Ali. 2018. Generative Adversarial Network for Medical Images (MI-GAN). *Journal of Medical Systems* 42, 11 (Nov. 2018), 231. https://doi.org/10.1007/s10916-018-1072-9

[27] Miso Jang, Hyun-jin Bae, Minjee Kim, Seo Young Park, A-yeon Son, Se Jin Choi, Jooae Choe, Hye Young Choi, Hye Jeon Hwang, Han Na Noh, Joon Beom Seo, Sang Min Lee, and Namkug Kim. 2023. Image Turing test and its applications on synthetic chest radiographs by using the progressive growing generative adversarial network. *Scientific Reports* 13, 1 (Feb. 2023), 2356. https://doi.org/10.1038/s41598-023-28175-1

[28] Jiwoong J. Jeong, Amara Tariq, Tobiloba Adejumo, Hari Trivedi, Judy W. Gichoya, and Imon Banerjee. 2022. Systematic Review of Generative Adversarial Networks (GANs) for Medical Image Classification and Segmentation. *Journal of Digital Imaging* 35, 2 (April 2022), 137152. https://doi.org/10.1007/s1027802100556w

[29] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training Generative Adversarial Networks with Limited Data. http://arxiv.org/abs/2006.06676 arXiv:2006.06676 [cs, stat].

[30] Tero Karras, Samuli Laine, and Timo Aila. 2019. A StyleBased Generator Architecture for Generative Adversarial Networks. http://arxiv.org/abs/1812.04948 arXiv:1812.04948 [cs, stat].

[31] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. http://arxiv.org/abs/1912.04958 arXiv:1912.04958 [cs, eess, stat].

[32] Parvinder Kaur, Baljit Singh Khehra, and Er. Bhupinder Singh Mavi. 2021. Data Augmentation for Object Detection: A Review. In *2021 IEEE International Midwest Symposium on Circuits and Systems (MWSCAS)*. IEEE, Lansing, MI, USA, 537543. https://doi.org/10.1109/MWSCAS47672.2021.9531849

[33] Justin Ker, Yeqi Bai, Hwei Yee Lee, Jai Rao, and Lipo Wang. 2019. Automated brain histology classification using machine learning. *Journal of Clinical Neuroscience* 66 (Aug. 2019), 239245. https://doi.org/10.1016/j.jocn.2019.05.019

[34] Byeongjoon Kim, Minah Han, Hyunjung Shim, and Jongduk Baek. 2019. A performance comparison of convolutional neural network-based image denoising methods: The effect of loss functions on low-dose CT images. *Medical Physics* 46, 9 (Sept. 2019), 39063923. https://doi.org/10.1002/mp.13713

[35] Hiba Kobeissi, Saeed Mohammadzadeh, and Emma Lejeune. 2022. Enhancing Mechanical Metamodels With a Generative Model-Based Augmented Training Dataset. *Journal of Biomechanical Engineering* 144, 12 (Dec. 2022), 121002. https://doi.org/10.1115/1.4054898

[36] Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot. 2021. SelfPath SelfSupervision for Classification of Pathology Images With Limited Annotations. *IEEE Transactions on Medical Imaging* 40, 10 (Oct. 2021), 28452856. https://doi.org/10.1109/TMI.2021.3056023

[37] Karol Kurach, Mario Lucic, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. 2019. A LargeScale Study on Regularization and Normalization in GANs. http://arxiv.org/abs/1807.04720 arXiv:1807.04720 [cs, stat].

[38] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. 2023. The Role of ImageNet Classes in Fr\'echet Inception Distance. http://arxiv.org/abs/2203.06026 arXiv:2203.06026 [cs, stat].

[39] Ying-Jia Lin and I-Fang Chung. 2019. Medical Data Augmentation Using Generative Adversarial Networks : X-ray Image Generation for Transfer Learning of Hip Fracture Detection. In *2019 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*. IEEE, Kaohsiung, Taiwan, 1–5. https://doi.org/10.1109/TAAI48200.2019.8959908

[40] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. 2017. A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis* 42 (Dec. 2017), 6088. https://doi.org/10.1016/j.media.2017.07.005 arXiv:1702.05747 [cs].

[41] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. 2018. Are GANs Created Equal? A LargeScale Study. http://arxiv.org/abs/1711.10337 arXiv:1711.10337 [cs, stat].

[42] Jose Maureira, Juan Tapia, Claudia Arellano, and Christoph Busch. 2021. Synthetic Periocular Iris PAI from a Small Set of Near Infrared Images. http://arxiv.org/abs/2107.12014 arXiv:2107.12014 [cs].

[43] Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. http://arxiv.org/abs/1411.1784 arXiv:1411.1784 [cs, stat].

[44] Takeru Miyato and Masanori Koyama. 2018. cGANs with Projection Discriminator. http://arxiv.org/abs/1802.05637 arXiv:1802.05637 [cs, stat].

[45] Anwesha Mohanty, Alistair Sutherland, Marija Bezbradica, and Hossein Javidnia. 2023. High Fidelity Synthetic Face Generation for Rosacea Skin Condition from Limited Data. http://arxiv.org/abs/2303.04839 arXiv:2303.04839 [cs, eess].

[46] Tony C. W. Mok and Albert C. S. Chung. 2019. Learning Data Augmentation for Brain Tumor Segmentation with CoarsetoFine Generative Adversarial Networks. Vol. 11383. 7080. https://doi.org/10.1007/97830301172387 arXiv:1805.11291 [cs].

