This PDF file includes:

- Preprocessing steps in CheXpert dataset
- Network training details
- Experimental subset selection details
- Diversity of generated radiographs
- Advantages of the proposed GAN-based approach
- Privacy concerns in imaging modalities other than x-rays
- Alternative GMs
- Adversarial robustness
- Domain adaption using a Cycle-GAN
- Figs. S1 to S5
- Tables S1 to S5
- Algorithm S1
- References
SUPPLEMENTARY MATERIALS

Preprocessing steps in CheXpert dataset

Frontal radiographs filtering

The Chexpert dataset consists of both frontal (191,027) radiographs and lateral radiographs. Only frontal thoracic radiographs were used for the training of the Stanford-GAN.

Generating labels

1. Pathology labels not mentioned (NaN) were assigned to 0.0.
2. According to the labeling performance comparison mentioned in [33], most of the uncertainty labels (U) were assigned to 1.0, except for the consolidation class.
3. Uncertainty labels for consolidation were assigned to 0.0.
4. A Python dictionary was used to record pathology labels of radiographs: the key was set to the radiograph filename and the corresponding value was set to the pathology vector.

Preprocessing of Radiographs

The aspect ratio of radiographs was rescaled to 1:1 by zero-padding. The radiographs’ intensity values were initially normalized to the range of [0, 255] to prepare the necessary files and radiographs in real batches were further scaled to the range of [-1, 1] during training.

Network training details

We trained two conditional GANs based on two separate datasets: the NIH ChestX-ray14 dataset and the Stanford CheXpert dataset. GANs involve both a generator and a discriminator that are trained to learn the underlying distribution of medical imaging data. Given a prior distribution and conditional information, the generator (G) maps low-dimensional input data, e.g. a vector of size (512, 1), to a particular image (X-ray in our case). As an adversary, the discriminator (D) tries differentiating between the generated and the true distribution of images. The training proceeds by letting G compete against D until an equilibrium state is reached.

In the original formulation of the GAN, such a competing game was mathematically formulated as a min-max objective function:

\[
\min_G \max_D \mathbb{E}_{x \sim P_r} [\log(D(x))] + \mathbb{E}_{\tilde{x} \sim P_g} [\log(1-D(\tilde{x}))]
\]

where \(P_r\) is the training distribution and \(P_g\) is the distribution learned by the generator. As shown in [11], the objective 3 can be further expressed by the Jensen-Shannon (JS) divergence between \(P_r\) and \(P_g\). The difficulty in training a GAN lies in the fact that the JS divergence yields zero when a minor overlap is found between \(P_r\) and \(P_g\). One historical remedy of increasing the overlap of \(P_r\) and \(P_g\) was to add random noise to real and generated samples, which proved to be generally harmful in terms of image quality [37].

Instead of optimizing the objective 3, we achieved robust training of our GANs with Wasserstein-distance based losses [37]:

\[
L_G = \mathbb{E}_{x \sim P_r} [D_w(x)]
\]

\[
L_D = \mathbb{E}_{x \sim P_r} [D_w(x)] + \mathbb{E}_{\tilde{x} \sim P_g} [D_w(\tilde{x})] + \lambda \mathbb{E}_{x \sim P_r} (\|\nabla_x D_w(x)\|_2 - 1)^2 + \epsilon \mathbb{E}_{x \sim P_r} [D_w(x)]^2
\]

where \(w\) is the model parameter, \(\lambda\) and \(\epsilon\) control the penalty term strength. The first and second term of \(L_D\) in equation 4 compute the Wasserstein distance between the true and the generated distributions. To avoid loss drifting, the score for real images was pinned around zero (Fig. S1A) by the epsilon penalty term [12].

In addition to using a meaningful Wasserstein loss, a progressive growing strategy was used to incrementally train the model towards higher spatial resolutions. This technique yields stable and high-quality results...
as we intentionally shifted the focus of the network from basic image contrast (learned from low-resolution images) to high frequency details (learned from high-resolution images). Such a progressive growing process is illustrated in Fig. S1A. Whenever the network doubled its spatial resolutions, i.e., from spatial resolution stage \(i-1\) to spatial resolution stage \(i\) (Fig. S1B), we linearly increased the weight \(\alpha\) to incorporate the new convolutional layer. In our setting, once training of one spatial resolution stage stabilized after being presented a total of 600,000 real radiographs (with repetitions), the layers responsible for the next spatial resolution stage were gradually faded in and training continued with another 600,000 radiographs during this fade-in stage (again with repetitions). In total, the GANs were each presented 12 million radiographs.

