



OPEN

## Image Turing test and its applications on synthetic chest radiographs by using the progressive growing generative adversarial network

Miso Jang<sup>1,2</sup>, Hyun-jin Bae<sup>3</sup>, Minjee Kim<sup>3</sup>, Seo Young Park<sup>4</sup>, A-yeon Son<sup>5</sup>, Se Jin Choi<sup>5</sup>, Jooae Choe<sup>5</sup>, Hye Young Choi<sup>5</sup>, Hye Jeon Hwang<sup>5</sup>, Han Na Noh<sup>6</sup>, Joon Beom Seo<sup>5</sup>, Sang Min Lee<sup>5</sup>✉ & Namkug Kim<sup>5,7</sup>✉

The generative adversarial network (GAN) is a promising deep learning method for generating images. We evaluated the generation of highly realistic and high-resolution chest radiographs (CXRs) using progressive growing GAN (PGGAN). We trained two PGGAN models using normal and abnormal CXRs, solely relying on normal CXRs to demonstrate the quality of synthetic CXRs that were 1000 × 1000 pixels in size. Image Turing tests were evaluated by six radiologists in a binary fashion using two independent validation sets to judge the authenticity of each CXR, with a mean accuracy of 67.42% and 69.92% for the first and second trials, respectively. Inter-reader agreements were poor for the first ( $\kappa = 0.10$ ) and second ( $\kappa = 0.14$ ) Turing tests. Additionally, a convolutional neural network (CNN) was used to classify normal or abnormal CXR using only real images and/or synthetic images mixed datasets. The accuracy of the CNN model trained using a mixed dataset of synthetic and real data was 93.3%, compared to 91.0% for the model built using only the real data. PGGAN was able to generate CXRs that were identical to real CXRs, and this showed promise to overcome imbalances between classes in CNN training.

The recent rapid development of artificial intelligence in medicine can be primarily attributed to advances in algorithms, the computing power of graphics processing units (GPUs), and the generation of healthcare bigdata<sup>1</sup>. The number of studies<sup>2</sup> applying deep learning techniques to medical imaging has increased significantly in recent years. Specifically, applications include lesion detection, image segmentation, classification, and image reconstruction<sup>3,4</sup>. However, there are several limitations, including a strong imbalanced dataset for specific diseases, expensive labels, and legal and ethical issues regarding patients' privacy concerns, in implementing deep learning techniques in medical imaging<sup>3,5</sup>. Given the rarity of some diseases and high dependency on vast amounts of good-quality labelled data, which requires considerable time input from experts and correspondingly high expenses, many medical datasets suffer from class imbalance and insufficient labeling<sup>6,7</sup>. Moreover, most supervised learning models exhibit optimal performance for specific tasks in narrow clinical settings, which in turn indicates "weak" artificial intelligence. However, the technique can be ineffective with limited coverage when used in real clinical settings, such as emergency departments, because of the diversity of clinical situations and imbalanced diseases. A potential approach for overcoming these issues involves applying unsupervised learning models for many tasks, including generating synthetic dataset and, anomaly detection.

<sup>1</sup>Department of Medicine, University of Ulsan College of Medicine, Asan Medical Center, Seoul, Republic of Korea. <sup>2</sup>Department of Biomedical Engineering, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea. <sup>3</sup>Promedius Inc., Seoul, Republic of Korea. <sup>4</sup>Department of Statistics and Data Science, Korea National Open University, Seoul, Republic of Korea. <sup>5</sup>Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine and Asan Medical Center, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Republic of Korea. <sup>6</sup>Department of Health Screening and Promotion Center, Asan Medical Center, Seoul, Republic of Korea. <sup>7</sup>Department of Convergence Medicine, University of Ulsan College of Medicine, Asan Medical Center, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Republic of Korea. ✉email: asellion@hanmail.net; namkugkim@gmail.com

Generative adversarial networks (GANs) are effective non-supervised learning method<sup>8,9</sup> that have gained popularity for their high performance in creating realistic images<sup>10</sup>. The generation of realistic medical images can lead to new opportunities for solving the problems involving class imbalance, data augmentation, and patients' privacy concerns<sup>5,8,9,11</sup>. GANs have been utilized in various medical imaging tasks to resolve the aforementioned problems with promising results<sup>12–18</sup>. Additionally, in some studies, attempts have been made to boost the performance of lesion detection by subtracting the most similar and GAN-generated normal image from a pathological real image<sup>19,20</sup>. Despite these promising results, the generation of synthetic medical images that are realistic to the maximum extent is a prerequisite for applying GANs to actual clinical settings. However, there is a scarcity of studies in which the evaluation of the perceived “realism” of GAN-generated medical images by radiologists is assessed<sup>21–23</sup>. Moreover, there have been no such studies focusing on the high-resolution chest radiographs (CXRs) using GANs. Recently, a progressive growing GAN (PGGAN)<sup>24</sup> was suggested in computer vision<sup>12,21,25–29</sup>. In some studies, a PGGAN model showed that synthetic body computed tomography images 512 × 512 pixels in size were highly realistic<sup>21</sup> and the generated cephalogram X-ray images could be helpful for training convolutional neural networks (CNN) in imbalanced dataset<sup>12</sup>.

Therefore, we used a PGGAN to generate highly realistic CXRs and performed image Turing tests and downstream tasks for classifying normal and abnormal CXRs. The classification was used to augment the realism of the synthetic images for evaluating the model. The contributions of our study are as follows:

1. We proposed a training method for producing highly realistic and high-resolution (1000 × 1000) synthetic CXR images with GAN.
2. Six thoracic radiologists evaluated these synthetic CXR images by visual Turing test.
3. The synthetic datasets may be used to train a downstream task to classify normal or abnormal CXR images without decreasing accuracy, which in turn can be used as an augmentation technique to overcome data imbalances.

## Results

**Results of the visual Turing test.** Table 1 summarizes the results of the first Turing test with ABN-PGGAN by the six readers. The mean accuracy, sensitivity, and specificity of the six readers were 67.4%, 57.3%, and 77.5%, respectively. Table 2 summarizes the results of the second Turing test with NOR-PGGAN by the six readers. The mean accuracy, sensitivity, and specificity of the six readers were 69.9%, 65.2%, and 74.7%, respectively. Inter-reader agreements of six radiologists were poor for first and second image Turing tests wherein the Kappa values (95% CI) were 0.10 (0.07–0.14) and 0.14 (0.10–0.18), respectively.

