OVERCOMING BARRIERS TO DATA SHARING WITH MEDICAL IMAGE GENERATION: A COMPREHENSIVE EVALUATION

August DuMont Schütte [1,2,*], **Jürgen Hetzel** [3], **Sergios Gatidis** [4], **Tobias Hepp** [1,4], **Benedikt Dietz** [2], **Stefan Bauer** [1], **Patrick Schwab** [5,†]
[1] Max Planck Institute for Intelligent Systems, [2] ETH Zurich,
[3] Department of Medical Oncology and Pneumology, University Hospital of Tübingen
[4] Department of Radiology, University Hospital of Tübingen
[5] GlaxoSmithKline, Artificial Intelligence & Machine Learning, Switzerland
augustschdmnt@gmail.com

ABSTRACT

Privacy concerns around sharing personally identifiable information are a major practical barrier to data sharing in medical research. We utilize Generative Adversarial Networks (GANs) to create medical imaging datasets consisting entirely of synthetic patient data. We assess the quality of synthetic data generated by two GAN models for chest radiographs with 14 different radiology findings and brain computed tomography (CT) scans with six types of intracranial hemorrhages. We measure the synthetic image quality by the performance difference of predictive models trained on either the synthetic or the real dataset. We find that synthetic data performance disproportionately benefits from a reduced number of unique label combinations and determine at what number of samples per class overfitting effects start to dominate GAN training. We conducted a reader study in which trained radiologists do not perform better than random on discriminating between synthetic and real medical images for both data modalities to a statistically significant extent. Our open-source benchmark offers valuable guidelines and outlines practical conditions under which insights derived from synthetic medical images are similar to those that would have been derived from real imaging data.

1 INTRODUCTION

Sharing sensitive data under strict privacy regulations remains a crucial challenge in advancing medical research (Lo, 2015), especially due to the potential abuse of personal information (Haas et al., 2011). While there have been several large-scale research collaborations to pool and de-identify medical data for open-source research purposes (Bycroft et al., 2018; Clark et al., 2013), most data is still isolated in hospitals and laboratories due to privacy concerns (van Panhuis et al., 2010).

In medical research, information is often analysed at the level of cohorts rather than individuals. A potential solution to the medical data sharing bottleneck, is therefore, the generation of synthetic patient data that, in aggregate, has similar statistical properties to those of a source dataset without revealing sensitive private information. Among other generative deep learning models (Vahdat & Kautz, 2020), GANs have demonstrated the capability to generate realistic, high-resolution imaging datasets (Karras et al., 2019; Brock et al., 2019). GANs have been utilized within several medical domains, among others, for the generation of retinal images (Costa et al., 2018), skin lesions (Ali et al., 2019), hematoxylin and eosin (H&E) stained breast cancer tissue (Quiros et al., 2020), x-ray mammographs (Zhou et al., 2020) or chest radiographs (Han et al., 2019).

To the best of our knowledge, there is no work aimed at providing a comprehensive benchmark analysis for the generation of synthetic medical images. Our contributions can be summarized as follows: 1) We develop an open benchmark to analyse the generation of synthetic medical images when varying the number of classes, the number of samples per class, and the spatial resolution. 2) We present valuable guidelines for the effective generation of medical image datasets by evaluating

*Corresponding author.
†Work partially done while at F. Hoffmann-La Roche Ltd.

our open-source benchmark. 3) We analyse privacy considerations, assess the causal contributions of predictive models trained on the synthetic datasets, and conduct a large-scale reader study in which trained radiologists discriminate between real and synthetic medical images.

2 APPROACH

We randomly split each patient cohort into training, validation and test set within strata of radiology findings. We developed all GAN models on the training datasets and stopped GAN training when the Fréchet Inception Distance (FID) score (Heusel et al., 2017) converged (A.5). Next, we generated synthetic data for the train and validation folds by conditioning on the respective labels. In all settings, we used a pre-trained densenet-121 Convolutional Neural Network (CNN) (Huang et al., 2017) as a predictive model, with the mean area under the receiver operating characteristics curve over all labels ($\overline{AUC}$) as the evaluation metric (A.6). For each classifier, we stopped training when the validation $\overline{AUC}$ converged, after which we evaluated both on the separate, real data test fold to compute the difference in performance: $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$. We repeated all experiments multiple times with varying random initialisation to perform statistical tests (A.7).

The prog-GAN model refers to the progressive GAN as a reference model (Karras et al., 2018), still commonly used in medical image generation (Ali et al., 2019; Han et al., 2019). The cpD-GAN refers to our novel model, based on the prog-GAN with changes to the architecture and the discriminator and generator conditioning (A.4.2). To assess the generalisation capabilities, there is no fine-tuning across different benchmark settings, only when increasing the resolution, we make the necessary changes to the network architectures.

The model performance is evaluated across three benchmark dimensions, detailed in Table 1. First, we varied the number of classes (unique binary label combinations) included in the dataset. Next, we fixed the present classes and assessed how changes in the number of samples per class impacted performance. While we evaluated the first two benchmark settings at a resolution of 32×32 pixels, we finally analysed how increasing the resolution to 64×64 and 128×128 pixels affected our scores. We only performed changes across a single dimension at a time to ensure no confounding factors can impact training.