To demonstrate the convergence process, we plotted the loss terms of the Stanford-GAN in Fig. S1C. Convergence is achieved by reaching a plateau for the real and synthesized score of the discriminator. It should be noted that this mathematical proceeding was in agreement with the visually assessed quality: we found that image quality saturated after the presentation of 10 million images, see also Fig. S1D.

**Fig. S1.** The training process of our progressive growing GAN. A: Starting from a basic \(4 \times 4\) spatial resolution, we smoothly increase the spatial resolution of both generator and discriminator by linearly fade in more convolutional layers. B: Technical details about how to incorporate a new convolutional layer for doubling network’s spatial resolution. When the network doubles its spatial resolutions, i.e., from spatial resolution stage \(i-1\) to spatial resolution stage \(i\), we linearly increase \(\alpha\) to sum outputs from a new convolutional layer and the old layer. C: To achieve conditional generation, we simply concatenate the conditioning representation, i.e., label vector, to the random input vector. The combined vector is passed to a fully-connected layer. Concatenation based conditioning can be viewed as a form of conditional biasing once we decompose the matrix-vector product of the fully-connected layer into two matrix-vector subproducts (see the inner plot). D: Scores for both real and generated samples during the training process of Stanford-GAN. Due to the epsilon penalty term, the real score fluctuates around zero and therefore the competition between G and D is reflected by the synthesized score. The convergence of the Stanford-GAN was indicated by a stable Wasserstein distance, i.e., real score - synthesized score, which was found after feeding 10 million radiographs to the network in the training process. E: Calculated FID for both Stanford- and NIH-GANs. The convergence of both GANs was demonstrated by the FID reaching a plateau after 10 million images.
Experimental subset selection details

In order to give a better overview over why we chose the numbers in our experiments, we have compiled the Table S1 as an overview and have added explanatory text for each experiment:

| Section                                      | Number of Subset | Justification                                                                 |
|----------------------------------------------|------------------|--------------------------------------------------------------------------------|
| Generation of synthesized Radiographs        | 1,000 fake       | It’s sufficiently high to represent the distribution of MS-SSIM values within one group as depicted in the histogram of Fig. S2A. |
| Ability of Human Readers to Distinguish      | 50 real, 50 fake | Power analysis                                                                  |
| Synthesized Radiographs from Real X-ray Images | 1,000 nearest neighbor | 99% certainty that fraction of images with private information is smaller than 0.005. |
| Performance of Classifiers Trained on        | 500 real + 500 fake | Estimation of typical data size in a realistic setting.                      |
| Synthesized Radiographs                      |                  |                                                                                |
| Federated averaging facilitates the training of local GANs | 20,000 real    | Estimation of typical data size in a setting with scarcity of labelled data. |

**Generation of synthesized Radiographs**

In this experiment we intended to demonstrate the distribution of SSIM for both the real and synthesized radiographs. As no reliable prior data were available for the performance of similar networks, no dedicated power analysis was used to predetermine the sample size. Rather, our goal was to give a faithful representation of the distribution and we checked whether the - admittedly extrapolated - distribution was visually represented by the histograms in Fig. S2A. We also performed an additional test by recalculating the same SSIM with $10^6$ samples and found that the relevant metrics (i.e. mean and standard deviation) differed by less than one percent.

**Ability of Human Readers to Distinguish Synthesized Radiographs from Real X-Ray Images**

In order to determine the number of needed samples, we used power analyses according to [36]. For the differentiation of synthesized against real radiographs by human readers we made the following assumptions: a priori ability to recognize a synthetic radiograph as synthetic: 0.5, a priori ability to recognize a real radiograph as real: 0.8, alpha error: 0.05, and power: 0.8. This resulted in a sample size of 38 for both the real and synthesized radiographs. To account for inevitable inaccuracies in the assumptions above, we chose to increase both numbers to 50.