[47] Alberto Montero, Elisenda BonetCarne, and Xavier Paolo BurgosArtizzu. 2021. Generative Adversarial Networks to Improve Fetal Brain FineGrained Plane Classification. *Sensors* 21, 23 (Nov. 2021), 7975. https://doi.org/10.3390/s21237975

[48] Mai Feng Ng and Carol Anne Hargreaves. 2023. Generative Adversarial Networks for the Synthesis of Chest X-ray Images. In *The 3rd International Electronic Conference on Applied Sciences*. MDPI, 84. https://doi.org/10.3390/ASEC2022-13954

[49] Milda Pocevičiūtė, Gabriel Eilertsen, and Claes Lundström. 2021. Unsupervised anomaly detection in digital pathology using GANs. http://arxiv.org/abs/2103.08945 arXiv:2103.08945 [cs, eess].

[50] Ben Poole, Alexander A. Alemi, Jascha SohlDickstein, and Anelia Angelova. 2016. Improved generator objectives for GANs. http://arxiv.org/abs/1612.02780 arXiv:1612.02780 [cs, stat].

[51] Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Günter Klambauer. 2018. Fr\'echet ChemNet Distance: A metric for generative models for molecules in drug discovery. http://arxiv.org/abs/1803.09518 arXiv:1803.09518 [cs, qbio, stat].

[52] Adalberto Claudio Quiros, Roderick Murray-Smith, and Ke Yuan. 2021. PathologyGAN: Learning deep representations of cancer tissue. http://arxiv.org/abs/1907.02644 arXiv:1907.02644 [cs, eess, stat].

[53] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. http://arxiv.org/abs/1511.06434 arXiv:1511.06434 [cs].

[54] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L. Ball, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. 2018. MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs. http://arxiv.org/abs/1712.06957 arXiv:1712.06957 [physics].

[55] Priyanka Rana, Arcot Sowmya, Erik Meijering, and Yang Song. 2022. *Data augmentation for imbalanced blood cell image classification.* preprint. Bioinformatics. https://doi.org/10.1101/2022.08.30.505762

[56] Jakob Fraes Rasmussen, Volkert Siersma, Jessica Malmqvist, and John Brodersen. 2020. Psychosocial consequences of false positives in the Danish Lung Cancer CT Screening Trial: a nested matched cohort study. *BMJ Open* 10, 6 (June 2020), e034682. https://doi.org/10.1136/bmjopen-2019-034682

[57] Hojjat Salehinejad, Shahrokh Valaee, Tim Dowdell, Errol Colak, and Joseph Barfett. 2018. Generalization of Deep Neural Networks for Chest Pathology Classification in XRays Using Generative Adversarial Networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Calgary, AB, 990994. https://doi.org/10.1109/ICASSP.2018.8461430

[58] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved Techniques for Training GANs. http://arxiv.org/abs/1606.03498 arXiv:1606.03498 [cs].

[59] August DuMont Schütte, Jürgen Hetzel, Sergios Gatidis, Tobias Hepp, Benedikt Dietz, Stefan Bauer, and Patrick Schwab. 2021. Overcoming Barriers to Data Sharing with Medical Image Generation: A Comprehensive Evaluation. http://arxiv.org/abs/2012.03769 arXiv:2012.03769 [cs, eess].

[60] Bradley Segal, David M. Rubin, Grace Rubin, and Adam Pantanowitz. 2021. Evaluating the clinical realism of synthetic chest x-rays generated using progressively growing Gans. *SN Computer Science* 2, 4 (2021). https://doi.org/10.1007/s42979-021-00720-7

[61] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. 2018. How good is my GAN? http://arxiv.org/abs/1807.09499 arXiv:1807.09499 [cs].

[62] Youssef Skandarani, PierreMarc Jodoin, and Alain Lalande. 2021. GANs for Medical Image Synthesis: An Empirical Study. http://arxiv.org/abs/2105.05318 arXiv:2105.05318 [cs, eess].

[63] Heithem Sliman, Imen Megdiche, Loay Alajramy, Adel Taweel, Sami Yangui, Aida Drira, and Elyes Lamine. 2023. MedWGAN based synthetic dataset generation for Uveitis pathology. *Intelligent Systems with Applications* 18 (May 2023), 200223. https://doi.org/10.1016/j.iswa.2023.200223

[64] Pooja Subramaniam, Tabea Kossen, Kerstin Ritter, Anja Hennemuth, Kristian Hildebrand, Adam Hilbert, Jan Sobesky, Michelle Livne, Ivana Galinovic, Ahmed A. Khalil, Jochen B. Fiebach, Dietmar Frey, and Vince I. Madai. 2022. Generating 3D TOFMRA volumes and segmentation labels using generative adversarial networks. *Medical Image Analysis* 78 (May 2022), 102396. https://doi.org/10.1016/j.media.2022.102396

[65] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going Deeper with Convolutions. http://arxiv.org/abs/1409.4842 arXiv:1409.4842 [cs].

[66] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. http://arxiv.org/abs/1512.00567 arXiv:1512.00567 [cs].

[67] Luke Taylor and Geoff Nitschke. 2018. Improving Deep Learning with Generic Data Augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, Bangalore, India, 15421547. https://doi.org/10.1109/SSCI.2018.8628742

[68] Lucas Theis, Aäron van den Oord, and Matthias Bethge. 2016. A note on the evaluation of generative models. http://arxiv.org/abs/1511.01844 arXiv:1511.01844 [cs, stat].