For further insights we compared the predictive models' feature importance when trained on either real or synthetic datasets, estimated with the method of Schwab & Karlen (2019) (A.9). We addressed privacy concerns by analysing differences between synthetic images and the most closely matching nearest neighbour images from the entire training dataset (A.8). Finally, we conducted a reader study in which we asked trained radiologists to label a mixed set of 100 images for both data modalities as real or synthetic (cpD-GAN) at a resolution of 128×128 pixels (A.10 and A.12).

3 RESULTS

Both models are stable across all benchmark settings with no collapses. The prog-GAN achieved an average $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ score of 0.0495 (± 0.0276) across all settings on chest x-rays and 0.1367 (± 0.0324) on brain scans. These scores were improved with the cpD-GAN that achieves 0.0206 (± 0.0100) on the chest x-rays and 0.0650 (± 0.0198) on the brain scans.

3.1 IMPACT OF NUMBER OF CLASSES

The classification performance on both real and synthetic data increased when we lowered the number of classes present in the dataset. We reason that the complexity of the predictive task decreases with fewer label combinations, resulting in higher $\overline{AUC}$ scores. However, as can be seen in Figure 1a and 1b, the differences in $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores also decreased when lowering the number of classes. For both datasets, the differences between the extreme cases (20 and 2 classes for the chest x-rays and 10 and 2 classes for the brain CT scans) for the cpD-GAN were statistically significant (p-values < 0.0001). The relative performance increase was even more pronounced for the prog-GAN. The trend of improvement in classifier performance when trained on synthetic data versus the performance when trained on real data shows that GAN models and the generated data quality disproportionately benefited from a smaller label space, thereby confirming the significance of the class conditioning methods. One crucial difference between the two evaluated GAN models is the improved label conditioning mechanism used with the cpD-GAN (Miyato & Koyama, 2018).

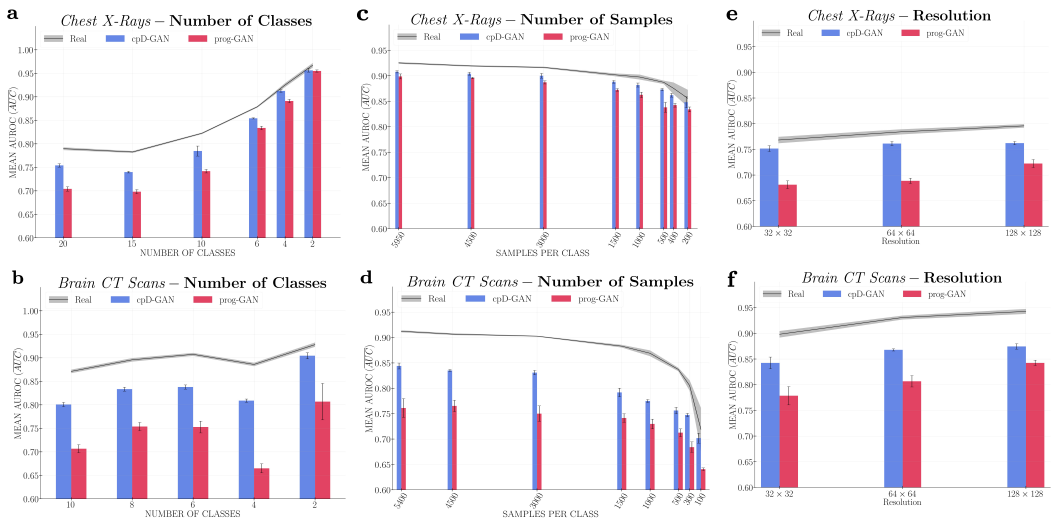

Figure 1: **Benchmark results.** In each figure, $\overline{AUC}_{\text{real}}$ scores are indicated by the black line with the shaded area representing the standard deviation across repeated experiments. The bar plots represent the $\overline{AUC}_{\text{syn}}$ scores generated by the cpD-GAN (blue) and prog-GAN (red), while the error bars indicate the standard deviation. The subfigures show the changes in predictive performance observed when varying the number of classes **a)** for chest x-rays and **b)** for brain CTs, the number of samples per class **c)** for chest x-rays and **d)** for brain CTs, and the image resolution **e)** for chest x-rays and **f)** for brain CTs.

Due to its inferior conditioning, the prog-GAN model benefited from a lower class number to a greater extent. Two important practical guidelines for synthetic data generation in medical imaging can be derived from these results: 1) Even though rare findings may be of particular medical interest including them could hurt performance and 2) research on the conditioning mechanism of GAN models should be prioritized as it is likely to lead to significant performance improvements, especially for downstream tasks.