**Ensuring Retainment of Private Information**

To determine how many test samples are needed for the privacy test, we made the following assumptions: If there is a fraction of synthesized images which contain private information we wanted to be able to discover a fraction of greater than 0.005 with a probability of at least 99 %. With a sample size of 1,000, the probability of drawing zero “hits” is 0.007. Thus, we can at least be 99 % sure that the fraction of synthesized images that contain private information is smaller than 0.005 when using a sample size of 1,000 for the visual reader test.

**Performance of Classifiers Trained on Synthesized Radiographs**

In this experiment our goal was to demonstrate the advantage of using synthesized X-rays in conjunction with real X-rays when manually labelled data is scarce. Manually labelling data is a tedious process and we estimated, that a typical size of labelled data for a specific task would have to be an order of magnitude higher than 100 (otherwise the classifier would have problems of being trained), but probably would have to be much smaller than 100,000 or even 10,000 due to the time it takes to label all the data. Our reasoning was as follows: to label a single x-ray accurately from our experience takes about 3 minutes when including the full workflow from opening the image to putting in the labels. For 10,000 X-rays, this would result in a net labelling time of 500 hours - so about 3 months of full-time labelling work. We therefore concluded that a suitable data size of images on which a classifier would be trained in this setting would be 1,000 as a compromise between the need for more data (more than 100) and the assignment of human resources.

**Federated averaging facilitates the training of local GANs**

Similarly to the reasoning in “Performance of Classifiers Trained on Synthesized Radiographs” we estimated, that small local hospitals might be able to label on the order of 1,000 images accurately. Our additional assumption was that a realistic collaboration of local hospitals might result in about 20 different entities. Thus
we arrived at a total of 20,000 X-rays being trained in a federated learning approach with each of the 20 entities contributing 1,000 labelled X-rays.

**Diversity of generated radiographs**

The SSIM difference in Fig. S2B and C is denoted as $\Delta$SSIM. For a total number of 100,000 repetitions, we randomly sampled two pairs of images from a pool containing 100 mixed real&synthesized radiographs (50% real and 50% synthesized), and obtained the SSIM difference between pairs via $\Delta$SSIM$_i = \text{SSIM}_{\text{pair}_i} - \text{SSIM}_{\text{pair}_{i+1}}$. As listed in Fig. S2D and E, we computed the p-value of the SSIM difference by determining what fraction of such found random differences were greater than the difference we actually found.

![Fig. S2. Diversity of generated radiographs](image)

**Advantages of the proposed GAN-based approach**

One additional advantage of the proposed concept of data sharing is the reduction in data storage requirements: storing a single radiograph image with a spatial resolution of $1024 \times 1024$ and a bit depth of 8 bits requires 1 megabyte of hard drive space. Considering that the number of radiographs acquired at a maximum care hospital easily reaches 100,000 per year, constructing a database of millions of radiographs (ImageNet comprises e.g. 14 million natural images) is not out of reach. While the actual images themselves were to take up terabytes of data, the storing of the weights of the GAN generators only amounts to a few hundred megabytes (in our case about 23 million weights with a precision of 32 bits: 80 MB of storage). The GANs thus also serve as a data content compression tool that might promote the development of CV algorithms at smaller institutions that do not have access to large data storage capabilities.
Privacy concerns in imaging modalities other than X-rays

The X-ray dataset used in this manuscript did not contain any name information pixel-hard-coded into the images. Therefore, there was no possibility that the GAN replicates private patient data through this route. However, other radiological images such as ultrasound datasets often have the patient data hard-coded into the pixel values. A failsafe procedure that prevents the replication of patient identifying text is the removal of any pixel-hardcoded information before the images are used for training of the GAN. For this, automated optical character recognition (OCR) to redact burnt-in text on images contained in DICOM instances could be used.

Alternative generative models

Alternative network architectures to the one chosen in this manuscript exist. Among those are flow models [38] and autoregressive models [39]. We have chosen the progressive growing GAN architecture as it is less demanding w.r.t. computational resources: both flow and autoregressive models demand resources that are currently outside of the possibility of most hospitals. However, with ever-increasing computational power, these approaches might equally well be suited for the creation of a database of anonymous radiological images and might need to be evaluated in the future.