[69] Anna N. A. Tosteson, Dennis G. Fryback, Cristina S. Hammond, Lucy G. Hanna, Margaret R. Grove, Mary Brown, Qianfei Wang, Karen Lindfors, and Etta D. Pisano. 2014. Consequences of False-Positive Screening Mammograms. *JAMA Internal Medicine* 174, 6 (June 2014), 954. https://doi.org/10.1001/jamainternmed.2014.981

[70] Lorenzo Tronchin, Rosa Sicilia, Ermanno Cordelli, Sara Ramella, and Paolo Soda. 2021. Evaluating GANs in Medical Imaging. In *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections*, Sandy Engelhardt, Ilkay Oksuz, Dajiang Zhu, Yixuan Yuan, Anirban Mukhopadhyay, Nicholas Heller, Sharon Xiaolei Huang, Hien Nguyen, Raphael Sznitman, and Yuan Xue (Eds.). Vol. 13003. Springer International Publishing, Cham, 112121. https://doi.org/10.1007/978303088210510 Series Title: Lecture Notes in Computer Science.

[71] Maximilian E. Tschuchnig, Gertie J. Oostingh, and Michael Gadermayr. 2020. Generative Adversarial Networks in Digital Pathology: A Survey on Trends and Future Potential. *Patterns* 1, 6 (Sept. 2020), 100089. https://doi.org/10.1016/j.patter.2020.100089

[72] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. 2018. Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience* 2018 (2018), 1–13. https://doi.org/10.1155/2018/7068349

[73] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. 2013. Regularization of Neural Networks using DropConnect. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 28)*, Sanjoy Dasgupta and David McAllester (Eds.). PMLR, Atlanta, Georgia, USA, 1058–1066. https://proceedings.mlr.press/v28/wan13.html

[74] Lilian Weng. 2019. From GAN to WGAN. http://arxiv.org/abs/1904.08994 arXiv:1904.08994 [cs, stat].

[75] Jelmer M. Wolterink, Anna M. Dinkla, Mark H. F. Savenije, Peter R. Seevinck, Cornelis A. T. van den Berg, and Ivana Isgum. 2017. Deep MR to CT Synthesis using Unpaired Data. http://arxiv.org/abs/1708.01155 arXiv:1708.01155 [cs].

[76] McKell Woodland, John Wood, Brian M. Anderson, Suprateek Kundu, Ethan Lin, Eugene Koay, Bruno Odisio, Caroline Chung, Hyunseon Christine Kang, Aradhana M. Venkatesan, Sireesha Yedururi, Brian De, YuanMao Lin, Ankit B. Patel, and Kristy K. Brock. 2022. Evaluating the Performance of StyleGAN2ADA on Medical Images. Vol. 13570. 142153. https://doi.org/10.1007/978303116980914 arXiv:2210.03786 [cs, eess].

[77] Yong Xia, Wenyi Wang, and Kuanquan Wang. 2023. ECG signal generation based on conditional generative models. *Biomedical Signal Processing and Control* 82 (April 2023), 104587. https://doi.org/10.1016/j.bspc.2023.104587

[78] Yawen Xiao, Jun Wu, and Zongli Lin. 2021. Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data. *Computers in Biology and Medicine* 135 (Aug. 2021), 104540. https://doi.org/10.1016/j.compbiomed.2021.104540

[79] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. 2015. Empirical Evaluation of Rectified Activations in Convolutional Network. http://arxiv.org/abs/1505.00853 arXiv:1505.00853 [cs, stat].

[80] Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. 2018. An empirical study on evaluation metrics of generative adversarial networks. http://arxiv.org/abs/1806.07755 arXiv:1806.07755 [cs, stat].

[81] Xin Yi, Ekta Walia, and Paul Babyn. 2019. Generative Adversarial Network in Medical Imaging A Review. *Medical Image Analysis* 58 (Dec. 2019), 101552. https://doi.org/10.1016/j.media.2019.101552 arXiv:1809.07294 [cs].

[82] Han Zhang, Zizhao Zhang, Augustus Odena, and Honglak Lee. 2020. Consistency Regularization for Generative Adversarial Networks. http://arxiv.org/abs/1910.12027 arXiv:1910.12027 [cs, stat].

[83] Jing Zhang, Lu Chen, Li Zhuo, Xi Liang, and Jiafeng Li. 2018. An Efficient Hyperspectral Image Retrieval Method: Deep Spectral-Spatial Feature Extraction with DCGAN and Dimensionality Reduction Using t-SNE-Based NM Hashing. *Remote Sensing* 10, 2 (Feb. 2018), 271. https://doi.org/10.3390/rs10020271

[84] Yu Zheng, Zhi Zhang, Shen Yan, and Mi Zhang. 2022. Deep AutoAugment. http://arxiv.org/abs/2203.06172 arXiv:2203.06172 [cs].

[85] Zhiming Zhou, Han Cai, Shu Rong, Yuxuan Song, Kan Ren, Jun Wang, Weinan Zhang, and Yu Yong. 2018. ACTIVATION MAXIMIZATION GENERATIVE ADVER SARIAL NETS. (2018).
[86] Jin Zhu, Guang Yang, and Pietro Lio. 2019. How Can We Make Gan Perform Better in Single Medical Image SuperResolution? A Lesion Focused MultiScale Approach. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, Venice, Italy, 16691673. https://doi.org/10.1109/ISBI.2019.8759517

A APPENDIX

In this section, we present an extensive overview of our MedFID metric, proposed as a medical-orientated FID adaption aimed at synthetic X-ray image evaluation, extending the application of the FID metric to a larger set of medical domains. We first present some background of the current metrics in section A.1 and A.2, highlighting a key limitation that questions their applicability to medicine in section A.3. Section A.4 presents a literature review of similar domain adaptations of FID Scores from which we derive our approach, presented in section A.5. We develop our approach using the dataset described in section A.6. Finally, in section A.7, we validate the metric by demonstrating a strong correlation to human perception.