3.2 Impact of number of samples per class

When we lowered the number of samples per label combination included in each dataset, the predictive performances obtained when training on real and synthetic data remained similar until approximately $3,000$ samples per class, see 1c and 1d. These results indicate that GAN model performance may be stable when the training data consists of at least $3,000$ samples per class. After that both the $\overline{AUC}_{\text{real}}$ and $\overline{AUC}_{\text{syn}}$ scores started to decrease substantially. However, we observed a relative performance improvement, meaning decreasing $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores for the cpD-GAN, when moving towards low numbers of samples. Despite the heightened variance in predictive performance, the difference in the scores between the extrema ($5,950$ and 200 samples per class for the chest x-rays and $5,400$ and 100 samples per class for the brain CT scans) was statistically significant (p-values < 0.001). The observed trend in performance in the low data regime indicates the growing effects of label overfitting during GAN training: Given a low number of samples, the variation within real images becomes too low, and the generative model may resort to memorizing label information in the training set instead of learning the real data distribution. This is supported by the fact that the FID scores increase towards low sample regimes indicating a decreasing image quality. In practice this means that no guarantees on the label and image consistency of synthetics from rare classes can be given.

3.3 Impact of resolution

When increasing the resolution from 32×32 pixels to 128×128 pixels, all $\overline{AUC}$ scores improved, as shown in Figure 1e and 1f. For both models the $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores either remain similar

or increase by a small margin on both datasets. This is in agreement with GAN model training at higher resolutions, which becomes more difficult due to the emergence of fine-scale details in the images. The analysed resolution levels are acceptable for evaluating the predictive models as most deep learning systems downsample medical images to reduce the computational requirements. In clinical practice, however, radiographs are analysed at a much higher resolution, and particularly fine-scaled details are essential for the accurate diagnosis of radiology findings. GAN models and other generative approaches can generate realistic images a higher spatial resolutions (Brock et al., 2019; Vahdat & Kautz, 2020), so scaling up GAN training for medical image generation is an important direction for future work.

3.4 Further evaluation

In Figure 2a and 2b , we show randomly sampled synthetic example images from the best performing cpD-GAN for both datasets at a resolution of 128×128 pixels. Below each synthetic image, we show the most similar real image (nearest neighbour) out of the entire training dataset. From a visual assessment, there appears to be no noticeable quality difference between the real and synthetic images. By comparing synthetics and nearest matching neighbours we demonstrate that the cpD-GAN model is not simply memorizing training data, and is therefore likely to preserve private, potentially sensitive information.

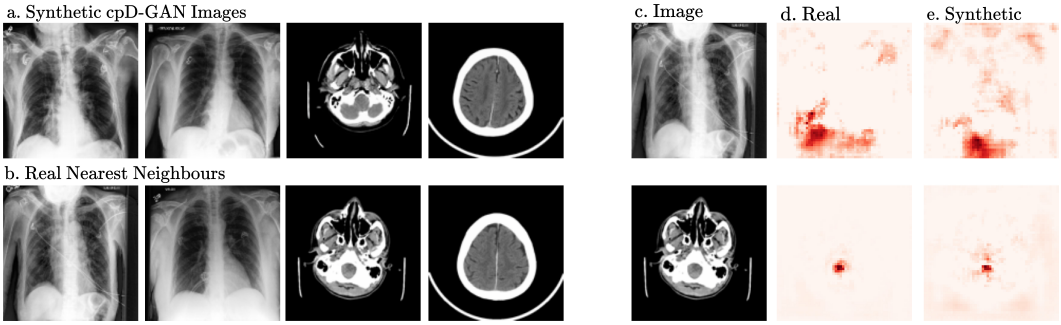

Figure 2: Synthetic images, real nearest neighbours and feature importance maps (deeper red colour indicates regions with larger causal contribution): **a)** cpD-GAN synthetics. **b)** Real nearest neighbours from training data. **c)** Real test images. **d)** Real images feature importance. **e)** Synthetic images feature importance.

To gain more interpretability, we show the feature importance at a resolution of 128×128 pixels of real test images in Figure 2c, 2e and 2f. Instead of randomly sampling real test images, we chose some of the real nearest neighbours from 2b. The observed results support the hypothesis that the predictive models trained on synthetic data from the cpD-GAN assign importance to similar image features as those trained on real data. We note that none of the feature importance maps were identical, which we expected given that the observed difference in predictive performance between classifiers trained on real and synthetic data was greater than zero at this resolution.

In our large scale reader study, we found that radiologists were unable to achieve a higher accuracy than a classifier assigning labels at random with an expected accuracy of 50% ($p < 0.05$ for chest radiographs, $p < 0.01$ for brain CT scans). The presented results indicate that trained clinicians cannot discriminate between real and synthetic images in the aforementioned setting, which further substantiates that both the general quality and label information in the synthetic images are realistic.

4 Conclusion

With our comprehensive benchmark evaluation we derived important insights and valuable guidelines for the generation of synthetic medical imaging data. While there remain open questions for further research, our results indicate that synthetic data sharing may in the future become an attractive and potentially privacy-preserving alternative to sharing real patient-level data in the right settings.

