ORIGINAL RESEARCH

Evaluating the Clinical Realism of Synthetic Chest X-Rays Generated Using Progressively Growing GANs

Bradley Segal[1] · **David M. Rubin[1]** · **Grace Rubin[2]** · **Adam Pantanowitz[1]**

Abstract

Chest X-rays are a vital diagnostic tool in the workup of many patients. Similar to most medical imaging modalities, they are profoundly multi-modal and are capable of visualising a variety of combinations of conditions. There is an ever pressing need for greater quantities of labelled images to drive forward the development of diagnostic tools; however, this is in direct opposition to concerns regarding patient confidentiality which constrains access through permission requests and ethics approvals. Previous work has sought to address these concerns by creating class-specific generative adversarial networks (GANs) that synthesise images to augment training data. These approaches cannot be scaled as they introduce computational trade offs between model size and class number which places fixed limits on the quality that such generates can achieve. We address this concern by introducing latent class optimisation which enables efficient, multi-modal sampling from a GAN and with which we synthesise a large archive of labelled generates. We apply a Progressive Growing GAN (PGGAN) to the task of unsupervised X-ray synthesis and have radiologists evaluate the clinical realism of the resultant samples. We provide an in depth review of the properties of varying pathologies seen on generates as well as an overview of the extent of disease diversity captured by the model. We validate the application of the Fréchet Inception Distance (FID) to measure the quality of X-ray generates and find that they are similar to other high-resolution tasks. We quantify X-ray clinical realism by asking radiologists to distinguish between real and fake scans and find that generates are more likely to be classed as real than by chance, but there is still progress required to achieve true realism. We confirm these findings by evaluating synthetic classification model performance on real scans. We conclude by discussing the limitations of PGGAN generates and how to achieve controllable, realistic generates going forward. We release our source code, model weights, and an archive of labelled generates.

Keywords Medical image generation · Generative adversarial networks · X-rays · Deep learning · Medicine · Radiology

Introduction

The chest radiograph (CXR) is the most common diagnostic radiological procedure [2] and is commonly used to screen, diagnose or monitor conditions across a variety of clinical contexts. This ubiquity has resulted in substantial research interest in the automated diagnosis of such films [21, 26, 37], with a recent surge in activity due to the COVID-19 pandemic [33, 46, 52]. State-of-the-art (SOTA) models are capable of radiologist-level performance across a subset of pathologies in only a fraction of the time needed for human review [15, 36, 37]. This activity in CXRs represents a section of the development of Computer Aided Diagnostic (CAD) systems that aim to provide tangible benefit to clinicians and patients alike by reducing diagnostic turnaround times, minimising errors, and supporting the clinical decision making process [36, 42]. Large scale, anonymised, public imaging datasets underscore these efforts by providing the necessary clinical data for the training of CAD models [15, 17, 53]. The development of these archives is both time consuming and costly, requiring extensive expert labelling and anonymisation of patient protected information prior to release. Radiology reports produced during clinical practice are typically used as surrogates for expert review and

✉ Bradley Segal
  brad@segal.co.za

1 School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg, 1 Jan Smuts Avenue, Braamfontein, South Africa

2 Department of Radiation Sciences, Division of Radiology, University of the Witwatersrand, Johannesburg, South Africa

are mined for diagnostic labels [15, 36, 37]. Anonymisation of images involves the detection of annotations containing protected patient information such as names that may be contained within the pixel data, or the removal of metadata that may be contained within digital imaging and communications in medicine (DICOM) files [17]. This process does not adjust the image data of a particular scan, which may result in patient re-identification if visual elements of the image are rare.

Medical imagery exhibits multi-modal long-tailed distributions with significant heterogeneity in the presentation of similar disease patterns due to equipment, scan technique, and patient variability [57]. This issue is demonstrated by most large CXR archives as they are limited to a single clinical site which reduces the captured variability of the factors of heterogeneity [57]. The resulting sparsity of training data has obvious implications for making the development of robust diagnostic models more challenging. In addition, alterations in protected patient information such as sex, age, ethnicity, and socioeconomic class have been demonstrated to produce biases in diagnostic performance not significantly explained by variations in disease prevalence with improvements only seen with usage of multi-source image datasets [45].

There is a clear need for methodologies capable of resolving this data disparity while retaining or even augmenting privacy. Generative adversarial networks (GANs) have received significant attention as a potential solution for their ability to synthesise medical images [4, 40, 51] without compromising patient confidentiality as they learn to replicate the source distribution without access to the underlying training data [54]. GANs capable of high fidelity image synthesis would address data challenges in machine learning and in medical education and training by:

- Creating datasets that do not compromise patient confidentiality, but provide the same diagnostic outcomes;
- Providing class-balanced datasets;
- Generating augmented images by semantic variation in visual content without altering the diagnosis; and
- Synthesising images with specified target pathologies

This study serves to comprehensively evaluate the medical plausibility of synthetic CXRs as well as their applications to diagnostic radiology. [1] We evaluate the Progressively Growing GAN (PGGAN) [18] methodology applied to generate multi-modal, megapixel resolution CXRs. We provide domain expert examination of the properties of generated images as well as a review of expert discrimination between real and generated samples. In addition, we provide the following contributions:

- An evaluation of the applicability of the Fréchet Inception Distance (FID) for automatically evaluating X-ray generates.
- A proposal for extracting images for a specific pathology through a modified latent space search.
- A proposal for generating patient image series through local pathology sampling.
- An evaluation of the performance of synthetic image datasets derived from multi-modal generators.

Source code is available [43], as well as model weights and a collection of labelled generated images [44].

Background

Generative Adversarial Networks

GANs are a form of implicit generative model rooted in game theory that learns to reproduce an unseen training distribution through competitive optimisation. The prototypical GAN consists of a pair of neural networks in opposition to one another: a generator (G) whose goal is to generate plausible samples, and a discriminator (D) whose goal is to distinguish such samples from real samples (x) drawn from a training distribution [10]. The generator learns to sample from a latent vector (z) to produce a sample ($G(z)$) of the generated ($\mathbb{P}_g$) distribution which is similar to a chosen reference distribution ($\mathbb{P}_r$). The training signal is provided by how effectively such a generate is classified as real by the discriminator ($D(G(z))$). This configuration is optimal when the generator produces samples indistinguishable from the reference set and the discriminator can no longer learn to detect generates [28, 41]. The configuration can be interpreted as a two-player minimax game with the following value function:

$$\min_G \max_D V(D, G) = \mathop{\mathbb{E}}_{x \sim \mathbb{P}_r} [\log D(x)] \\ + \mathop{\mathbb{E}}_{x \sim \mathbb{P}_g} [\log(1 - D(G(x)))] \tag{1}$$

wherein the discriminator (D) attempts to maximise the value function by separating real from generated images and the generator (G) attempts to minimise it by producing samples that cannot be detected by the discriminator.

The GAN formulation has found utilisation for a variety of tasks, namely image-to-image translation [16, 58], image super-resolution [24, 55], and semantic image editing [47] to name just a few. Perhaps the most common usage is that of

---

[1] An ethics waiver was issued for this work by the Human Research Ethics Committee (Medical) of the University of the Witwatersrand, Johannesburg on 11/08/2020.

Fig. 1 Training configuration for PGGAN. The network trains until convergence and then doubles the spatial resolution. This process is repeated until the desired resolution is achieved
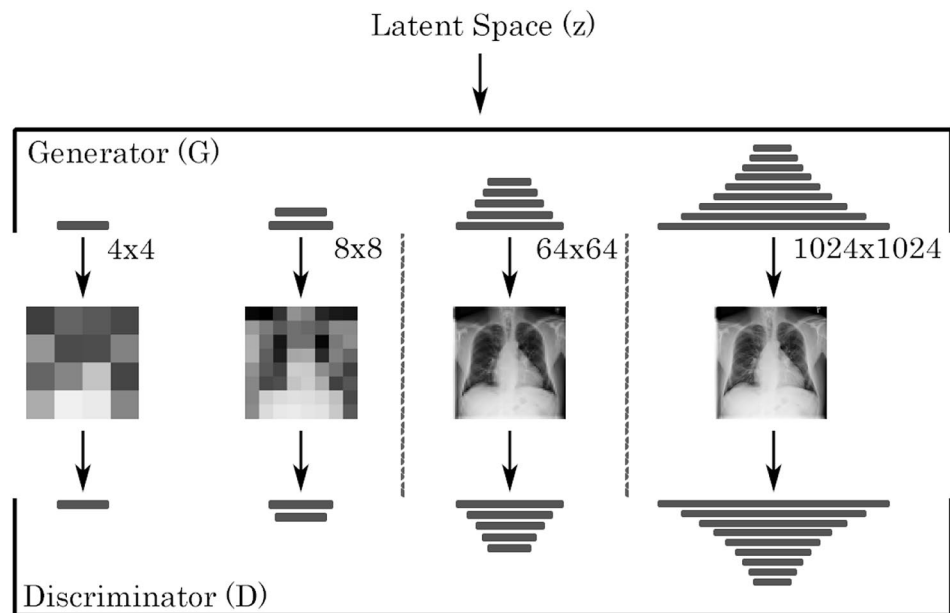

image synthesis, wherein a GAN attempts to produce images that mirror a reference image set. The Deep Convolutional GAN [35] was an initial step towards improving resolutions by making use of convolutional layers as opposed to the fully connected approach of the original GAN.