As shown in Table 3, more experienced readers exhibited a higher probability of guessing the correct answer in the first Turing test dataset. However, there was no statistical difference between two groups in the second test dataset. As shown in Table 4, the reading time was higher when readers answered correctly in only synthetic images irrespective of whether the correction for the reader effect was considered. However, it was observed that the reading time was shorter when the reader effect was corrected in only real images.

As shown in Fig. 1, six synthetic images were judged as synthetic by all the readers and one synthetic image was judged as real by all the readers.

Reader	Accuracy (% , 95% CI)	Sensitivity (% , 95% CI)	Specificity (% , 95% CI)
R01	46.00 (38.95, 53.17)	59.00 (48.71, 68.74)	33.00 (23.92, 43.12)
R02	56.50 (49.33, 63.48)	76.00 (66.43, 83.98)	37.00 (27.56, 47.24)
R03	48.00 (40.90, 55.16)	59.00 (48.71, 68.74)	37.00 (27.56, 47.24)
R04	89.50 (84.40, 93.38)	90.00 (82.38, 95.10)	89.00 (81.17, 94.38)
R05	73.50 (66.81, 79.48)	86.00 (77.63, 92.13)	61.00 (50.73, 70.60)
R06	91.00 (86.15, 94.58)	95.00 (88.72, 98.36)	87.00 (78.80, 92.89)

**Table 1.** Performance of the first image Turing test.

Reader	Accuracy (% , 95% CI)	Sensitivity (% , 95% CI)	Specificity (% , 95% CI)
R01	52.00 (44.84, 59.10)	63.00 (52.76, 72.44)	41.00 (31.26, 51.29)
R02	52.50 (45.34, 59.59)	59.00 (48.71, 68.74)	46.00 (35.98, 56.26)
R03	90.50 (85.56, 94.18)	96.00 (90.07, 98.90)	85.00 (76.47, 91.35)
R04	83.00 (77.06, 87.93)	67.00 (56.88, 76.08)	99.00 (94.55, 99.97)
R05	50.50 (43.36, 57.63)	67.00 (56.88, 76.08)	34.00 (24.82, 44.15)
R06	91.00 (86.15, 94.58)	96.00 (90.07, 98.90)	86.00 (77.63, 92.13)

**Table 2.** Performance of the second image Turing test.

Datasets	Applied data	Odds ratio of correct answer (95% CI)	p-value
First Turing test dataset	All	6.65 (3.29, 13.47)	< 0.001
	Only real images	5.30 (2.75, 10.22)	< 0.001
	Only synthetic images	7.69 (3.29, 17.98)	< 0.001
Second Turing test dataset	All	1.67 (0.33, 8.42)	0.54
	Only real images	1.22 (0.17, 8.90)	0.85
	Only synthetic images	3.95 (0.31, 51.34)	0.29

**Table 3.** Mixed effect logistic regression model for correcting answers in the image Turing test. Regression model: log of correct answer ~ expert or novice + image random effect + reader random effect.

Datasets	Applied data	Regression model 1		Regression model 2	
		Beta coefficient of reading time (95% CI)	p-value	Beta coefficient of reading time (95% CI)	p-value
First Turing test dataset	All	- 0.027 (- 0.088, 0.033)	0.376	0.202 (0.125, 0.278)	< 0.001
	Only real images	- 0.196 (- 0.283, - 0.109)	< 0.001	0.105 (- 0.017, 0.227)	0.091
	Only fake images	0.090 (0.003, 0.177)	0.044	0.277 (0.175, 0.379)	< 0.001
Second Turing test dataset	All	- 0.052 (- 0.131, 0.027)	0.195	0.074 (- 0.002, 0.150)	0.058
	Only real images	- 0.003 (- 0.116, 0.111)	0.962	0.127 (0.015, 0.238)	0.027
	Only fake images	- 0.024 (- 0.137, 0.089)	0.674	0.020 (- 0.084, 0.124)	0.706

**Table 4.** Mixed effect logistic regression model for reading time in the image Turing test. Regression model 1:  $\log(\text{reading time} + 1) \sim \text{correct} (y/n) + \text{image random effect} + \text{reader random effect}$ . Regression model 2:  $\log(\text{reading time} + 1) \sim \text{correct} (y/n) + \text{image random effect}$ .

**Normal probability scores in real and synthetic images and Grad-CAM.** Figure 2 shows the normal probability scores of the known CNN classifier with real and synthetic images in the abnormal dataset (a) and normal dataset (b). Additionally, the Grad-CAMs of synthetic images that decisions of all readers were “synthesized” are shown in the Fig. 3. The Grad-CAMs highlighted abnormal lung lesions in three synthetic images, which were not considered real by all the readers.

**Performances of models using real and synthetic mixed datasets..** Figure 4 shows the confusion matrices of the two models. The AUROCs of the same real test set were 0.96 and 0.98 for the models trained using the real and synthetic mixed datasets, respectively. For the model trained using the real dataset, the overall accuracy was 91.0%, sensitivity was 87.0%, and specificity was 95.0%, while for the model trained using the synthetic mixed dataset, the overall accuracy was 93.3%, sensitivity was 90.5%, and specificity was 95.6%. The results of supplemental experiments, which were performed with more than 6000 images of training data, were similar. The AUROCs of the same real test set were 0.98 and 0.98 for the models trained using the real and synthetic mixed datasets, respectively. The confusion matrices are presented in supplement Fig. S1.

## Discussion

In this study, we demonstrated the generation of high-resolution CXRs using PGGAN and two visual Turing tests of synthetic and real CXRs of abnormal and normal patients performed by six readers with different experience levels. The test results indicate that readers, excluding most experienced radiologists exhibited different performances in terms of sensitivity and specificity, with extremely low Kappa values in the two Turing tests<sup>30</sup>. Therefore, the radiologists were not able to distinguish synthetic CXRs from real CXRs. The first step in applying GAN to the development of models applicable to medical fields depends on the generation of highly realistic and high-resolution images. At the beginning of the study, we questioned whether realistic images  $1024 \times 1024$  pixels in size can be generated by ABN-PGGAN and NOR-PGGAN. Although several reports have investigated the benefits of GANs in various medical imaging fields<sup>7,9,31,32</sup>, evaluation of the realistic nature of GAN-generated synthetic images via a comparison of normal and abnormal CXRs has not been reported to the best of the author’s knowledge. Therefore, the realistic nature of synthetic images has not been well validated.