REFERENCES

Ibrahim Saad Ali, Mamdouh Farouk Mohamed, and Yousef Bassyouni Mahdy. Data Augmentation for Skin Lesion using Self-Attention based Progressive Generative Adversarial Network, 2019.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Clare Bycroft et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

Kenneth Clark et al. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, 2013.

P. Costa et al. End-to-End Adversarial Retinal Image Synthesis. *IEEE Transactions on Medical Imaging*, 37(3):781–791, 2018.

Harm de Vries et al. Modulating early visual processing by language. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6597–6607. Curran Associates Inc., 2017.

Adam E Flanders et al. Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. *Radiology: Artificial Intelligence*, 2(3):e190211, 2020.

Sebastian Haas et al. Aspects of privacy for electronic health records. *International Journal of Medical Informatics*, 80(2):e26–e31, 2011.

Tianyu Han et al. Breaking medical data sharing boundaries by employing artificial radiographs. *bioRxiv*, 2019. doi: 10.1101/841619.

Martin Heusel et al. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pp. 6626–6637, 2017.

Gao Huang et al. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Jeremy Irvin et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019.

T. Karras, S. Laine, and T. Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2019.

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

Tero Karras et al. Analyzing and Improving the Image Quality of StyleGAN. *arXiv preprint preprint arXiv:1912.04958*, 2019.

Inc Labelbox. Labelbox: The leading training data platform for data labeling. URL `https://labelbox.com`.

Bernard Lo. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. *JAMA*, 313(8): 793–794, 02 2015.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for GANs do actually converge? In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3481–3490, 10–15 Jul 2018.

Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations*, 2018.

Adalberto Claudio Quiros, Roderick Murray-Smith, and Ke Yuan. Pathology GAN: Learning deep representations of cancer tissue. In *Medical Imaging with Deep Learning*, 2020.

Patrick Schwab and Walter Karlen. CXPlain: Causal Explanations for Model Interpretation under Uncertainty. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. *arXiv preprint arXiv:12007.03898*, 2020.

Willem G van Panhuis et al. A systematic review of barriers to data sharing in public health. *Bulletin of the World Health Organization*, 88(6):468–468, 2010.

Xiaolong Wang et al. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018.

Han Zhang et al. Consistency regularization for generative adversarial networks. In *International Conference on Learning Representations*, 2020.

Yuanpin Zhou et al. Generating high resolution digital mammogram from digitized film mammogram with conditional generative adversarial network. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, pp. 508 – 513. International Society for Optics and Photonics, SPIE, 2020.

A Appendix

A.1 Acknowledgements

PS is an employee and shareholder of GlaxoSmithKline plc.

A.2 Data availability

Both datasets are publicly available and free to download for any registered user. The CheXpert chest radiograph dataset (Irvin et al., 2019) can be accessed at https://stanfordmlgroup.github.io/competitions/chexpert/ and the RSNA Intracranial Hemorrhage dataset (Flanders et al., 2020) is available at https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection.

A.3 Datasets and pre-processing

The CheXpert dataset consists of $224,316$ chest radiographs of $65,240$ patients, collected from radiographic examinations of the chest at the Stanford Hospital, between October 2002 and July 2017 (Irvin et al., 2019). In the dataset study, an automatic labeling tool was used to identify and classify the certainty of the presence of 14 observations from the radiology report. We turned uncertain labels into positives, to make use of all data, resulting in a binary multi-label dataset, where a large number of label combinations can co-occur.

The RSNA Intracranial Hemorrhage Dataset is composed of computed tomography studies supplied by four research institutions and labeled with the help of The American Society of Neuroradiology (Flanders et al., 2020). It consists of $752,803$ CT scan slices of the head from $18,938$ unique patients and the corresponding probabilities for the presence of 5 different hemorrhage types and the no finding label. For consistency, we turned any probability $p_{y_i} > 0$ into a positive label $y_i = 1$ and else $y_i = 0$, also resulting in a binary multi-label dataset. Since $644,874$ (85.7%) of CT scans are without any intracranial hemorrhage, we undersampled the no-finding class, resulting in a balanced dataset where at least 50% of images show some form of hemorrhage.

We randomly split the entire patient cohort into training (80%), validation (10%), and test folds (10%) within strata of radiology findings for each dataset. We excluded chest x-rays of classes with fewer then 256 samples, resulting in $117,168$ train images $(44,153$ patients), $15,318$ validation images $(5,519$ patients) and $14,687$ test images $(5,520$ patients). For the hemorrhage dataset we removed label combinations below a frequency of 100, resulting in $173,271$ train images $(15,133$ patients), $22,095$ validation images $(1,892$ patients), and $20,500$ test images $(1,892$ patients).