A.1 FID Scores

As a primarily visual technique, GANs are best evaluated by the manual evaluation of images [10, 13, 17, 68]. This technique evaluates the visual quality of the image but does not take into consideration the ability of the GAN to capture the nuances of the data used in training the networks, for example, it may fail to detect mode collapse. Additionally, manually evaluating an image is cumbersome, subjective and variant [8, 46]. *Inception Scores* have long been the standard in unbiased, quantitative GAN image evaluation. Inception Scores use an Inception Network [65], trained on the ImageNet dataset, to determine the probability that an image belongs to each class defined in ImageNet. The higher the probability (that is, the more similar an image is to other images in the predicted class), the higher the inception score and thereby the quality of the image. However, the statistics of real-world samples are not compared to those of generated samples [24].

Heusel et al. (2018) improved upon Inception Scores [24]. The underlying idea is that a GAN trained to convergence should generate data that aligns with the distribution of data in the training samples such that the distance between the distributions can quantify the dissimilarity between real and generated samples. They use the *Fréchet Distance*, alternatively known as the *Wasserstein-2 Distance*, to calculate the distance between generated data $p(.)$ and real data $p_w(.)$. FID scores use the same pre-trained Inception Network used in Inception Scores however the terminal layer (a softmax classification layer) is removed and the final coding layer $pool_3$ is used to generate 2048 vision-relevant features of an image. The mean and covariance of the feature vectors are used to parameterize a multidimensional Gaussian distribution such that $p(.) \sim \mathcal{N}(m, C)$ and $p_w(.) \sim \mathcal{N}(m_w, C_w)$. The FID score is the distance between these distributions as given in Equation 5 where Tr is the trace linear algebra operation.

$$d^2((m, C), (m_w, C_w)) = ||m - m_w||_2^2 + Tr(C + C_w - 2(C \cdot C_w)^{1/2}) \quad (5)$$

A.2 InceptionNet

InceptionNet [65] is a deep convolutional neural network for classification developed for the ImageNet Large-Scale Visual Recognition Challenge 2014 [65]. The CNN uses *Inception Modules* that use a combination of 1x1, 3x3 and 5x5 convolutions that allow the network to learn feature maps at different scales. These are then concatenated together allowing the network to learn both local and global features. FID scores use InceptionNet V3 which replaces the 5x5 convolution with two 3x3 convolutions, decreasing computation time and adding in Batch Normalization for the auxiliary classifiers [66]. This is implemented with an RMSProp optimizer.

A.3 Lack of adequate medical training data

Several GAN measures exist however there is no agreement on the *best* metric. Ideally, a metric should measure the strengths and expose weaknesses in models and serve as a basis for model selection. Arguably, FID scores have become the standard for objective GAN synthetic image evaluation however it does contain intrinsic biases with effect analysis. The Inception Network it is based on is trained on Imagenet, which notably contains no medical data. The features extracted from the network focus on properties of real objects and natural images [1] and are, therefore, not representative of medical data possibly leading to misleading results [70]. An empirical study [76] has demonstrated a negative correlation between FID scores and human judgement, however the subjectivity and bias towards realistic-looking pictures in manual evaluation, as opposed to accurate images, [61] questions the validity of the metric [81]. Additionally, further research is to be conducted on the validity of this observation. As such, our aim is to propose a suitable alternative that accounts for the unique characteristics of X-ray images (shape, texture, pixel distribution, and so on) allowing extracted features to be relevant to the problem domain.

A.4 Proposed solutions in literature

Due to the bias towards natural images, there has been research into variations of FID for niche domains to which FID scores may not be suited. We evaluate two common approaches: we study approaches that generate features by applying transfer learning and fine-tuning on a pre-trained Inception V3 network in section A.4.1. In section A.4.2, we look at some approaches involving the development of new encoder networks from which features are extracted. Note that in all cases, unless otherwise specified, extracted features undergo the formula in equation 5. Section A.4.3 serves as a preamble to section A.5 by examining recent approaches of transfer learning on a pre-trained InceptionV3 network for classification networks.

A.4.1 Transfer Learning and Fine Tuning. We define *transfer learning* to be the use of a pre-trained network which is either used as is or augmented with additional layers that are trained whilst keeping the pre-trained network frozen. Han et al. (2020) [23] used an Inception V3 network pre-trained on ImageNet, much like the original FID, but extracted features from the 3rd pooling layer to get more general feature representations for an image. Andreou et al. (2023) [4] created an FID-like metric for evaluating a series of GANs generated to create sequences of background hits in a Cylindrical Drift Chamber (the reader is referred to the paper [4] for more

information on the topic). Using the same Inception V3 network as FID scores,the last pooling layer was removed and three additional layers were added: one convolutional layer and two fully connected layers. The weights of the base model were frozen allowing only the new layers to be trained. Features were then extracted from the last fully connected layer.

A.4.2 New encoder. Kynkäänniemi et al. (2023) [38] evaluated the standard Inception-based FID metric against a ResNet-50 classifier, amongst others, trained on ImageNet. Features were generated by removing the classification layer of the ResNet. Interestingly, the FID scores obtained from the ResNet-50 approach were similar to those of the Inception approach. Preuer et al. [2018] [51] proposed an FID variant for drug design called the *Fréchet ChemNet Distance* (FCD). Features are extracted from ChemNet, a long short-term memory recurrent neural network trained to predict bioactivities of 6000 assays, allowing the features to encode chemical and biological information in samples. This allows the metric to be used to evaluate GAN-based drug design models as it encapsulates the defining and desired attributes of assays. Features are extracted from the final coding layer of the network. Subramaniam et al. (2022) [64] created an FID variant for use on 3D datasets by extracting features from MedicalNet, a 3D ResNet model trained on medical datasets, optimized for segmentation tasks.