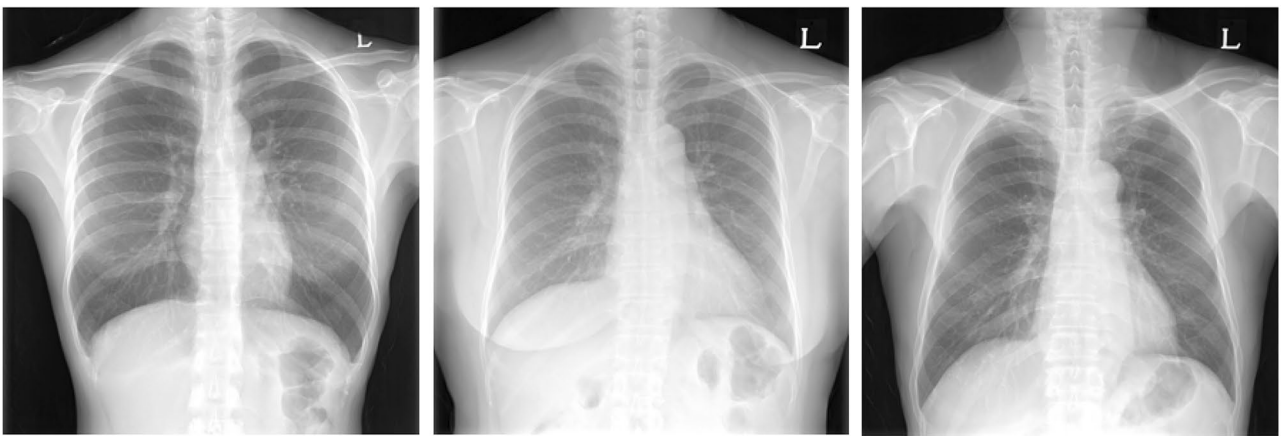
There were different points between the two Turing tests. Odds ratios in the first Turing test statistically exceeded those in the second Turing test in correcting answers, specifically when only synthetic images were used, thereby indicating that synthetic images generated from ABN-PGGAN were less real. Nevertheless, the results of beta coefficient on reading time of abnormal images in the first Turing test suggested it was difficult to judge synthetic images as synthetic. Furthermore, more time was required to obtain the correct answer in the first Turing test when compared to that in the second Turing test.

In this study, radiologists were not able to distinguish normal synthetic CXRs from real CXRs, irrespective of their expertise; however, more experienced radiologists were able to differentiate abnormal synthetic chest CXRs from abnormal synthetic CXRs, whilst ABN-PGGAN trained more CXRs than NOR-PGGAN. However, the

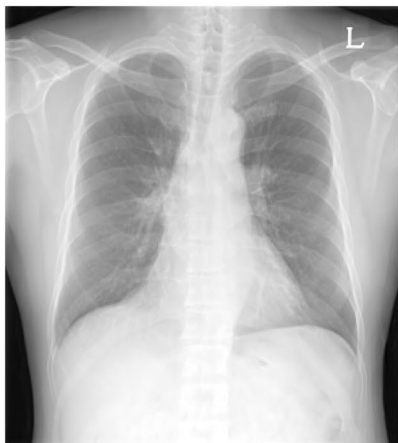
(a)



(b)

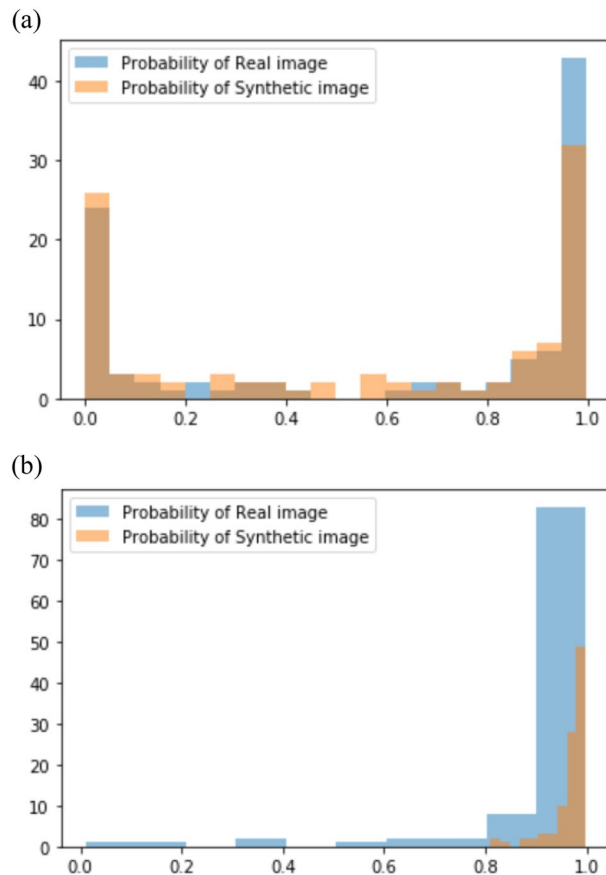


(c)



**Figure 1.** Several cases of the Turing test. **(a)** Synthetic images wherein the decisions of all the readers are synthetic in the first Turing test. **(b)** Synthetic images wherein the decisions of all the readers are synthetic in the second Turing test. **(c)** Synthetic images wherein the decisions of all the readers are real in the first Turing test.

range of abnormal CXRs is suspected to vary significantly when trained by GAN. Therefore, more experienced readers can detect artificial findings with abnormal lesions in synthetic images. In the analysis of reading time, when considering reader effect, reading time was short when readers determined real images as real in the first Turing test dataset. Conversely, NOR-PGGAN trained normal CXRs features relatively well. In the present study, inter-reader agreement was poor for the entire image set consisting of synthetic and real images, which indicates that it is difficult to distinguish between synthetic and real images (i.e., identifying realistic synthetic images).



**Figure 2.** Probability scores of the CNN classifier in Turing test datasets. **(a)** First image Turing test dataset. **(b)** Second image Turing test dataset.