This, however, did not resolve fundamental problems with training instability that caused models to fail to converge or resulted in mode dropping, a phenomenon, where image variability is sacrificed for image quality and at worst can result in a the generation of only a handful of images [3]. These issues become more prominent as image resolution increases as the discriminator is able to more easily detect generated samples [18]. This places significant constraints on the upper limit of resolutions that a DCGAN model can achieve.

Recent work has adapted the loss functions employed [3, 11, 27], progressively scaled generate size [18], or employed style transfer techniques [19, 20] to improve training stability. These techniques have enabled megapixel resolution generation while retaining much of the quality and diversity of reference images.

We focus on the Progressively Growing GAN (PGGAN) methodology employed by Karras et al. [18]. This technique trains on progressively larger images as each resolution converges, stabilising the training process and enabling initial training to progress faster as batch sizes can be larger. It is an attractive option for high-resolution image synthesis as it has far more moderate computational requirements compared to its later variants [19, 20]. The training methodology for PGGAN additionally combats the effects of generator modal collapse through the inclusion of minibatch discrimination, which operates by considering the variability of samples it is trained on. This promotes the generator to maintain sample

diversity as a reduction would aid the discriminator in identifying samples as fake. Despite these benefits, there are important theoretical limitations for the PGGAN that must be considered. The later StyleGAN variant by Karras et al. [19] discusses that the network's only source of stochastic variation arises from the initial latent space sample which results in the consumption of network synthetic capacity to preserve variation for later image scales. This results in repetitive patterns in images and a loss of variability at greater resolutions. In addition, the progressively growing technique itself is known to result in phase artifacts with sections of images becoming fixed in preferred locations from training at lower resolutions [20] (Fig. 1).

Synthetic Medical Data

There is a multitude of work that aims to adapt GANs for resolving data disparities via data generation while improving both patient confidentiality and model performance. In the pursuit of this, synthetic medical data aims to capture information of diagnostic utility while eliminating the possibility of patient re-identification [34]. Park et al. utilise GANs to produce anonymised clinical data tables that are interoperable with model architectures applied to standard tables, yet can be shared without concern of violating patient privacy [34].

Other works have successfully applied DCGANs for dataset augmentation in various CXR classification tasks. Moradi et al. focus on generating both normal and cardiac abnormality images, they find that GAN-based augmentation outperforms traditional training augmentation of flipping, cropping, or scaling images [30]. Salehinejad et al. expand upon this concept by producing a per-class DCGAN and

utilising generates to balance classes for a more diverse performance improvement [40]. Both approaches demonstrate the potential benefits GANs offer for dataset augmentation, however, are handicapped by the need for per-class models. Training a model per class is computationally prohibitive and fails to capture the interactions between pathologies that may be exploited to improve diagnostic accuracy.

Beyond CXRs, PGGANs have been applied to generate other high resolution medical images. Beers et al. demonstrate the applicability of the method for creating high fidelity reproductions of the retinal fundus and MRI slices of gliomas [4], while Togo et al. produce patches of X-rays of gastritis [51]. Both implementations introduce domain-specific modifications. Beers et al. include segmentation maps as additional channels for fundal images to enhance the generation of features relevant to diagnosis of retinopathy, while Togo et al. introduce a conditional loss at higher resolutions to promote the distinction between gastritis and normal tissues. Korkinof et al. synthesise high resolution full field mammograms and provide comparisons of generated and source images, noting that the images appear similar with the preservation of several common tissue artifacts [23]. Bowles et al. augment CT and MRI data with synthetic slices and observe improvements in segmentation performance [5].

Despite significant variation in both domain and task, most medical image GAN implementations are constrained to produce individual classes or random samplings from the latent space with little control over generated content.

Image Quality Metrics

The automated evaluation of the quality of medical imagery is a diverse field with a variety of available methods dependent on the evaluation task at hand. These methods are typically divided into full, reduced, and no reference assessment methods. Full reference methods typically evaluate degradation of an image against a source. This may take many forms, from directly comparing pixel values, signal to noise ratios, or to assessing structural or feature similarities. Reduced reference methods evaluate alterations in images against natural image statistics and are typically used in finding distortions in transmitted images. No reference methods only compute elements of a given image to produce a quality assessment and are similarly utilised for transmitted images, where no reference may exist. Utilising standard full reference methods for synthetic image evaluation is difficult as while the images may closely resemble the source image distribution, the generator does not have direct access to the distribution and as such does not reproduce any particular image. This precludes the use of standard comparison techniques as even structural or feature similarity metrics are intended for comparison of matched images [50].

The Fréchet Inception Distance (FID) is a metric intended to provide a solution for evaluating the quality of generated images [13]. It functions by embedding a set of real and synthetic images in the final average pooling layer of an Inception Net [49] pre-trained on ImageNet [39]. The sets are assumed to be multivariate Gaussian distributions with the average and covariance of each utilised to calculate the Fréchet distance (2), also known as the Wasserstein-2 distance. $(\mu_{P_r}, \sum_{P_r})$ and $(\mu_{P_g}, \sum_{P_g})$ refer to the mean and covariance of the real and generated distributions, respectively.

$$\text{FID}(P_r, P_g) = \left\| \mu_{P_r} - \mu_{P_g} \right\|_2^2 \\ + \text{Tr} \left(\sum P_r + \sum P_g - 2 \left(\sum P_r \sum P_g \right)^{\frac{1}{2}} \right) \tag{2}$$

This distance reflects the difference in the average features extracted from each image set based on the learned kernels of the Inception Net model. This is broadly similar to the feature-based full reference quality assessment metric, yet is capable of application to unpaired images by considering the average combination of features in both sets. The distance has been demonstrated to be consistent with human judgement of visual quality and more resistant to noise than prior approaches [13, 25]. The FID is sensitive to class mode dropping, with distances increasing with greater class discrepancies between the two sets as the average set of features begins to differ. Despite this sensitivity, the FID reports ideal results if a model reproduces the training samples perfectly [25]. The use of the FID for medical images requires the underlying Inception model be capable of extracting features to adequately represent and compare sets against one other. This may be problematic given that X-rays are distinct from the classes found in ImageNet.

Human eYe Perceptual Evaluation (HYPE) [56] in contrast to the aforementioned automated approaches, standardises the human evaluation of model generates by considering either the time necessary to discriminate between real and fake images (HYPE_{time}) or the average error rate given unlimited evaluation time (HYPE_∞). The evaluation methodology is demonstrated to produce reliable results that are capable of separating out differences in model performance.

In addition to the already mentioned quality measures for generated images, a significant element to examine is the performance of models developed using synthetic data applied to the original task of interest. Shmelkov et al. [48] propose the GAN-train metric, which evaluates the classification performance of a model trained exclusively on synthetic data and then run on the original test set. This method quantifies the effective difference between synthetic and real datasets on a benchmark task of interest and provides important information regarding the extent to which

the model captures significant features of the underlying data distribution.

X-Ray Synthesis Experiments

Dataset

The ChestX-ray14 dataset is an update to the ChestX-ray8 dataset [53], a large, open medical X-ray dataset published by the National Institutes of Health (NIH) Clinical Center. The dataset comprises 112120 X-rays from 30805 unique patients. The images comprise 15 classes, viz.:

– No Finding
– Atelectasis
– Cardiomegaly
– Consolidation
– Edema
– Effusion
– Emphysema
– Fibrosis
– Hernia
– Infiltration
– Mass
– Nodule
– Pleural Thickening
– Pneumonia
– Pneumothorax

These images were collected from a clinical archive and should broadly reflect the typical clinical prevalence of these conditions within the community served by the NIH. 75% of images are normal investigations. The remainder are made up of the various labels ranging from the most prevalent, infiltration (10%), to the least, hernia (0.5%). The diagnostic labels are accompanied by bounding boxes for feature localisation. These labels were created through natural language processing (NLP) of the associated radiology and were initially estimated to be over 90% accurate. This accuracy has been disputed, with the visual content reviewed to not adequately match the proposed labels for a number of investigations [31]. A modified set of labels was made available by Rajpurkar et al. [36] as part of their work on pathology detection, wherein a network was trained to classify images based on the original labels and then subsequently used to relabel the original dataset. These labels are available for the majority of images with a residual manually labelled test set which has not been released publicly.

The NIH dataset, license and publication can be found here: https://nihcc.app.box.com/v/ChestXray-NIHCC

Progressively Growing GAN

We implement a modified PGGAN model [18] in Pytorch and perform training via Pytorch Lightning [9] based on the open-source implementation produced by Facebook Research and available through the Pytorch Model Zoo. [2] We start with a randomly initialised model and initially generate 4x4 images with progressively doubling of the spatial resolution after convergence at each scale. We allow mixing of the prior trained layers with the newly added layers through upsampling with the proportion of the previous layer considered reducing in linearly as training progresses.

