Synthesising Clinically Realistic Chest X-Rays using Generative Adversarial Networks

Bradley Segal

Supervised by Professor David M. Rubin

School of Electrical and Information Engineering

University of the Witwatersrand

Johannesburg, South Africa

A dissertation submitted to the Faculty of Engineering and the Built Environment, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science in Engineering.

Johannesburg, 2021

Declaration

I declare that this dissertation is my own unaided work. It is being submitted to the Master of Science in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination to any other University.

Bradley Segal

Date: 18/11/2021

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.

This project addresses these concerns by introducing latent class optimisation which enables efficient, multi-modal sampling from a GAN and with which a large archive of labelled generates is synthesised. This project applies a Progressive Growing GAN (PGGAN) to the task of unsupervised x-ray synthesis and has radiologists evaluate the clinical realism of the resultant samples. 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 is provided. 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 is also validated. X-ray clinical realism is quantified by asking radiologists to distinguish between real and synthesised scans with the finding that generates are more likely to be classed as real than by chance, however, there remains progress required to achieve true realism. These findings are confirmed through the evaluation of a synthetic classification model performance on real scans. The project concludes with a discussion the limitations of PGGAN generates and how to achieve controllable, realistic generates in future work.

Acknowledgements

I thank my supervisors David Rubin, Adam Pantanowitz, and Grace Rubin for their oversight, guidance, and support throughout my studies. More specifically, I acknowledge David Rubin for convening this project and enabling me to pursue my interests at the intersection of medicine and machine learning. I thank Adam Pantanowitz for his expertise in machine learning and cloud infrastructure as well as his feedback and discussion throughout. In addition, I am grateful to Grace Rubin for her radiology expertise and time in reviewing study samples.

I thank Michael Klipin for starting me on the pathway of looking beyond the expected pathway for medicine to pursue that which interests me.

Contents

List of Figures

List of Tables

List of Symbols

Symbol	Description
α	Pixel Value
λ	Penalty Coefficient
$\mathbb{P}$	Probability Distribution
$\mathbb{P}_g$	Generated Probability Distribution
$\mathbb{P}_r$	Real Probability Distribution
$\mathcal{N}$	Normal Distribution
μ	Mean
σ	Standard Deviation
σ^2	Variance
θ	Interpolation Fraction
C	Image Channels
c	Transport Cost
γ	Transport Plan
D	Discriminator
$\mathcal{D}$	1-Lipschitz Constrained Discriminator
F	Image Classifier
G	Generator
I	Image
K	Kernel
L	Loss Value
N_N	Number of Negative Samples
N_P	Number of Positive Samples
T	Threshold
W_N	Negative Class Weight
W_P	Positive Class Weight
x	Sample
$\hat{x}$	Interpolated Sample
χ	Dataset
y	Class Ground Truth
$\hat{y}$	Class Predicted Probability
S	Class Predicted Logits
Z	Latent Vector
z	Latent Vector Sample

Nomenclature

Term/Abbreviation	Definition
ANN	Artificial Neural Network
AP	Anteroposterior
AUROC	Area Under the Receiving Operator Characteristic
AWS	Amazon Web Services
BN	Batch Normalisation
CAD	Computer Aided Diagnosis
CNN	Convolutional Neural Network
CV	Computer Vision
CXR	Chest X-Ray
DCGAN	Deep Convolutional GAN
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
EMD	Earth Mover Distance
FID	Fréchet Inception Distance
GAN	Generative Adversarial Network
GP	Gradient Penalty
GPU	Graphics Processing Unit
HoG	Histograms of Orientated Gradients
HYPE	Human eYe Perceptual Evaluation
JSD	Jensen-Shannon Divergence
KLD	Kullback-Leibler Divergence
ML	Machine Learning
NIH	National Institutes of Health
NLP	Natural Language Processing
PA	Posteroanterior
PGGAN	Progressive Growing GAN
ROC	Receiver Operating Characteristic
SIFT	Scale-Invariant Feature Transform
SOTA	State of the Art
TPU	Tensor Processing Unit
VAE	Variational Autoencoder
WGAN	Wasserstein GAN

Chapter 1

Introduction

The chest radiograph (CXR) is the most common diagnostic radiological procedure [1] 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 [2, 3, 4], with a recent surge in activity due to the COVID-19 pandemic [5, 6, 7]. 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 [4, 8, 9]. This activity in CXRs represents a section of the development of Computer Aided Diagnosis (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 [9, 10]. Large scale, anonymised, public imaging datasets underscore these efforts by providing the necessary clinical data for the training of CAD models [8, 11, 12]. 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 are mined for diagnostic labels [4, 8, 9]. 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 [11]. 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 [13]. 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 [13]. 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 [14].

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 [15, 16, 17] without compromising patient confidentiality as they learn to replicate the source distribution without access to the underlying training data [18]. 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 project serves to comprehensively evaluate the medical plausibility of synthetic CXRs as well as their applications to diagnostic radiology [1]. In pursuit of this, the dissertation evaluates the Progressively Growing GAN (PGGAN) [19] methodology applied to generate multi-modal, megapixel resolution CXRs. Domain expert examination of the properties of generated images as well as a review of expert discrimination between real and generated samples is provided. In addition, this project includes the following research 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; and
- an evaluation of the performance of synthetic image datasets derived from multi-modal generators.

Project source code is available [20], as well as model weights and a collection of labelled generated images [21].

[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.

1.1 Problem Statement

The advancement of CAD systems requires the development of techniques capable of resolving the seemingly incompatible tasks of increased data availability and improved patient privacy. GANs offer a potential solution to both problems through the synthesis of high resolution anonymous surrogates. Current approaches to the implementation of GANs are limited by an inability to control multi-modal conditional image generation as well as the resolution of current systems. CXRs, as well as many other medical image domains, are inherently multi-modal and require a broad solution capable of producing high fidelity samples from a single GAN, as the training computational requirements render it infeasible to produce class-specific networks.

1.2 Research Objectives

This dissertation aimed to evaluate the realism of synthetic chest x-rays produced via a PGGAN architecture. This overarching goal is achieved through the following objectives:

1. Train a PGGAN in an unsupervised manner to generate x-ray images;
2. Train an x-ray diagnostic model for identifying pathological features present in x-ray images;
3. Evaluate the diversity of pathological features identified within randomly sampled generated x-rays relative to the source dataset;
4. Review of the realism of features seen in generated x-rays relative to their original, real counterparts by a domain expert;
5. Conduct an experimental study to evaluate the degree to which synthetic x-rays are classified as real by domain experts;
6. Develop a method for broad multi-model class generation from a single trained GAN; and
7. Evaluate the performance of a diagnostic model trained using synthetic class generates compared to the baseline trained with the real dataset.

1.3 Dissertation Outline

This dissertation is composed of five chapters each of which are described here:

Chapter 1: This chapter provides a background on CXRs, GANs, and the necessity for methods capable of improving data access while improving patient protection. In addition, it includes a description of the problem statement as well as the research objectives.

Chapter 2: The literature review is contained herein and provides a comprehensive overview of the field of deep learning as well as its applications for image generation and classification. Particular focus is given to contextualising the techniques employed throughout the dissertation.

Chapter 3: This chapter explains the data source used for all experiments as well as outlining the methodology employed in constructing the computational models used for both generation and classification. Hyperparameter and training configuration details are included for model development.

Chapter 4: The experimental outcomes are included in this chapter. The contents each discuss the results in the context of an element of the research objectives. A demonstration of the various properties of generated images are included. The chapter concludes with recommendations for future work.

Chapter 5: This chapter concludes the work performed.

Chapter 2

Background

2.1 Deep Learning

Machine learning (ML) is the study of algorithms and techniques that enable computer programs to automatically improve through experience [22]. Deep learning (DL) is a subset of machine learning that makes use of multiple processing layers to facilitate programs adapting to utilise multiple levels of abstraction. DL methods are a form of representation learning that constructs representations from raw data through the composition of layers of differentiable, non-linear functions. These representations allow for the achievement of a particular goal, such as object classification, in relation to the raw data that may not be readily achieved with conventional machine learning techniques. These functions are parameterised by the use of the backpropagation algorithm which allows for the optimisation of functions with respect to a task of interest via gradient descent [23]. The functions utilised are typically artificial neural networks (ANN) [23] which are computational frameworks inspired by the connectivity inherent to the biological networks of animal brains that encode representations in the interactions between neural nodes [24]. The exact architecture of ANNs is largely specified by the task of interest and is typically informed by the properties of the data used in the task [23]. The advent of DL has produced significant advances in machine performance across a variety of domains such as speech recognition, visual object recognition, object detection, and many others [23].

2.2 Generative Adversarial Networks

Generative Adversarial Networks are a class of implicit generative models rooted in game theory, with the objective of training a generator to produce samples indistinguishable from a source distribution [25]. GANs are typically formulated by two networks in opposition to one another: a generator (G) whose goal is to generate plausible samples, and a discriminator (D), whose goal is to detect such samples. The generator operates by transforming a sample from a noise vector (z) into a sample ($G(z)$) which is similar to a chosen reference distribution, with the training signal provided by how effectively such

a generate can fool the discriminator into classifying it as real. The discriminator operates by determining whether a sample (x) belongs to the real ($\mathbb{P}_r$) or the generated ($\mathbb{P}_g$) distributions. This configuration optimally reaches a Nash equilibrium when the generator produces samples indistinguishable from the reference set and the discriminator can no longer learn to detect generates [26, 27]. A Nash equilibrium is a steady state solution in game theory wherein each player knows the actions of the other players and cannot increase their own returns by changing their chosen action [28].

The training of GANs to reach the Nash equilibrium remains an evolving problem as the optimisation of such networks requires the minimisation of a cost function in the hopes that a solution will be similar to the equilibrium. Numerous alterations to the training setup and loss functions have been utilised to improve sample quality and variability as well as the reliability of GAN training convergence. These improvements have enabled networks to reliably learn more complex distributions, convergence to the Nash equilibrium however, is still not guaranteed [26, 27].

GANs have been utilised for a variety of tasks, namely image-to-image translation [29, 30], image super-resolution [31, 32], and semantic image editing [33] to name just a few. Perhaps the most common usage is that of image synthesis, wherein a GAN attempts to produce images that mirror a reference image set and can be applied to producing novel mixtures of elements from the set.

The technique was first proposed by Goodfellow *et al.* [25], with the formulation described as a two-player minimax game with the following value function:

$$\min_G \max_D \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r}[\log D(x)] + \mathop{\mathbb{E}}_{z\sim\mathbb{P}_g}[\log(1 - D(G(z)))] \tag{2.1}$$

A minimax game refers to the objective of a player in such a game, wherein they are tasked with finding an action that minimises their losses based on the actions that their counterparts may take, or maximises their returns if the actions of their counterparts are known. In games that are strictly competitive, such as the GAN optimisation construct, a Nash equilibrium may be achieved if and only if the chosen actions of all players are minimaximisers [28]. The GAN value function constructs a strictly competitive interaction that similarly may only achieve equilibrium if the individual networks select optimal minimax actions. The discriminator's optimal play is to adequately separate out generates from reals, this minimaxes the return value by ascribing a high value for reals in the first term while minimising the ability of the generator to reduce its reward in the second term by producing a low value for the generated outputs. In contrast, the generator's objective is reversed and must minimise the value function. This necessitates producing more samples that the discriminator would classify as real and less that are clearly false. This would minimise the second term and reduce the overall function as a result. In addition to this more obvious goal, the generator has a more subtle ability to manipulate the overall

value function due to the discriminator adapting to the properties of synthetic samples. As the quality of samples more closely mirror the real distribution, the training dynamics for the discriminator become increasingly erratic which worsens its ability to confidently separate out the distributions, reducing the value of the first term. These objectives reach an equilibrium when generates provide no information that can be utilised to discriminate them from real samples and the discriminator simply ascribes equal likelihood for samples to belong to either set. This state simultaneously minimises the potential losses for the discriminator in the case of perfect generates and for the generator for an ideal discriminator. This can be seen to be a Nash equilibrium as neither network can improve their returns by deviating from the given strategies.

The GAN value function for the discriminator is equivalent to the standard cross entropy loss utilised in most binary classification tasks. This constrains the score in the range $[0, 1]$, and can be thought of as the discriminator being tasked with binary classification of a sample between being real $\{1\}$ or synthetic $\{0\}$. This converges fairly easily for the discriminator, however, as mentioned by Goodfellow *et al.* [25], this often results in the generator's gradient for the term $\log(1-D(G(z)))$, saturating with a resultant failure of convergence. Goodfellow *et al.*, instead propose training the generator to maximise $\log D(G(z))$, as the gradient it provides allows for more stable training when the discriminator significantly outperforms the generator. Despite this attempt to circumvent gradient saturation, the overall training process remains remarkably unstable, with generator updates generally worsening as the discriminator improves [34].

Arjovsky & Bottou [34], discuss this instability as a result of the chosen value function in association with the underlying distributions $\mathbb{P}_r$ and $\mathbb{P}_g$ being potentially disjointed or existing on non-overlapping lower dimensional manifolds. This formulation would then allow for the discriminator to find a method for perfectly separating the distributions, despite the manifolds potentially lying arbitrarily close to one another. As a result, the samples produced, no matter how broadly similar in appearance, would be discriminated and cause unstable generator training as the gradient would saturate despite the prior restatement of the optimisation problem. Arjovsky & Bottou [34], demonstrate this phenomenon experimentally via a Deep Convolutional GAN (DCGAN) model [35], trained to produce sharp, realistic images. The DCGAN [35] was a step towards improving generated resolutions by making use of convolutional layers as opposed to the fully connected approach of the original GAN [25]. Despite scale improvements, this modification did not resolve fundamental problems with training instability that caused converge failures or mode dropping [34]. These issues become more prominent with increasing resolutions as the discriminator is more readily capable of detecting generated samples [19]. This significantly constrains the upper limit of resolutions that a DCGAN can achieve. Despite the image quality achieved by Arjovsky & Bottou [34] with the DCGAN trained using the standard value function, a new discriminator can be trained to separate real from generated samples with perfect accuracy,

despite the samples being grossly indistinguishable at the chosen resolution.

The restatement of the generator's cost function to maximise the discriminator's scoring of synthetic images, instead of minimising the detection of such images has been demonstrated to alter training dynamics beyond the quality of gradient information. The function penalises obviously synthetic samples, however, it does not ascribe a significantly cost to mode dropping [34]. This is an often seen property of GAN training wherein the variety of generated samples decreases as training progresses as the generator selects for images that appear the most realistic at the expense of class diversity. In extreme cases this may lead to mode collapse, where only an extremely limited number of samples can be produced despite a variety of latent space samples.

The overall optimisation problem is related to the original GAN formulation (2.1). This cost function equates to minimising the Jensen–Shannon Divergence (JSD) (2.3) of $\mathbb{P}_r$ and $\mathbb{P}_g$ [25]. The JSD is a symmetrical, finite metric derived from the Kullback-Leibler Divergence (KLD) (2.2) for each distribution. The KLD measures the amount of information lost from a distribution compared to a reference distribution. Minimising the KLD is in keeping with the goals of GAN training, as it would promote one distribution to approximate another, however, the divergence itself is not symmetrical as it provides different values dependent on which distribution is the reference [36]. The JSD resolves this by taking the sum of KLD from each distribution against an average reference distribution, effectively producing a symmetrical metric that can be utilised as a distance.

$$D_{\mathrm{KL}}(P\|Q) = \sum_{x\in\chi} P(x_i)\log\frac{P(x_i)}{Q(x_i)} = \int_{\chi} P(x)\log\frac{P(x)}{Q(x)}dx \tag{2.2}$$

$$\begin{aligned} D_{\mathrm{JS}}(P\|Q) = \frac{1}{2}D_{\mathrm{KL}}(P\|M) + \frac{1}{2}D_{\mathrm{KL}}(Q\|M), \\ M = \frac{1}{2}(P+Q) \end{aligned} \tag{2.3}$$

Both divergence values tend towards their upper bounds if the supporting manifolds of the distributions do not perfectly align. This occurs as there is a complete distinction between the distributions that may be exploited by the discriminator which will result in complete information loss [34].

Demonstrating how a perfect discriminator would cause the metrics to tend towards their upper bounds: Given the aforementioned separation, a real image would have a non-zero probability assigned by the discriminator as belonging to the real set ($\mathbb{P}_r(x) > 0$) and a zero probability of belonging to the generated set ($\mathbb{P}_g(x) = 0$). When the considered distribution, $\mathbb{P}_r$, is non-zero and the reference, $\mathbb{P}_g$, approaches zero, $\mathrm{KL}(\mathbb{P}_r(x)\|\mathbb{P}_g(x)) = +\infty$. The same would

be true for a generated image, as a discriminator would provide a non-zero probability of belonging to the generated set ($\mathbb{P}_g(x) > 0$), and a zero probability of belonging to the real set ($\mathbb{P}_r(x) = 0$), producing a similarly infinite result, $\mathrm{KL}(\mathbb{P}_g(x)\|\mathbb{P}_r(x)) = +\infty$. The JS divergence as a result can be calculated given that $M = \frac{1}{2}(\mathbb{P}_r(x) + \mathbb{P}_g(x))$, and $D_{\mathrm{KL}} = \int_{\chi} \mathbb{P}(x) \log \frac{\mathbb{P}(x)}{\frac{1}{2}(\mathbb{P}_r(x)+\mathbb{P}_g(x))} dx$, where $\mathbb{P}(x)$ is either the real or generated distribution based on the term of the distance calculation. Irrespective of which is selected, the distinction between the two necessitate that one remain non-zero, while the other is zero. Simplifying the expression produces: $D_{\mathrm{KL}} = \log(2) \int_{\chi} \mathbb{P}(x) dx$. The integral of a probability distribution is one, as such $D_{\mathrm{KL}} = \log 2$, with the same result for the other term, with a resultant: $D_{\mathrm{JS}} = \log 2$. This demonstrates that a superior discriminator, able to readily separate the distributions, provides no useful information to the generator that can be utilised for resolving the distinction between sets, and necessitates that an alternative value function be utilised.