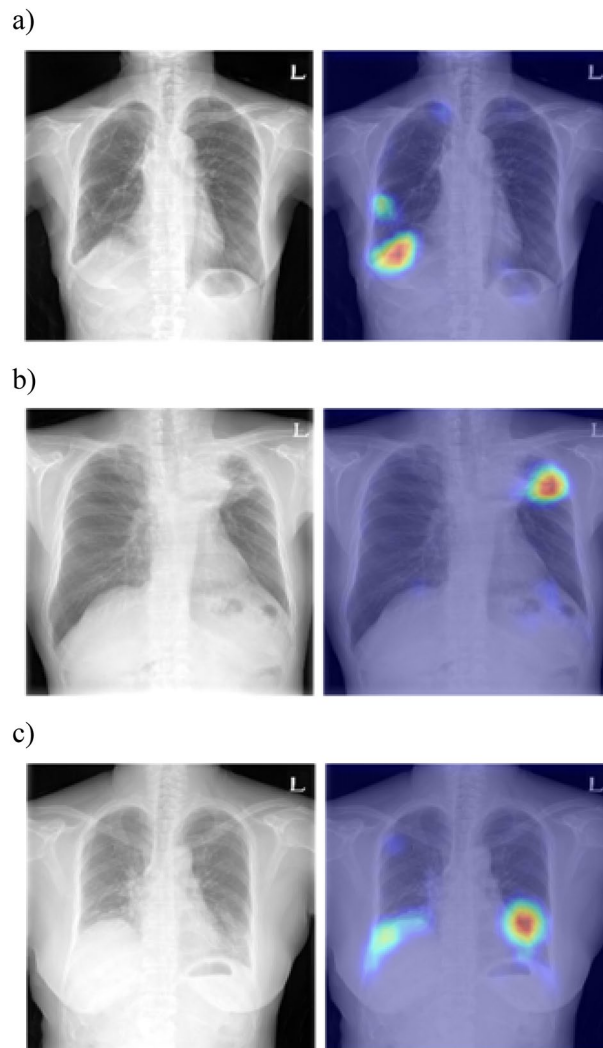
In CXRs, several lines and stripes are observed from the anatomic structures in the chest. Radiologists are trained to recognize their normal and abnormal appearances of lines and stripes on CXRs. In certain synthetic images, some discontinuities are present in the lines and stripes including interface of lung, vessels, and ribs, this tends to be more pronounced in abnormal synthetic images. By competing for the generator and discriminator, the GAN converges to an appropriate local minimum, and the generator produces a realistic image with a given latent vector<sup>33</sup>. The latent space of GANs is the result of learning the mapping from a latent distribution to the real images via adversarial training<sup>34</sup>. The latent spaces of ABN-PGGAN included pathologic features and less normal features when compared with those of NOR-PGGAN.

From the CNN viewpoint, the CNN was unable to distinguish between synthetic and real images, with the synthetic images considered more appropriate for training based on the purpose. All normal synthetic images generally exhibited normal probability scores of close to 1. The borderline of normal and abnormal CXRs was not evident, and thus the ABN-PGGAN trained dataset included normal CXRs. Furthermore, the visual Turing test of the abnormal dataset included normal real CXRs. The probability scores in the visual Turing test of abnormal dataset varied. However, the Grad-CAMs of abnormal synthetic images, which all readers determined as fake, highlighted pathologic lesions as opposed to artificial regions.

Furthermore, synthetic images were used to develop the classification task. The effects of class imbalance on the performance of CNNs were examined, where it was determined to decrease the performance of CNNs<sup>35,36</sup>. The CNN model was trained on two datasets, namely only real and mixed synthetic datasets. There is always a shortage of disease data in clinical situation, meaning that the mixed synthetic dataset included abnormal synthetic CXRs. The classification performance of two models were comparable and synthetic images improved the performance of CNN classifiers. In general, 1000 images were used as a rough criterion for training a model per class. In a recent study, approximately 50,000 images per class was deemed necessary for acceptable performance<sup>37</sup>. In our current study, enough images were generated for further studies. Additionally, disentangling the latent space of GAN leads to controllable abnormal CXR features<sup>38</sup>, indicating that synthetic images generated by GAN can lead to a solution for training CNN in rare diseases. Since a method for evaluating the quality of the generated data has not yet been established, when synthetic data alone are used, performance may be lower than when using real data.

The potential clinical applicability of useful GANs can be image reconstruction and denoising<sup>16,39</sup>, translation between different radiologic modalities<sup>14,40,41</sup>, and anomaly detection<sup>19,29</sup>. Detecting abnormalities is predicated on learning the probability distribution of normal training data, unlike other GAN-based tasks. Any image data





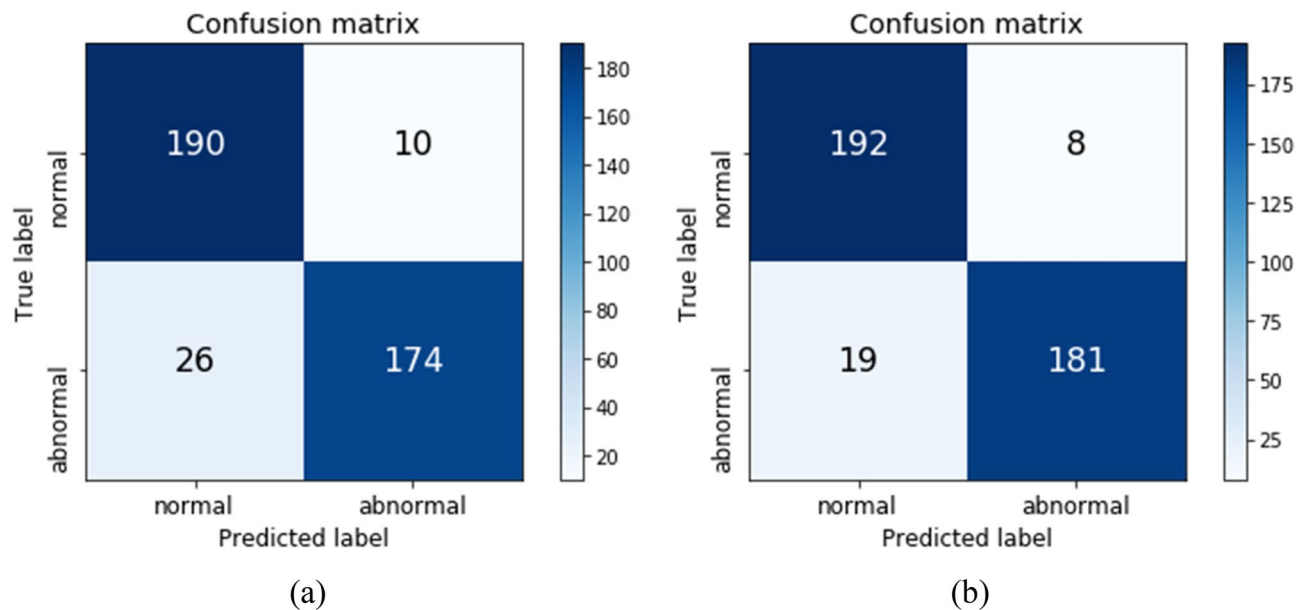
**Figure 3.** Grad-CAMs of synthetic images in the first Turing test dataset wherein the decisions of all readers are synthetic.

that deviate from this distribution are regarded as abnormal. In daily clinical situations, diagnostic images are clinically acquired for patients with a variety of diseases. If GAN models can filter out normal chest x-rays well, doctors can focus more on chest images with abnormal findings. However, since the normal range may vary depending on the clinical situation and the age of the patient, many additional studies are needed to actually utilize it.

