공학석사 학위논문

생성적 적대 신경망을 이용한 흉부 방사선 영상 합성 및 이상 탐지

Image Synthesis and Anomaly Detection on Chest Radiographs using Generative Adversarial Networks

울산대학교 대학원
의 과 학 과
김 민 지

생성적 적대 신경망을 이용한 흉부 방사선 영상 합성 및 이상 탐지

지도교수 김 남 국, 홍 길 선

이 논문을 공학석사 학위 논문으로 제출함

2021년 08월

울산대학교 대학원
의 과 학 과
김 민 지

김민지의 공학석사학위 논문을 인준함

심사위원 배 현 진 (인)

심사위원 김 남 국

심사위원 홍 길 선 (인)

울 산 대 학 교 대 학 원

2021 년 08 월

감사의 글

의료인공지능에 대한 열망을 품고 입학한지 어느덧 2 년 남짓이 되었습니다. 기쁨과 즐거움, 슬픔과 노여움이 공존했던 연구실 생활을 되돌아봅니다. 망망대해에 정처없이 떠다니는 배와 같았던 저에게 나침반이 되어준 많은 분들께 감사를 올립니다.

우선, 석사 기간 동안 날카로운 비평과 진심 어린 조언으로 많은 가르침을 주신 김남국 교수님 그리고 홍길선 교수님께 감사의 인사를 올립니다. 융합 연구자로서의 자질에 돛을 달아주신 덕분에 바다를 갈라 무사히 항해의 마침표를 찍을 수 있었습니다. 바쁘신 와중에도 심사위원장을 맡아 주신 영원한 사수, 프로메디우스의 배현진 대표님, 공학도보다 더 공학적인 관점으로 부족한 부분을 파악하고 지도해주신 서준범 교수님, 임상의 시점에서 여러 의견을 주시고 응원해주신 이상민 교수님, 김성훈 교수님, 조준영 교수님께도 감사의 글을 올립니다.

저에게 큰 귀감이 되었던 연구실 동료들, 특히 수학적으로 많은 영감을 주신 장령우 선생님, 다양한 아이디어로 해결책을 제시해주신 조성만 선생님, 연구에 대한 열정으로 항상 동기를 부여해 주시는 은다인 선생님, 졸업 과정을 같이 밟으며 힘이 되어주신 김성철 선생님, 묵묵하고 성실한 모습을 가진 김다은 선생님, 박주영 선생님, 조경진 선생님, 박승주 선생님께도 감사의 인사를 전합니다. 덕분에 다양한 주제에 흥미를 잃지 않고 즐겁게 연구할 수 있었습니다. 그리고 학문적인 글쓰기 능력을 끌어주신 최준명 박사님, 이경화 선생님, 장미소 선생님, 김민규 박사님께도 고개 숙여 감사드립니다.

무엇보다 사랑하는 어머니(이순학)와 아버지(김영섭), 언니이자 동료인 김민경 선생님, 사랑하는 동생 현준이의 무한한 응원과 배려에 감사의 인사를 전합니다.

Abstract

Deep learning, one of the artificial intelligence technologies in the spotlight recently, has offered promising results in numerous computer vision tasks. Within the advances of convolutional neural networks (CNN), deep learning has been rapidly adopted in medical imaging such as classification and detection of lesion patterns, automated segmentation of organs, medical image reconstruction, etc. Extensive datasets with high-quality images and their precise annotations are required for the traditional approaches with supervised learning methods. However, constructing a high-quality dataset is challenging in the field of medicine. Limited data access, unbalanced datasets, and expensive annotation processes could limit the prediction power and induce biased results in deep learning models.

Amid the increasing need for high-quality medical image datasets, the emergence of generative adversarial networks (GAN)[1] has provided a new breakthrough. GANs learn an ability to generate new plausible samples from an existing dataset through an adversarial process. GANs have demonstrated potentials in various tasks, including domain adaptation, super-resolution, image-to-image translation, image style transfer, and anomaly detection.

Despite the promising results of GANs, it has been less explored in the medical field. In this study, we suggested several unsupervised methods using progressive growing of GANs (PGGAN)[3] for potential applications in medical imaging, especially on chest X-ray (CXR) images. We address the following tasks: (a) evaluating the fidelity of synthetic CXR images generated from PGGAN, (b) generating synthetic CXR images with a desired pulmonary disease pattern by disentangling semantic representations in the latent space of PGGAN learned in (a), and (c) developing an anomaly detection system that identifies anomalous patterns in CXR images with an unsupervised scheme using PGGAN. In the first topic of (a), we proposed a 3-step method that utilized a deep learning-based classification network (classifier). We compared the performances of two classifiers which have separately trained on real and synthetic CXR images (step 1 and 2) and evaluated on the identical test dataset of real CXR images in terms of a binary classification: normal or abnormal (step 3). We have found that synthetic CXR images generated from PGGAN preserved radiologic informatics as much as the real ones. In the second topic of (b), we explored and discovered semantic representations of predefined pulmonary disease patterns in the latent space of PGGAN.