Dataset Information	Benchmark	Resolution	m_{labels}	$m_{label\ comb}$	n_{tr}	$n_{val/te}$	$n_{per\ label\ comb}$
Chest Radiographs	Classes	32×32	9	20	29000	3800	1450
			8	15	24000	2850	1600
Data Pool:			5	10	20000	1900	2000
n_{tr} $(n_{pat}) = 117168$ (44153)			5	6	13800	1140	2300
n_{vl} $(n_{pat}) = 15318$ (5519)			5	4	15600	760	3900
n_{te} $(n_{pat}) = 14687$ (5520)			4	2	12600	380	6300
	Samples	32×32	4	3	17850	2250	5950
All Labels:			4	3	13500	2250	4500
Enlarged Cardiomediastinum,			4	3	9000	2250	3000
Cardiomegaly, Lung Opacity,			4	3	4500	2250	1500
Lung Lesion, Edema,			4	3	3000	2250	1000
Consolidation, Pneumonia,			4	3	1500	2250	500
Atelectasis, Pneumothorax,			4	3	1200	2250	400
Pleural Effusion, Pleural Other,			4	3	600	2250	200
Fracture, Support Device,	Resolution	32×32	14	138	117168	8000	$256 - 7586$
No Finding	(pixels)	64×64	14	138	117168	8000	$256 - 7586$
		128×128	14	138	117168	8000	$256 - 7586$
Brain Hemorrhage CTs	Classes	32×32	5	10	25000	3000	2500
			5	8	24960	2400	3120
Data Pool:			5	6	25020	1800	4170
n_{tr} $(n_{pat}) = 173271$ (15133)			4	4	25000	1200	6250
n_{vl} $(n_{pat}) = 22095$ (1892)			2	2	25000	600	12500
n_{te} $(n_{pat}) = 20500$ (1892)	Samples	32×32	5	6	32400	3000	5400
			5	6	27000	3000	4500
All Labels:			5	6	18000	3000	3000
Epidural, Subarachnoid,			5	6	9000	3000	1500
Subdural, Intraparenchymal,			5	6	6000	3000	1000
Intraventricular, No Finding			5	6	3000	3000	500
			5	6	1800	3000	300
			5	6	600	3000	100
	Resolution	32×32	5	20	117168	8000	$155 - 85876$
	(pixels)	64×64	5	20	117168	8000	$155 - 85876$
		128×128	5	20	117168	8000	$155 - 85876$

Table 1: **All benchmark settings.** *Dataset information* summarizes the total available data for each dataset after preprocessing. m_{labels} refers to the number of labels. $m_{label\ comb}$ refers to the number of unique label combinations. n_{tr}, n_{vl}, n_{te} refers to the total number of training, validation and test samples. $n_{per\ label\ comb}$ refers to the number of training samples per unique label combination.

We developed the resolution benchmark for both datasets on the aforementioned setting. For the class benchmarking, we gradually reduced the number of clinical finding combinations present in the dataset, while keeping the total number of training images constant via over-sampling. When benchmarking the effect of samples per clinical finding, we fixed the number of classes and gradually decreased each class's frequency. Table 1 gives a complete summary of all dataset settings, the entire set of labels, the size of training, validation and test sets, and information on remaining labels and samples per class.

A.4 GAN MODEL DEVELOPMENT

A.4.1 PROG-GAN

We used the prog-GAN model as originally proposed in (Karras et al., 2018), as it is still regularly used for generating medical images (Ali et al., 2019; Han et al., 2019). We analysed several hyper-parameter settings, mainly different batch sizes, learning rates, number of feature channels and optimiser settings, but we determined that the original parameters proposed in (Karras et al., 2018) performed best. We began training at a spatial resolution of 8×8 pixels, which we determined to be the lowest resolution at which meaningful information is still visually apparent in downsampled images. Each transition and stabilisation phase at a resolution of 32×32 pixels lasted until the discriminator had seen 1.4M real images, which corresponded to 1.4M fake images as the number of discriminator updates per generator step is $n_{critic} = 1$. At a resolution of 64×64 and 128×128 pixels, we reduced the number of real images per phase to 1M.

A.4.2 CPD-GAN

We developed the cpD-GAN based on the prog-GAN with several important improvements that we highlight below. Please see above or (Karras et al., 2018) for details on the architecture and methods if not explicitly stated. Inspired by Style-GAN (Karras et al., 2019; Karras et al., 2019), we dropped progressive growth as we observed that it was not necessary for stable training. This allowed us to experiment with new architectures, where output skip connections within the image feature space

of the generator and standard residual connections in the discriminator improved the performance the most. We achieved significantly lower $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores when replacing the auxiliary classifier conditioning with a projection based discriminator: In the last discriminator layer, the inner product between the label vector y and the feature vector is computed as the final output, resulting in a conditioning mechanism that respects the role of the conditional information in the underlining probabilistic model (Miyato & Koyama, 2018). Inspired by conditional batch normalisation (de Vries et al., 2017), we modified the pixel-wise feature vector normalisation after each generator convolution by conditioning it on a label and noise dependent scaling and bias parameter:

$$b^i_{x,y} = \frac{a^i_{x,y}}{\sqrt{1/N \sum_{j=0}^{N-1} (a^j_{x,y})^2 + \epsilon}} \cdot \gamma^i + \beta^i \tag{1}$$

where $a^i_{x,y}$ and $b^i_{x,y}$ are the original and normalised feature of channel i in pixel (x, y) and $\epsilon = 10^{-8}$. The scaling parameter is defined as $\gamma = W_1[z; y] + b_1$ and the bias parameter as $\beta = W_2[z; y] + b_2$, where W_i and b_i are trainable weight matrices and vectors, while $[z; y]$ refers to the vector concatenation of the random normal input noise z and label y. Figure 3 shows the overall model structure and a detailed description of a generator resolution block.