Although several evaluation metrics were used to measure the quality of generated images using GAN models, such as Inception Score (IS)<sup>42</sup> or Fréchet Inception Distance (FID)<sup>43</sup>, the metrics did not fully explain the extent of failure or success of the generated synthetic images in medical images. This is because the IS and FID metrics solely focus on the distributions of synthetic images using a CNN network (Inception V3), as well as ignore the semantics of the images. Therefore, it was concluded that, to date, a visual Turing test by human experts is the only viable solution to fairly evaluate the quality of generated synthetic medical images.

This study included several limitations. First, for ABN-GAN training, the ratio of abnormal and normal CXRs was approximately 8:2, which was emphatically determined. However, it is necessary to further train the model to generate more realistic abnormal images. Second, there are various types of GAN architectures including StyleGAN2<sup>44</sup>, StyleGAN2-Adaptive Discriminator Augmentation (ADA)<sup>45</sup>, and score-based generative model<sup>46</sup>, which can be evaluated in this manner. We have used StyleGAN2-ADA model<sup>45</sup> for training with the learning rate of 0.002 and  $r1\_gamma$  of 26.2144 without mirror augmentation. The details of generated CXRs seem to fall off in minute parts such as ribs and pulmonary blood vessels until now. Further experiments and research are currently ongoing. Third, given the need to further examine the latent space of GAN proceeds, additional research must focus on determining the amount of and range of training data. Finally, the number of readers of the visual Turing test was relatively low, meaning that it is not possible to generalize factors related to the experience of readers. Hence, this should be examined further.

In conclusion, ABN- and NOR-PGGAN models were able to generate highly realistic and high-resolution CXRs that were validated by radiologists with different levels of expertise and a previously trained CNN classifier.



**Figure 4.** Confusion matrices of the two models. (a) Performance of the trained model using only the real dataset. (b) Performance of the trained model using the synthetic mixed dataset.

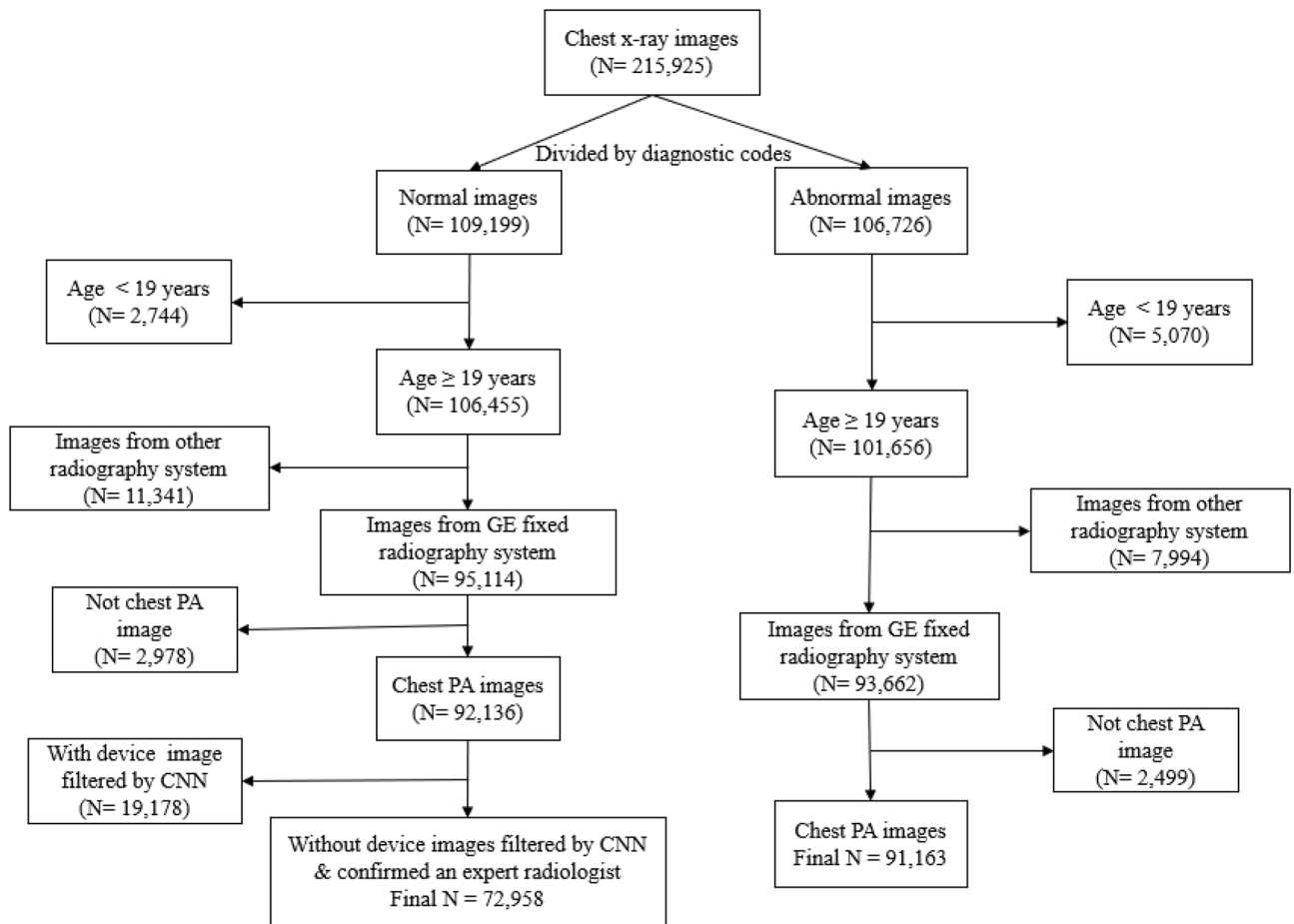
The generation power of NOR-PGGAN was considered to exceed that of ABN-PGGAN, indicating that GAN generated abnormal images require more data. Nevertheless, the proposed models imply significant value for the development of CNN models using GAN-based data augmentation. Further research will also be able to show the utilization of GAN-generated data for developing anomaly detection for avoiding expensive labels, overcoming strong imbalanced datasets for rare diseases, and avoiding legal and ethical issues related to privacy concerns.