A.4.3 Transfer learning on InceptionV3 for classification tasks. Ker et al. (2019) [33] used a pre-trained Inception V3 network to classify histology slides for brain and breast tissues, leading to a $\sim$ 80% improvement in the F1 score over a model trained from scratch. This was done by removing the last 4 layers of InceptionV3, adding global average pooling along with 4 fully connected layers with ReLU activation functions. Finally, a softmax layer was added to perform the classification. Chen et al. (2020) [12] took a similar approach to classify histopathological H&E [10] stain images by removing the final inception layer, adding in a single fully connected layer along with a softmax output. Quan et al. (2019) [21] created a classification model for four different types of cervical lymphadenopathy. They used an Inception V3 network with weights initialized to the model trained on ImageNet. Three additional fully connected layers were added along with a softmax classification output. The entire model is fine-tuned on their dataset.

A.5 Approach

To create a suitable FID variant for medical data (MedFID), we assimilate the approaches used above and selectively combine elements to derive a novel approach closely following that of [33] and [21]. A pre-trained Inception V3 network, trained on the ImageNet dataset, is used as a base. We remove the last layer of the model but keep the $pool_3$ layer which generates the 2048 features generally used in the FID computation. These are then frozen. We attach 4 fully connected layers with a ReLU activation function maintaining 2048 features at each additional layer. Dropout was also applied at a rate of 20% to avoid over-fitting. Lastly, since this is a multi-class classification problem, a softmax classification layer was added. Our architecture is presented in table 6. The model was trained for 50 epochs using an RMSProp optimizer (learning rate = 0.0001) and categorical cross-entropy loss. This resulted in an accuracy of 0.347. To improve performance, the top 2 inception blocks were made trainable and the entire model was fine-tuned for 20 epochs resulting in a new accuracy of 0.518 on the test set. Given the time constraints of this research, this was acceptable however continuing with fine-tuning for more epochs would considerably improve these results. Our approach is set up as a multiclass classification problem attempting to predict the type of extremity in an image as opposed to a specific pathology. Thereby, we keep the approach in line with that of FID scores which were set up to identify ImageNet classes.

MedFID Fine-Tuned Inception		
Layer (type)	Output Shape	Param #
Inception_v3 (functional)	(none, 8, 8, 2048)	21 802 784
Global_average_pooling2d	(none, 2018)	0
Dense	(None, 2048)	4 196 352
Dense_1	(None, 2048)	4 196 352
Dense_2	(None, 2048)	4 196 352
Dense_3	(None, 2048)	4 196 352
Dropout	(None, 2048)	0
Dense_4	(None, 7)	14 343
Total params:		38 602 535
Trainable params:		16 799 751
Non-trainabe params:		21 802 784

Table 4: Architecture of the MedFID Inception Network

Inception V3 requires all images to be of size $3x299x299$ hence all training data is rescaled to fit these dimensions. To prevent overfitting on the classifier, geometric data augmentations are used (horizontal flips and width and height flips)
The output of $dense_3$ is used to generate feature vectors of 2048 features for an image. The feature vectors of two sets of images undergo the FID calculation described in Equation 5.

A.6 Dataset

The pre-trained Inception V3 network was fine-tuned using a merged dataset consisting of images from two publicly available X-ray datasets. The dataset described in section (2.1) was not for two reasons: Firstly, the metric should not contain any bias towards the data used to train the GAN it will evaluate data and, secondly, using a publically available dataset keeps the metric generic for future research, circumnavigating any patient confidentiality and ethical clearance issues needed to reproduce the metric.
The following publically available, anonymized datasets were consolidated to produce our Inception training dataset.

(1) **MURA-v1.1** MURA [54] is a collection of 40 561 bone X-rays of the upper extremities where each X-ray is classified as either *normal* or *abnormal*. Each image is also labelled with the type of body part being studied e.g. finger, elbow, forearm and so on.
(2) **LERA** (Lower Extremity RAdiographs [11]) is a collection of radiographs of the foot, ankle, hip and knee of 182 patients at

[10] H&E - hematoxylin and eosin

[11] https://aimi.stanford.edu/lera-lower-extremity-radiographs

Distortion	Correlation	P-Value	Figure
Gaussian Blur	0.837	1.889e-07	figure 12
Noise	0.940	2.535e-10	figure 13
Salt & Pepper Noise	0.826	4.034e-06	figure 14
Blocks	-0.312	0.5474	figure 15

Table 5: Correlations of the intensity of distortions to MedFID Scores for CelebA

the Stanford University Medical Centre whereby each image is labelled as either *normal* or *abnormal*.

Merging the two datasets creates a dataset containing radiographs for the upper and lower extremities keeping the metric generalizable for non-elbow synthetic image analysis. We used an 80/20 split for training and test data. To balance the dataset and to remove minority classes, the hip xrays were dropped (appropriate since the other classes deal with extremities) and each class was reduced to a test set of 100 images and a training set of 400 images (maintaining the 80/20 split). We selected 100 images for the test set as this is the maximum allowable number of items to ensure uniformity in the dataset. 400 images were chosen to be part of the training set to maintain the 80/20 split. Images were randomly dropped. Training images undergo vertical and horizontal shifts and horizontal flips as a form of data augmentation. No augmentations are applied on the test set.