2.2.1 Wasserstein GAN

Given the aforementioned difficulties with the standard JSD for training GANs, Arjovsky *et al.* [37] propose the use of the Wasserstein metric (2.4), also known as the Earth Mover Distance (EMD), as an alternative cost function, a formulation they refer to as the Wasserstein GAN (WGAN). The EMD is a distance metric comparing two distributions and is intuitively understood as the optimal transport plan for converting the one distribution to the other. The metric considers some transport cost, $c(x, y)$, which computes the cost in the range, $[0, \infty)$, to transfer a unit from point x to point y. A transport plan, $\gamma(x, y)$, returns the quantity to be moved from x to y that will be incurred at the per unit cost. The EM distance considers the greatest lower bound on the set of the costs of the possible transport plans that convert one distribution to another. This restricts the set of plans to only those which transfer the entire mass of the distribution. This necessitates that the mass shifted from x be the total mass of x, $\int \gamma(x, y) dy = \mathbb{P}_x$, and the total mass transferred to y be the total mass of y, $\int \gamma(x, y) dx = \mathbb{P}_y$. This is equivalent to defining a bivariate probability distribution with marginals equivalent to the two distributions, $\prod(\mathbb{P}_x, \mathbb{P}_y)$ [37, 38]. The overall cost is then the sum of transport costs and the amount to be transferred at each location, $\int c(x, y) d\gamma(x, y)$. In terms of the EM distance, the cost is set as the Euclidean distance between points and the mass as the probability of a given point. The metric therefore computes the weighted sum of the probability and distance of each data point from the optimal locations necessary to recreate a reference distribution. This allows the restatement of the distance in terms of the Euclidean norm as used by Arjovsky *et al.* [37]:

$$D_W(\mathbb{P}_x, \mathbb{P}_y) = \inf_{\gamma \in \prod(\mathbb{P}_x, \mathbb{P}_y)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|_2] \tag{2.4}$$

Arjovsky & Bottou [34] initially propose the application of the metric to the development of GANs as a potential remedy to minimizing the distance between

real and generated distributions in the context of the shortcomings of the JS divergence. The construction of such a model is however only provided in a later work by Arjovsky *et al.* [37]. In contrast to the aforementioned cost functions, the EM distance is shown under mild assumptions to be continuous everywhere and differentiable almost everywhere, while remaining a weaker divergence than the other common functions. This allows the convergence of approximations that would fail to converge under conventional value functions [37, 39], and would enable the training of a generator even with imperfectly aligned distribution manifolds and a superior discriminator.

The overall benefits of the EM distance are limited by the original formulation being potentially intractable and an optimisation problem in itself [37] as this prohibits its application as cost function for model training. To resolve this, the distance is rewritten under the Kantorovich-Rubinstein duality (2.5), wherein the initial infimum problem (2.4), is restated as the supremum over the difference between all 1-Lipschitz functions for both distributions. In other words, the duality restates the problem as the least upper bound of the difference of all functions with a derivative of less than one that operate on both distributions. This alternate formulation has been shown to be almost equivalent to the original formulation in Equation 2.4 [37]. Utilising this calculation for the training of GANs relies on the discriminator's parameters being enforced to be Lipshcitz constrained. The method employed by Arjovsky *et al.* [37] clips the weights of the discriminator in the range $[-c, +c]$, with the value of c being included as a network hyperparameter. The discriminator is additionally modified to provide the final logit scores instead of a binary classification for each sample. This alteration is called the critic as it is no longer trained to classify and instead provides an unconstrained score for a particular image. This alteration is achieved by removing the final sigmoid layer from the discriminator [37]. The modified value function of the WGAN is shown in Equation 2.6, with $\mathcal{D}$ referring to the set of 1-Lipschitz functions [39].

$$D_W(\mathbb{P}_x, \mathbb{P}_y) = \sup_{\|f\|_L \leq 1} \mathop{\mathbb{E}}_{x \sim \mathbb{P}_x} [f(x)] - \mathop{\mathbb{E}}_{y \sim \mathbb{P}_y} [f(y)] \tag{2.5}$$

$$\min_G \max_{D \in \mathcal{D}} \mathop{\mathbb{E}}_{x \sim \mathbb{P}_r} [D(x)] - \mathop{\mathbb{E}}_{z \sim \mathbb{P}_g} [D(G(z))] \tag{2.6}$$

The WGAN formulation allows for far more stable training of GANs, however, the choice of c can significantly impair training and requires careful selection based on the critic's architecture. In addition, very deep WGAN models often fail to converge as the capacity of the network is underused due to the weight clipping limitation [39]. Searching for an optimal value for c is similarly difficult as the training time and costs generally scale with network depth which results in significant barriers to performing multiple runs to evaluate potential values. As a solution to this, Gulrajani *et al.* suggest a modified loss function utilising a gradient penalty (GP) multiplied by an alternate hyperparameter, the penalty coefficient (λ). The gradient is calculated between random samples ($\hat{x}$) drawn along straight lines from the generator and data distribution. This sampling

procedure is typically achieved by producing a batch of real and generated images and combining them by randomly interpolating each pixel between the two sets. The degree of interpolation (θ) is chosen for each image in the range $[0, 1]$, with the resultant sample ($\hat{x}$) being the sum of the interpolation and its inverse multiplied by the real and generated images: $\hat{x} = \theta \times x_r + (1 - \theta) \times x_g$. The critic is used to score this resulting batch with the Jacobian then determined from this score with respect to the input samples. The Euclidean norm of this Jacobian is then utilised as a proxy for direct Lipschitz constraints of the magnitude of the derivative, as for a 1-Lipschitz function, the norm can be at most 1 [39]. This so-called WGAN-GP method preserves much of the networks capability and results in improved training time and sample quality compared to the baseline WGAN [39]. The value function of the WGAN-GP is shown in Equation 2.7.

$$\min_G \max_D \mathop{\mathbb{E}}_{x \sim \mathbb{P}_r} [D(x)] - \mathop{\mathbb{E}}_{z \sim \mathbb{P}_g} [D(G(z))] + \lambda \mathop{\mathbb{E}}_{\hat{x} \sim \mathbb{P}_{\hat{x}}} [(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] \tag{2.7}$$

2.2.2 Progressively Growing GANs

The WGAN-GP method allows for the training of deeper adversarial pairs that are able to produce higher resolution samples at greater quality than prior approaches. However, practical issues remain in that while it may be possible to train such a network, the time to convergence may be excessive. Karras *et al.* [19], present a methodology to train a GAN by progressively expanding the scale of the network outputs. The network initially generates 4x4 images until convergence, this is then followed by a doubling of the spatial resolution of generates. Every scale increase is accompanied by a period of mixing a nearest-neighbour upsampled version of the previous layer with the new layer output. The prior layer is mixed in a linearly decreasing fashion until the new layer is used in its entirety. The new layer is then trained exclusively until stability and then the scale increased again until the final desired resolution is attained.

In addition to scaling the generation process, there are several other techniques also used to stabilise training and promote image variability. As part of these, the network employs layers modified to utilise equalized learning rates and normalises feature vectors in the generator to limit excessive signal amplification. Normalisation is applied after each convolution and activation function pair in the generator. This is performed by 'Local Response Normalization' (2.8) [40], wherein the value of a pixel in a feature map is normalised with respect to the other maps in the same layer. This is referred to as a form of brightness normalisation which prevents signal escalation in the generator that may dominate the resulting images. The exact implementation by Karras *et al.* [19] is simplified from the original which maintains several hyperparameters to tune the normalisation process. Equation 2.8 instead, simply divides the pixel values (a) at a particular co-ordinate (x, y) by the root of the average

value across all layers (N) with an additional small value (ϵ) utilised to prevent potential zero division.

$$b_{x,y} = \frac{a_{x,y}}{\sqrt{\frac{1}{N}\sum_{j=0}^{N-1}(a_{x,y}^{j})^2 + \epsilon}} \tag{2.8}$$

Equalised learning rates are enforced by the use of He's Initialisation [41]. This initialises the weights of a particular layer to a zero-centered Gaussian distribution with a standard deviation seen in Equation 2.9, the biases are initialised to zero. The standard deviation is determined by the number of incoming connections to a particular layer (n_l). In terms of a convolutional layer, this is the dimensions of the kernel squared (k^2) multiplied by the number of incoming channels (C_{in}). For a linear layer, this would just be the output dimension of the layer. Initialising weights in this manner improves ReLU-based network convergence time and improves the training of very deep networks. Networks with improperly configured weights may exhibit difficulties with signal propagation due to uncontrolled attenuation or amplification of input values, a problem that worsens as the number of layers in the network increases. This phenomenon may impair or, in some cases, prevent the training of such networks [41]. He's Initialisation therefore improves signal propagation at the onset of training, but does not guarantee this relationship beyond the initial weight update as it provides no constraints on how the network updates its values. In response to this, Karras *et al.* instead implement the relationship as a constraint on the weights of an individual layer, a process they refer to as learning rate equalisation. This modification scales the weights of a layer on each forward pass to maintain an equivalent dynamic range by dividing the weights by σ (2.9). The weights are initially sampled from a zero-centered normal distribution with a standard deviation of one, $\mathcal{N}(0, 1)$. This at the onset is equivalent to a standard He's Initialisation as the unit normal distribution is scaled to be a zero-centered Gaussian with standard deviation σ. As training progresses, the weight values will be updated and continuously scaled by this constant, ensuring that the dynamic range of each layer remains equivalent. This improves the performance of adaptive optimisation techniques [42] to update the network, as it ensures that that the scale of update remains the same across layers, therefore equalising their learning rates [19].

$$\sigma = \sqrt{\frac{2}{n_l}} \tag{2.9}$$

Image variation, while partially addressed through the selection of a cost function that minimises mode dropping [37, 39], is improved upon by introducing minibatch discrimination to the critic. This relatively simple concept, calculates the standard deviation of features extracted by the critic across a minibatch as an additional channel to be included in image scoring. This promotes the creation of images with a similar level of variation as that found in real image samples and in so doing, penalises aberrant behaviour such as mode dropping [19].

The final element of the PGGAN formulation is the modified cost function, which is the same as the WGAN-GP function (2.7) with an additional drift term utilised to discourage the critic's score from becoming excessively large. This is included as a small weight (ϵ_{drift}) on the square of discriminators output for real images.

$$\min_G \max_D \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r}[D(x)] - \mathop{\mathbb{E}}_{z\sim\mathbb{P}_g}[D(G(z))] + \lambda \mathop{\mathbb{E}}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2] + \epsilon_{drift} \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r}[D(x)^2] \tag{2.10}$$

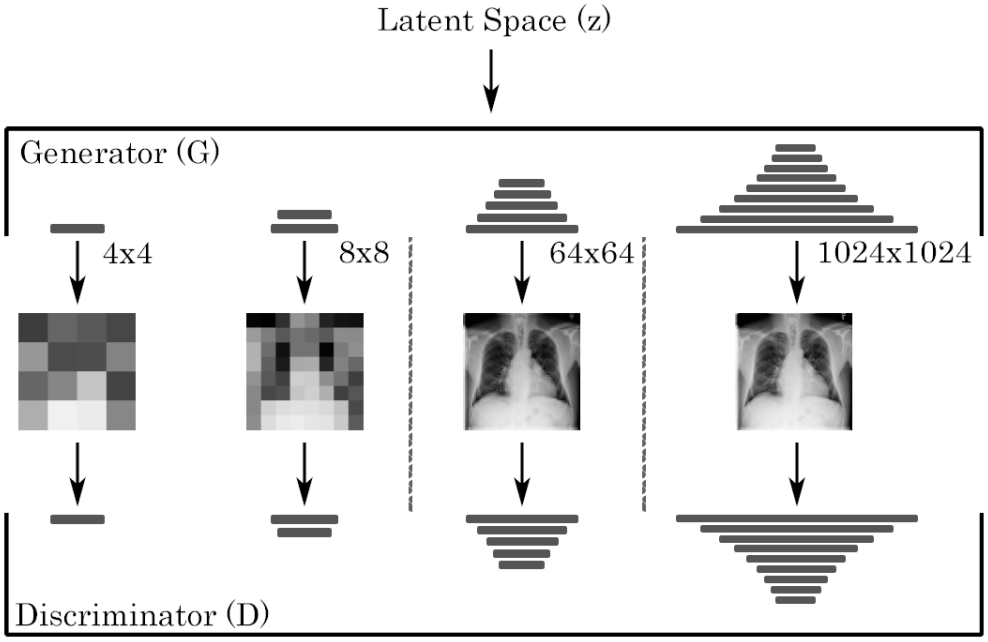


Figure 2.1: Training configuration for the PGGAN. The network trains until convergence on each scale, then doubles the spatial resolution. Each upscale operation has a period of time wherein the output is a mixture of the lower resolution output nearest-neighbour upsampled and mixed with the newly scaled output. The proportion of overlap is decreased linearly until only the scaled output is considered. This stabilises the growing process by preventing a period of training where generator outputs suddenly become significantly worse and may lead to a discriminator advantage.

Limitations of the PGGAN framework are discussed in later work by Karras *et al.* [43, 44], in which they modify the structure of the PGGAN to utilise style-based techniques to produce additional improvements in image quality and variation. As part of this work, Karras *et al.* note that the typical network's only source of stochastic variation arises from the initial latent space

sample. This results in the consumption of network synthetic capacity to preserve variation for later image scales, causing repetitive patterns in images with 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 [44].

Phase artifacts are a direct ramification of the assumption allowing the technique to function. The step-wise enlargement of spatial resolution allows the network to progressively learn scale-specific features and can be considered a contemporary to a design heuristic of starting with the broadest level of detail and slowly including finer properties. Implicit to this, is the understanding that the larger details define regions where smaller details can be included, and as result, if there are limitations to the placement of broader details, this propagates to the placement of smaller features even if there are no such limitations on these features. Phase artifacts result from location selection limitations in the early generation process, as the initial model is able to define a handful of broad regions for subsequent layers to utilise. This initial selection results in regional biases in the positioning of image details which may become apparent in produced images that do not demonstrate full spatial variability. Examples given by Karras *et al.* [44] when applied to synthetic human face generation, reveal fixed zones where the model places properties such as eyes or mouths across a number of different images. This at times produces challenges to image coherency when regional preferences limit the degree to which the model is capable of adapting to alterations in perspective. Similar examples show fixation of teeth despite positional shifts of the head, resulting in a subtle implausible distinction between the orientation of parts of the image.

2.2.3 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 [45]. 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.

The FID is a metric intended to provide a solution for evaluating the quality

of generated images [46]. It functions by embedding a set of real and synthetic images in the final average pooling layer of an Inception Net [47] pre-trained on ImageNet [48]. The sets are assumed to be multivariate Gaussian distributions with the average and covariance of each utilised to calculate the Fréchet distance (2.11), 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 co-variance of the real and generated distributions respectively.

$$\mathrm{FID}(P_r, P_g) = \left\|\mu_{P_r} - \mu_{P_g}\right\|_2^2 + \mathrm{Tr}\left(\sum P_r\right) + \sum P_g) - 2(\sum P_r \sum P_g)^{\frac{1}{2}}) \tag{2.11}$$

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 [46, 49]. 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 [49]. 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) [50] 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 synthetic images (HYPE_{time}) or the average error rate given unlimited evaluation time (HYPE_{∞}). The HYPE_{time} metric evaluates the time threshold necessary to distinguish images by altering the amount of time an image is shown based on prior performance. Correct responses reduce the allocated time while incorrect answers have the opposite effect, the net result being convergence towards the average time threshold necessary to maintain a 50% error rate for the detection of generates. HYPE_{∞}, in contrast, provides an evaluator unlimited time to make a determination and scores them on the average error rate across both real and generated sets. Both methods have been demonstrated to produce results that are capable of reliably separating out differences in model performance, however, are limited in their ability to quantify broader properties of generates, such as overall variability.

In addition to the already mentioned quality measures for generated images, an important element to examine, beyond quality estimates, is the performance of models developed using synthetic data applied to the original task of interest. Shmelkov *et al.* [51] 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. This inherently requires a model to be capable of creating data in the same manner as the classes used for the original classification task.

2.2.4 Synthetic Medical Data

There is an abundance 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 [52]. 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 [52].

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 [53] . 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 [16]. Both approaches demonstrate the potential benefits GANs offer for dataset augmentation, however they 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 [15], while Togo *et al.* produce patches of x-rays of gastritis [17]. 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 the 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 [54]. Bowles *et al.* augment CT and MRI data with synthetic slices and observe improvements in segmentation performance [55].

Despite significant variation in both domain and task, most medical image GAN implementations are constrained to the production of individual classes

or random samplings from the latent space with little control over generated content. Tasks that involve images simultaneously belonging to multiple classes are typically resolved by splitting the task into a number of binary classification exercises with a GAN trained and utilised for each class.