We utilise WGAN-GP loss [11], equalised learning rates, minibatch discrimination, and pixel normalisation [18]. We also implement an exponential moving average of generator weights which we use for evaluation. We train up to a resolution of 1024x1024 with 800000 images shown during each period of layer mixing and a further 800000 for training of the added layers. All training was performed on an Amazon Web Services (AWS) p3.8xlarge instance with 4 V100 GPUs. Model training took 6 days.

Examples of images generated by the model can be seen in Fig. 2.

Image Classification

For labelling the various CXRs, we implement a Densenet-121 [14] pre-trained on ImageNet [39] and replace the final fully connected layer with the number of classes in the ChestX-ray14 dataset. The model is trained end-to-end in a multi-label configuration to predict all classes simultaneously while making use of the weighted cross entropy function (3) implemented by Guendel et al. [12]. The modified loss balances the frequency of positive (N_p) and negative (N_n) labels per class (c) based on the overall frequency within the training dataset. This adjusts the loss to require the model to discriminate equally between all classes which improves performance for rarer classes.

$$D_{CE}(y_c, \hat{y}_c) = w_P \cdot y_c \log(\hat{y}_c) + w_N \cdot (1 - y_c) \log(1 - \hat{y}_c),$$
$$w_P = \frac{N_p + N_n}{N_p}, \quad w_N = \frac{N_p + N_n}{N_n}$$

(3)

For training, we extract a subset of the full dataset to include only images with modified labels from Rajpurkar et al. [36], we group images at the patient level and average the labels across all images for an individual to produce a summary of the average set of conditions per patient. The dataset is then split into training, validation and test sets through iterative
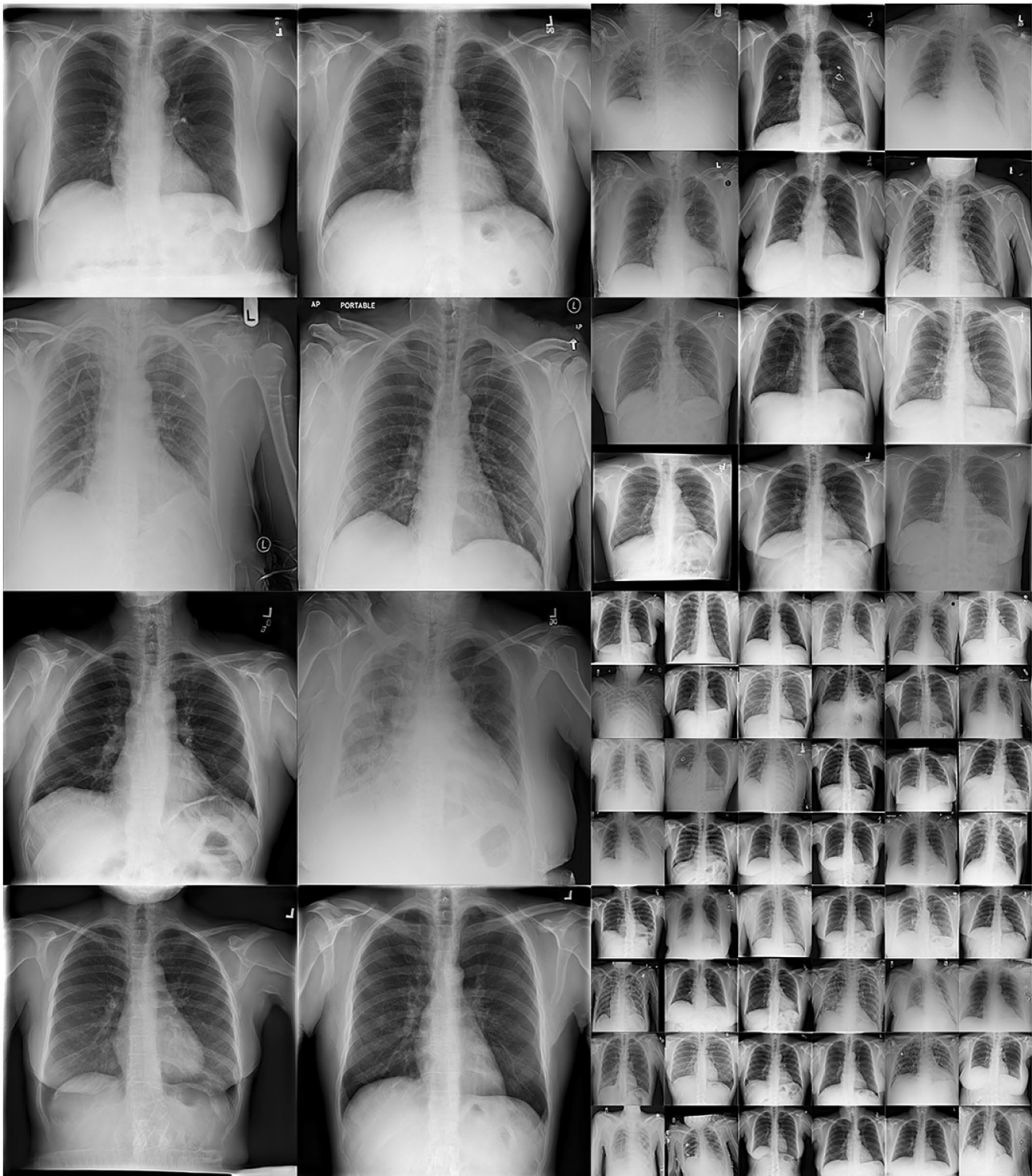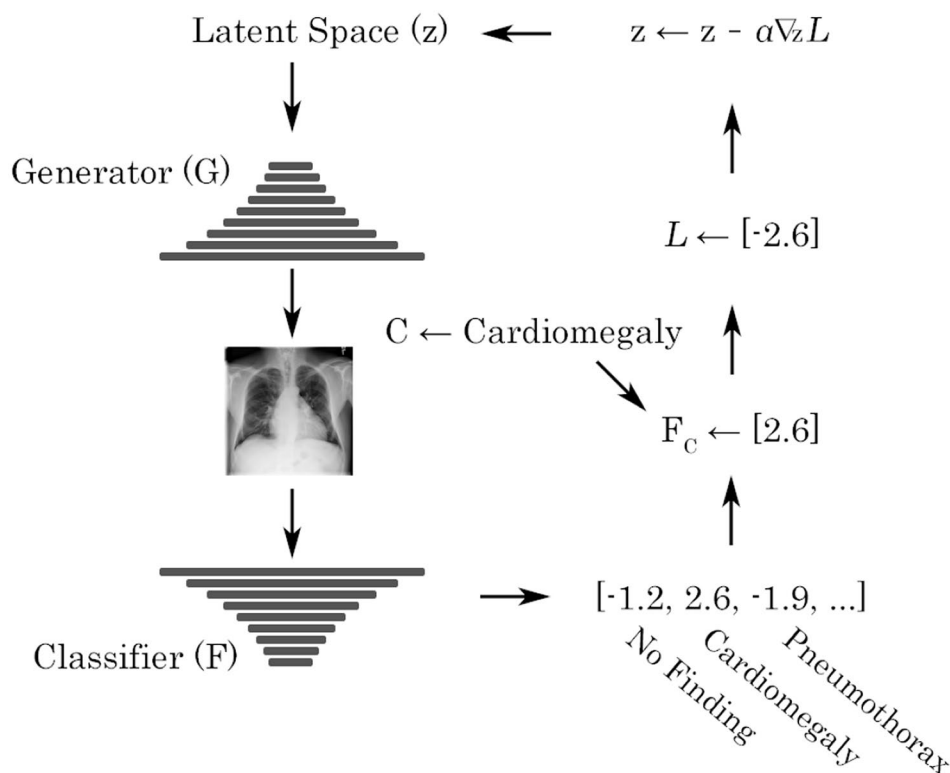

Fig. 2 Examples of uncurated, random samples from the exponential moving average of the generator. Samples are drawn at random from a $\mathcal{N}(0, 1)$ distribution without use of the truncation trick

stratification of the average patient labels to maintain label proportions across sets while ensuring no patient overlap occurs. We augment images with a 10° rotation, random horizontal flip probability of 50%, and colour jitter for brightness, contrast, saturation, and hue of 0.1. The model is trained using an ADAM optimiser [22] using the default settings with an initial learning rate of 10^{-3}, which we reduce by a factor of ten if the validation macro-averaged Receiver

Fig. 3 Visual representation of class generation method shown in Algorithm 1

$$\text{Latent Space (z)} \longleftarrow \quad z \leftarrow z - \alpha \nabla_z L$$

Generator (G)

$$L \leftarrow [\text{-2.6}]$$

$$C \leftarrow \text{Cardiomegaly}$$

$$F_C \leftarrow [2.6]$$

Classifier (F)

$$[\text{-1.2, 2.6, -1.9, ...}]$$

No Finding   Cardiomegaly   Pneumothorax

Operating Characteristic (ROC) Area Under Curve (AUC) fails to improve for several epochs. Once model performance plateaus, it is evaluated on the test set. We repeat the same process for the final trained discriminator, replacing the final linear scoring layer with the number of classes and train the subsequent classification model end-to-end.