We evaluated various loss functions, such as the logistic GAN loss with and without R1 or R2 regularisation, the hinge loss with and without gradient penalty, or the non-saturating GAN loss (Mescheder et al., 2018), but the Wasserstein loss with gradient penalty worked best. Replacing a specific convolutional layer in both the generator and discriminator by a normal, sparse, or non-local self-attention layer (Wang et al., 2018) did not improve performance. Neither consistency regularisation (Zhang et al., 2020), nor matching the gradients of an auxiliary classifier by minimisation of the cosine-distance when predicting the labels of fake and real images resulted in better scores. We analysed many hyper-parameters, among others the number of feature channels, batch sizes, and learning rates. The performance peaked for 512 feature channels, a batch size of 256 and learning rates of 0.005 with one discriminator update per generator update $(n_{critic} = 1)$.

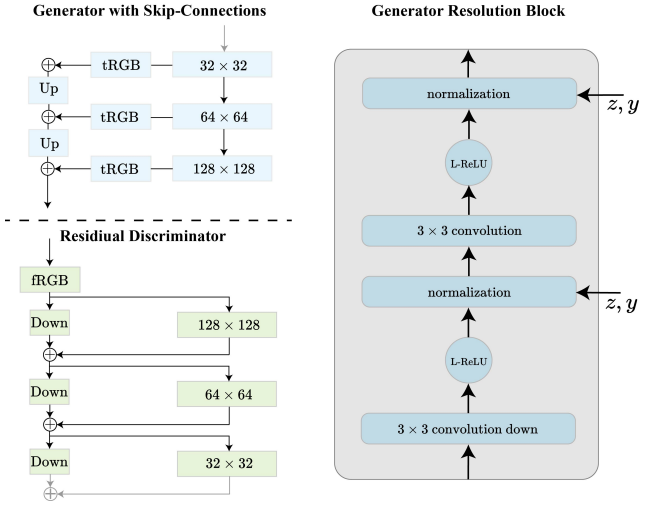

Figure 3: **Network architecture and generator block. Left:** In the generator, output skip connections in the image feature space are included after each resolution block, while the discriminator blocks have residual connections. *Up* and *Down* refer to nearest neighbour upsampling and downsampling by average pooling while *tRGB* and *fRGB* refer to the 1×1 convolution mappings to and from the image space. **Right:** The first convolution in each generator block doubles the spatial resolution via nearest neighbour upsampling and reduces the number of feature channels (if needed). Each pixel-wise feature vector normalisation layer is conditioned on the label information $\mathbf{y}$ and random normal noise vector $\mathbf{z}$. The Leaky-ReLU non-linearity is used as an activation function.

A.5 GAN TRAINING

We used the Adam optimiser for all GAN models and the hyper-parameters as proposed in (Karras et al., 2018; Brock et al., 2019), except for the learning rates that we fine-tuned as mentioned in A.4.2. We stopped training in all settings when the Fréchet Inception Distance (FID) between $10,000$ real and synthetic images converged. The FID score is a commonly used metric to compare the visual quality between two sets of images and allows for an unbiased GAN training evaluation (Heusel et al., 2017). We ran all models for a minimum number of steps, until the discriminator had seen as many real images as the prog-GAN discriminator after the final stabilisation phase. At a resolution of 32×32 pixels, each progressive phase lasted until the discriminator had seen 1.4M real images, resulting in a minimum number of 7M real images for the other models. For 64×64 and 128×128 pixels, we lowered the number of images per phase to 1M, resulting in a minimum number of images of 7M and 9M, respectively. At this point, we computed the FID score after every 400T real images, and if there was no improvement for two consecutive evaluations, we stopped training. We stopped all repetitions for each experiment at the same step as the first model to get comparable results.

A.6 PREDICTIVE MODEL DEVELOPMENT AND TRAINING

In all settings, we used a pre-trained densenet-121 convolutional neural network as the predictive model (Huang et al., 2017). We added a randomly initialised fully connected layer with sigmoid activation to the pre-trained model for classification with the binary cross-entropy loss. We resized the input images to match the densenet-121 spatial input resolution of 224×224 pixels. To make training as similar as possible across different benchmark settings, we used a maximum number of 5000 images per epoch with a batch size of 48. In settings where the total number of samples is below 5000, the number of images per epoch is accordingly lower. After each epoch, we computed the area under the receiver operating characteristic curve (AUROC) for each label in all validation data samples. We reduce the initial learning rate of 0.0001 by a factor of 10 if the mean validation AUROC ($\overline{AUC}_{val}$) across all labels does not improve after two consecutive epochs (patience of 2). If the $\overline{AUC}_{val}$ does not improve for a patience of 3 epochs, we stopped training. To compute delta scores, we tested all models on the held-out, real data test set.