2.3 Image Classification

Image classification is a computer vision (CV) task wherein a set of labels are assigned to an image based on the recognition of semantic content from within the frame [56]. Historically this was performed using hand-crafted feature extraction techniques such as histograms of orientated gradients (HoG) [57] or scale-invariant keypoints (SIFT) [58]. Features are small descriptive properties of images that are used to identify the classes to which an image may belong [59]. Traditional feature engineering requires extensive per-class experimentation to isolate elements useful for class identification, a process that scales exceedingly poorly with the number of classes in the set [59]. DL, in contrast, offers an end-to-end approach to image classification. This involves providing an ANN with images and labels, then training the network via backpropagation [23]. The backpropagation algorithm enables an ANN to automatically discover features useful for class discrimination without human oversight [23]. This mechanism is capable of producing models that outperform human evaluators in substantially less time than that taken with conventional approaches [59].

Taking an image as an input compared to already extracted features may result in millions of pixels needing to be addressed. Standard fully-connected ANNs become unwieldy if applied directly, as they would require even more parameters than the input dimensions to be useful for classification. Modern classification networks instead typically employ convolutional neural networks (CNN) [60] as a means of efficient classification without excessive parameterisation. These introduce inductive biases inherent to the 2D topographical structure of images to constrain the network scale while preserving or augmenting performance [61].

A vital component of CNNs is the exploitation of translation invariance and local connectivity. Translation invariance means that the semantic content of an image is unaffected by shifts in the pixel coordinates, i.e. a picture of a cat remains one no matter where in the image the cat appears [62]. Locality uses only regional information in the vicinity of each pixel in the early layers of the CNN, as these usually form highly correlated local motifs that may be easily recognised [23]. Numerous layers then allow for these local features to be combined across a range of scales and promotes the hierarchical representation of features within the CNN. The combination of these properties mean that it is unnecessary to have a unique node for each input pixel, and instead make it possible to have a set of weights that act locally and repeat across the entire image [62, 63]. This is achieved through a convolution operation (2.12) using a parameterised kernel (K) that is applied to each pixel (i, j) in an image (I) and

encodes information from the surrounding region, with the size of the region determined by the kernel's dimensions (m, n).

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n) \tag{2.12}$$

The repeated application of the kernel results in a substantial reduction in parameters as the network shares a single set of weights across multiple locations. Stacked kernels with intermediate non-linear functions allows CNNs to progressively encode more complicated structures until they are capable of adequately recognising elements necessary to classify an image [61]. Pooling operations may also be included after several convolutional layers and serve to augment translation invariance. These operations subsample the spatial dimensions of the extracted features within a region. The exact downsampling mechanism varies, however, often involves selecting the maximum value (max pooling) or taking the mean (average pooling) from within an area. This process combines the contribution from several related regions of the source image and results in further invariance to local translations [61].

Despite significant modifications to improve the efficiency of CNN training, the computational resources needed are still immense compared to traditional CV techniques. This overhead typically requires specialised hardware, such as graphics (GPU) or tensor processing units (TPU), to allow for reasonable development times [59]. In addition, DL techniques are famously opaque and suffer from limited interpretability in determining the cause for various predictions. This may cause unexpected model behaviour or the unconscious introduction of predictive biases that may be difficult to detect [64].

Modern CNNs saw a resurgence with the advent of AlexNet [40], an extension of the original CNN architecture [60] applied to the ImageNet challenge [48], a classification task featuring over 15 million images. AlexNet demonstrated substantial performance improvements over conventional CV methods and was made possible by improvements in GPU performance that had hindered previous CNN usage [40]. Further performance advances were achieved with increased network depth alongside a switch in paradigm to constructing CNNs from composable computational blocks. These blocks typically consist of a stack of convolutional, non-linear, and pooling layers that perform progressive feature extraction with each repeated block [65]. The number of blocks as well as the number of convolutional layers within each block defines the overall structure of the network and allows for a more efficient network description [62]. At the same time, others were experimenting with alternative feature extraction kernel combinations. The then top performing variant, GoogLeNet, utilised several parallel convolutions of varying size per block. This allowed for different feature scales to be examined at once and overall lead to a reduction in parameters while maintaining performance [66]. Despite these alterations in model construction, the overall training of these networks remained slow and unstable, an issue that only worsened as network depth increased. He initialisation [41] improved

network training dynamics by providing a more stable starting point from which to start optimising. This is achieved by randomly sampling weights in a manner that prevents excessive alterations in signal magnitude within the early network. He *et al.* [41] show that such initialisation improves convergence time and overall performance, as they demonstrate the first model to surpass human-level ImageNet classification performance. In a similar vein, Ioffe & Szegedy [67] propose a batch normalisation (BN) layer (2.13) for reducing the variability in network dynamic ranges that may slow training by requiring per-layer learning rates. This attempts to solve the same problem as He initialisation except BN is able to function beyond the onset of model initialisation and can act as a training regulariser [62]. BN operates at each intermediate network state by normalising the output for an example (x), by the mean ($\hat{\mu}_B$) and variance ($\hat{\sigma}_B^2$) estimated from the overall batch (B). γ and β act as a learnable element-wise ($\odot$) scalar and bias respectively. A small value (ϵ) is included for numerical stability when dividing. These parameters allow for the model to alter an output from having a mean of zero and unit variance to whichever variant is most appropriate for the task at hand [62].

$$BN(x) = \gamma \odot \frac{x - \hat{\mu}_B}{\sqrt{\hat{\sigma}_B^2 + \epsilon}} + \beta \tag{2.13}$$

Even with these improvements to neural network construction and training, He *et al.* [68] note that standard network depth does not directly correlate to overall performance. They find that while deep networks clearly surpass conventional results, the optimal layer depth is arbitrary and often degrades after a point. This process is not due to network over-specification as the reduction is seen in training as well as evaluation dynamics. The networks are simply less performant than shallower variants. They attribute this phenomenon to the baseline stacked convolutional structure not enforcing an identity mapping between layers, this results in each layer developing a variant of the prior layer's function without encapsulating the features it extracts. As a result these networks are more difficult to train and do not guarantee performance retention when layers are added. They resolve this through the introduction of a residual block which adds the output of a layer to the identity of the previous layer. These so named ResNets promote additional layers to learn residual features that are not considered by earlier layers. This maintains performance with greater depth and improves training by allowing for improved gradient information for earlier network layers [68]. Huang *et al.* [69] expand upon this technique in their DenseNet model by concatenating the outputs of all prior layers as an input to each additional layer. This process allows for later layers to learn relationships between features extracted at varying receptive levels while further augmenting gradient information flow to the early network. This reuse of earlier features allows DenseNet models to outperform ResNets with several times less parameters [69].

2.3.1 X-Ray Classification

Significant work of late has gone into utilising deep learning architectures for screening and diagnosing diseases from medical images [2, 3, 4, 9, 70, 71]. Radiological studies are a core focus of a number of works with recent advances surpassing radiologist-level accuracy in the classification and localisation of certain pathologies when considering an image without any additional information [4, 9]. In addition, feature heatmaps produced from deep learning models have been demonstrated to augment radiologist performance [72].

Guendel *et al.* [73], achieved some of the highest published results on the chestxray14 dataset [12] by re-purposing a pre-trained DenseNet121 model [69] to diagnose x-rays. They train the model using a weighted cross entropy term (2.14), with the weighting based on a positive class weight, w_P, and a negative class weight, w_N. These weights are calculated based on the number of samples with a specified label, N_p, and those without, N_n. This may be calculated across the entire training set or per training batch, which given the method of sampling, may differ in prevalence from the overall training set. Guendel *et al.* [73], calculate weights per training batch and as a result, utilise a larger training size of 128 images to increase the likelihood of rarer conditions being drawn. They utilise high resolution (1024x1024) input images that are downsampled prior to classification. The downsampling is achieved by initialising two strided convolutional layers with Gaussian filters which maintain the number of colour channels, but each halve the spatial resolution. It is worth noting that these results were achieved by training on an additional closed image set to boost performance.

$$\begin{aligned} D_{CE}(y_c,\ \hat{y}_c) = w_P \cdot y_n \log(\hat{y}_c) + w_N \cdot (1 - y_c) \log(1 - \hat{y}_c), \\ w_P = \frac{N_p + N_n}{N_p}, \\ w_N = \frac{N_p + N_n}{N_n} \end{aligned} \tag{2.14}$$

The weight calculations by Guendel *et al.* [73] are performed per batch, and as a result the values fluctuate with respect to the random sampling from the training set. If one were to sample at random without replacement, as is the norm, one would expect the overall weights to approximate the average ratio found within the training set. This is the approach used by Rajpurkar *et al.* [4, 9] in their CheXNet [4] and CheXNeXt [9] models. They instead employ the overall proportion of labels in the training set to produce a weighted cross entropy loss per class as seen through the altered weight formulas in Equation 2.15.

$$w_P = \frac{N_n}{N_p + N_n}, \quad w_N = \frac{N_p}{N_p + N_n} \tag{2.15}$$

The initial CheXNet model performed on par with radiologists for pneumonia detection and the subsequent CheXNeXt ensemble performed at or above radiologist level for the majority of conditions evaluated. The first work made use of a single DenseNet-121 model [69] to binary classify images into pneumonia and non-pneumonia. The follow-up utilised an ensemble of ten such models, one per condition, with a total of 69 million parameters. Both models binarise each condition into an individual diagnostic problem by considering images with a particular label as positive and all others, irrespective of co-morbid labels, negative. This ignores the multi-modal distribution of labels across images and prevents the networks from exploiting information from the relationships between various conditions. Guendel *et al.* [73], in contrast, use a single network to classify all conditions. This multi-label scheme includes separate labels for the two included datasets despite an overlap of certain labels. Their results show a significant improvement over the baseline set by Wang *et al.* [12] established with multi-label classification when releasing the dataset. Their training on an additional closed dataset serves as evidence that such an improvement is in part due to common features being extracted by the network when attempting to classify multiple conditions at once, as despite not collating related labels, the overall classification performance improved significantly.

The different weighting approaches appear quite similar yet have different training dynamics. Equation 2.14 applies a weight that increases the value of each class in an inverse proportion to its prevalence. This scales each class to have the same overall value for both positive and negative versions of the same class and makes different classes overall equivalent as well. Equation 2.15, in contrast, multiplies each class by the inverse's prevalence. This balances the weight of positive and negative samples within a class but only ensures that the positive and negative labels for each class are weighted equally, however, this does not account for the difference in prevalence for each label. This approach is ideal for discrimination between binary classes, yet is problematic for multilabel configurations as the weights will be skewed by the different prevalence figures for each label.

Overall, the state of deep learning has advanced rapidly within recent years. This has resulted in a diverse method toolbox capable of tackling tasks from classification, to clustering, to synthesis. SOTA classification models are capable of performing at superhuman levels if provided with sufficient data, while generative networks can now reliably produce multi-megapixel, high-quality synthetic images. Healthcare applications tend to exploit the difficulties in the deployment of these systems by having limited data for designing solutions to diverse, multi-modal challenges. The volume of available medical data is constrained for a variety of reasons, however, an important consideration are the numerous legal and ethical safeguards that restrict the sharing of patient information. Furthermore, distributed data suffers from deficient annotation diversity owing to the time, cost, and difficulty of domain-expert labelling.

GANs have for some time been thought of as a potential salve for these data pain points, however, the method of application has not as yet proven clear. Despite producing high-quality images and having demonstrated performance improvements in medical image classification tasks, these two benefits are more often than not mutually exclusive for GANs. Surveyed methods are woefully computationally inefficient, utilising a full GAN architecture per data class, a process that cannot be sustainable given the number of potential labels of which a healthcare-associated dataset may consist. Alternatives attempting to condition GAN synthesis have found difficulties with sustaining quality for multilabel tasks, a common feature of the medical domain. Ultimately, there is a need for the development of a technique capable of sustaining synthetic visual performance across a range of classes without the need to train multiple generative networks while allowing for the preservation of inter-class relationships.

Chapter 3

Experimental Methodology

3.1 Data Source

The ChestX-ray14 dataset is an update to the ChestX-ray8 dataset [12], a large, open medical X-Ray dataset published by the National Institutes of Health (NIH) Clinical Center. The dataset comprises 112 120 x-rays from 30 805 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 reports 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 [74]. A modified set of labels was made available by Rajpurkar *et al.* [9] 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`

3.2 GAN Architecture

This work implements a modified PGGAN model [19] in Pytorch [75] and performs training via Pytorch Lightning [76] based on the open-source implementation produced by Facebook Research and available through the Pytorch Model Zoo [1]. The model is initially randomly instantiated and generates 4x4 images at the onset with progressively doubling of the spatial resolution after convergence at each scale. Prior trained layers are mixed with the newly added layers through upsampling with the proportion of the previous layer considered reducing linearly as training progresses.

WGAN-GP loss [39], equalised learning rates, minibatch discrimination, and pixel normalisation [19] are all utilised as part of model training. An exponential moving average of the generator weights is maintained and used for evaluation of synthetic performance. The model is trained up to a resolution of 1024x1024 with 800 000 images shown during each period of layer mixing and a further 800 000 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.

3.3 X-Ray Classification

For labelling the various CXRs, a Densenet-121 [69] pre-trained on ImageNet [48] is implemented with the final fully connected layer replaced with a new 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 (3.1) implemented by Guendel *et al.* [73]. 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.

$$\begin{aligned} D_{CE}(y_c,\ \hat{y}_c) = w_P \cdot y_c \log(\hat{y}_c) + w_N \cdot (1 - y_c) \log(1 - \hat{y}_c), \\ w_P = \frac{N_p + N_n}{N_p}, \quad w_N = \frac{N_p + N_n}{N_n} \end{aligned} \tag{3.1}$$

[1] `https://github.com/facebookresearch/pytorch_GAN_zoo`

For training, a subset of the full dataset is extracted which includes only images with modified labels from Rajpurkar *et al.* [9], These images are grouped at the patient level and with the labels averaged 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 stratification of the average patient labels to maintain label proportions across sets while ensuring no patient overlap occurs. Images are augmented 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 [42] using the default settings with an initial learning rate of 10^{-3}, which is reduced by a factor of ten if the validation macro-averaged Area Under the Receiver Operating Characteristic (AUROC) curve fails to improve for several epochs. Once model performance plateaus, it is evaluated on the test set. This process is then repeated for the final trained discriminator, replacing the final linear scoring layer with the number of classes and training the subsequent classification model end-to-end.

3.4 Pathology Optimisation

Given the multi-modal nature of pathology contained within x-ray images, it is not possible to utilise standard conditional GANs [77, 78] 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 [79], 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 [77]. 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.

These issues in multi-modal conditional training are sidestepped by instead drawing inspiration from work on the embedding of images into the latent space [80]. 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 [81]. This work opts to find a latent representation that maximises the corresponding image's classification score for a label of interest. Optimisation is performed using a single label at a time to allow for inclusion of related labels that would be consistent with

the presentation of real multi-modal x-rays. The implementation is based on the work on generator inversion for encoding images into the latent space by Creswell and Bharath [80]. Pseudocode for such a method can be seen in Algorithm 1 with a visual representation of this technique in Figure 3.1. The process proceeds to produce class-representative images by optimising the latent code for a generate that achieves a particular logit threshold. The logits of the other classes are ignored during this process to allow for natural combinations of labels. The threshold is selected based on the class-specific J statistic determined from the validation AUROC value and may be varied to produce alterations in pathology severity. The model weights are frozen for both the generator and classifier throughout the process.

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

$z \sim P_z(Z)$; `// Initialise` z `from prior distribution`
$S \leftarrow F_c(G(z))$;
while $S < T_c$ **do**
    $x \leftarrow G(z)$; `// Generate image proposal`
    $S \leftarrow F_c(x)$; `// Obtain class score from classifier`
    if $S \geq T_c$ **then**
        break;
    end
    else
        $L \leftarrow -S$;
        $z \leftarrow z - \alpha \nabla_z L$; `// Apply gradient descent`
    end
end
return z

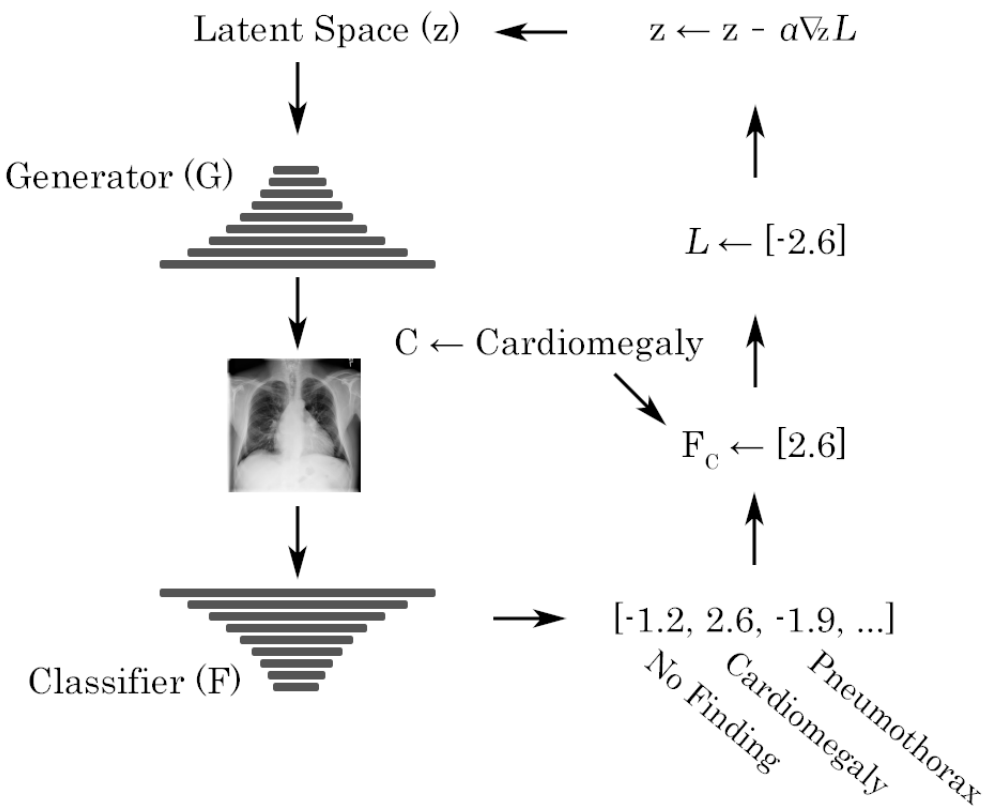


Figure 3.1: Visual representation of the class generation method shown in Algorithm 1.