Pathology Optimisation

Given the multi-modal nature of pathology contained within X-ray images, it is not possible to utilise standard conditional GANs [29, 32] for label-specific image synthesis, as these typically require independent classes that can be encoded and fed to the generator to control synthesis. Multi-modal constraints are more consistent with text-to-image generation [38], which considers a vector representation of the conditional text input that is semantically meaningful and capable of providing a useful distance based on model predictions [29]. This enables the discriminator to evaluate both image quality and similarity to the conditional text input in scoring generates. These techniques are likely to be problematic in the progressively growing formulation as the disease patterns vary considerably in terms of scale of finding and influence over the

image as a whole. As an example, *Cardiomegaly* may be evaluated simply by having the cardiac border enlarged relative to the overall width of the chest cavity, whereas *Emphysema* in contrast may have enlarged lung fields, flattened diaphragms, and a reduced apparent cardiac size. These features become apparent at differing resolutions and as such, attempting to condition this information throughout training is likely to hamper performance.

We sidestep these issues in multi-modal conditional training by instead drawing inspiration from work on the embedding of images into the latent space [8]. Embedding images requires the inversion of the generator. This is typically performed either by training an encoder network that maps an image to a location in the latent space or by gradient descent optimisation of a random sample to minimise reconstruction loss [1]. We opt to find a latent representation that maximises the corresponding image's classification score for a label of interest. We optimise using a single label at a time to allow for inclusion of related labels that would be consistent with the presentation of real X-rays. We base the implementation on the work by Creswell et al. [8]. Pseudocode for such a method can be seen in Algorithm 1 with a visual representation of this technique in Fig. 3.

Algorithm 1: Algorithm for inferring $z^* \in Z$, a latent representation for a generator G that produces an image $x \in R^{m \times n}$ representative of class $c \in C$ determined by a classifier F exceeding a class specific threshold T_c

```
1  z* ~ P_z(Z) ;                    // Initialise z from prior distribution
2  S ← F_c(G(z*));
3  while S < T_c do
4  |    x ← G(z*) ;                  // Generate image proposal
5  |    S ← F_c(x) ;                 // Obtain class score from classifier
6  |    if S ≥ T_c then
7  |    |    break;
8  |    end
9  |    else
10 |    |    L ← −S;
11 |    |    z* ← z* − α∇_z L ;       // Apply gradient descent
12 |    end
13 end
14 return z*
```

Results and Analysis

Image Quality

The generated images, as shown in Fig. 2, appear to broadly reflect the NIH source images with global features varying similarly across both sets. Sex, posture, exposure, and positioning (AP vs PA) are all represented within the generates. The images accurately reflect the standard anatomical features found in CXRs with realistic alterations in perspective seen with changes in patient posture or positioning of the X-ray detector. Soft tissues such as the heart, liver, and stomach are faithfully reproduced with normal variability in their relative positioning. Bony structures are correctly placed but suffer from inconsistent profiles, with the ribs in particular tending to reveal a degree of undulation. Closer inspection often reveals slight curves or alterations in calibre that are seldom explained by the perspective of the image. Beyond the X-ray itself, the model has learnt to include various markup elements included in the reference images. Most images have a symbol or tag demonstrating the left side of the image, in addition, projection descriptors such as 'PORTABLE' and 'AP' are included on numerous images. Images are typically sided correctly and tend to have visual projections similar to included labels, although at times this is difficult to confirm. Some images do not match the included text, with some having multiple copies of a particular label. A white arrow typically used to demonstrate that a patient was upright for portable scan has also been reproduced, tends to co-occur with text denoting the scan type as portable and is often in similar positions to

the reference images. Elements such as white borders, poor exposure, cropping, or rotation are all present in the source dataset and reproduced in some samples by the PGGAN model (Fig. 4).

The main distinction between distributions is that the generated images broadly lack certain smaller scale features of the source set. Jewelry, ECG leads, pacemakers, and IV catheters are largely absent with occasional partially formed objects in locations, where these should appear. The elements of the missing structures form only a fraction of each overall image and are often fairly detailed objects themselves. A potential explanation for this phenomenon is the progressive growing technique itself, as these objects would only be generated near the end of the training process as the resolution of the images allowed for the details to be appreciated. Examining the generator in the final model without the moving weight average, supports this as a larger proportion of the images possess partially formed objects, implying the network was slowly incorporating these features. The appearance of these elements failed to improve despite training beyond the recommended total number of images. If this is the case, it is probably due to the reduced batch size at higher resolutions which would result in exceedingly slow convergence for rarer, more subtle features. An alternate explanation is the aforementioned capacity loss that prompted the shift to style-based networks as networks consume their own synthetic ability to retain variability in later layers. The effects of this are likely to be more pronounced in CXRs compared to other domains, like facial data, due to the presence of distinct small scale structures such as IV lines or ECG leads that are readily discernible on standard images.
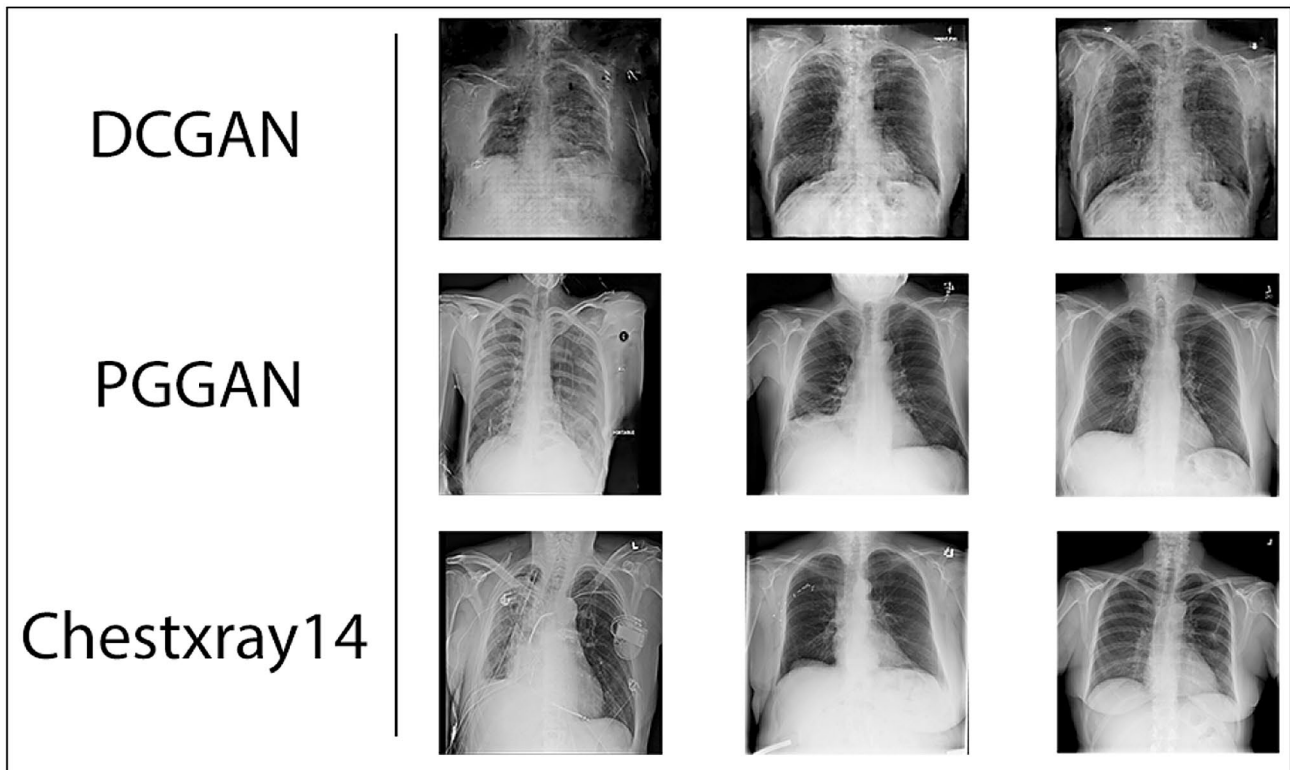
Fig. 4 Comparison of sample quality between different GAN architectures and the source dataset

To evaluate the applicability of the FID for automatic X-ray quality assessment, we first attempt to produce a meaningful zero measure by splitting the NIH dataset at the patient level and using the average labels from each patient to iteratively stratify the patients into two similar groups. The concept being that the groups should have similar disease distribution and be free of patient overlap. We calculate a low value of 0.53 for the FID, indicating that the underlying Inception network extracts similar features on average from both sets. These features may be sub-optimal given that the underlying network has not been trained on biomedical data, yet the small distance demonstrates that the images are embedded in a similar manner. We now need to demonstrate whether the metric can in fact distinguish between X-rays.