A.7 STATISTICAL TESTS FOR BENCHMARK

We repeated every experiment of our benchmark with at least four different random initialisation of the entire training and evaluation pipeline, allowing us to compute the standard deviation for each setting across repetitions. This is necessary as different parameter initialisation resulting from different random seeds can substantially impact the training of deep learning systems. For the number of classes benchmark, we repeated the cpD-GAN training and subsequent synthetic classification as well as the real data classification for 10 different random initialisation for both datasets at the extrema: For 20 and 2 classes for the chest x-rays and 10 and 2 classes for brain CT scans. Subsequently we performed the one-sided, parametric-free, Mann–Whitney U test on the $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores between the extrema to determine whether there is a statistically significant difference. We followed the same approach for the samples per class benchmark with 20 repetitions at different random initialisation: For $5,950$ and 200 samples per class for the chest x-rays and $5,400$ and 100 samples per class for brain CT scans. Here we repeated both the early-stopping cpD-GAN experiments, as well as the overfitting version. We once again performed the one-sided, parametric-free, Mann–Whitney U test on the $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores between the extreme settings to determine the statistical significance.

A.8 NEAREST NEIGHBOURS

To analyse differences between our generated medical images when compared to the training data, we computed the nearest neighbours for a set of randomly sampled synthetics. For both datasets, we used the synthetic images generated by the cpD or prog-GAN model at a resolution of 128×128 pixels with the lowest $\overline{AUC}_{\text{real}} - \overline{AUC}_{\text{syn}}$ scores. We used the predictive model trained on real data at the same resolution level to find the final dense layer representation for each synthetic image; a $1,024$ dimensional vector. We compute the same representation for all real training images and determine the pair of synthetic and real images for which the cosine distance between the final

densenet representations is minimal. Using a measure of similarity in the predictive model's feature space results in a more reliable determination of nearest neighbours that exploits invariances to shifts and rotations within the image space of the chest radiographs or brain scans.

A.9 FEATURE IMPORTANCE

We computed the causal contribution of image neighbourhoods towards the label prediction with the method of Schwab & Karlen (2019). More precisely, we successively zero masked regions of 2×2 pixels in the input image and computed the new, increased predictive model loss. If the masking of a particular neighbourhood resulted in a significant loss increase, the region had accordingly higher importance. All regions were masked for each input image of 224×224 pixels after $12,544$ repetitions. To determine the feature importance we subtracted the original model loss and normalised the attribution map. Similar causal contribution maps indicate a similar quality and structure between real and synthetic images, leading to predictive models that attribute the same regions with high feature importance.

A.10 READER STUDY

We conducted the reader study by asking trained radiologists to label a set of 100 images for both data modalities as real or synthetic at a resolution of 128×128 pixels with a web-based labeling tool (Labelbox). Each set consisted of 50 randomly sampled real and synthetic images, generated by the best performing cpD-GAN. Participants were told that each individual image was sampled at random to avoid any bias during evaluation, without knowledge about the total number of reals and synthetics. For the chest x-rays, 11 radiologists participated, while 9 radiologists labeled the brain CT sets. From each labeled set we computed the values for true reals (TR), false reals (FR), true synthetics (TS) and false synthetics (FS), to determine the classification accuracy $\text{acc} = \frac{TR+TS}{TR+TS+FR+FS}$. Next, we performed the one-sided, non-parametric Wilcoxon signed-rank test to assess whether the distribution of accuracies is equal or less than the mean accuracy of a fully random classifier with $\overline{acc} = 0.5$ (50%).

A.11 MORE FIGURES

A.11.1 CPD-GAN

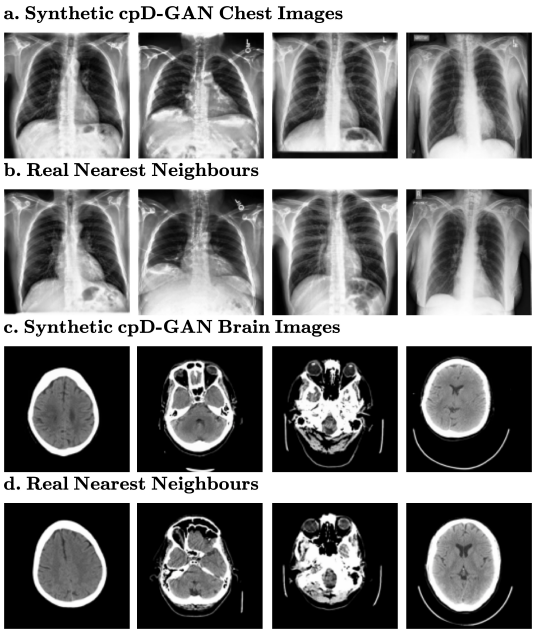

Figure 4: **More randomly sampled synthetic images from the cpD-GAN and nearest neighbours from all real training images at a resolution of** 128×128 **pixels. a)** Synthetic chest radiographs. **b)** Nearest matching real images found in the chest radiograph training set. **c)** Synthetic brain computed tomography (CT) scans. **d)** Nearest matching real images found in the brain CT training set.