Chapter 4

Results and Discussion

4.1 Image Quality

The generated images, as seen in Figure 4.1, 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.

A comparison between DCGAN, PGGAN, and the source images can be seen in Figure 4.2. DCGAN generates have obvious difficulties in producing large segments of the x-rays, preserving only the general structure. Generates show significant checkerboard artifacts alongside poorly resolved organ and bony structures. PGGAN images in contrast, show far more faithful reproductions of both global and regional features from the source dataset, yet still lack

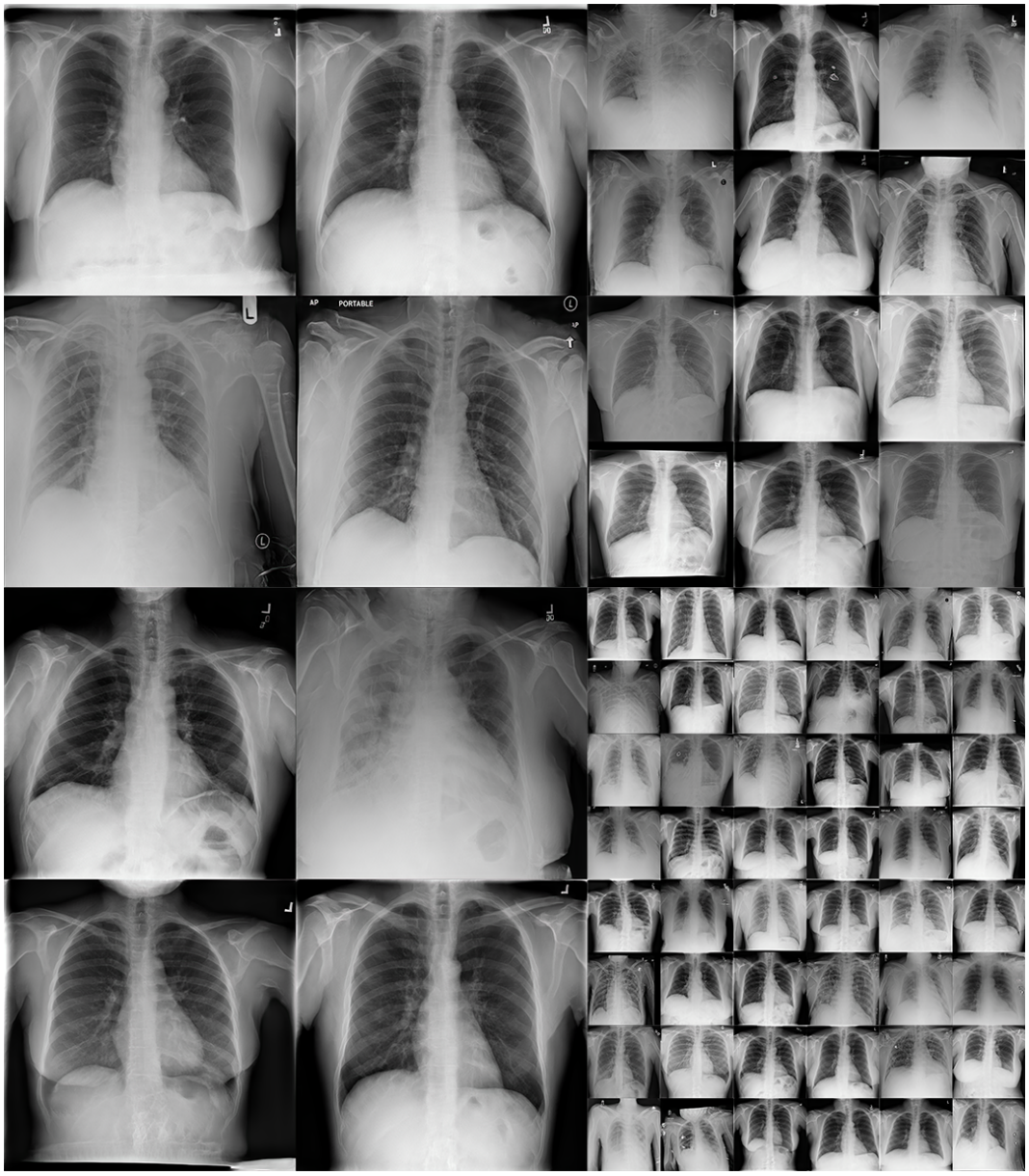


Figure 4.1: Examples of uncurated, random samples from the exponential moving average of the generator. Samples are drawn at random from a $\mathcal{N}(0, 1)$ distribution without use of the truncation trick.

small-scale details such as pacemakers and IV catheters.

The main distinction between real and PGGAN generated image sets is that the synthetic 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

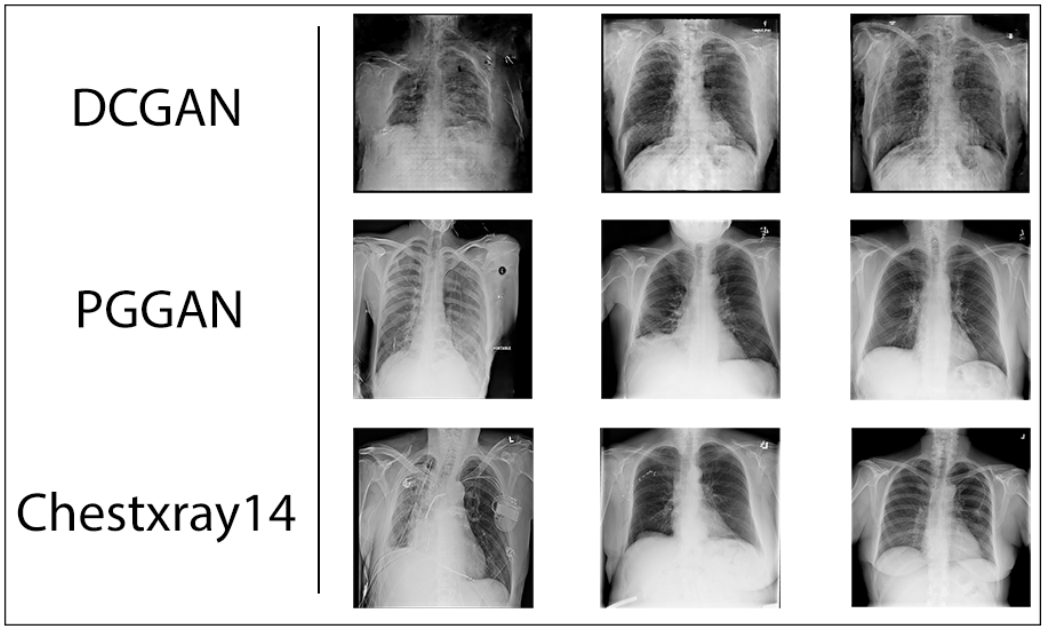


Figure 4.2: Comparison of sample quality between the DCGAN and PGGAN architectures as well as the source dataset at 256x256 resolution.

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, a potential 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.

To evaluate the applicability of the FID for automatic x-ray quality assessment, an attempt is first made 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 images into two similar groups. The underlying concept being that the groups should have similar disease distributions and be free of patient overlap. A low value of 0.53 for the FID is calculated through this method. This indicates that the underlying Inception network extracts similar features from both sets on average. These features may be sub-optimal for comparison given that the underlying network has not been trained on biomedical data, yet the small distance demonstrates that the groups of images are embedded in a similar manner within the network's final layer. This demonstrates that the FID has a reasonable lower bound on comparing similar images, however, the degree to which the metric can in fact distinguish between distinct x-rays requires validation.

Given that the images are known to vary by pathology label, this work argues that if the distance metric can reliably separate disease classes, it must have filters that are capable of extracting features that are meaningful for evaluating the clinical plausibility of generates, as the underlying network must be able to evaluate elements of studies relevant to the diagnosis. Similarly, if one 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 the *No Finding* label is utilised 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. The distance between it and the set of x-rays containing each other label is evaluated, the results of which can be seen in Table 4.1 with the addition of a split along sex and the overall comparison between real and synthetic image groups. 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 pathological change on the image. Exceedingly large distances are associated with findings that may distort major segments of the lung fields such as the *Edema* and *Consolidation* labels. In contrast, significantly reduced distances are associated with findings that generally occupy only small segments of a study, such as *Mass* and *Nodule* labels. A pairwise distance matrix between all classes, followed by the application of multidimensional scaling is implemented to plot the relative 2D positions determined by the FID in Figure 4.3. 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, seen in Figure 4.4, successfully groups sets of related or antecedent findings. The cluster demonstrates that the features extracted by the network underlying the FID are likely to be of semantic value with an adequate discriminative ability to group sets of related findings. Overall the technique appears to be applicable to CXRs despite the absence of explicit training to detect features of such biomedical images.

The quality of synthetic x-rays compared to the NIH source images is assessed via the FID given the evidence in favour of the metric for the evaluation of CXRs. The FID between the full NIH source dataset and 100 000 random, uncurated PGGAN generates is 8.02. This value is comparable with the FIDs reported for the PGGAN architecture on other non-biomedical, high resolution image generation tasks [19].

4.2 Pathological Variability

Synthetic CXR label prevalence was estimated to quantify the extent to which the PGGAN model captures the pathological variability of the various disease

Table 4.1: 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

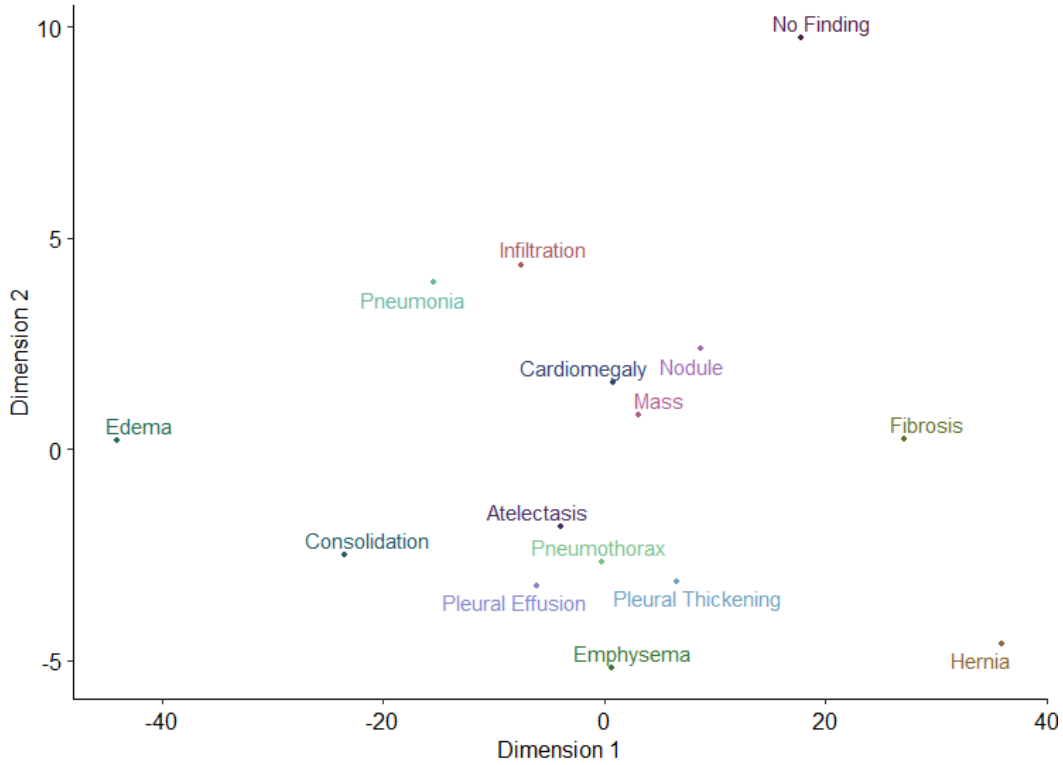


Figure 4.3: Classes plotted according to the multidimensional scaling of the pairwise distance matrix determined through the FID between all pathology classes.

entities contained within the source CXRs. The label proportions were determined via a common classifier for both the NIH and generated image sets with confidence intervals derived through bootstrapping the predictions 10 000 times. The proportion of the class labels from the NIH dataset and a random selection

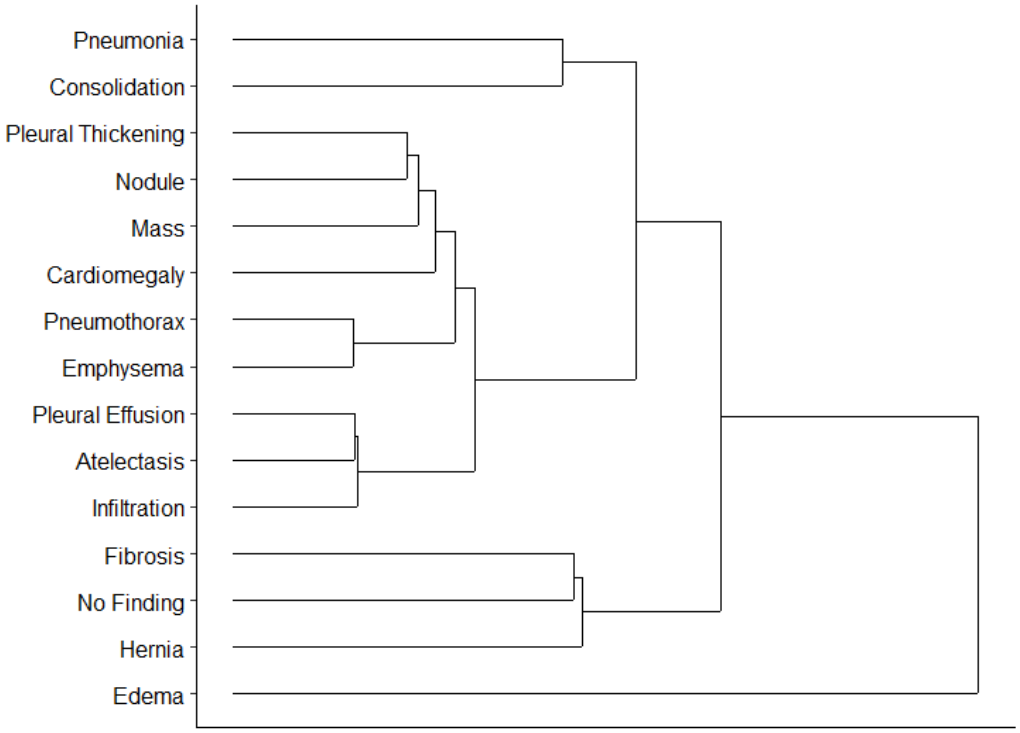


Figure 4.4: Hierarchical cluster of related pathologies generated from the pairwise distance matrix between all labels determined through the FID.

of 130 000 generated images can be seen in Figure 4.5. 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*. This alteration is not adequately explained by the PGGAN model simply selecting for the most common conditions as *Consolidation*, *Infiltration*, *Nodule*, *Pleural Effusion*, *Pleural Thickening*, *Pneumonia*, and *Pneumothorax* are all more prevalent than *Cardiomegaly* in the labelled samples, yet all demonstrate retained or reduced proportions in the generated distribution, while *Cardiomegaly* shows increased prevalence. It was hypothesized that this is likely to be due to a continuation of the phenomenon noted with the absence of certain smaller objects in scans, the objects in this case being features of various diseases. This is supported by the label distribution, as more common conditions with more prominent 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.