Adversarial robustness

The brittleness of deep learning models introduced by adversarial attacks has attracted significant attention in the scientific community. Particularly, an adversarial threat is less detectable in the standard setting of a federated classifier as the server has no visibility into training data. A remedy to this issue is the use of adversarial training in which the model parameters were optimized by performing gradient descent over adversarial examples. Using the Stanford CheXpert dataset, we trained a robust classifier against an iterative fast gradient sign (FGSM) adversary with $l_\infty$ norm constrain. We set the maximum perturbation ($\varepsilon$) to 0.005, the step-size to 0.0025, and the number of iterations to 10 for generating adversarial examples. When applying $\varepsilon = 0.01$ attack on the test samples, we observed a significant amount of ROC-AUC decrease in standard models (2nd row in Fig. S3) whereas robust models were less perturbed (1st row in Fig. S3).

Fig. S3. Performance of our adversarially trained classifier against iterative FGSM adversaries. 1st row: the CheXpert classifier was trained against $\varepsilon = 0.005$ perturbation imperceptible to humans. 2nd row: a classifier was trained on clean samples from the CheXpert dataset. In the test phase, we used either clean ($\varepsilon = 0$) or perturbed ($\varepsilon = 0.01$) samples to quantify the influence of adversarial attacks. We observed that our robust model was resistant to $\varepsilon = 0.01$ attacks, however, the attacks were able to fool the standard model.
Fig. S4. Generated radiographs with a spatial resolution of $256 \times 256$ and $1024 \times 1024$. A: Real thoracic radiographs from the database with labels healthy, cardiomegaly, effusion, and hernia are displayed in the upper row. Synthesized radiographs as generated by the generator with the same labels are shown in the bottom row. Note that consistent changes are seen in both the real and synthesized radiographs: enlargement of the heart (2nd column), opacification of the left lung (3rd column) and mediastinal enlargement/shift (4th column). B: Randomly generated radiographs at a spatial resolution of $256 \times 256$ (left column) and their three nearest neighbors (NN) counterparts (i.e. most similar radiographs) in the real dataset in columns two to four. As implied by our network architecture, no duplication of an existing radiograph was found neither by visual inspection nor numerically by a high similarity measure. C: Quantitative pathology transitions. Transition from healthy to diseased generated radiographs (top row: cardiomegaly, bottom row: effusion), when the latent vector that characterizes a specific pseudo-patient is held fixed and the entry in the one-hot encoding vector for that specific disease is increased from 0 (1st column), via 0.3 (2nd column) and 0.6 (3rd column) to 1.0 (4th column). Both cardiomegaly and effusion are realistically depicted: the heart shape grows, while the increasing opacification in the left lower lung reflects the accumulating amount of fluid which is less transparent for X-rays than lung tissue. Note that the left lung corresponds to the right image side as radiographs are depicted with the patient facing the reader. Cardiomegaly was marked by red arrows whereas effusion was marked by blue asterisks. Corresponding animations for both of these as well as the remaining 12 diseases are given as an online supplement. D: Handpicked examples where the generation of pneumonia by the generator seemed to be difficult. 1st row: simulated pneumonia examples from the NIH-GAN. 2nd row: simulated pneumonia examples from the Stanford-GAN. Note that the pneumonia features in NIH-GAN generated samples are less realistic due to the limited number of pneumonia examples in the training set. E: Generated high-resolution images with a resolution of $1024 \times 1024$ look real initially. However, details reveal the synthesized origin of the data. From left to right: dense line in unphysiological orientation adjacent to mediastinum (the 1st column), a bizarre configuration of the humeral head (the 2nd column), an unphysiological configuration of right pulmonary vessels (the 3rd column) and periodic, wavelike pattern superimposed on the lung parenchyma (the 4th column). Note, that those artifacts do not appear consistently and even though the GAN has difficulties reproducing all details of a radiograph collectively, it often generates them in a realistic matter: the humeral head (artifact in the 2nd column) is correctly depicted in the 3rd column, while the aberrant vessel structure of the right hilus (artifact in the 3rd column) is correctly depicted in the 2nd column.
Domain Adaption using a Cycle-GAN

Each hospital has its scan protocol and different vendor and scan parameters. Normalization of each data set is of immense importance to get consistent results for the trained classifiers. In the experiments described in section Performance of Classifiers Trained on Synthesized Radiographs, we employed a mere simple mean to account for this by cutting the X-rays to a standard size and normalizing the dynamic range to (0, 255). However, not all differences between scan protocols can thus be caught (e.g. a different geometry setup of the X-ray tube during image acquisition). We, therefore, performed additional experiments, in which we employed Cycle-GANs [40]. Cycle-GANs are a powerful tool to transfer the “style” of one set of images (hospital A) to the second set of images (hospital B) without giving detailed instructions on what properties have to be changed. We trained a Cycle-GAN to transfer the general appearance of the Stanford X-rays to the NIH X-rays. The results of this training are shown in Fig. S5 The visual impression indeed confirms that the general appearance of the Stanford X-rays can be transferred to that of the NIH images and that this is a viable method that warrants future research into this direction.