Given that the images are known to vary by pathology label, we argue that if the distance metric can reliably separate distinct classes, it must have filters that are capable of extracting features that are meaningful for evaluating the clinical plausibility of generates. If we were to re-train an Inception network purely to calculate the FID for medical imagery, the network would be trained to classify an image according to its class labels. If the current weights of the network are already able to provide such a separation, then there is no need to re-train and the metric can in fact be applied to the problem at hand.

To evaluate the extent to which the FID can discriminate between individual pathologies we utilise the *No Finding* label as a baseline as it principally should be absent of any significant changes while still varying similarly along protected patient parameters such as sex and age. We then evaluate the distance between it and the set of X-rays containing each other label, the results of which can be seen in table with the addition of a split along sex and the overall comparison between real and synthetic images 1. The results provide evidence that the FID is able to distinguish between the labelled pathologies and a baseline X-ray with the distance seemingly recapitulating the relative scale of the

**Table 1** Fréchet Inception Distance (FID) per Dataset Split

Split	FID	Split	FID
Stratified	0.53	Hernia	15.96
Sex	7.87	Infiltration	20.05
No Finding	0.00	Mass	10.06
Atelectasis	19.90	Nodule	6.22
Cardiomegaly	14.23	Pleural Effusion	23.90
Consolidation	42.45	Pleural Thickening	13.05
Edema	59.40	Pneumonia	32.05
Emphysema	19.56	Pneumothorax	18.00
Fibrosis	9.72	Synthetic	8.02
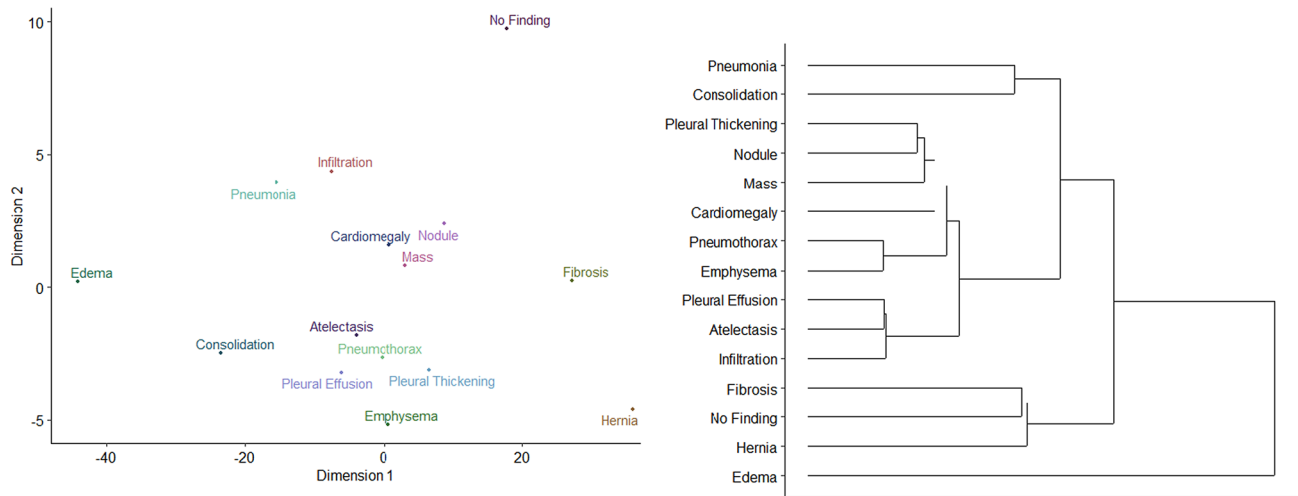
Fig. 5 NIH Class Label FIDs. Left: Classes plotted according to the multidimensional scaling of the pairwise distance matrix calculated by selecting each class as a baseline and determining the distance to all other classes in turn. Right: A hierarchical cluster based on the distance matrix showing the grouping of similar classes based on the features extracted by the Inception network when calculating the FID

Fig. 6 Label prevalence of the original NIH dataset and a random selection of 130,000 PGGAN generates. Point estimates are the average number of each label across all images. 95% Confidence intervals are provided by re-sampling the full set of labels with replacement 10,000 times. The network demonstrates a degree of mode dropping of classes with finer details that it finds difficult to reproduce
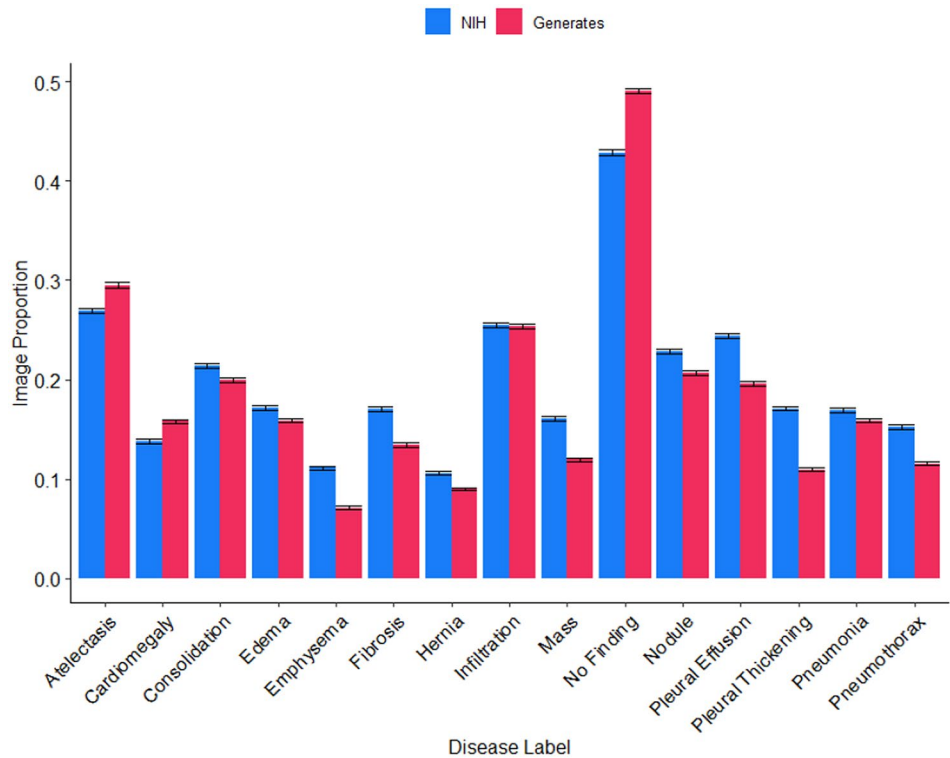
pathological change on the image. Very large distances are associated with the *Edema* and *Consolidation* labels, which are findings that may distort major segments of the lung fields, while *Mass* and *Nodule* labels produce significantly reduced distances as they generally occupy only small segments of the study. To provide more substantial evidence, we calculate a pairwise distance matrix between all classes and utilise multidimensional scaling to plot the relative 2D positions in Fig. 5. The various class locations show that the FID is able to cluster similar labels together. In addition, a hierarchical cluster based on the distance matrix, also shown in Fig. 5, successfully groups sets of related findings. The cluster demonstrates that the features extracted by the network underlying the FID are likely to be of a high enough quality to group sets of related findings and that the

technique is likely applicable to CXRs despite not being explicitly trained to detect features of these images.

We as a result evaluate the quality of synthetic X-rays compared to the NIH source images. The FID between the full source dataset and 100000 random PGGAN generates is 8.02. This value is comparable with FIDs reported for the PGGAN architecture on other non-biomedical high-resolution image generation tasks [18].

Pathological Variability

We estimate the pathological variability of the PGGAN generates in an effort to quantify the extent to which model capture the variability of the various disease entitites contained within the source CXRs. The labels were estimated by utilising a common classifier to label both the NIH and generated iamge sets with confidence intervals derived by bootstrapping the predictions 10,000 times. The proportion of the class labels from the NIH dataset and a random selection of 130,000 generated images can be seen in Fig. 6. Overall it can be seen that the generates cover the range of labels with each represented in broadly similar proportions to the NIH set. Despite the similarity, there appears to be a degree of mode dropping with a majority of labels being significantly less prevalent in the generates compared to the reference set with the exception of *Atelectasis*, *Cardiomegaly*, *Infiltration* and *No Finding*. We hypothesize that this is likely to be due to a continuation of the phenomenon noted with the absence of certain smaller objects, the objects in this case being features of disease. This is supported by the label distribution, as more common conditions with larger disease features tend to be over represented, while diseases with progressively finer features are increasingly sparse in the generated samples. *No Finding* for example, is largely defined by the absence of smaller features and can be seen to be significantly in excess beyond the NIH set. *Atelectasis* and *Cardiomegaly* in comparison, define several characteristics that are present at a moderate resolution and see a smaller increase in prevalence. This is in stark contrast to *Emphysema* and *Pleural Thickening*, which may be quite fine on X-ray and similarly show the greatest reduction in prevalence.