A.7 Validation

Original validation of the FID metric involved proving the correlation to image quality [24]. The idea is to demonstrate a negative correlation between FID and human perception such that the more distorted an image is, the higher the FID score. To validate MedFID, we follow a similar approach: apply a set of distortions to images and measure the MedFID score at varying degrees of distortions. We applied the following distortions: Gaussian noise for various kernel sizes, Noise with varying means (standard deviations = 0.1), Salt&Pepper noise for various noise levels and we added an increasing number of blocks to random positions in an image to simulate portions of the image not being seen. This process was done using a sample image, chosen at random, from our dataset as well as using a random sample from the CelebA dataset as done in [24].

A.7.1 CelebA Data. Table 5 presents a summary of the correlation between image quality and degree of distortion. As expected, there is a strong correlation with an increasing MedFID score, which is a negative correlation with image quality, as expected. Notably, the block distortion has a negative correlation with an increasing MedFID score. We hypothesize that since the MedFID metric is trained on black-and-white images with space around the bone structure, adding black blocks to an image creates some similarity to X-ray images as we reduce the amount of RGB space in the image.

A.7.2 Medical Data. Table 6 shows a similar analysis for the MedFID metric used on medical data. As expected, there is a negative correlation between image quality and degree of distortion. Notably,

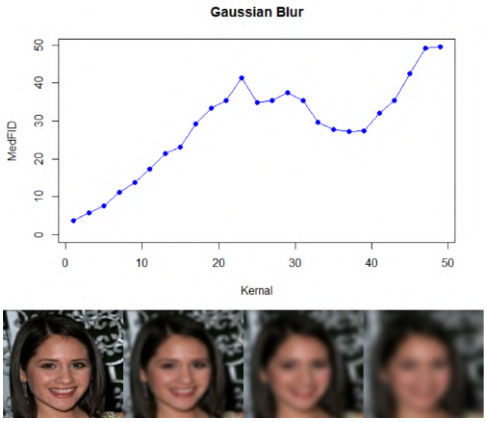


Figure 12: MedFID vs kernel size for image blurring with selected image examples for kernel = {0, 15, 31, 49}.

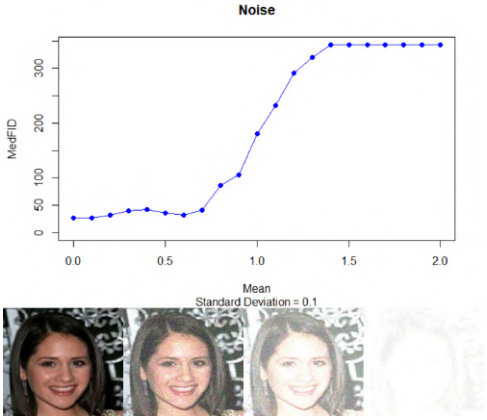


Figure 13: MedFID vs noise disturbances with selected image examples for mean = {0, 0.3, 0.6, 1}. Noise was applied with a standard deviation of 0.1.)

the p-value of the block distortion suggests that there is no correlation. We hypothesise, that the black space in the image increases as blocks are added but the quality of the visible sections remains unchanged. This mimics the black space around x-ray imagery hence this is being caught as a feature by our inception network whereas in the noise distortions, the removal of black sections increases the MedFID score as expected. Interestingly, however, for less than 10,000 blocks, the correlation is 0.764 (p-value = 0.076).

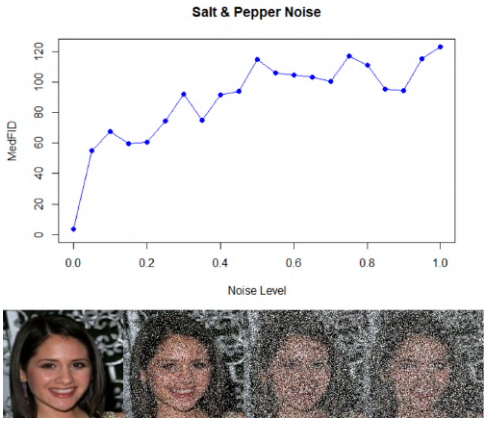


Figure 14: MedFID vs noise level for salt&pepper disturbances with selected image examples for noise level = {0, 0.3, 0.6, 0.9}.

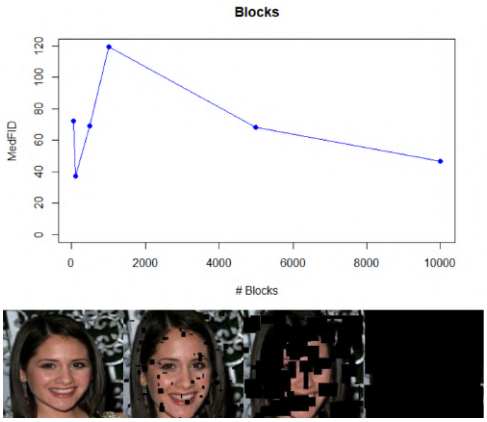


Figure 15: MedFID vs the amount of blocks (a portion of image missing) with selected image examples for 0, 50, 1000, 10000 blocks.