![Diagram](image.png)

Fig. S5. Radiographs from Stanford dataset can be transferred into a NIH style by using a Cycle-GAN.

Table S2: Performance measures of participants in the GAN real/synthesized test. Note that the reading conditions for Radiologist (Rad) 4 were different from the remaining readers: Radiologist 4 performed the test at a dedicated radiological working station and was allowed to first analyze the high-resolutions (1024×1024) radiographs for telltale signs of the GAN before going back to the low-resolution examples.

|        | 256×256 | 512×512 | 1024×1024 |
|--------|---------|---------|-----------|
|        | Accu. % | Sens. % | Spec. %  | Accu. % | Sens. % | Spec. %  | Accu. % | Sens. % | Spec. %  |
| CV 1   | 53      | 52      | 54       | 90      | 94      | 86       | 86      | 88      | 84       |
| CV 2   | 62      | 70      | 54       | 62      | 74      | 50       | 77      | 80      | 74       |
| CV 3   | 66      | 76      | 56       | 51      | 78      | 24       | 84      | 96      | 72       |
| Rad 1  | 45      | 50      | 40       | 66      | 66      | 66       | 69      | 74      | 64       |
| Rad 2  | 58      | 66      | 50       | 53      | 68      | 38       | 68      | 76      | 60       |
| Rad 3  | 50      | 48      | 52       | 77      | 90      | 64       | 96      | 98      | 94       |
| Rad 4  | 86      | 76      | 96       | 96      | 92      | 92       | 92      | 92      | 92       |

Abbreviations: CV, computer vision experts; Rad, radiologists with more than 5 years experiences; Aver., average; Accu., accuracy; Sens., sensitivity; Spec., specificity.
The seven overlapping labels between the NIH and Stanford dataset are listed in the table above. Note that pathological radiographs seem to be more common in the Stanford CheXpert dataset due to the differences in the labeling process [33]. Abbreviations: NIH, National Institutes of Health Clinical Center; SD, standard deviation.

Table S4: p-values for quantities in Fig. 2 (left) and Fig. 3 (right).

Table S5: Generator and discriminator that we use to generate 1024×1024 radiographs. The total number of trainable parameters for the generator and the discriminator are 23.2M and 23.1M, respectively.
Algorithm S1 Federated Averaging GAN. We use default values of $K = 20$, $C = 0.1$, $T = 5,000$, $b = 32$, $E = 10$, $\lambda = 10$, $n_{\text{discriminator}} = 10$, and $\alpha = 0.0001$.

1: **Aggregation server executes:**
2: **Require:** Total number of $K$ local clients are indexed by $i$; a fraction of clients $C$; total number of iterations $T$.
3: initialize global generator $w_{G,0}$
4: initialize global discriminator $w_{D,0}$
5: **for** each round $t$ from 0 to $T$ do
6: $m \leftarrow \max(C \cdot K, 1)$
7: $S_t \leftarrow$ (random set of $m$ clients, $n$ data items)
8: **for** each local client $i \in S_t$ in parallel do
9: $w_{i,G,t+1}, w_{i,D,t+1} \leftarrow \text{LocalUpdate}(k, w_{i,G,t}, w_{i,D,t})$
10: **end for**
11: $w_{G,t+1} \leftarrow \sum_i \frac{n_i}{n} w_{i,G,t+1}$
12: $w_{D,t+1} \leftarrow \sum_i \frac{n_i}{n} w_{i,D,t+1}$
13: **end for**