Pathology Generation

We utilise class optimisation to generate pathological images based on the method described in Algorithm 1. We sample the initial latent code from a truncated normal distribution with a threshold of 0.7 [6]. We attempt the method with both a Densenet-121 and a re-purposed Discriminator as classifiers. We generally found optimising the Densenet-121 model proved to be slower, more difficult and less reliable overall. The process frequently plateaus and subsequently fails to produce a sample for a particular class. Successes often rely on favourable sampling from the latent space, with such samples typically converging rapidly. The same process applied to the re-purposed discriminator proved more successful as samples converged more regularly with greater ease in achieving higher logit values for a particular class. This discrepancy is due to the input size of the different models. The Densenet as a result of being trained on a smaller resolution requires images to be resized prior to classification and as such results in a worse estimation of the gradient. Replication of the process with a Densenet trained on a larger input size produced significant improvements to both the reliability and rapidity of convergence. Examples of successful pathology generates can be seen in Fig. 7. Many more examples of multi-modal pathology can be seen in the linked image archive.

Attempts to modify the optimisation process to produce examples of isolated labels often fails to converge, with optimisation by minimising other class scores alongside the maximisation of the class of interest typically resulting in a significant limitation to the score that can be achieved. This phenomenon is explained by the multi-modal nature of the disease classes and the underlying biomedical relationship between classes that results in label co-occurrence. These relationships can be seen with increases to the logit threshold when optimising for a particular class. Clearer examples of particular conditions tend to co-occur with their medical complications. Increasing class scores for *Cardiomegaly* begin to include a greater proportions of *Pleural Effusion* labels, this may be related to the complications that occur in heart failure that may cause images with these labels to co-occur. Similarly, increasing *Emphysema* class values produces associated *Pneumothorax* labels, while *Nodule* findings produce associated *Mass* labels. These findings reflect that the underlying unsupervised training methodology has captured the relationship between various X-ray features and optimising for the pathology enables us to tease out these properties.

As part of class optimisation, we're required to set a score threshold for determining when a proposed image has converged to be representative of a particular class. We utilise Youden's J statistic to determine the optimum cut-off based on the validation ROC curve per class when training the classifier. To simulate degrees of disease severity in produced images, we randomly increase this threshold by the absolute value of the distribution of $\mathcal{N}(1, 1)$ when optimising. It is quite seldom that a dataset includes only a single image of a patient. Typically there are series of patient images, where individuals are re-scanned several times. There are a multitude of potential reasons for this, follow-up on patient condition, assessment of medical intervention, or routine imagery prior to surgery, are just a handful of potential reasons. The end result is that most datasets have a
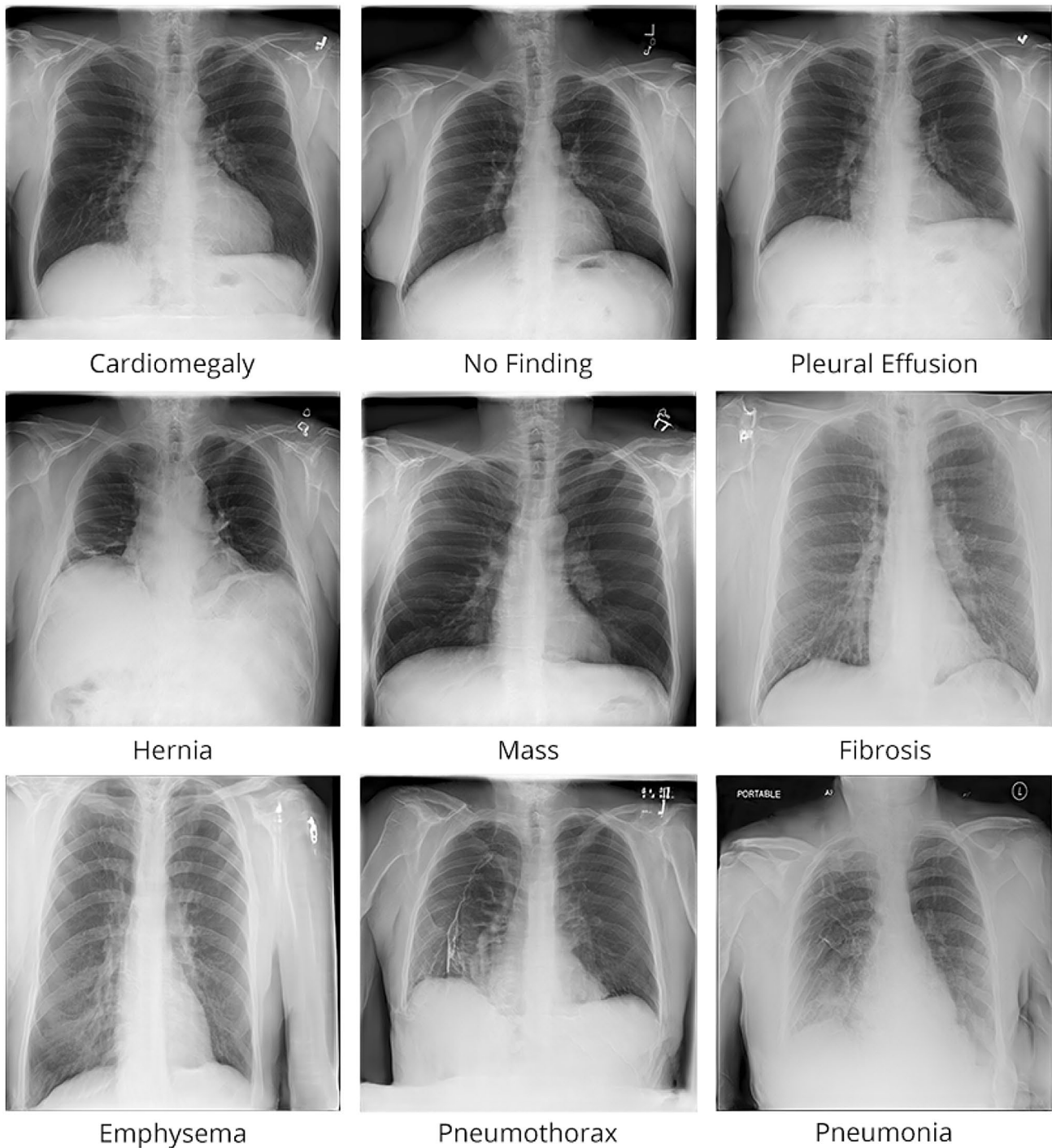
Fig. 7 Example classes generated by maximising the classifier class logit. Images often have multiple findings, the finding optimised is the label given

significant proportion of very similar images, wherein position or disease severity may be altered slightly but the overall image is largely unchanged. We simulate this process by sampling in the vicinity of a particular pathology-optimised latent code. We treat the optimised location as a centre and sample the surrounding hypersphere within $\mathcal{N}(0, 0.2)$. We

find that this enables the creation of images with the same set of diagnoses but minor changes in orientation, exposure, or disease severity. The degree of variability can be tuned by adjusting the standard deviation; however, we find a value of 0.2 tends to produce minor variations without compromising the diagnostic label (Fig. 8).

Fig. 8 Example of local sampling after class generation. Images show similar anatomical and pathological features with variations in image contrast, text markup, and patient positioning

The fact that the latent space may be optimised with regard to disease classification at all implies a degree of orientation with respect to pathology which may hold semantic value. We leave a full exploration of the properties of the latent space to future work.

Radiologist Review

We evaluate the plausibility of generated X-rays to domain experts through a series of image reviews. We ask a group of practising radiologists to identify a set of real images out of a grid of mixed X-rays as well as to select a real image out of a pair of real and generated images. We received feedback from 3 consultant (equivalent to board certified) radiologists, and 5 radiology registrars (equivalent to residents in North America). Image grids consist of six X-rays that are produced with a 50% probability of being real or generated, we evaluate six such grids. Real images were identified as such by radiologists 73% (95% CI 63, 82) of the time, while generates were identified as real 61% (95% CI 51, 70) of the time, with both groups more likely than chance to be identified as real, with reals $p < 0.0001$ and generates $p = 0.0155$ for a one-sided t-test. Radiologist performance on X-ray discrimination equates to a HYPE$_\infty$ score of 33% (95% CI 24, 43). Discrimination of image pairs is clearly in favour of real images, with radiologists correctly identifying the true image of a pair 71% (95% CI 55, 86) of the time. We evaluate 20 such image pairs. Comments from respondents mentioned bony abnormalities as the main feature used to identify samples as being synthetic. Images were embedded as options within a Google Form and shown at a 260×260 resolution with unlimited time to review images prior to making a decision.