## Materials and methods

**Ethical approval.** This retrospective study was conducted according to the principles of the Declaration of Helsinki and in accordance with current scientific guidelines. The study protocol was approved by the Institutional Review Board Committee of Asan Medical Center (AMC), Seoul, Korea (No. 2019-0321). The requirement for informed patient consent was waived by the Institutional Review Board Committee of Asan Medical Center because of the retrospective nature of this study.

**Data collection.** A large number of chest X-ray images were collected in the department of radiology of AMC between January 2011 and December 2018. The original dataset was cleaned as illustrated in Fig. 5. Normal and abnormal chest x-ray images were classified using diagnostic codes. This study was conducted on chest x-rays of adults aged 19 years and older, and the images were included solely from fixed radiography systems of GE Healthcare. The age range generally agreed upon as an adult was selected because body parts included in chest X-ray images may vary due to differences in body size according to age, and differences in prevalent diseases also exist in children and adolescents. This ensured the control of domain shift due to various types of x-ray equipment. Furthermore, CXR posteroanterior (PA) images were acquired by removing other chest images because the original dataset contained various types of chest images, such as chest lateral images and chest decubitus images, which can be only differentiated by using DICOM fields. For selecting a normal CXR PA image, classified normal CXRs with many devices or wires, such as central venous catheters and ECG lines, were excluded from the normal group via the simple convolutional neural network (CNN) classifier. This was further confirmed by an expert radiologist. Finally, the number of normal group chest images was 72,958 and abnormal group chest images was 91,163. The DICOM files of CXRs were converted into  $1024 \times 1024$  pixel-sized 8-bit PNG format with normal or abnormal labels.

**Training PGGAN models.** PGGAN is a GAN variant and consists of two networks including a generator and discriminator. The PGGAN model was selected to generate high resolution CXRs because this model exhibits better performance in reconstructing a global structure and fine details with high resolution when compared with other GAN variant models<sup>47–49</sup>. Furthermore, PGGAN learns to generate images starting from a low resolution of  $4 \times 4$  pixels to a high resolution of  $1024 \times 1024$  pixels, by progressively growing generator and discriminator networks<sup>24</sup>. The general characteristics of the training images are trained through progressive learning and detail-oriented characteristics are trained in addition to layer growth. The output of the lower resolution layer has an impact on the high-resolution output due to the fade-in type used while raising resolution. PGGAN gradually creates a high-resolution image from a large-scale low-resolution image while considering the learning outcome of a previous layer. A publicly available official website of PGGAN was implemented using TensorFlow in Python (Tensorflow-gpu 1.6.0, Python 3.4.0). GAN training is defined by a game theory in which



**Figure 5.** Overview of collection of normal and abnormal datasets collection.

two players compete against each other. The generator network learns to map a noise to the input space, and the discriminator network learns to distinguish between the generated and true samples. Formally, the loss function is defined by minimax objective:

$$\min_G \max_D \mathbb{E}_{x \sim \mathbb{P}_r} [\log D(x)] + \mathbb{E}_{\tilde{x} \sim \mathbb{P}_g} [\log(1 - D(\tilde{x}))]$$

where  $\mathbb{P}_r$  is the data distribution of images (real),  $\mathbb{P}_g$  is the model distribution implicitly defined by  $\tilde{x} = G(z)$ ,  $z \sim p(z)$  ( $p(z)$  is Gaussian distribution), and  $\mathbb{P}_{\tilde{x}}$  is defined by uniformly sampled along straight lines between point pairs sampled from  $\mathbb{P}_r$  and  $\mathbb{P}_g$ .

PGGAN use the improved Wasserstein GAN loss<sup>50</sup> (WGAN-GP loss), which perform better than Wasserstein GAN (WGAN)<sup>51</sup> by virtue of gradient penalty.

In this study, we selected PGGAN with the improved Wasserstein GAN (WGAN-GP) loss, as it stabilizes the training sufficiently to synthesize high-resolution images. The equation of WGAN-GP loss is defined as:

$$L = \mathbb{E}_{\tilde{x} \sim \mathbb{P}_g} [D(\tilde{x})] - \mathbb{E}_{x \sim \mathbb{P}_r} [D(x)] + \lambda \mathbb{E}_{\tilde{x} \sim \mathbb{P}_{\tilde{x}}} \left[ \left( \|\nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1 \right)^2 \right]$$

A value of  $\lambda = 10.0$  was used in the experiments.

In this study, the PGGAN model was trained with 91,163 abnormal CXRs and randomly sampled 20,000 normal CXRs as abnormal-PGGAN (ABN-PGGAN) for training spectrum from normal to abnormal CXRs. Because there is a region in which the boundary between normal and abnormal cannot be clearly divided. Furthermore, the PGGAN model was trained solely with 72,958 normal CXRs as normal-PGGAN (NOR-PGGAN) as a control study for generating normal CXRs. The variations in normal CXRs were regarded as smaller than those of abnormal CXRs. ABN-PGGAN was trained for 130 epochs and required approximately 12.2 days with two Nvidia Titan RTX GPUs, and NOR-PGGAN was trained for 160 epochs, which required approximately 12.5 days with two Nvidia P40 GPUs. Additionally, two distinct sets of 50,000 synthetic CXR images were generated by using the trained generators of ABN-PGGAN and NOR-PGGAN.

**Image Turing test.** The image Turing test was conducted twice to assess the realistic nature of synthetic CXRs, specifically, 400 CXRs were selected for the test. Fifty percent of these CXRs were randomly selected



from the two sets of 50,000 synthesized images while the other fifty percent of the images were real images that were randomly selected from the training set. The first set of Turing test images consisted of 100 real normal and abnormal chest images, randomly sampled from ABN-PAGGAN trained set, and 100 synthetic images randomly sampled from 50,000 images generated by ABN-PGGAN. The second set of Turing test images consisted of 100 real normal chest images, randomly sampled from NOR-PAGGAN trained set, and 100 synthetic images randomly sampled from 50,000 images generated by NOR-PGGAN.

To avoid selection bias, not all the synthetic images were individually selected by the researchers. The image Turing test was conducted with six readers (radiology residents and four thoracic radiologists) by displaying images one-by-one via a web-based interface (supplement Fig. S5). The readers comprised of one-year and three-year radiology residents and one-year, three-year, ten-year, and twenty-year radiology specialists. To reduce environmental variability during the Turing test, the images were displayed in the same order, and any previous answers could not be modified. All readers successfully performed the test and decided whether each image was real or synthetic without any time limit and no prior information on the number of real or synthetic images. Additionally, sensitivity, specificity, accuracy and reading time were derived after the image Turing tests were completed.