1: **LocalUpdate**($i, w_{D}, w_{G}$): Run on client $i$
2: **Require:** The gradient penalty coefficient $\lambda$; the number of discriminator iterations per generator iteration $n_{\text{discriminator}}$; the local batch size $b$; the number of local generator update iterations $E$; local learning rate $\alpha$.
3: **Require:** initial discriminator parameters $w_{D,0}$, initial generator parameters $w_{G,0}$.
4: **for** each local iteration $n$ from 1 to $E$ do
5: Sample a batch of latent variables $\{z^{(j)}\}_{j=1}^b \sim p(z)$ and class labels $\{c^{(j)}\}_{j=1}^b \sim p_c$.
6: $w_{G} \leftarrow w_{G} - \alpha \nabla_{w_{G}} \sum_{j=1}^{b} (-D_{w_{D}}(G_{w_{G}}(z,c)) + L_{\text{synthesized}})$
7: **for** $i = 1, \ldots, n_{\text{discriminator}}$ do
8: **for** $j = 1, \ldots, b$ do
9: Sample real data $x \sim p_r$, latent variable $z \sim p(z)$, a random number $\varepsilon \sim U[0,1]$, $c \sim p_c$.
10: $\tilde{x} \leftarrow G_{w_{G}}(z,c)$
11: $\hat{x} \leftarrow \varepsilon x + (1-\varepsilon)\tilde{x}$
12: $L^{(j)} \leftarrow D_{w_{D}}(\hat{x}) - D_{w_{D}}(x) + \lambda (\|\nabla_{x} D_{w_{D}}(\hat{x})\|_2 - 1)^2 + L_{\text{real}}$
13: **end for**
14: $w_{D} \leftarrow w_{D} - \alpha \nabla_{w_{D}} \sum_{j=1}^{b} L^{(j)}$
15: **end for**
16: **end for**
17: return $w_{G}, w_{D}$ to server
REFERENCES AND NOTES

1. S. Bickelhaupt, P. F. Jaeger, F. B. Laun, W. Lederer, H. Daniel, T. A. Kuder, L. Wuesthof, D. Paech, D. Bonekamp, A. Radbruch, S. Delorme, H.-P. Schlemmer, F. H. Steudle, K. H. Maier-Hein, Radiomics based on adapted diffusion kurtosis imaging helps to clarify most mammographic findings suspicious for cancer. *Radiology* **287**, 761–770 (2018).

2. A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).

3. K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: surpassing human-level performance on ImageNet classification, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 7 to 13 December 2015.

4. A. Hosny, C. Parmar, T. P. Coroller, P. Grossmann, R. Zeleznik, A. Kumar, J. Bussink, R. J. Gillies, R. H. Mak, H. J. W. L. Aerts, Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study. *PLOS Med.* **15**, e1002711 (2018).

5. V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros, Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).

6. D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, L. van der Maaten, Exploring the limits of weakly supervised pretraining, in *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 8 to 14 September 2018.

7. M. M. Mello, V. Lieou, S. N. Goodman, Clinical trial participants’ views of the risks and benefits of data sharing. *N. Engl. J. Med.* **378**, 2202–2211 (2018).

8. L. M. Prevedello, S. S. Halabi, G. Shih, C. C. Wu, M. D. Kohli, F. H. Chokshi, B. J. Erickson, J. Kalpathy-Cramer, K. P. Andriole, A. E. Flanders, Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions. *Radiol. Artif. Intell.* **1**, e180031 (2019).
9. J. Xu, F. Wang, Federated learning for healthcare informatics (2019); arXiv:1911.06270.

10. W. Li, F. Milletari, D. Xu, N. Rieke, J. Hancox, W. Zhu, M. Baust, Y. Cheng, S. Ourselin, M. J. Cardoso, A. Feng, Privacy-preserving Federated Brain Tumour Segmentation, in *International Workshop on Machine Learning in Medical Imaging* (Springer, 2019), pp. 133–141.

11. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in *Advances in Neural Information Processing Systems* (Neural Information Processing Systems Foundation, Inc., 2014), pp. 2672–2680.

12. T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation (2017); arXiv:1710.10196.

13. P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, Image-to-image translation with conditional adversarial networks, *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, 21 to 26 July 2017.

14. J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, *Proceedings of the 2017 IEEE International Conference on Computer Vision* (*ICCV*), Venice, Italy, 22 to 29 October 2017.

15. J. M. Wolterink, A. M. Dinkla, M. H. F. Savenije, P. R. Seevinck, C. A. T. van den Berg, I. Işgum, Deep MR to CT synthesis using unpaired data, in *Simulation and Synthesis in Medical Imaging* (Springer, 2017), pp. 14–23.