Distortion	Correlation	P-Value	Figure
Gaussian Blur	0.816	6.542e-07	figure 16
Noise	0.898	3.355e-08	figure 17
Salt & Pepper Noise	0.998	2.2e-16	figure 18
Blocks	0.438	0.2773	figure 19

Table 6: Correlations of the intensity of distortions to MedFID Scores on medical data

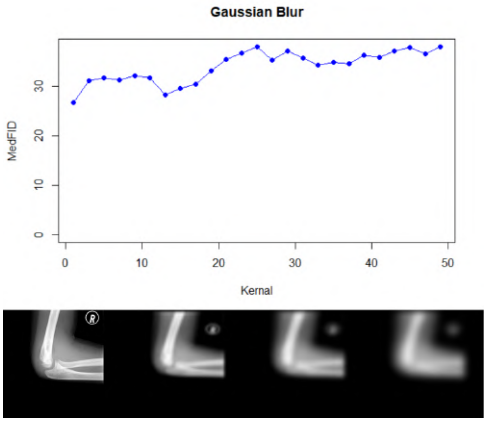


Figure 16: MedFID vs kernel size for image blurring with selected image examples for kernel = {0, 15, 31, 49}

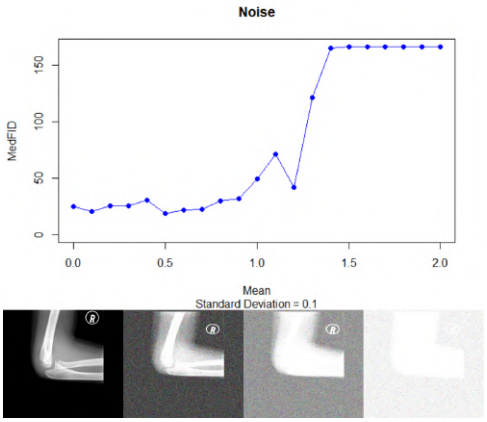


Figure 17: MedFID vs noise disturbances with selected image examples for mean = {0, 0.3, 0.6, 1}. Noise was applied with a standard deviation of 0.1.

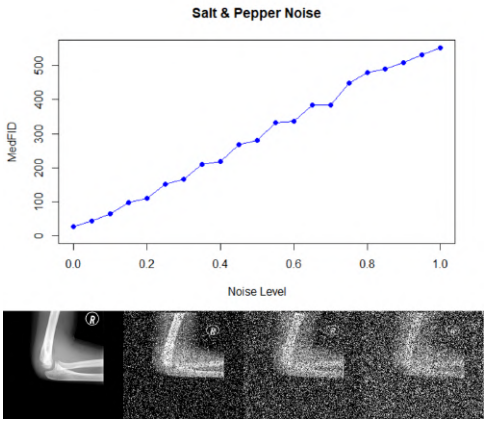


Figure 18: MedFID vs noise level for salt&pepper disturbances with selected image examples for noise level = {0, 0.3, 0.6, 0.9}.

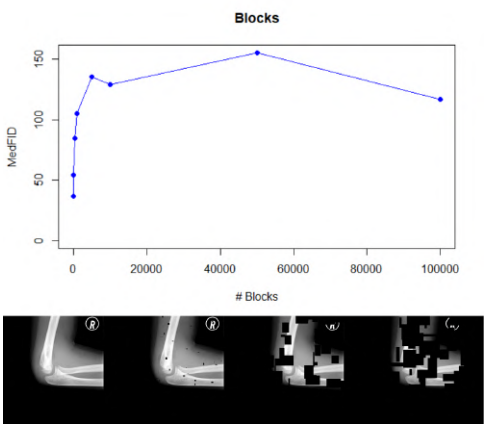


Figure 19: MedFID vs the amount of blocks (a portion of image missing) with selected image examples for 0, 50, 1000, 10000 blocks.

B TRAINING DATA

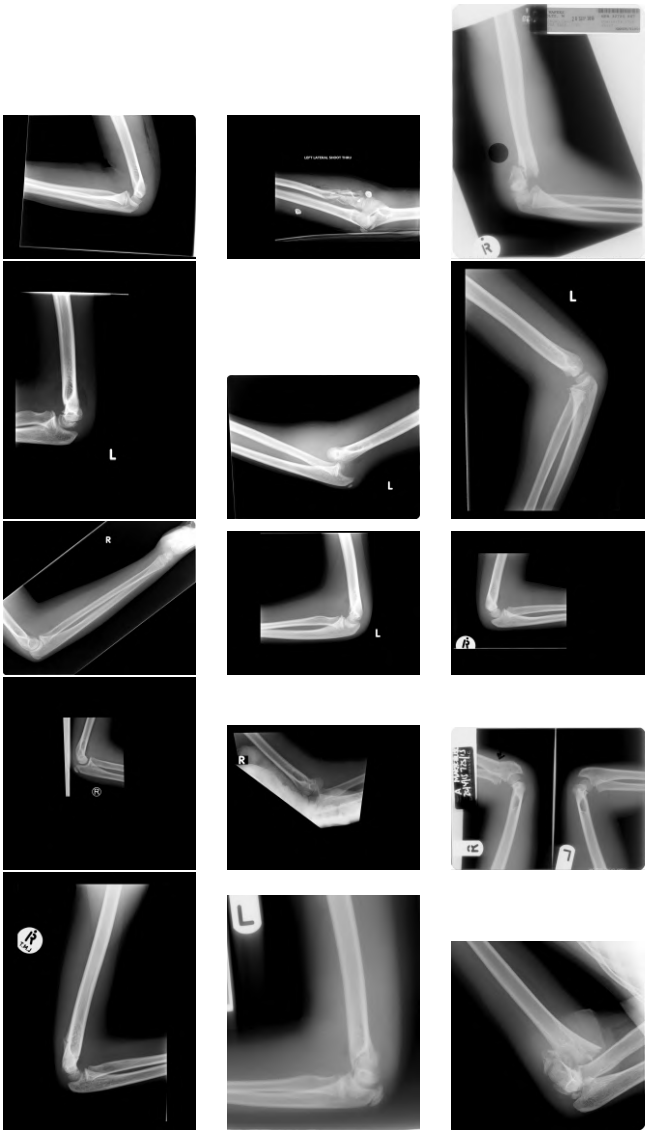


Figure 20: A sample of 15 images from the original training dataset.