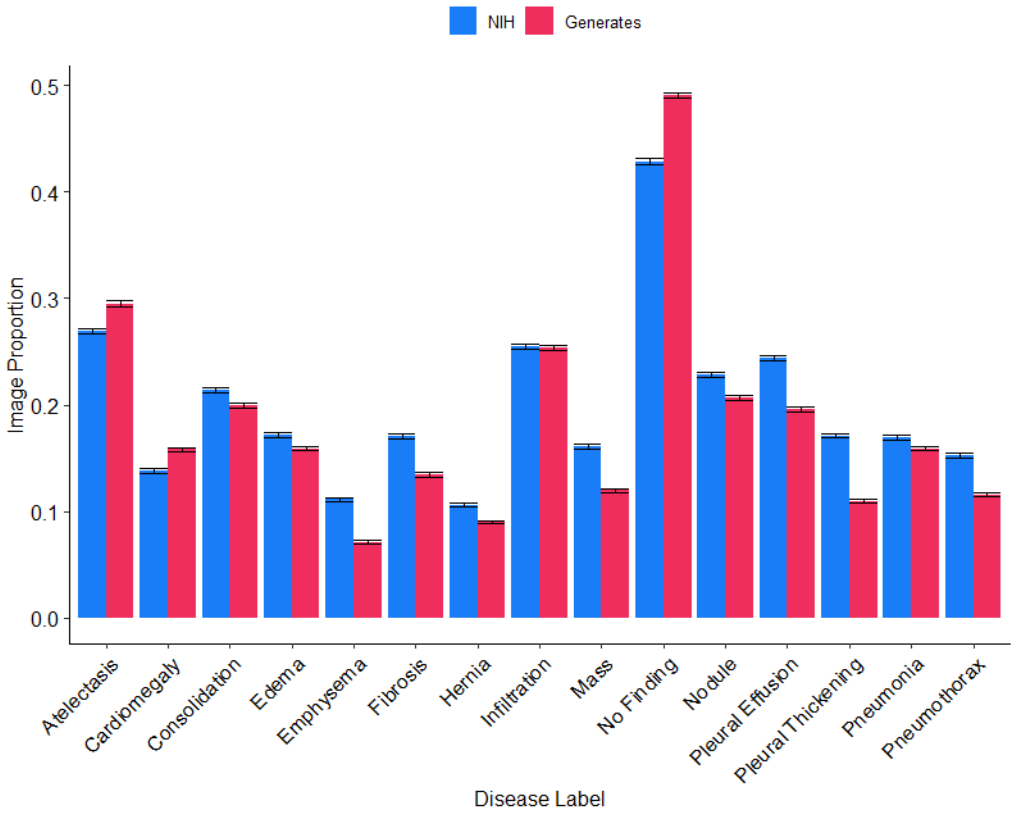


Figure 4.5: Proportion of labels in the original NIH dataset and 130 000 random PGGAN generates with 95% confidence intervals.

4.3 Pathology Generation

Pathological images are generated utilising class optimisation based on the method described in Algorithm 1. The initial latent code is sampled from a truncated normal distribution with a threshold of 0.7 [82]. The optimisation method is attempted with both a Densenet-121 and a re-purposed Discriminator as classifiers. It was generally found that optimising classes with the Densenet model proved to be a slower, more challenging, and less reliable overall process that frequently plateaus with a subsequent failure to produce an adequate sample for a particular class. Successes often relied 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 model as a result of being trained on smaller resolution images, requires inputs to be resized prior to classification. This creates an informational bottleneck from the classifier to the latent space and

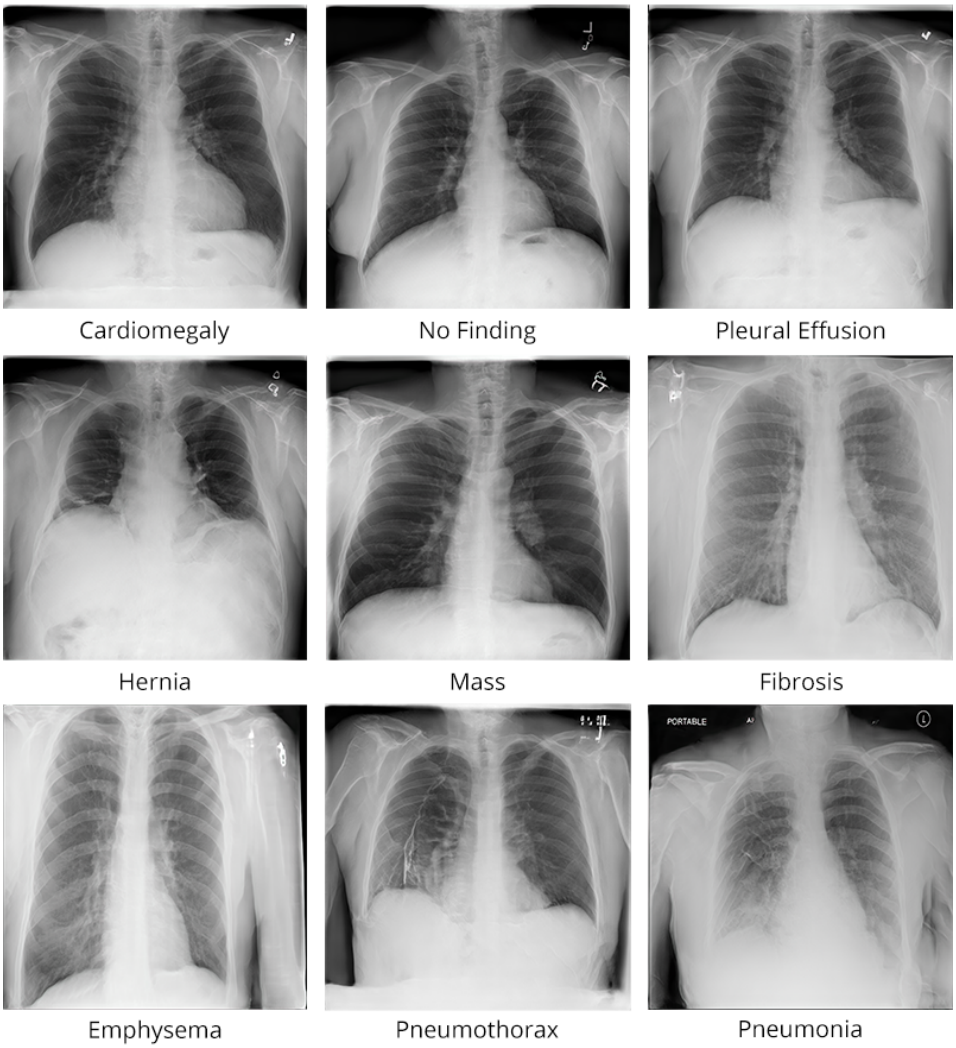


Figure 4.6: Example classes generated by maximising the classifier class logit. Images often have multiple findings, the finding optimised is the label given.

subsequently produces an inferior estimation of the gradient for optimisation. Replication of the process with a Densenet trained on a larger input resolution produced significant improvements to both the reliability and rapidity of convergence. Examples of successful pathology generates can be seen in Figure 4.6. Many more examples of multi-modal pathology can be seen in the linked image archive [21].

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, a score threshold is required for determining when a proposed image has converged to be representative of a particular class. Youden's J statistic [83] is utilised 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, this threshold is randomly increased by the absolute value of the distribution of $\mathcal{N}(1, 1)$ when optimising.

It is quite seldom that a dataset includes only single scans of a patient. Typically, patients have an image series wherein they are re-scanned several times. There are a multitude of potential reasons for this: follow-up on patient condition, assessment of the results of a medical intervention, routine imagery prior to surgery, to name a few. The end result is that most datasets have a significant proportion of very similar images wherein position or disease severity may be altered slightly but the overall image is largely unchanged. This process is simulated by sampling in the vicinity of a particular pathology-optimised latent code. The optimised latent location is treated as a centre with samples drawn from the surrounding hypersphere within a radius of $\mathcal{N}(0, 0.2)$. This was found empirically to enable the creation of images with the same set of diagnoses but minor changes in patient orientation, exposure, or disease severity. The degree of variability can be tuned by adjusting the width of the distribution, however, a value of 0.2 was found to produce minor variations without compromising the diagnostic label.

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. A full exploration of the properties of the CXR latent space is left to future work.

4.4 Radiologist Review

The plausibility of generated x-rays to domain experts was evaluated through a series of comparative image reviews. A group of practising radiologists were asked to identify a set of real images out of a grid of mixed real and synthetic x-rays as well as to select a real image out of a pair of real and generated images. Feedback was received from 3 consultant (equivalent to

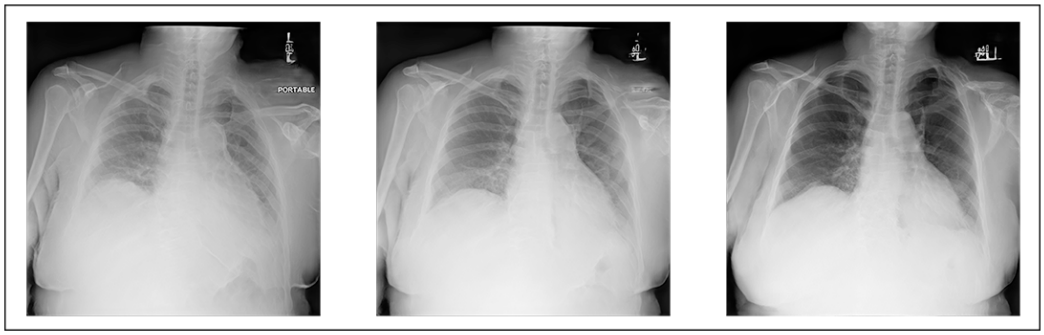


Figure 4.7: Example of local sampling after class generation. Images show similar anatomical and pathological features with variations in image contrast, text markup, and patient positioning.

board certified) radiologists, and 5 radiology registrars (equivalent to residents in North America). Image selection grids consisted of six x-rays produced with a 50% probability of being real or generated, six such grids were evaluated by each radiologist, the results of which can be seen in Table 4.2. Statistical tests evaluated the likelihood of identifying the given image variety at a rate greater than by the baseline 50% probability. 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 $p < 0.001$ and generates $p = 0.015$ for a one-sided t-test. Radiologist performance on x-ray discrimination equates to a $HYPE_{\infty}$ score of 33% (95% CI: 24, 43).

Table 4.2: Comparison of Radiologist Accuracy by Level of Experience

	Real	p	Synthetic	p
Registrar	71.25 (59.60 - 82.90)	<0.001	67.00 (54.75 - 79.25)	0.005
Consultant	75.00 (61.24 - 88.76)	<0.001	50.00 (38.13 - 61.87)	0.500
Average	72.66 (63.17 - 82.14)	<0.001	60.63 (51.08 - 70.18)	0.015

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 across 20 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 260x260 resolution with unlimited time to review images prior to making a decision.

A copy of the feedback form can be seen in Appendix A.2

4.5 Synthetic Model Evaluation

For evaluating the classification performance of a model trained only on synthetic data, An archive of CXRs of comparable size and with labels in a similar proportion to the source NIH dataset was generated. A set of 3 to 6 related images per class optimisation operation were produced through local pathology sampling and were treated as belonging to the same 'patient' as the original generated image. The subsequent synthetic patients are iteratively stratified into training and validation sets in the same manner as during initial classifier training, retaining only the original test set of real images, which was used for all comparisons. A Densenet-121 classifier was trained in the same configuration as described in Section 3.3. Performance was evaluated relative to real images as well as images embedded into the generator's latent space. Directly embedding images in the latent space was considered as the current alternative to individual class generation, whereby a network predicts the latent encoding of a given image. To achieve this, the discriminator was retrained 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. Classification performace was assessed with the proposed image reconstruction via the predicted embedding alongside the labels of each original image. A table of comparative AUROC values can be seen in Table 4.3

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 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 [17]. 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. Comparisons were attempted with generating classes through a simpler DCGAN architecture similarly trained on the full NIH dataset. The generation process was found to be largely intractable as the model often, despite extensive optimisation, failed to produce scores of a sufficient magnitude to be considered examples of a class. This failing was attributed to a combination of generator mode dropping as well as insufficient image quality to reproduce important features of disease

Table 4.3: 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	0.510	0.901	0.972
Fibrosis	0.630	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

which precluded the formation of certain classes for comparison.

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.

4.6 Recommendations for Future Work

Future work should seek to expand the applications of generative modelling for medical imagery given the aforementioned benefits of improved individual privacy and augmented datasets. This should at a minimum evaluate the improvements made through advances in synthetic image generation when applied to biomedical imagery. Potential avenues of exploration include: GAN architectural advances made in subsequent models within the PGGAN lineage [43, 44], usage of techniques that foster GAN coherence at varying scales [84], and assessment of other generative techniques able to produce high resolution images such as Very Deep Variational Autoencoders (VAEs) [85]. Focus should be directed to expert evaluation across related biomedical imaging modalities covering a diverse range of image content within each.

Beyond improving the quality of synthetic images, future work should explore applications of the class generation technique. The utility of the technique is not limited to any particular model and should offer greater control over fine-grained semantic content of images when paired with the appropriate generative model. The StyleGAN [43] model, as an example, possesses an extended latent space with multiple, scale-specific samples which offers control over varying image scales. Selective application of class optimisation in such a configuration should allow for significant control over the content of generates, allowing for profound dataset augmentation capabilities. This may take the form of demonstrating potential exacerbation or resolution of pathological features without altering any features of an individual 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 to identify such support devices. Continued development along this route should enable task-specific image extraction from broad image generators trained on large corpora of diverse images. These generators should allow for the extraction of a multitude of classes based on the capabilities of an associated classifier. This should allow for the creation of image classes that may not have been labelled within the source dataset. This would provide a vital tool to resolving model shortcomings when utilising data produced from another institution, as disease features could be produced alongside the properties of the scanning equipment of the alternate clinical site. Datasets could similarly be balanced along a variety of disease features and protected patient details to assist in reducing model diagnostic bias.

Chapter 5

Conclusion

Within this work, the PGGAN model was applied to the task of synthesising high-resolution CXR 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 was to employ such a model to improve the protection of patient privacy without compromising data availability.

This project overall demonstrated that it is possible to produce realistic, clinically plausible images that capture much of the variation in standard CXRs, however, there remains a significant need for improvement in the reproduction of small-scale details to achieve truly indistinguishable samples. The evaluated generative model is demonstrated to be capable of visually reproducing all pathological classes of interest in a similar proportion to the source image distribution despite some challenges with small-scale features. The work introduces a methodology for the extraction of class representative images from the generator's latent space through the optimisation of a classifier score. Evidence is provided that such a method is capable of constructing a fully-labelled, multi-modal, synthetic dataset from an unsupervised generator that achieves classification performance on par with focused tasks using the same model for individual class generation.

The applicability of the Fréchet Inception Distance to the evaluation of synthetic chest x-rays was evaluated and found that network underlying the calculation is capable of providing a meaningful metric for generate quality, despite the difference in data distribution from its training source.

Potential avenues of improvement for generate quality and improved semantic control of generates are included along with the anticipation that such improvements will enable the production of high quality teaching and training images without the concern of breaching patient confidentiality.

The source code, final model weights and a large archive of labelled synthetic x-rays are made available along with this project [20, 21].

References

[1] Andy Adam et al. “Current Status of Thoracic Imaging”. en.
In: *Grainger & Allison's Diagnostic Radiology: A Textbook of Medical Imaging*. Elsevier, 2021, p. 3. ISBN: 978-0-7020-7562-9 978-0-7020-7561-2.

[2] Daniel S. Kermany et al. “Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning”. en.
In: *Cell* 172.5 (Feb. 2018), 1122–1131.e9. ISSN: 00928674.
DOI: 10.1016/j.cell.2018.02.010.

[3] Anna Majkowska et al. “Chest Radiograph Interpretation with Deep Learning Models: Assessment with Radiologist-Adjudicated Reference Standards and Population-Adjusted Evaluation”. en.
In: *Radiology* 294.2 (Feb. 2020), pp. 421–431. ISSN: 0033-8419, 1527-1315.
DOI: 10.1148/radiol.2019191293.

[4] Pranav Rajpurkar et al. “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning”. en.
In: *arXiv:1711.05225 [cs, stat]* (Dec. 2017).
arXiv: 1711.05225 [cs, stat].

[5] Tulin Ozturk et al. “Automated Detection of COVID-19 Cases Using Deep Neural Networks with X-Ray Images”. en.
In: *Computers in Biology and Medicine* 121 (June 2020), p. 103792.
ISSN: 00104825. DOI: 10.1016/j.compbiomed.2020.103792.

[6] Farah E. Shamout et al.
“An Artificial Intelligence System for Predicting the Deterioration of COVID-19 Patients in the Emergency Department”. en.
In: *arXiv:2008.01774 [cs, eess]* (Nov. 2020).
arXiv: 2008.01774 [cs, eess].

[7] Linda Wang and Alexander Wong.
“COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images”. en.

In: *arXiv:2003.09871 [cs, eess]* (May 2020).
arXiv: 2003.09871 [cs, eess].

[8] Jeremy Irvin et al. "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison". en.
In: *arXiv:1901.07031 [cs, eess]* (Jan. 2019).
arXiv: 1901.07031 [cs, eess].

[9] Pranav Rajpurkar et al.
"Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists". en.
In: *PLOS Medicine* 15.11 (Nov. 2018). Ed. by Aziz Sheikh, e1002686.
ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1002686.

[10] Marcel Koenigkam Santos et al.
"Artificial Intelligence, Machine Learning, Computer-Aided Diagnosis, and Radiomics: Advances in Imaging towards to Precision Medicine". en.
In: *Radiologia Brasileira* 52.6 (Dec. 2019), pp. 387–396. ISSN: 1678-7099.
DOI: 10.1590/0100-3984.2019.0049.

[11] Alistair E. W. Johnson et al. "MIMIC-CXR, a de-Identified Publicly Available Database of Chest Radiographs with Free-Text Reports". en.
In: *Scientific Data* 6.1 (Dec. 2019), p. 317. ISSN: 2052-4463.
DOI: 10.1038/s41597-019-0322-0.

[12] X. Wang et al. "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)*. July 2017, pp. 3462–3471.
DOI: 10.1109/CVPR.2017.369.