**Comparison of the CNN classifier on real and synthetic images.** We measured the normal probability score of the 200 synthetic and 200 real images by using our previously trained classifier with an accuracy of 94.7 to differentiate between normal and abnormal CXRs<sup>52</sup>. As the probability score tends to 1, the probability of a normal CXR increases. To determine the decision-making process of the model and identify the most important regions of the model for classifying abnormal CXRs in the abnormal dataset, the gradient-weighted class activation mapping technique (Grad-CAM)<sup>53</sup> was used by overlaying the most significant regions of abnormal lesions in the images with red color.

**Efficacy of synthetic images by comparing the performance of models trained solely on real and/or synthetic mixed datasets.** To verify the utility of the synthetic images, a CNN-based classification was performed as a downstream task. The task involved classifying normal or abnormal images in only the real chest radiographs dataset and adding synthetic images generated by ABN-PGGAN. The synthetic images were added, because normal images significantly exceeded abnormal chest radiographs in the real world.

Therefore, the real dataset consisted of 1000 normal and 1000 abnormal chest radiographs, wherein the latter included 200, 200, 200, 200, and 200 images with nodules, consolidation interstitial opacity, pleural effusion, and pneumothorax, respectively<sup>52</sup>. Normal and abnormal datasets with nodule[s], including mass/consolidation or interstitial opacities, were confirmed via chest CT and pleural effusion. Furthermore, pneumothorax on CXRs were determined via consensus of two thoracic radiologists with corresponding chest CT images. The real dataset was randomly split into 80% for training and 20% for testing. The test set was fixed, and half of the abnormal chest radiographs from the training dataset were randomly sampled for use in the synthetic mixed dataset.

The synthetic mixed dataset was composed of 800 normal and 800 abnormal chest radiographs. Specifically, 800 normal chest radiographs were from the training dataset of the real dataset, 400 abnormal chest radiographs were from the training dataset of the real dataset, and 400 abnormal chest images were ABN-PGGAN-generated images with high abnormal probability score according to the known CNN classifier<sup>52</sup>.

In addition, supplemental experiments were conducted with the more training dataset than the previous training dataset using 3269 normal and 3269 patients with including 904, 510, 240, 1324, and 291 CXRs with nodules, consolidation, interstitial opacity, pleural effusion, and pneumothorax, respectively<sup>44</sup>. The synthetic mixed dataset was constructed by including only 1635 synthetic abnormal images from the abnormal data and subtracting the same number of real data. The 1635 synthetic images were randomly extracted from 50,000 images generated by ABN-PGGAN.

The training using real and synthetic mixed datasets was performed via vanilla ResNet-50<sup>54</sup>, which was chosen as a baseline for training classification models and its performance was excellent. To compare performance, two models were tested on the fixed real test set, and an area under the receiver operating characteristic curve (AUROC) was drawn.

**Statistical analysis.** The sensitivity, specificity, accuracy, and reading time of the six readers were calculated for the image Turing test. Inter-reader agreement was evaluated using Fleiss Kappa. The 95% confidence intervals (CI) of accuracy, sensitivity, and specificity were computed using binomial distribution<sup>55</sup>. To investigate how experience influences the probability of the correct answer, readers were classified into two groups based on their work experience. Specifically, R04, R05, and R06 were part of a more-experienced group and R01, R02, and R03 were part of a less-experienced group. Given that the results of the image Turing test were in a binary format for each image and the data were correlated with each individual and each image, mixed effect logistic regression, which models each reader and each image as random effects, was used to test whether the more-experienced group exhibited higher probability for correct answers than the less-experienced group. To evaluate the effect of experience on reading time, linear mixed models were used with each image and with/without each reader as random effects<sup>56</sup>.