Synthetic Model Training

For evaluating the classification performance of a model trained only on synthetic data, we generate an archive of CXRs of comparable size and with labels in a similar proportion to the source NIH dataset. We uniformly sample between 3 and 6 related images per class optimisation operation and treat them as belonging to the same 'patient'. We

similarly iteratively stratify the synthetic patients into training and validation sets in the same manner as during initial classifier training, retaining only the original test set of real images, which we use for all comparisons. We train a Densenet classifier in the same configuration as described in Sect. 3.3. We evaluate performance relative to real images as well as images embedded into the generator's latent space. We consider embedding images as the current alternative to individual class generation, whereby a network predicts the latent encoding of a given image. To achieve this, we retrain the discriminator to predict such a representation given a random sample from the generator and minimising the mean squared error between the predicted and actual locations in the latent space. We train with the proposed embedded image alongside the original labels. A table of comparative ROC AUC values can be seen in Table 2

The class generation results provide further evidence that the PGGAN model has learnt to replicate all the classes within the original NIH dataset, as the synthetic classifier is able to detect each pathology from the source to a

Table 2 Comparison of Performance of Classification Models Trained on Synthetic vs Real Data

Class Label	Embedded	Class Generated	Real
Atelectasis	0.646	0.851	0.923
Cardiomegaly	0.602	0.907	0.968
Consolidation	0.675	0.928	0.963
Edema	0.746	0.941	0.985
Emphysema	0.51	0.901	0.972
Fibrosis	0.63	0.868	0.961
Hernia	0.611	0.951	0.986
Infiltration	0.596	0.796	0.847
Mass	0.516	0.825	0.936
No finding	0.605	0.861	0.932
Nodule	0.543	0.748	0.909
Pleural effusion	0.636	0.903	0.954
Pleural thickening	0.583	0.881	0.954
Pneumonia	0.595	0.932	0.975
Pneumothorax	0.568	0.877	0.947
Average	0.604	0.878	0.947

degree significantly greater than expected by chance, despite never actually seeing an image from the original dataset. Furthermore, the generated classes demonstrate that the method reliably produces examples of each class, with a performance reduction similar to that seen with previous PGGAN generation tasks [51]. These results should scale with improvements made to the quality of generated images, as augments to the underlying GAN will reduce the distinction between the synthetic and real sets, which alongside class generation, will allow for even better representations for training. The poor performance of embedded samples are indicative of such a method failing to preserve class-specific information, and show it to be less suitable of a method for producing anonymised clinical archives. A potential alternative to predictive embedding would be to optimise the latent space to minimise image reconstruction loss. Our experiments with this showed promise for more faithful representations; however, the performance was profoundly worse than alternatives, requiring days to convert an archive compared to only a handful of hours for competing approaches. We attempted comparisons with generating classes through a simpler DCGAN architecture similarly trained on the full NIH dataset. We found the generation process intractable as the model often failed to produce class scores sufficient to be considered examples of that class despite extensive optimisation. We attribute this to a combination of mode dropping and lower image quality which precludes the formation of certain classes.

These results provide substantial evidence that class optimisation is effective at producing images that are representative of a particular disease label and can enable a single unsupervised GAN to produce a fully labelled cohort of classes from an anonymised dataset.

Discussion

Through both automated and expert review of the medical plausibility of generated X-rays as well an evaluation of synthetic model performance we have outlined the properties of the PGGAN architecture applied to CXRs. Expert review demonstrated global coherence of the imagery with reproduction of numerous features associated with varying disease classes. Deficiencies have been noted with a broad absence of detailed, small scale features that are typically easily discernible on X-ray films. Automated comparisons of image features by means of the FID provide evidence that the quality of generates is of a similar standard to typical high-resolution tasks, where PGGAN is applied.

These limitations are likely explained by stochastic bottlenecks and slower convergence of novel features at higher resolutions and are inherent to the design and training formulation employed by the baseline PGGAN methodology.

Future work should follow improvements made to subsequent GAN architectures in the PGGAN lineage [19, 20], or alternatively examine alternate approaches such as Very Deep Variational Autoencoders (VAEs) [7] which have similarly shown promise for high-resolution image synthesis. Irrespective of the exact mechanism used to generate images, the implementation of multi-modal class optimisation allows for the extraction of disease-representative classes from the latent space that can be utilised to construct labelled synthetic datasets of arbitrary size.

We suspect this method will prove to be of even greater utility with GAN architectures that possess less entangled latent spaces and could allow for disease optimisation at varying resolutions by considering scale-specific elements of the extended latent space seen in StyleGAN and later works [19]. We envision that such techniques should allow significant control over the severity of disease present in generates and may provide profound capabilities for dataset augmentation. This would allow for applications such as demonstrating potential progression or resolution of disease or retaining disease presentation but shifting patient particular parameters such as age or sex. Scale specific modifications may enable the synthesis of support devices such as IV lines that may be differed against the source image and used for augmentation in segmentation tasks. We envision that continued development along this route will enable task-specific image extraction from broad image generators, as unsupervised GANs trained on large corpora of diverse images should be able to extract a multitude of classes based on the capabilities of the associated classifier, even if such classes were not labelled in the original individual datasets.

Conclusion

We have applied a progressively growing GAN model to the task of synthesising high-resolution chest X-ray images for the evaluation of their suitability as a replacement for standard images for the tasks of model development and student education. The overarching goal of this investigation being to improve the protection of patient privacy without compromising data availability. We evaluate the applicability of the Fréchet Inception Distance to the evaluation of synthetic chest X-rays and find that the underlying network is capable of providing a meaningful metric for generate quality despite the difference in data distribution. We demonstrate that it is possible to produce realistic, clinically plausible images that capture much of the variation in standard X-rays; however, there remains a significant need for improvement in the reproduction of small-scale details to achieve truly indistinguishable samples. We demonstrate that the model is capable of reproducing all abnormalities of interest and in

similar proportions to the source image distribution despite differences in small scale features. We describe a methodology for the extraction of class representative images from the generator's latent space by optimisation of a classifier score and demonstrate that such a method is capable of constructing a fully labelled synthetic dataset from an unsupervised generator. We describe potential avenues of improvement for generate quality and anticipate that such improvements will enable the production of high quality teaching and training images without the concern of breaching patient confidentiality. We make the source code, final model weights and a large archive of labelled generates used in this study available to the broader research community.

Declarations

Conflicts of Interest The authors declare that they have no conflict of interest.

Ethics Approval An ethics waiver was issued for this work by the Human Research Ethics Committee (Medical) of the University of the Witwatersrand, Johannesburg on 11/08/2020.

References

1. Abdal R, Qin Y, Wonka P. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space? In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019;pp. 4431–4440. IEEE, Seoul, Korea (South) . https://doi.org/10.1109/ICCV.2019.00453

2. Adam A, Dixon AK, Gillard JH, Schaefer-Prokop C, Grainger RG. Current status of thoracic imaging. In: Grainger & Allison's diagnostic radiology: a textbook of medical imaging, 2021;p. 3. Elsevier.

3. Arjovsky M, Bottou L. Towards principled methods for training generative adversarial networks. arXiv:1701.04862 [cs, stat]. 2017.

4. Beers A, Brown J, Chang K, Campbell JP, Ostmo S, Chiang MF, Kalpathy-Cramer J. High-resolution medical image synthesis using progressively grown generative adversarial networks. arXiv:1805.03144 [cs] 2018.

5. Bowles C, Chen L, Guerrero R, Bentley P, Gunn R, Hammers A, Dickie DA, Hernández MV, Wardlaw J, Rueckert D. GAN augmentation: augmenting training data using generative adversarial networks. arXiv:1810.10863 [cs] 2018.

6. Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. In: International conference on learning representations. 2019.

7. Child R. Very deep VAEs generalize autoregressive models and can outperform them on images. In: International conference on learning representations. 2021.

8. Creswell A, Bharath AA. Inverting the generator of a generative adversarial network. IEEE Trans Neural Netw Learn Syst. 2019;30(7):1967–74. https://doi.org/10.1109/TNNLS.2018.2875194.

9. Falcon W. PyTorch lightning. GitHub. https://github.com/PyTorchLightning/pytorch-lightning 2021;**3**.

10. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Proceedings of the 27th international conference on neural information processing systems - Volume 2, NIPS'14, 2014;pp. 2672–2680. MIT Press, Cambridge, MA, USA.

11. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. Improved training of wasserstein GANs. In: Proceedings of the 31st International conference on neural information processing systems, NIPS'17, pp. 5769–5779. Curran Associates Inc., Red Hook, NY, USA. 2017.

12. Gündel S, Grbic S, Georgescu B, Liu S, Maier A, Comaniciu D. Learning to recognize abnormalities in chest x-rays with location-aware dense networks. Lecture Notes Comput Sci. 2019;pp. 757–765. https://doi.org/10.1007/978-3-030-13469-3_88

13. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the 31st international conference on neural information processing systems, NIPS'17, pp. 6629–6640. Curran Associates Inc., Red Hook, NY, USA. 2017.

14. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), 2017;pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243

15. Irvin J, Rajpurkar P, Ko M, Yu Y., Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K, Seekins J, Mong DA, Halabi SS, Sandberg JK, Jones R, Larson DB, Langlotz CP, Patel BN, Lungren MP, Ng AY. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. arXiv:1901.07031 [cs, eess] 2019.

16. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), 2017;pp. 5967–5976. IEEE, Honolulu, HI. https://doi.org/10.1109/CVPR.2017.632

17. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng Cy, Mark RG, Horng S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data. 2019;6(1):317. https://doi.org/10.1038/s41597-019-0322-0.

18. Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. In: International conference on learning representations. 2018.