[13] S. Kevin Zhou et al. "A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises".
In: *Proceedings of the IEEE* (2021), pp. 1–19.
ISSN: 0018-9219, 1558-2256. DOI: 10.1109/JPROC.2021.3054390.
arXiv: 2008.09104.

[14] Laleh Seyyed-Kalantari et al.
"CheXclusion: Fairness Gaps in Deep Chest X-Ray Classifiers". en.
In: *Biocomputing 2021*.
Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC, Nov. 2020,

pp. 232–243. ISBN: 9789811232695 9789811232701.
DOI: 10.1142/9789811232701_0022. arXiv: 2003.00827.

[15] Andrew Beers et al. "High-Resolution Medical Image Synthesis Using Progressively Grown Generative Adversarial Networks". en.
In: *arXiv:1805.03144 [cs]* (May 2018). arXiv: 1805.03144 [cs].

[16] Hojjat Salehinejad et al. "Synthesizing Chest X-Ray Pathology for Training Deep Convolutional Neural Networks". In: *IEEE Transactions on Medical Imaging* 38.5 (May 2019), pp. 1197–1206.
ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2018.2881415.

[17] Ren Togo, Takahiro Ogawa, and Miki Haseyama. "Synthetic Gastritis Image Generation via Loss Function-Based Conditional PGGAN". en.
In: *IEEE Access* 7 (2019), pp. 87448–87457. ISSN: 2169-3536.
DOI: 10.1109/ACCESS.2019.2925863.

[18] Xin Yi, Ekta Walia, and Paul Babyn.
"Generative Adversarial Network in Medical Imaging: A Review". en.
In: *Medical Image Analysis* 58 (Dec. 2019), p. 101552. ISSN: 13618415.
DOI: 10.1016/j.media.2019.101552.

[19] Tero Karras et al. "Progressive Growing of GANs for Improved Quality, Stability, and Variation".
In: *International Conference on Learning Representations*. 2018.

[20] Bradley Segal et al. "CXR PGGAN Code".
In: *GitHub. Note: https://github.com/BradSegal/CXR_PGGAN* (2021).

[21] Bradley Segal et al. "CXR PGGAN Model and Library". In: *Kaggle. Note: https://kaggle.com/bradsegal/synthetic-pggan-chest-xrays* (2021).

[22] Tom M. Mitchell. *Machine Learning*. en.
McGraw-Hill Series in Computer Science. New York: McGraw-Hill, 1997.
ISBN: 978-0-07-042807-2.

[23] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning". en.
In: *Nature* 521.7553 (May 2015), pp. 436–444.
ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature14539.

[24] F. Rosenblatt. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." en.
In: *Psychological Review* 65.6 (1958), pp. 386–408.
ISSN: 1939-1471, 0033-295X. DOI: 10.1037/h0042519.

[25] Ian J. Goodfellow et al. “Generative Adversarial Nets”.
In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS’14.
Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680.

[26] Lars M. Mescheder, Andreas Geiger, and Sebastian Nowozin.
“Which Training Methods for GANs Do Actually Converge?” In: *ICML*.
2018.

[27] Tim Salimans et al. “Improved Techniques for Training GANs”.
In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS’16.
Red Hook, NY, USA: Curran Associates Inc., 2016, pp. 2234–2242.
ISBN: 978-1-5108-3881-9.

[28] Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. en.
Cambridge, Mass: MIT Press, 1994.
ISBN: 978-0-262-15041-5 978-0-262-65040-3.

[29] Phillip Isola et al.
“Image-to-Image Translation with Conditional Adversarial Networks”.
In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, July 2017, pp. 5967–5976.
ISBN: 978-1-5386-0457-1. DOI: 10.1109/CVPR.2017.632.
arXiv: 1611.07004.

[30] Jun-Yan Zhu et al. “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks”. en.
In: *arXiv:1703.10593 [cs]* (Nov. 2018). arXiv: 1703.10593 [cs].

[31] Christian Ledig et al. “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
Honolulu, HI: IEEE, July 2017, pp. 105–114. ISBN: 978-1-5386-0457-1.
DOI: 10.1109/CVPR.2017.19. arXiv: 1609.04802.

[32] Chenyu You et al. “CT Super-Resolution GAN Constrained by the Identical, Residual, and Cycle Learning Ensemble(GAN-CIRCLE)”. en.
In: *IEEE Transactions on Medical Imaging* 39.1 (Jan. 2020), pp. 188–203.
ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2019.2922960.
arXiv: 1808.04256.

[33] Yujun Shen et al.
"Interpreting the Latent Space of GANs for Semantic Face Editing". en.
In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, June 2020, pp. 9240–9249. ISBN: 978-1-72817-168-5.
DOI: 10.1109/CVPR42600.2020.00926.

[34] Martin Arjovsky and Léon Bottou. "Towards Principled Methods for Training Generative Adversarial Networks". en.
In: *arXiv:1701.04862 [cs, stat]* (Jan. 2017).
arXiv: 1701.04862 [cs, stat].

[35] Alec Radford, Luke Metz, and Soumith Chintala.
"Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". en.
In: *arXiv:1511.06434 [cs]* (Jan. 2016). arXiv: 1511.06434 [cs].

[36] Christopher D. Manning and Hinrich Schütze.
Foundations of Statistical Natural Language Processing.
Cambridge, Mass: MIT Press, 1999. ISBN: 978-0-262-13360-9.

[37] Martin Arjovsky, Soumith Chintala, and Léon Bottou.
"Wasserstein Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*.
Ed. by Doina Precup and Yee Whye Teh. Vol. 70.
Proceedings of Machine Learning Research.
International Convention Centre, Sydney, Australia: PMLR, Aug. 2017, pp. 214–223.

[38] Soheil Kolouri et al. "Optimal Mass Transport: Signal Processing and Machine-Learning Applications". en.
In: *IEEE Signal Processing Magazine* 34.4 (July 2017), pp. 43–59.
ISSN: 1053-5888. DOI: 10.1109/MSP.2017.2695801.

[39] Ishaan Gulrajani et al. "Improved Training of Wasserstein GANs".
In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17.
Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 5769–5779.
ISBN: 978-1-5108-6096-4.

[40] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". en. In: *Communications of the ACM* 60.6 (May 2017), pp. 84–90. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3065386.

[41] Kaiming He et al. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. ICCV '15. USA: IEEE Computer Society, 2015, pp. 1026–1034. ISBN: 978-1-4673-8391-2. DOI: 10.1109/ICCV.2015.123.

[42] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *ICLR (Poster)*. 2015.

[43] Tero Karras, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, June 2019, pp. 4396–4405. ISBN: 978-1-72813-293-8. DOI: 10.1109/CVPR.2019.00453.

[44] Tero Karras et al. "Analyzing and Improving the Image Quality of StyleGAN". en. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, June 2020, pp. 8107–8116. ISBN: 978-1-72817-168-5. DOI: 10.1109/CVPR42600.2020.00813.

[45] Rohit Thanki et al. "Medical Imaging and Its Objective Quality Assessment: An Introduction". en. In: *Classification in BioApps*. Ed. by Nilanjan Dey, Amira S. Ashour, and Surekha Borra. Vol. 26. Cham: Springer International Publishing, 2018, pp. 3–32. ISBN: 978-3-319-65980-0 978-3-319-65981-7. DOI: 10.1007/978-3-319-65981-7_1.

[46] Martin Heusel et al. "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. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 6629–6640. ISBN: 978-1-5108-6096-4.

[47] Christian Szegedy et al.
"Rethinking the Inception Architecture for Computer Vision".
In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 2818–2826.

[48] Olga Russakovsky et al.
"ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 115.3 (Dec. 2015), pp. 211–252.
ISSN: 0920-5691. DOI: 10.1007/s11263-015-0816-y.

[49] Mario Lucic et al. "Are GANs Created Equal? A Large-Scale Study".
In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18.
Red Hook, NY, USA: Curran Associates Inc., 2018, pp. 698–707.

[50] Sharon Zhou et al. "HYPE: A Benchmark for Human eYe Perceptual Evaluation of Generative Models". en.
In: *arXiv:1904.01121 [cs]* (Oct. 2019). arXiv: 1904.01121 [cs].

[51] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari.
"How Good Is My GAN?" In: *Computer Vision – ECCV 2018*.
Ed. by Vittorio Ferrari et al. Vol. 11206.
Cham: Springer International Publishing, 2018, pp. 218–234.
ISBN: 978-3-030-01215-1 978-3-030-01216-8.
DOI: 10.1007/978-3-030-01216-8_14. arXiv: 1807.09499.

[52] Noseong Park et al.
"Data Synthesis Based on Generative Adversarial Networks". en. In: *Proceedings of the VLDB Endowment* 11.10 (June 2018), pp. 1071–1083.
ISSN: 2150-8097. DOI: 10.14778/3231751.3231757. arXiv: 1806.03384.

[53] Mehdi Moradi et al. "Chest X-Ray Generation and Data Augmentation for Cardiovascular Abnormality Classification".
In: *Medical Imaging 2018: Image Processing*.
Ed. by Elsa D. Angelini and Bennett A. Landman.
Houston, United States: SPIE, Mar. 2018, p. 57.
ISBN: 978-1-5106-1637-0 978-1-5106-1638-7. DOI: 10.1117/12.2293971.

[54] Dimitrios Korkinof et al.
"Mammo{GAN: High-Resolution Synthesis of Realistic Mammograms".
In: *International Conference on Medical Imaging with Deep Learning –*

Extended Abstract Track. London, United Kingdom, July 2019. arXiv: 1807.03401.

[55] Christopher Bowles et al. "GAN Augmentation: Augmenting Training Data Using Generative Adversarial Networks". en. In: *arXiv:1810.10863 [cs]* (Oct. 2018). arXiv: 1810.10863 [cs].

[56] Tanishq Abraham et al. "Chapter 7 - Applications of Artificial Intelligence for Image Enhancement in Pathology". In: *Artificial Intelligence and Deep Learning in Pathology*. Ed. by Stanley Cohen. Elsevier, 2021, pp. 119–148. ISBN: 978-0-323-67538-3. DOI: 10.1016/B978-0-323-67538-3.00007-5.

[57] N. Dalal and B. Triggs. "Histograms of Oriented Gradients for Human Detection". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. 2005, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177.

[58] David G. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints". en. In: *International Journal of Computer Vision* 60.2 (Nov. 2004), pp. 91–110. ISSN: 0920-5691. DOI: 10.1023/B:VISI.0000029664.99615.94.

[59] Niall O'Mahony et al. "Deep Learning vs. Traditional Computer Vision". en. In: *Advances in Computer Vision*. Ed. by Kohei Arai and Supriya Kapoor. Vol. 943. Cham: Springer International Publishing, 2020, pp. 128–144. ISBN: 978-3-030-17794-2 978-3-030-17795-9. DOI: 10.1007/978-3-030-17795-9_10.

[60] Y. Lecun et al. "Gradient-Based Learning Applied to Document Recognition". In: *Proceedings of the IEEE* 86.11 (Nov./1998), pp. 2278–2324. ISSN: 00189219. DOI: 10.1109/5.726791.

[61] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[62] Aston Zhang et al. *Dive into Deep Learning*. 2020.

[63] Athanasios Voulodimos et al.
"Deep Learning for Computer Vision: A Brief Review". en.
In: *Computational Intelligence and Neuroscience* 2018 (2018), pp. 1–13.
ISSN: 1687-5265, 1687-5273. DOI: 10.1155/2018/7068349.

[64] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". en. In: *International Journal of Computer Vision* 128.2 (Feb. 2020), pp. 336–359.
ISSN: 0920-5691, 1573-1405. DOI: 10.1007/s11263-019-01228-7.

[65] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". en.
In: *arXiv:1409.1556 [cs]* (Apr. 2015). arXiv: 1409.1556 [cs].

[66] Christian Szegedy et al. "Going Deeper with Convolutions". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA: IEEE, June 2015, pp. 1–9. ISBN: 978-1-4673-6964-0.
DOI: 10.1109/CVPR.2015.7298594. arXiv: 1409.4842.

[67] Sergey Ioffe and Christian Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift".
In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15.
Lille, France: JMLR.org, 2015, pp. 448–456. arXiv: 1502.03167.

[68] Kaiming He et al. "Deep Residual Learning for Image Recognition".
In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, June 2016, pp. 770–778.
ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.90.
arXiv: 1512.03385.

[69] G. Huang et al. "Densely Connected Convolutional Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017, pp. 2261–2269. DOI: 10.1109/CVPR.2017.243.

[70] F. Pasa et al. "Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization". en.
In: *Scientific Reports* 9.1 (Dec. 2019), p. 6268. ISSN: 2045-2322.
DOI: 10.1038/s41598-019-42557-4.

[71] Pranav Rajpurkar et al.
"CheXpedition: Investigating Generalization Challenges for Translation of Chest X-Ray Algorithms to the Clinical Setting". en.

In: *arXiv:2002.11379 [cs, eess]* (Mar. 2020).
arXiv: 2002.11379 [cs, eess].

[72] Nan Wu et al. "Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening". en.
In: *arXiv:1903.08297 [cs, stat]* (Mar. 2019).
arXiv: 1903.08297 [cs, stat].

[73] Sebastian Gündel et al. "Learning to Recognize Abnormalities in Chest X-Rays with Location-Aware Dense Networks".
In: *Lecture Notes in Computer Science* (2019), pp. 757–765.
ISSN: 1611-3349. DOI: 10.1007/978-3-030-13469-3_88.

[74] Luke Oakden-Rayner.
"Exploring Large-Scale Public Medical Image Datasets". en.
In: *Academic Radiology* 27.1 (Jan. 2020), pp. 106–112. ISSN: 10766332.
DOI: 10.1016/j.acra.2019.10.006.

[75] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library".
In: *Advances in Neural Information Processing Systems 32*.
Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035.

[76] WA Falcon. "PyTorch Lightning". In: *GitHub.*
https://github.com/PyTorchLightning/pytorch-lightning 3 (2019).

[77] Mehdi Mirza and Simon Osindero.
"Conditional Generative Adversarial Nets". en.
In: *arXiv:1411.1784 [cs, stat]* (Nov. 2014).
arXiv: 1411.1784 [cs, stat].

[78] Augustus Odena, Christopher Olah, and Jonathon Shlens.
"Conditional Image Synthesis with Auxiliary Classifier GANs". In:
Proceedings of the 34th International Conference on Machine Learning.
Ed. by Doina Precup and Yee Whye Teh. Vol. 70.
Proceedings of Machine Learning Research.
International Convention Centre, Sydney, Australia: PMLR, Aug. 2017,
pp. 2642–2651. arXiv: 1610.09585.

[79] Scott Reed et al. "Generative Adversarial Text to Image Synthesis". In:
Proceedings of the 33rd International Conference on Machine Learning.
Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48.

Proceedings of Machine Learning Research.
New York, New York, USA: PMLR, June 2016, pp. 1060–1069.

[80] Antonia Creswell and Anil Anthony Bharath.
"Inverting the Generator of a Generative Adversarial Network".
In: *IEEE Transactions on Neural Networks and Learning Systems* 30.7 (July 2019), pp. 1967–1974. ISSN: 2162-237X, 2162-2388.
DOI: 10.1109/TNNLS.2018.2875194. arXiv: 1611.05644.

[81] Rameen Abdal, Yipeng Qin, and Peter Wonka. "Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space?" In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.
Seoul, Korea (South): IEEE, Oct. 2019, pp. 4431–4440.
ISBN: 978-1-72814-803-8. DOI: 10.1109/ICCV.2019.00453.
arXiv: 1904.03189.

[82] Andrew Brock, Jeff Donahue, and Karen Simonyan.
"Large Scale GAN Training for High Fidelity Natural Image Synthesis".
In: *International Conference on Learning Representations*. 2019.
arXiv: 1809.11096.

[83] WJ Youden. "Index for Rating Diagnostic Tests".
In: *Cancer* 3.1 (Jan. 1950), pp. 32–35. ISSN: 0008-543X.
DOI: 10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3.

[84] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva.
"A U-Net Based Discriminator for Generative Adversarial Networks".
In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020. arXiv: 2002.12655.

[85] Rewon Child. "Very Deep {VAEs Generalize Autoregressive Models and Can Outperform Them on Images".
In: *International Conference on Learning Representations*. 2021.
arXiv: 2011.10650.

Appendix A

Additional Materials

A.1 Human Research Ethics Waiver

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG	HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

Office of the Deputy Vice-Chancellor (Research & Post Graduate Affairs)

TO: Mr B Segal
School: Electrical & Information Engineering
University

E-mail: 1377977@students.wits.ac.za

CC: Supervisor: Mr A Pantanowitz <Adam.Pantanowitz@wits.ac.za>
and <HREC-Medical.ResearchOffice@wits.ac.za>

FROM: Iain Burns
Human Research Ethics Committee (Medical)
Tel: 011 717 1252

E-mail: Iain.Burns@wits.ac.za

DATE: 11/08/2020

REF: *R14/49*

PROTOCOL NO: W-CBP-200811-01 (This is your ethics application study reference number. Please quote this reference number in all correspondence relating to this study)

PROJECT TITLE: *Hierarchical learning and generative networks for X-ray diagnostics*

Please find attached the Ethics Waiver Certificate for the above project. I hope it goes well and that an article in a recognized publication comes out of it. This will reflect well on your professional standing and contribute to the Government funding of the University.

MSWorks2000/Iain0007/ClearScanWaiver.wps

UNIVERSITY OF THE
WITWATERSRAND
JOHANNESBURG

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

11/08/2020

Ref: W-CBP-200811-01

TO WHOM IT MAY CONCERN:

Waiver: This certifies that the following research does not require clearance from the Human Research Ethics Committee (Medical).