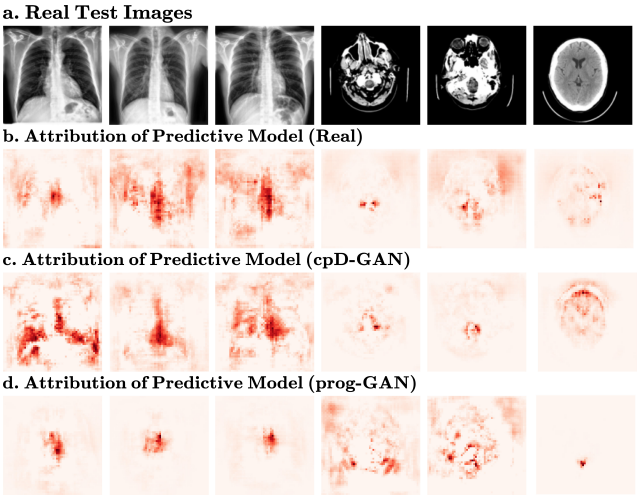

Figure 5: **More feature importance maps of predictive models.** Deeper red colour indicates regions that have a larger causal contribution to the label prediction. **a)** Real test images at 128×128 resolution. All displayed images are without any clinical finding. **b)** Feature importance of predictive model trained on real data. **c)** Feature importance of predictive model trained on synthetic data generated by the cpD-GAN. **d)** Feature importance of predictive model trained on synthetic data generated by the prog-GAN.

A.11.2 PROG-GAN

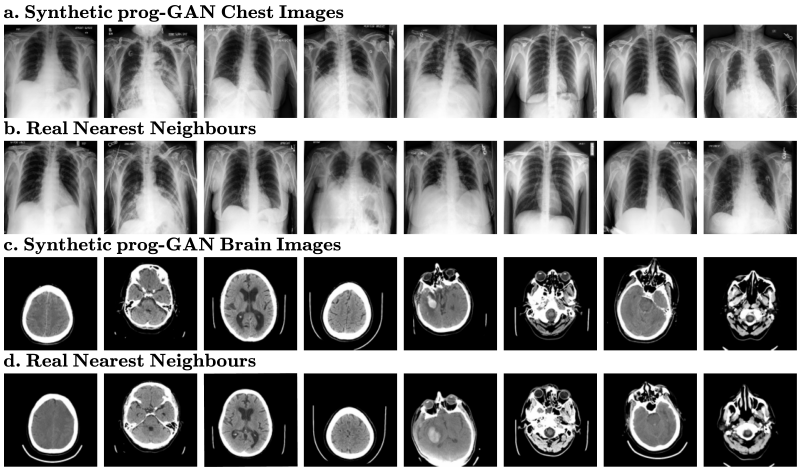

Figure 6: **Randomly sampled synthetic images from the prog-GAN and nearest neighbours from all real training images at resolution of** 128×128 **pixels. a)** Synthetic chest radiographs. **b)** Nearest matching real images found in the chest radiograph training set. **c)** Synthetic brain computed tomography (CT) scans. **d)** Nearest matching real images found in the brain CT training set.

A.12 DETAILS ON READER STUDY

The detailed confusion matrices and accuracies from the reader study are shown in Table 2 and 3.

		Radiologist	
		Real	Synthetic
Actual	Real	$TR = 25.6\ (\pm 7.1)$	$FS = 24.4\ (\pm 7.1)$
	Synthetic	$FR = 31.0\ (\pm 8.2)$	$TS = 19.0\ (\pm 8.2)$

Accuracies of Radiologist Labels

0.45	0.52	0.45	0.52	0.50	0.25	0.49	0.39	0.40	0.55	0.39

Table 2: **Chest radiographs reader study. Top:** Means and standard deviation from 11 trained radiologists for real and synthetic images at 128×128 resolution: TR = True Reals, FR = False Reals, TS = True Synthetics, FS = False Synthetics. **Bottom:** Computed accuracies from radiologist labels.

		Radiologist	
		Real	Synthetic
Actual	Real	$TR = 25.2\ (\pm 3.5)$	$FS = 24.8\ (\pm 3.5)$
	Synthetic	$FR = 30.3\ (\pm 5.8)$	$TS = 19.7\ (\pm 5.8)$

Accuracies of Radiologist Labels

0.51	0.45	0.44	0.44	0.46	0.50	0.48	0.36	0.40

Table 3: **Brain CT scans reader study. Top:** Means and standard deviation from 9 trained radiologists for real and synthetic images at 128×128 resolution: TR = True Reals, FR = False Reals, TS = True Synthetics, FS = False Synthetics. **Bottom:** Computed accuracies from radiologist labels.