19. Karras T, Laine S, Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 4396–4405. IEEE, Long Beach, CA, USA, 2019. https://doi.org/10.1109/CVPR.2019.00453

20. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and Improving the Image Quality of StyleGAN. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), 2020;pp. 8107–8116. IEEE, Seattle, WA, USA. https://doi.org/10.1109/CVPR42600.2020.00813

21. ...Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F, Dong J, Prasadha MK, Pei J, Ting MY, Zhu J, Li C, Hewett S, Dong J, Ziyar I, Shi A, Zhang R, Zheng L, Hou R, Shi W, Fu X, Duan Y, Huu VA, Wen C, Zhang ED, Zhang CL, Li O, Wang X, Singer MA, Sun X, Xu J, Tafreshi A, Lewis MA, Xia H, Zhang K. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell. 2018;172(5):1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010.

22. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: ICLR (Poster) 2015.

23. Korkinof D, Heindl A, Rijken T, Harvey H, Glocker B. MammoGAN: High-resolution synthesis of realistic mammograms. In: International conference on medical imaging with deep learning - Extended Abstract Track. London, United Kingdom. 2019.

24. Ledig C, Theis L, Huszar F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Tot, J, Wang Z, Shi W. Photo-realistic single image super-resolution using a generative adversarial network. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), 2017;pp. 105–114. IEEE, Honolulu, HI. https://doi.org/10.1109/CVPR.2017.19

25. Lucic M, Kurach K, Michalski M, Bousquet O, Gelly S. Are GANs created equal? A large-scale study. In: Proceedings of the 32nd international conference on neural information processing systems, NIPS'18, pp. 698–707. Curran Associates Inc., Red Hook, NY, USA. 2018.

26. Majkowska A, Mittal S, Steiner DF, Reicher JJ, McKinney SM, Duggan GE, Eswaran K, Cameron Chen PH, Liu Y, Kalidindi SR, Ding A, Corrado GS, Tse D, Shetty S. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. Radiology. 2020;294(2):421–31. https://doi.org/10.1148/radiol.2019191293.

27. Mao X, Li Q, Xie H, Lau RY, Wang Z, Smolley SP. Least squares generative adversarial networks. In: 2017 IEEE International conference on computer vision (ICCV), 2017;pp. 2813–2821. IEEE, Venice. https://doi.org/10.1109/ICCV.2017.304

28. Mescheder LM, Geiger A, Nowozin S. Which training methods for GANs do actually converge? In: ICML. 2018.

29. Mirza M, Osindero S. Conditional generative adversarial nets. arXiv:1411.1784 [cs, stat] 2014.

30. Moradi M, Madani A, Karargyris A, Syeda-Mahmood TF. Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In: E.D. Angelini, B.A. Landman (eds.) Medical Imaging 2018: Image Processing, p. 57. SPIE, Houston, United States. 2018. https://doi.org/10.1117/12.2293971

31. Oakden-Rayner L. Exploring large-scale public medical image datasets. Acad Radiol. 2020;27(1):106–12. https://doi.org/10.1016/j.acra.2019.10.006.

32. Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier GANs. In: D. Precup, Y.W. Teh (eds.) Proceedings of the 34th international conference on machine learning, *Proceedings of Machine Learning Research*, 2017;vol. 70, pp. 2642–2651. PMLR, International Convention Centre, Sydney, Australia.

33. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med. 2020;121:103792. https://doi.org/10.1016/j.compbiomed.2020.103792.

34. Park N, Mohammadi M, Gorde K, Jajodia S, Park H, Kim Y. Data synthesis based on generative adversarial networks. Proceedings of the VLDB Endowment. 2018;11(10):1071–83 (**10.14778/3231751.3231757**).

35. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434 [cs] 2016.

36. ...Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz CP, Patel BN, Yeom KW, Shpanskaya K, Blankenberg FG, Seekins J, Amrhein TJ, Mong DA, Halabi SS, Zucker EJ, Ng AY, Lungren MP. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLOS Med. 2018;15(11):e1002686. https://doi.org/10.1371/journal.pmed.1002686.

37. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K, Lungren MP, Ng AY. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv:1711.05225 [cs, stat] 2017.

38. Reed S, Akata Z, Yan X, Logeswaran L, Schiele B, Lee H. Generative adversarial text to image synthesis. In: M.F. Balcan, K.Q. Weinberger (eds.) Proceedings of the 33rd international conference on machine learning, *Proceedings of Machine Learning Research*, vol. 48, pp. 1060–1069. PMLR, New York, New York, USA. 2016.

39. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet large scale visual recognition challenge. Int J Comput Vis. 2015;115(3):211–52. https://doi.org/10.1007/s11263-015-0816-y.

40. Salehinejad H, Colak E, Dowdell T, Barfett J, Valaee S. Synthesizing chest X-Ray pathology for training deep convolutional neural networks. IEEE Trans Med Imaging. 2019;38(5):1197–206. https://doi.org/10.1109/TMI.2018.2881415.

41. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. In: Proceedings of the 30th international conference on neural information processing systems, NIPS'16, pp. 2234–2242. Curran Associates Inc., Red Hook, NY, USA. 2016.

42. Santos MK, Ferreira Júnior JR, Wada DT, Tenório APM, Barbosa MHN, Marques PMA. Artificial intelligence, machine learning, computer-aided diagnosis, and radiomics: advances in imaging towards to precision medicine. Radiologia Brasileira. 2019;52(6):387–96. https://doi.org/10.1590/0100-3984.2019.0049.

43. Segal B, Rubin D, Rubin G, Pantanowitz A. CXR PGGAN Code. GitHub. Note: https://github.com/BradSegal/CXR_PGGAN. 2021.

44. Segal B, Rubin D, Rubin G, Pantanowitz A. CXR PGGAN model and library. Kaggle. Note: https://kaggle.com/bradsegal/synthetic-pggan-chest-xrays. 2021.

45. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: Fairness gaps in deep chest X-ray classifiers. In: Biocomputing 2021, 2020;pp. 232–243. WORLD SCIENTIFIC, Kohala Coast, Hawaii, USA. https://doi.org/10.1142/9789811232701_0022

46. Shamout FE, Shen Y, Wu N, Kaku A, Park J, Makino T, Jastrzębski S, Witowski J, Wang D, Zhang B, Dogra S, Cao M, Razavian N, Kudlowitz D, Azour L, Moore W, Lui YW, Aphinyanaphongs Y, Fernandez-Granda C, Geras KJ. An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department. arXiv:2008.01774 [cs, eess]. 2020.

47. Shen Y, Gu J, Tang X, Zhou B. Interpreting the Latent Space of GANs for Semantic Face Editing. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 9240–9249. IEEE, Seattle, WA, USA. 2020. https://doi.org/10.1109/CVPR42600.2020.00926

48. Shmelkov K, Schmid C, Alahari K. How Good Is My GAN? In: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (eds.) Computer Vision—ECCV 2018, vol. 11206, pp. 218–234. Springer International Publishing, Cham. 2018. https://doi.org/10.1007/978-3-030-01216-8_14

49. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. IEEE Conf Comput Vis Pattern Recognit (CVPR). 2016;2016:2818–26.

50. Thanki R, Borra S, Dey N, Ashour AS. Medical Imaging and Its Objective Quality Assessment: An Introduction. In: N. Dey, A.S. Ashour, S. Borra (eds.) Classification in BioApps, vol. 26, pp. 3–32. Springer International Publishing, Cham. 2018. https://doi.org/10.1007/978-3-319-65981-7_1

51. Togo R, Ogawa T, Haseyama M. Synthetic gastritis image generation via loss function-based conditional PGGAN. IEEE Access. 2019;7:87448–57. https://doi.org/10.1109/ACCESS.2019.2925863.

52. Wang L, Wong A. COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images. arXiv:2003.09871 [cs, eess] 2020.

53. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-Ray8: Hospital-scale chest x-ray database and benchmarks on

weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), 2017;pp. 3462–3471. https://doi.org/10.1109/CVPR.2017.369

54. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review. Med Image Anal. 2019;58:101552. https://doi.org/10.1016/j.media.2019.101552.

55. You C, Li G, Zhang Y, Zhang X, Shan H, Ju S, Zhao Z, Zhang Z, Cong W, Vannier MW, Saha PK, Wang G. CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble(GAN-CIRCLE). IEEE Trans Med Imaging. 2020;39(1):188–203. https://doi.org/10.1109/TMI.2019.2922960.

56. Zhou S, Gordon ML, Krishna R, Narcomey A, Fei-Fei L, Bernstein MS. HYPE: A Benchmark for human eYe perceptual evaluation of generative models. http://arxiv.org/abs/1904.01121arXiv:1904.01121 [cs] 2019.

57. Zhou SK, Greenspan H, Davatzikos C, Duncan JS, van Ginneken B, Madabhushi A, Prince JL, Rueckert D, Summers RM. A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. Proc IEEE. 2021;1–19. https://doi.org/10.1109/JPROC.2021.3054390.

58. Zhu JY, Park T, Isola P, Efros AA. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. arXiv:1703.10593 [cs] (2018)