C STYLEGAN GENERATIONS FOR > 9000 ITERATIONS

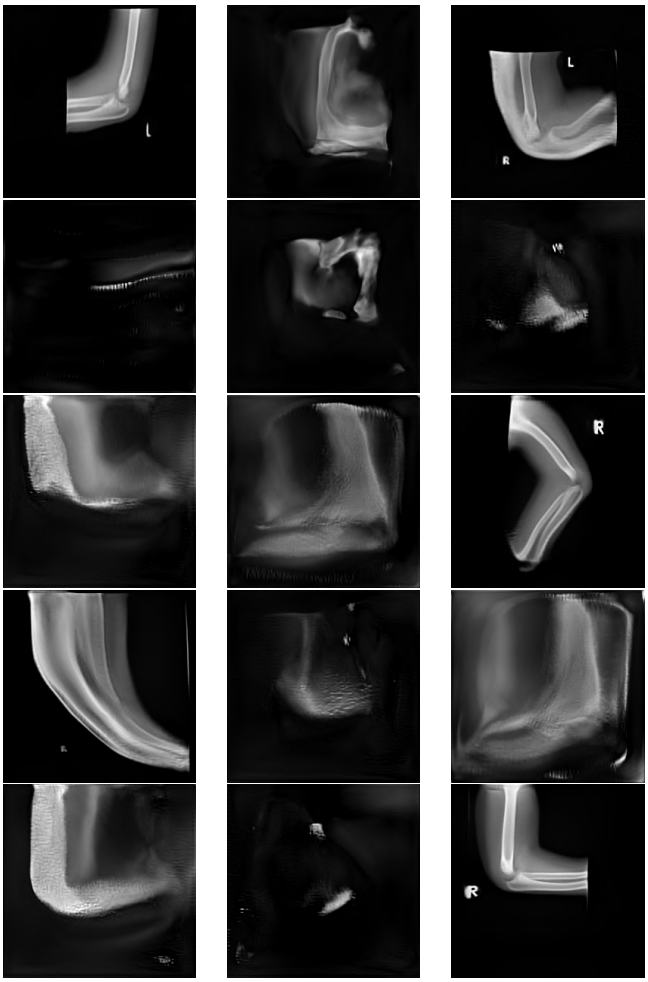


Figure 21: A selection of outputs from larger StyleGAN2 iterations demonstrating the instabilities learnt in the model. Note that not all images are affected indicating that StyleGAN2 has retained initial information however the degree of variation within images is more extreme.

D LAT AND AP ELBOW: COMBINED GENERATIONS

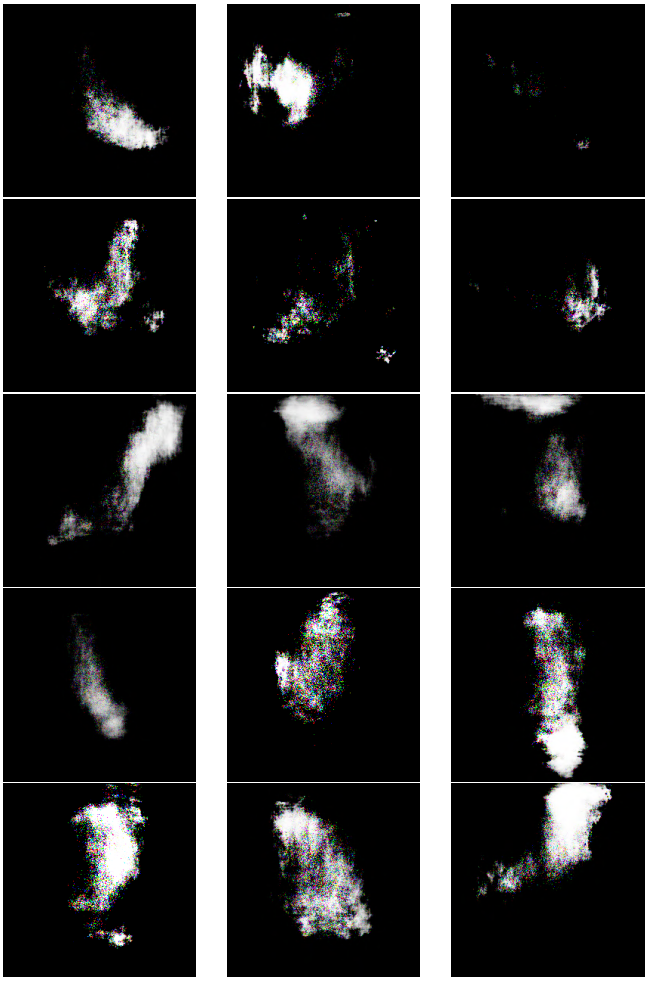


Figure 22: A selection of outputs from WGAN5 ($\beta_1 = 0.5, \beta_2 = 0.9, lr = 0.002$) with a dataset consisting of LAT and AP elbow images. These demonstrate some of the merging of LAT and AP image features, most noticeably the white patches generated apart from the main image which we predict is an attempt to move between the two image types.