Received: 24 October 2021; Accepted: 13 January 2023

Published online: 09 February 2023

## References

- Kim, M. *et al.* Deep learning in medical imaging. *Neurospine* **17**, 471–472. <https://doi.org/10.14245/ns.1938396.198.c1> (2020).
- Soffer, S. *et al.* Convolutional neural networks for radiologic images: a radiologist's guide. *Radiology* **290**, 590–606 (2019).
- Chartrand, G. *et al.* Deep learning: a primer for radiologists. *Radiographics* **37**, 2113–2131 (2017).
- Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
- Miotto, R., Wang, F., Wang, S., Jiang, X. & Dudley, J. T. Deep learning for healthcare: Review, opportunities and challenges. *Brief Bioinf.* **19**, 1236–1246 (2018).
- Bria, A., Marrocco, C. & Tortorella, F. J. C. Addressing class imbalance in deep learning for small lesion detection on medical images. *Comput. Biol. Med.* **120**, 103735 (2020).
- Rezaei, M. *et al.* in *Medical Imaging 2020: Computer-Aided Diagnosis*. 113140E (International Society for Optics and Photonics).
- Frid-Adar, M. *et al.* GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* **321**, 321–331 (2018).
- Sandfort, V., Yan, K., Pickhardt, P. J. & Summers, R. M. J. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Sci. Rep.* **9**, 1–9 (2019).
- Goodfellow, I. *et al.* Generative adversarial networks. *Commun. ACM* **63**, 139–144 (2020).
- Ker, J., Wang, L., Rao, J. & Lim, T. J. I. A. Deep learning applications in medical image analysis. *IEEE Access* **6**, 9375–9389 (2017).
- Kim, M. *et al.* Realistic high-resolution lateral cephalometric radiography generated by progressive growing generative adversarial network and quality evaluations. *Sci. Rep.* **11**, 1–10 (2021).
- Kazuhiro, K. *et al.* Generative adversarial networks for the creation of realistic artificial brain magnetic resonance images. *Tomography* **4**, 159–163 (2018).
- Ben-Cohen, A. *et al.* Cross-modality synthesis from CT to PET using FCN and GAN networks for improved automated lesion detection. *Eng. Appl. Artif. Intell.* **78**, 186–194 (2019).
- Jiang, J. *et al.* Cross-modality (CT-MRI) prior augmented deep learning for robust lung tumor segmentation from small MR datasets. *Med. Phys.* **46**, 4392–4404 (2019).
- Kang, E., Koo, H. J., Yang, D. H., Seo, J. B. & Ye, J. C. Cycle-consistent adversarial denoising network for multiphase coronary CT angiography. *Med. Phys.* **46**, 550–562 (2019).
- Liu, X. *et al.* Accurate colorectal tumor segmentation for CT scans based on the label assignment generative adversarial network. *Med. Phys.* **46**, 3532–3542 (2019).
- Seah, J. C., Tang, J. S., Kitchen, A., Gaillard, F. & Dixon, A. F. J. R. Chest radiographs in congestive heart failure: Visualizing neural network learning. *Radiology* **290**, 514–522 (2019).
- Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U. & Langs, G. in *International conference on information processing in medical imaging*. 146–157 (Springer).
- Alex, V., KP, M. S., Chennamsetty, S. S. & Krishnamurthi, G. in *Medical Imaging 2017: Image Processing*. 101330G (International Society for Optics and Photonics).
- Park, H. Y. *et al.* Realistic High-resolution body computed tomography image synthesis by using progressive growing generative adversarial network: Visual Turing test. *JMIR Med. Inf.* **9**, e23328 (2021).
- Chuquicusma, M. J., Hussein, S., Burt, J. & Bagci, U. in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. 240–244 (IEEE).
- Zakka, C., Saheb, G., Najem, E. & Berjawi, G. J. a. p. a. MammoGANesis: Controlled generation of high-resolution mammograms for radiology education. (2020).
- Karras, T., Aila, T., Laine, S. & Lehtinen, J. J. Progressive growing of gans for improved quality, stability, and variation. (2017).
- Alam, S. B., Hossain, M. & Kobashi, S. in *International Symposium on Affective Science and Engineering ISASE2020*. 1–5 (Japan Society of Kansei Engineering).
- Abdelhalim, I. S. A., Mohamed, M. F. & Mahdy, Y. B. Data augmentation for skin lesion using self-attention based progressive generative adversarial network. *Expert. Syst. Appl.* **165**, 113922 (2021).
- Teramoto, A. *et al.* Deep learning approach to classification of lung cytological images: Two-step training using actual and synthesized images by progressive growing of generative adversarial networks. *PLoS ONE* **15**, e0229951 (2020).
- Kokomoto, K., Okawa, R., Nakano, K. & Nozaki, K. Intraoral image generation by progressive growing of generative adversarial network and evaluation of generated image quality by dentists. *Sci. Rep.* **11**, 1–10 (2021).
- Kwon, H.-J., Shin, D.-H. & Chung, K. J. I. A. PGGAN-based anomaly classification on chest x-ray using weighted multi-scale similarity. *IEEE Access* **9**, 113315–113325 (2021).
- McHugh, M. L. J. B. Interrater reliability: The kappa statistic. *Biochem. Med.* **22**, 276–282 (2012).
- Sun, L. *et al.* An adversarial learning approach to medical image synthesis for lesion detection. *IEEE J. Biomed. And Health. Inf.* **24**, 2303–2314 (2020).
- Salehinejad, H., Valaee, S., Dowdell, T., Colak, E. & Barfett, J. in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 990–994 (IEEE).
- Goodfellow, I. J. *et al.* Generative adversarial networks. (2014).
- Shen, Y., Gu, J., Tang, X. & Zhou, B. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9243–9252.
- Buda, M., Maki, A. & Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.* **106**, 249–259. <https://doi.org/10.1016/j.neunet.2018.07.011> (2018).
- Thabtah, F., Hammoud, S., Kamalov, F. & Gonsalves, A. Data imbalance in classification: Experimental evaluation. *Inf. Sci.* **513**, 429–441. <https://doi.org/10.1016/j.ins.2019.11.004> (2020).
- Learning, Goodfellow, D. I., Bengio, Y., & Courville, A. *The reference book for deep learning models* (2016).
- Tran, L., Yin, X. & Liu, X. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1415–1424.
- You, C. *et al.* Structurally-sensitive multi-scale deep neural network for low-dose CT denoising. *IEEE Access* **6**, 41839–41855 (2018).
- Park, J. E., Vollmuth, P., Kim, N. & Kim, H. S. Research highlight: use of generative images created with artificial intelligence for brain tumor imaging. *Korean J. Radiol.* **23**, 500 (2022).
- Lei, Y. *et al.* MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. *Med. Phys.* **46**, 3565–3581 (2019).
- Szegedy, C. *et al.* in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.
- Karras, T., Laine, S. & Aila, T. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4401–4410.
- Karras, T. *et al.* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8110–8119.
- Karras, T. *et al.* Training generative adversarial networks with limited data. (2020).
- Song, Y. *et al.* Score-based generative modeling through stochastic differential equations. (2020).
- Albahli, S. J. I. Efficient GAN-based Chest Radiographs (CXR) augmentation to diagnose coronavirus disease pneumonia. *Int. J. Med. Sci.* **17**, 1439 (2020).
- Salehinejad, H., Colak, E., Dowdell, T., Barfett, J. & Valaee, S. J. Synthesizing chest x-ray pathology for training deep convolutional neural networks. *IEEE Trans. Med. Imaging* **38**, 1197–1206 (2018).

49. Madani, A., Moradi, M., Karargyris, A. & Syeda-Mahmood, T. in *Medical Imaging 2018: Image Processing*. 105741M (International Society for Optics and Photonics).
50. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. & Courville, A. C. in *NIPS*.
51. Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein GAN. *ArXiv abs/1701.07875* (2017).
52. Park, B. *et al.* A curriculum learning strategy to enhance the accuracy of classification of various lesions in chest-PA X-ray screening for pulmonary abnormalities. *Sci. Rep.* **9**, 1–9 (2019).
53. Nogueira, K., Penatti, O. A. & Dos Santos, J. A. J. P. R. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognit.* **61**, 539–556 (2017).
54. He, K., Zhang, X., Ren, S. & Sun, J. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
55. Dorai-Raj, S. & Dorai-Raj, M. S. J. C. Package 'binom'. (2009).
56. Skrondal, A. & Rabe-Hesketh, S. *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. (Crc Press, 2004).

## Acknowledgements

This study included results of the "HPC Support" Project, supported by the 'Ministry of Science and ICT' and NIPA.

## Author contributions

M.J. wrote the original draft, analyzed the data, and performed formal analysis. H.J.B. contributed to data acquisition and analysis and provided technical guidance for the project. M.K. contributed to the classification task and interpretation of results. S.Y.P. provided statistical guidance for the project. A.S., S.J.C., J.C., H.Y.C., H.J.H., and H.N.N. participated as readers and contributed to interpretation of results. J.B.S. conceptualized the project and contributed to data acquisition. N.K. and S.M.L. conceptualized the project and developed the methodology for the project. All the authors reviewed the final manuscript.

## Funding

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HI18C0022, HI18C2383).

## Competing interests

HJB and N.K. are shareholders of Promedius Inc. The other authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-28175-1>.

**Correspondence** and requests for materials should be addressed to S.M.L. or N.K.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023