16. A. Chartsias, T. Joyce, R. Dharmakumar, S. A. Tsaftaris, Adversarial image synthesis for unpaired multi-modal cardiac data, in *Simulation and Synthesis in Medical Imaging* (Springer, 2017), pp. 3–13.

17. Z. Zhang, L. Yang, Y. Zheng, Translating and segmenting multimodal medical volumes with cycle- and shape-consistency generative adversarial network, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, 18 to 22 June 2018.
18. A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks (2015); arXiv:1511.06434.

19. A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier GANs, \textit{Proceedings of the 34th International Conference on Machine Learning-Volume 70} (JMLR.org, 2017), pp. 2642–2651.

20. P. Rui, K. Kang, National Hospital Ambulatory Medical Care Survey: 2015 Emergency Department Summary Tables (2015).
https://www.cdc.gov/nchs/data/nhamcs/web_tables/2015_ed_web_tables.pdf [accessed 16 January 2020].

21. S. Conjeti, A. Katouzian, A. G. Roy, L. Peter, D. Sheet, S. Carlier, A. Laine, N. Navab, Supervised domain adaptation of decision forests: Transfer of models trained in vitro for in vivo intravascular ultrasound tissue characterization. \textit{Med. Image Anal.} \textbf{32}, 1–17 (2016).

22. B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. y Arcas, Communication-Efficient Learning of Deep Networks from Decentralized Data, in \textit{Proceedings of the 20th International Conference on Artificial Intelligence and Statistics}, Fort Lauderdale, FL, USA, 20 to 22 April 2017.

23. A. N. Bhagoji, S. Chakraborty, P. Mittal, S. Calo, Analyzing federated learning through an adversarial lens. arXiv:1811.12470 (2018).

24. D. Stutz, M. Hein, B. Schiele, Confidence-calibrated adversarial training: Generalizing to unseen attacks. arXiv:1910.06259 (2019).

25. S. Augenstein, H. B. McMahan, D. Ramage, S. Ramaswamy, P. Kairouz, M. Chen, R. Mathews, B. A. y Arcas, Generative models for effective ML on private, decentralized datasets. arXiv:1911.06679 (2019).

26. I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville, Improved training of Wasserstein GANs, in \textit{Advances in Neural Information Processing Systems}, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, Eds. (Curran Associates, Inc., 2017), pp. 5767–5777.
27. H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, J. Barfett, Generalization of deep neural networks for chest pathology classification in X-rays using generative adversarial networks, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2018), pp. 990–994.

28. M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, GANs trained by a two time-scale update rule converge to a local nash equilibrium, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2017), pp. 6626–6637.

29. S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, I. S. Kohane, Adversarial attacks on medical machine learning. *Science* **363**, 1287–1289 (2019).

30. Y. Li, Y. Gal, Dropout inference in Bayesian neural networks with alpha-divergences, in *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (JMLR.org, 2017), pp. 2052–2061.

31. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083 (2017).

32. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R. M. Summers, ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017), pp. 2097–2106.

33. J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Hagghoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, A. Y. Ng, CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. arXiv:1901.07031 (2019).

34. E. Perez, F. Strub, H. de Vries, V. Dumoulin, A. Courville, FiLM: Visual reasoning with a general conditioning layer. arXiv:1709.07871 (2017).

35. P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, B. N. Patel, K. W. Yeom, K. Shpanskaya, F. G. Blankenberg, J. Seekins, T. J. Amrhein, D.
A. Mong, S. S. Halabi, E. J. Zucker, A. Y. Ng, M. P. Lungren, Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLOS Med.* **15**, e1002686 (2018).

36. B. Rosner, *Fundamentals of Biostatistics* (Nelson Education, 2015).

37. M. Arjovsky, S. Chintala, L. Bottou, Wasserstein GAN. arXiv:1701.07875 (2017).

38. D. P. Kingma, P. Dhariwal, Glow: Generative flow with invertible 1x1 convolutions, in *Advances in Neural Information Processing Systems* (2018), pp. 10215–10224.

39. T. Salimans, A. Karpathy, X. Chen, D. P. Kingma, PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications. arXiv:1701.05517 (2017).

40. J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, T. Darrell, CyCADA: Cycle-consistent adversarial domain adaptation. arXiv:1711.03213 (2017).