Investigator: Mr B Segal
Student No. (if appropriate): 1377977
Staff No. (if appropriate):

Supervisor: Mr A Pantanowitz

School: Electrical & Information Engineering
University

Project title: *Hierarchical learning and generative networks for X-ray diagnostics*

Reason: Review of information in the public domain.
No human participants will be involved in the study.

Dr CB Penny
Co-Chairperson: Human Research Ethics Committee (Medical)

Research Office Secretariat:
Physical address: Phillip Tobias Building, 3rd Floor, Office 302, Corner York Road and Princess of Wales Terrace, Parktown, Johannesburg 2193.
Postal address: Private Bag 3, Wits 2050
Tel Nos. +27 (0)11-717-1234/2656/2700/1252
Office E-mail: HREC-Medical.ResearchOffice@wits.ac.za.
Website:http://www.wits.ac.za/research/about-our-research/ethics-and-research-integrity/

A.2 X-Ray Review Form

Synthetic X-Ray Evaluation

*Required

Form Purpose:

This form is a follow-up to a project that sought to create a method of generating synthetic chest x-rays. The algorithm was trained to attempt to generate images that appear similar to x-ray images, while another system was trained to detect if a fake image was present. These two are trained in opposition to one another to build the optimal x-ray generation and detection systems. The final generation system can be used to produce high quality x-ray images without concerns of the ethics of releasing private medical data.

This form is intended to be completed in less than 15 minutes.

This form is divided into four sections:
1) Real Image Selection
2) Image Discrimination
3) Pathology Identification
4) Comments & Feedback

The first section will ask you to pick images you believe to be possible chest x-rays that could exist, i.e. that an image does not appear to be obviously flawed or impossible.

The second section will present you with randomised pairs of images where one is real and the other fake with the goal of selecting the image you believe to be a true x-ray.

The third section presents a handful of generated images and asks you to select checkboxes that match featues seen in the image.

The fourth section asks if you have any comments or feedback based on the images.

Email addresses are not collected, however, you are required to sign in to prevent multiple form submissions.
All information regarding responses will be aggregated.

If you have any questions about this project or the images, please contact Brad Segal:
1377977@students.wits.ac.za

1. Level of Qualification in Radiology *

Mark only one oval.

- ◯ Consultant
- ◯ Registrar
- ◯ Medical Officer
- ◯ Intern
- ◯ Medical Student
- ◯ Other

Real Image Selection

Goal:

Please select all the images that appear to you as if they are clinically plausible. In other words, select the images which are not impossible or obviously flawed. You are free to select any number of images with no selection implying no images in the set are possible and all images selected meaning that all the images in the set are possible.

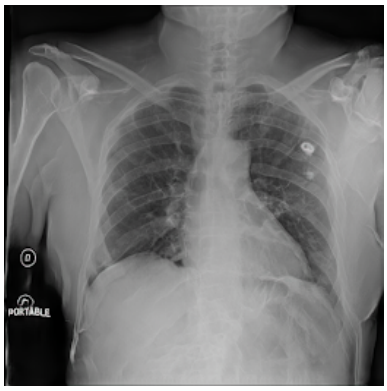
The sampled images are a mix of normal and pathological images that would be encountered in normal practice. All are frontal chest x-rays and may have a variety of support devices, markers, or text annotations present. The generated images are intended to replicate this and as such should be expected to have similar feautres.

Images are randomly selected for both model and algorithm sets without any form of curation. You are free to select no images from the set if you believe none are real.

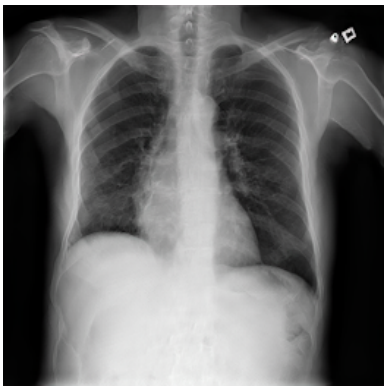
We realise the image size limits on the form are sub-optimal for close inspection, however, the goal is to gauge the overall impression of the image.

2. 1. Which X-Rays Appear Real?

Tick all that apply.

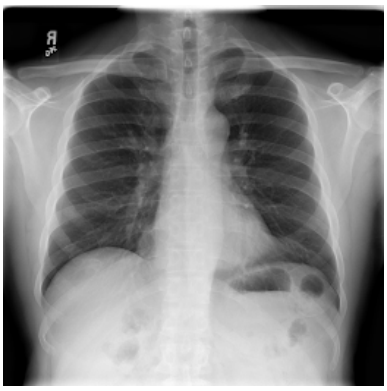


A



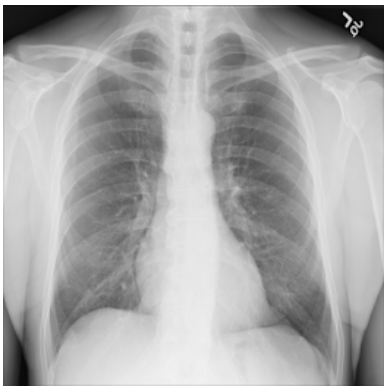
B

C



D

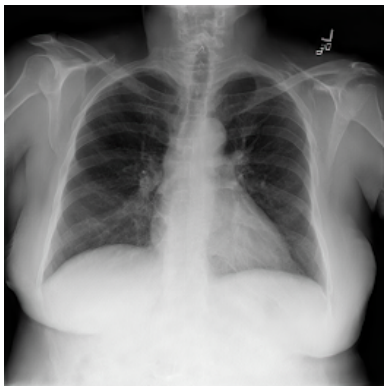
E



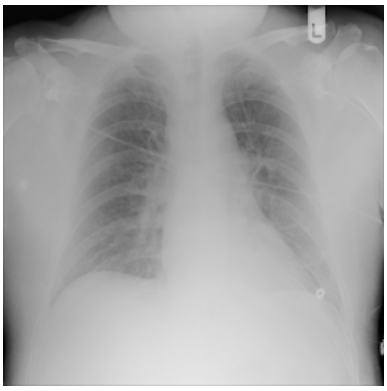
F

3. 2. Which X-Rays Appear Real?

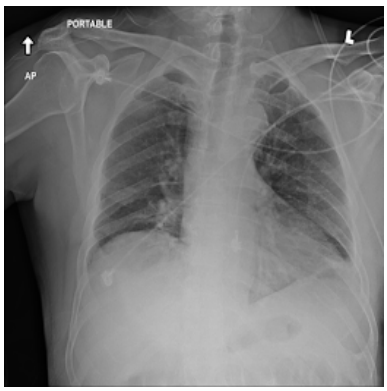
Tick all that apply.



☐ A

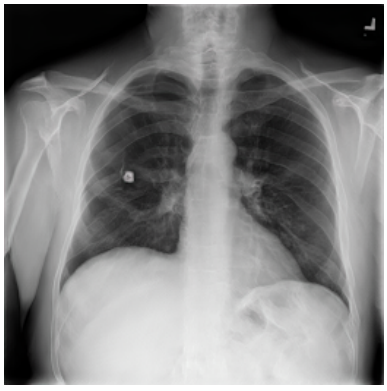


☐ B

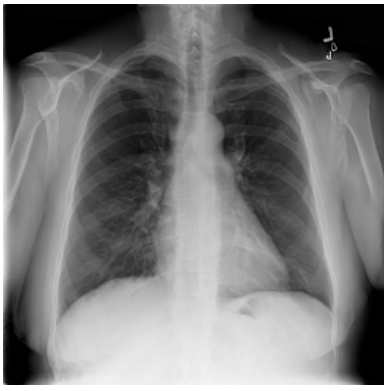


☐ C

☐ D



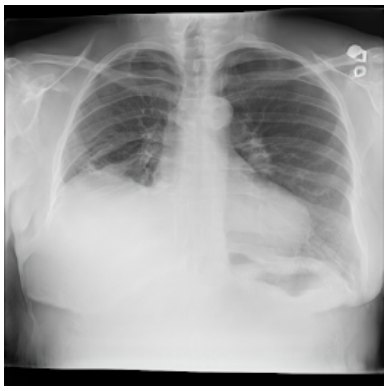
☐ E



☐ F

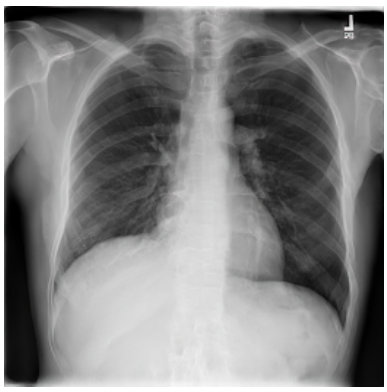
4. 3. Which X-Rays Appear Real?

Tick all that apply.



☐ A

☐ B



☐ C

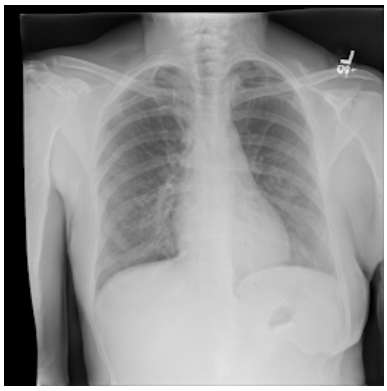
☐ D

☐ E

☐ F

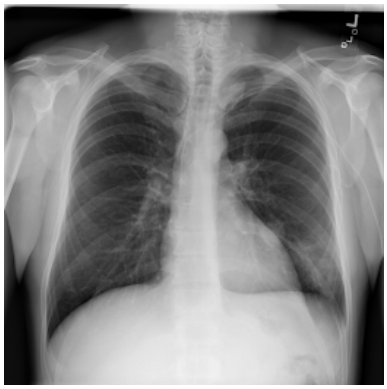
5. 4. Which X-Rays Appear Real?

Tick all that apply.



A

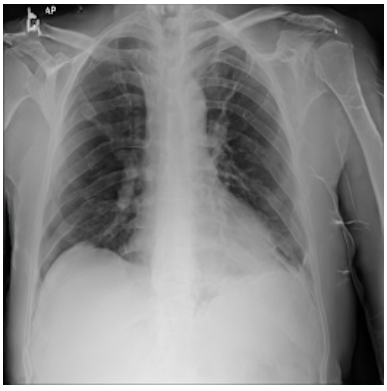
B



C

D

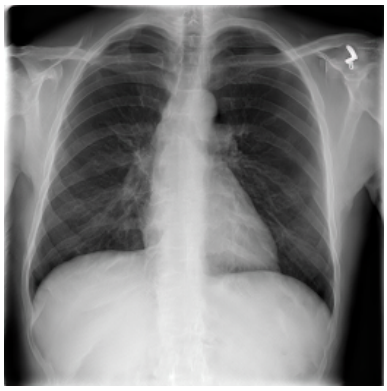
E



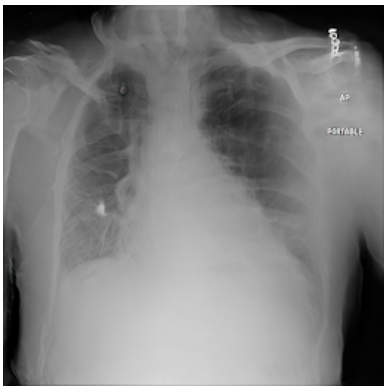
F

6. 5. Which X-Rays Appear Real?

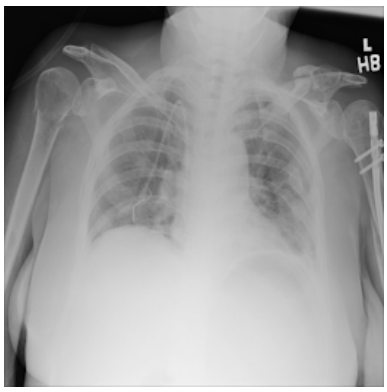
Tick all that apply.



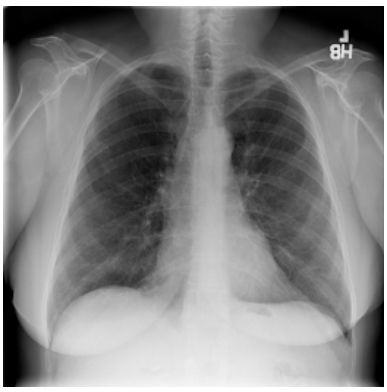
☐ A



☐ B



☐ C



☐ D

☐ E

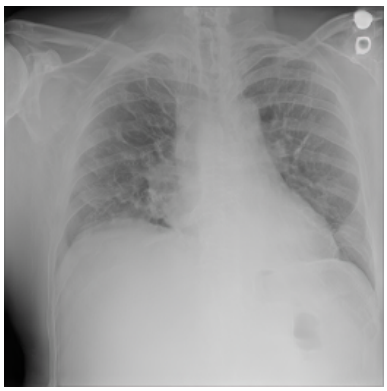
☐ F

7. 6. Which X-Rays Appear Real?

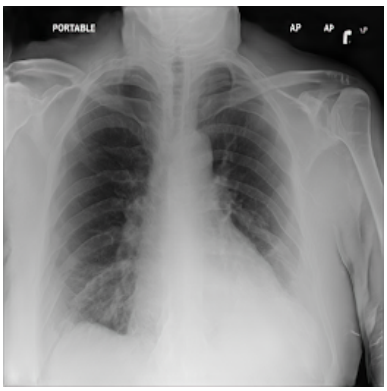
Tick all that apply.

A

B



C



D

E

F

Image Discrimination

Goal:

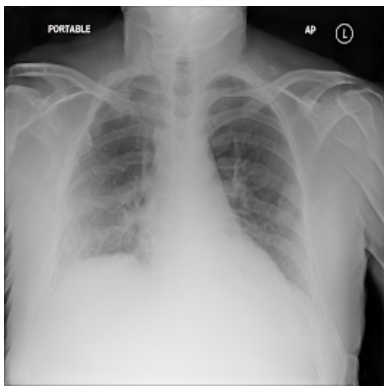
Please select the image which you believe to be a real chest x-ray. All the properties of the previous samples apply, with images being selected at random. Images may be normal or pathological and may have various markup features.

The order in which you see the images is randomised and as such the question numbers may appear to be unordered.

One and only one image is real, please indicate which you think it is.

8. 1. Which X-Ray Appears To Be Real? *

Mark only one oval.



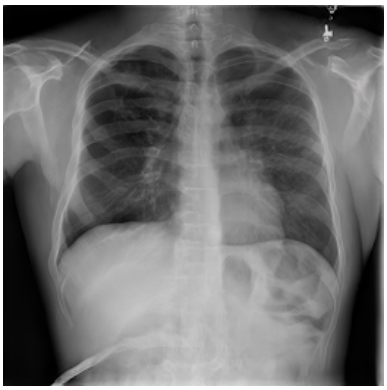
A

B

9. 2. Which X-Ray Appears To Be Real? *

Mark only one oval.

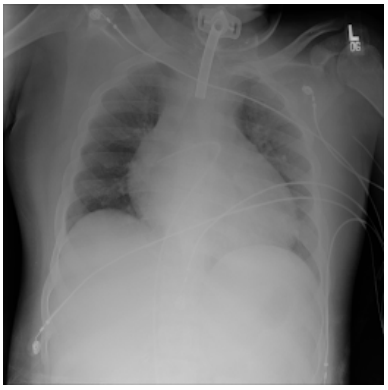
A



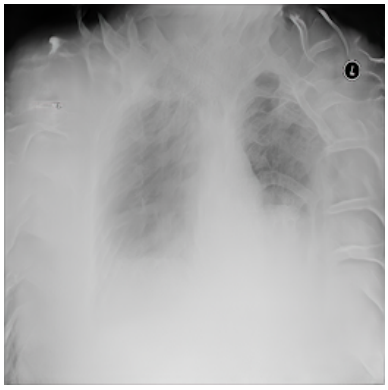
B

10. 3. Which X-Ray Appears To Be Real? *

Mark only one oval.



A



B

11. 4. Which X-Ray Appears To Be Real? *

Mark only one oval.

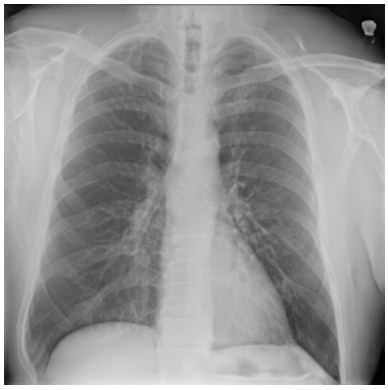
A

B

12. 5. Which X-Ray Appears To Be Real? *

Mark only one oval.

A

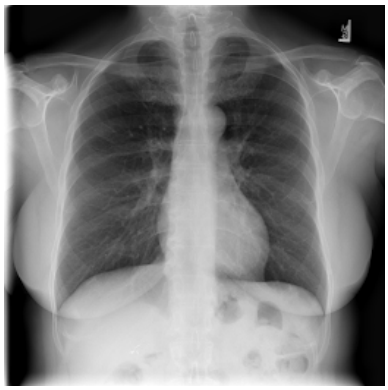


B

13. 6. Which X-Ray Appears To Be Real? *

Mark only one oval.

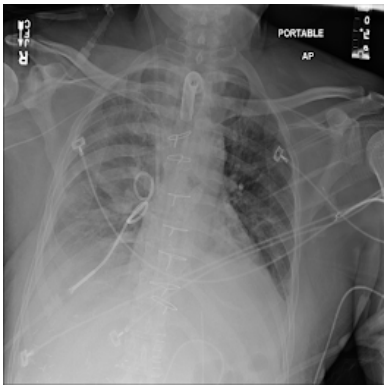
A



B

14. 7. Which X-Ray Appears To Be Real? *

Mark only one oval.

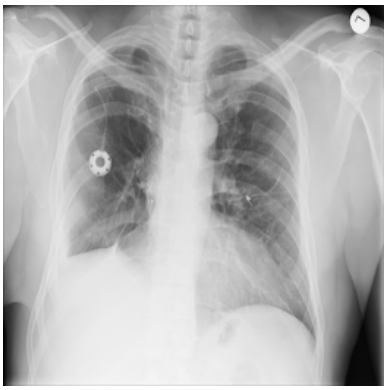


A

B

15. 8. Which X-Ray Appears To Be Real? *

Mark only one oval.

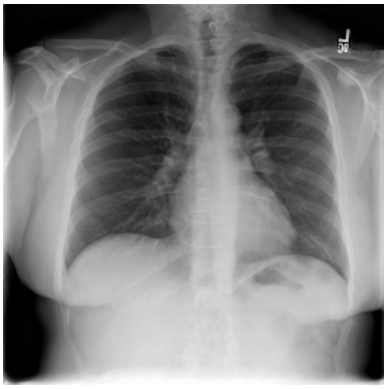


A

B

16. 9. Which X-Ray Appears To Be Real? *

Mark only one oval.



A

B

17. 10. Which X-Ray Appears To Be Real? *

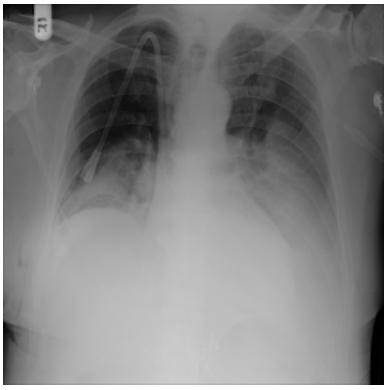
Mark only one oval.

A

B

18. 11. Which X-Ray Appears To Be Real? *

Mark only one oval.

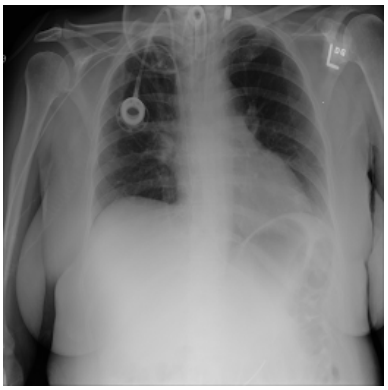


A

B

19. 12. Which X-Ray Appears To Be Real? *

Mark only one oval.

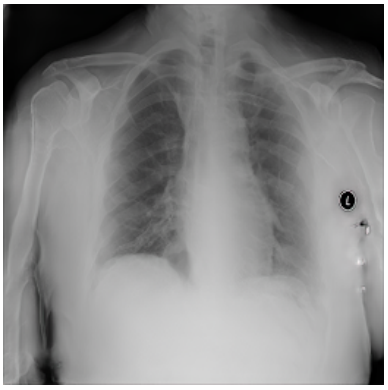


A

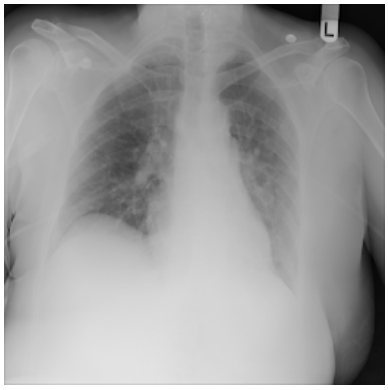
B

20. 13. Which X-Ray Appears To Be Real? *

Mark only one oval.



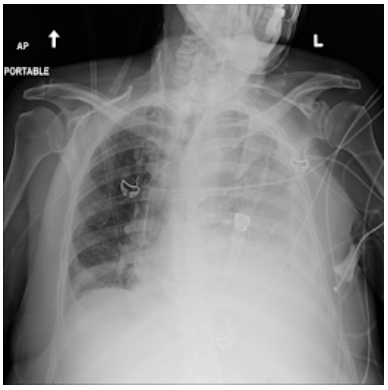
A



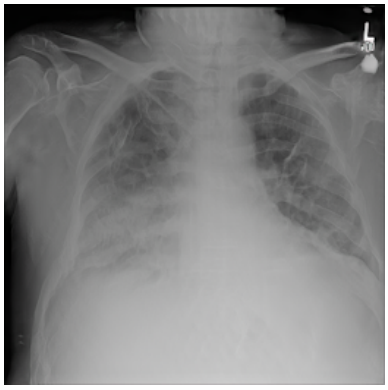
B

21. 14. Which X-Ray Appears To Be Real? *

Mark only one oval.



A

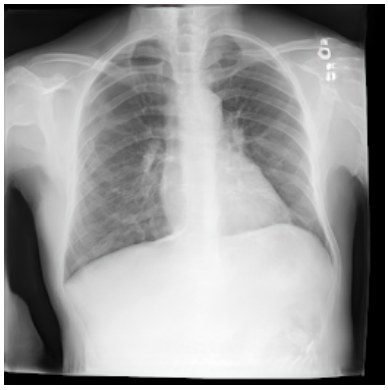


B

22. 15. Which X-Ray Appears To Be Real? *

Mark only one oval.

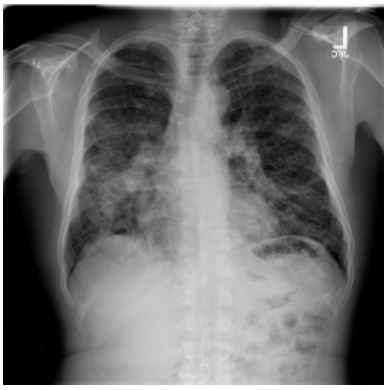
A



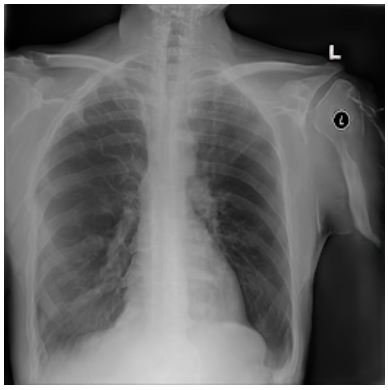
B

23. 16. Which X-Ray Appears To Be Real? *

Mark only one oval.



A



B

24. 17. Which X-Ray Appears To Be Real? *

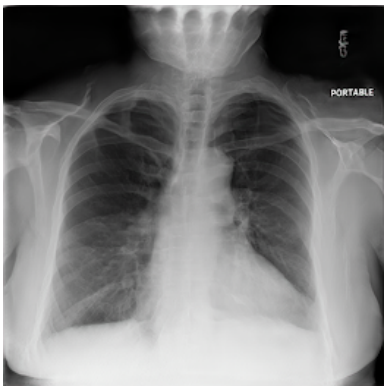
Mark only one oval.

A

B

25. 18. Which X-Ray Appears To Be Real? *

Mark only one oval.

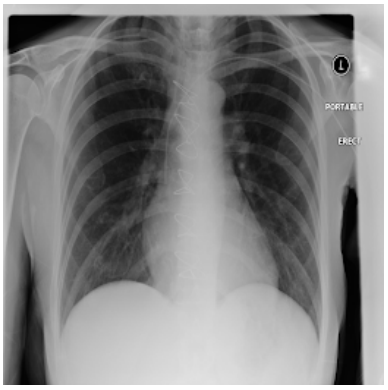


A

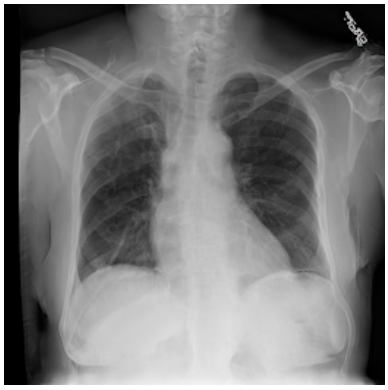
B

26. 19. Which X-Ray Appears To Be Real? *

Mark only one oval.



A

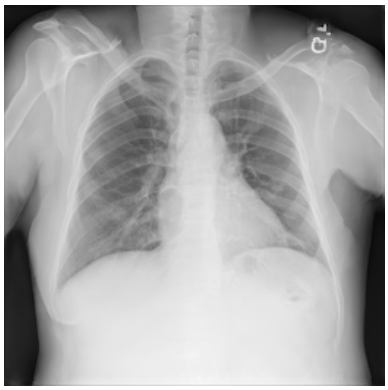


B

27. 20. Which X-Ray Appears To Be Real? *

Mark only one oval.

A



B

Pathology Identification

Goal:

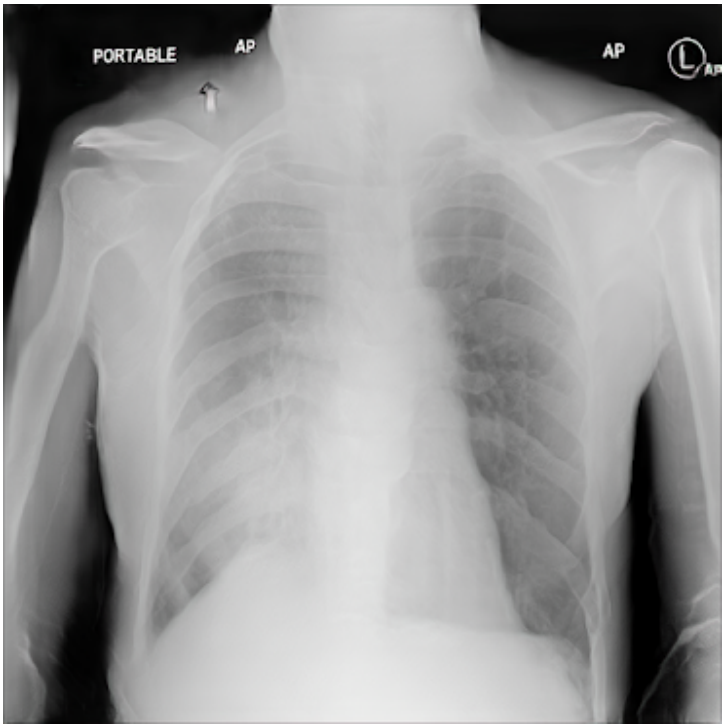
The following images are generates. Please select from the checkboxes below each image based on which features you are able to identify in the image.
If the image looks normal or has no identifiable findings, please select No Findings and leave the other options blank.

28. 1. Please identify which pathological features (if any) the image possesses *

Tick all that apply.

- [ ] Atelectasis
- [ ] Cardiomegaly
- [ ] Consolidation
- [ ] Emphysema
- [ ] Fibrosis
- [ ] Hernia
- [ ] Infiltration
- [ ] Lymph Adenopathy
- [ ] Mass
- [ ] Pleural Effusion
- [ ] Pleural Thickening
- [ ] Pneumonia
- [ ] Pneumothorax
- [ ] Pulmonary Oedema
- [ ] No Findings

29. 2. Please identify which pathological features (if any) the image possesses *



Tick all that apply.

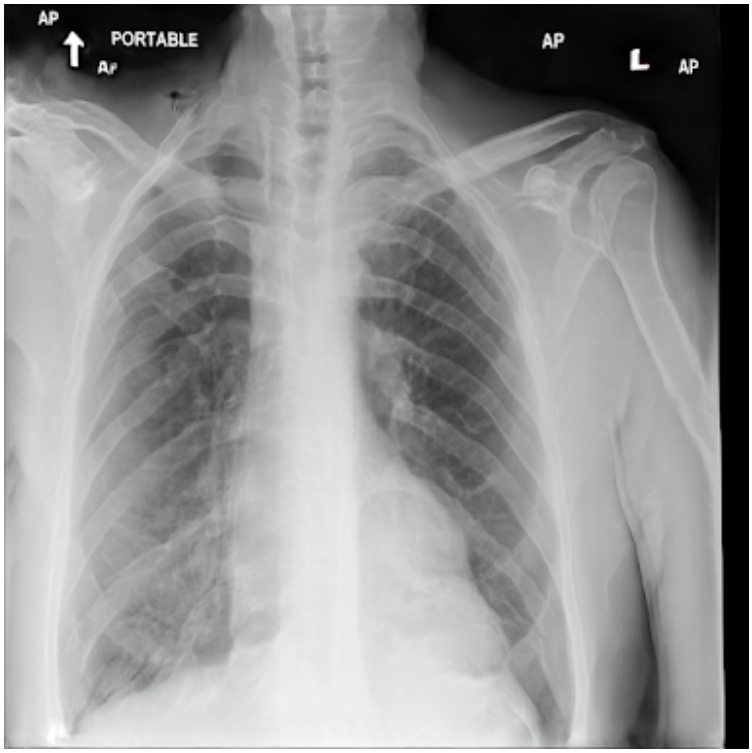
- [ ] Atelectasis
- [ ] Cardiomegaly
- [ ] Consolidation
- [ ] Emphysema
- [ ] Fibrosis
- [ ] Hernia
- [ ] Infiltration
- [ ] Lymph Adenopathy
- [ ] Mass
- [ ] Pleural Effusion
- [ ] Pleural Thickening
- [ ] Pneumonia
- [ ] Pneumothorax
- [ ] Pulmonary Oedema
- [ ] No Findings

30. 3. Please identify which pathological features (if any) the image possesses *

Tick all that apply.

- [ ] Atelectasis
- [ ] Cardiomegaly
- [ ] Consolidation
- [ ] Emphysema
- [ ] Fibrosis
- [ ] Hernia
- [ ] Infiltration
- [ ] Lymph Adenopathy
- [ ] Mass
- [ ] Pleural Effusion
- [ ] Pleural Thickening
- [ ] Pneumonia
- [ ] Pneumothorax
- [ ] Pulmonary Oedema
- [ ] No Findings

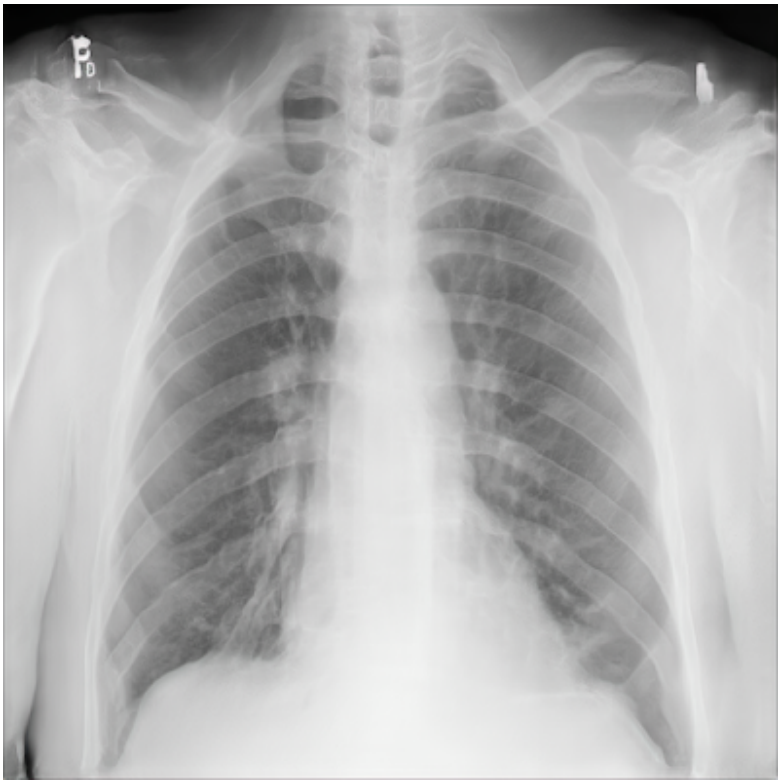
31. 4. Please identify which pathological features (if any) the image possesses *



Tick all that apply.

- [ ] Atelectasis
- [ ] Cardiomegaly
- [ ] Consolidation
- [ ] Emphysema
- [ ] Fibrosis
- [ ] Hernia
- [ ] Infiltration
- [ ] Lymph Adenopathy
- [ ] Mass
- [ ] Pleural Effusion
- [ ] Pleural Thickening
- [ ] Pneumonia
- [ ] Pneumothorax
- [ ] Pulmonary Oedema
- [ ] No Findings

32. 5. Please identify which pathological features (if any) the image possesses *



Tick all that apply.

- [ ] Atelectasis
- [ ] Cardiomegaly
- [ ] Consolidation
- [ ] Emphysema
- [ ] Fibrosis
- [ ] Hernia
- [ ] Infiltration
- [ ] Lymph Adenopathy
- [ ] Mass
- [ ] Pleural Effusion
- [ ] Pleural Thickening
- [ ] Pneumonia
- [ ] Pneumothorax
- [ ] Pulmonary Oedema
- [ ] No Findings

Comments

33. How difficult did you feel it was in general to identify fake x-rays? *

Mark only one oval.

	1	2	3	4	5	6	7	8	9	10	
Easy	◯	◯	◯	◯	◯	◯	◯	◯	◯	◯	Difficult

34. Did you notice any features that made it easy to identify fake x-rays? (Optional)