With a simple linear regression, we demonstrated that controllable generation of CXR images with desired disease pattern is possible. The evaluation was performed qualitatively and quantitatively, by a visual scoring from an expert radiologist with more than 20-year-experience and by a metric using the classifier suggested in (a), respectively. In the third topic of (c), we proposed an anomaly detection system based on CXR images with an unsupervised method using PGGAN. We trained PGGAN with a normal CXR dataset to identify anomaly CXR samples. Given a CXR image with abnormality from a real dataset, we approximated the most analogous normal CXR image by optimizing a latent vector with an iterative algorithm. In the evaluation, we have demonstrated that anomalous patterns in CXR could sensitively be detected without the need for disease annotations.

This study has shown the potentials of GANs in developing unsupervised deep learning-based applications in medical imaging. Each result of high-fidelity image synthesis, controllable image synthesis (image manipulation), and anomaly detection on CXR images can be exploited to address existing problems in supervised learning such as patient privacy, unbalanced dataset, and expensive annotations in medical imaging.

Abbreviations

CNN (Convolutional Neural Networks)

CXR (Chest X-Ray)

CAD (Computer Aided Diagnosis)

CT (Computed Tomography)

MRI (Magnetic Resonance Imaging)

GAN (Generative Adversarial Networks)

VAE (Variational Auto-Encoder)

AMC (Asan Medical Center)

SNUBH (Seoul National University Bundang Hospital)

PA (Postero-Anterior)

ROC (Receiver Operating characteristic)

AUROC (Area Under the Receiver Operating Characteristic)

DICOM (Digital Imaging and Communications in Medicine)

PNG (Portable Network Graphics)

IS (Inception Score)

FID (Fréchet Inception Distance)

P&R (Precision and Recall)

PPL (Perceptual Path Length)

Grad-CAM (Gradient-weighted Class Activation Mapping)

Contents

Tables

Figures

Introduction

Background

Deep learning, a part of machine learning algorithms based on artificial neural networks, has provided an unprecedented performance in various computer vision tasks. Especially, deep learning with convolutional neural networks (CNN) has rose its popularity after the winning to the famous challenge, ImageNet Large-Scale Visual Recognition Challenge in 2012, by reducing the error rate by half in image classification task. The use of CNN has shown high performances not only in image classification task but also in object detection, semantic segmentation, image reconstruction, depth estimation, visual question answering, etc.

These promising results had led to a lot of attentions from medical domain. Deep learning-based computer aided diagnosis (CAD) systems have been rapidly adapted as providing complementary to physicians to improve diagnostic reliability. In chest X-ray images (CXR), the most commonly used diagnostic imaging modality, studies have demonstrated that CNN-based classification networks can achieve radiologist-level performances on lesion detection of pulmonary diseases[4-6]. Another study has shown combined performance of a radiologist and CAD system outperformed a single radiologist, CAD only, and two readers[7]. Also, applications to other famous modalities were developed and embedded in medical equipment to help clinical practice such as automatic liver tumor segmentation in computed tomography (CT) scans[8,9] and a reconstruction of magnetic resonance imaging (MRI) scans[2,10-12].

The rapid growth of deep learning applications has been driven by virtue of the accessibility of enormous datasets, expanded computational power, and the development of deeper neural networks. Supervised learning, the most widespread approach with discriminative models, also requires a large number of data samples and their labels. High performances in the above studies could be achieved in virtue of large, high-quality medial datasets with corresponding annotations[6,13,14]. However, medical images are not always available in real world settings. Data access to medical images is highly restricted by ethical considerations to protect patient privacy. Also, the process of annotation is time-consuming, tedious, and expensive as it requires an expertise in medical domain. Furthermore, data imbalance problem often exists with different size of class samples, especially with a number of common diseases and a paucity of rare ones. These low accessibility in medical datasets

hinder broad applications of deep learning techniques in medicine.

Generative model approach

A generative model can generate new plausible data samples. Generative models learn the data distribution and estimate how likely a given sample is, while discriminative models differentiate between data samples with different categories. The emergence of generative adversarial networks (GAN)[1], by Ian Goodfellow in 2014, has marked a new era in deep generative networks. GAN provided a new approach in constructing training datasets by generating realistic synthetic images. GAN estimates implicit generative models with an adversarial training of two networks: a generator and a discriminator. A generator creates synthetic data similar to the input (the real one) and a discriminator distinguishes between the real and synthetic data created from the generator. The two networks are trained simultaneously through an adversarial process of which the generator is trained to maximize the probability of a discriminator making mistakes. In the meanwhile, the generator understands a distribution of training dataset and learns meaningful representations of the dataset. The rationale behind GAN is to learn a nonlinear mapping from a randomly sampled variable, i.e. latent vector, in the latent space to a specific output image. Instead of the extensive sampling procedure for estimating the real data distribution, GAN chose to sample from a simple distribution such as Gaussian distribution and then transform the latent vector (or the random noise) using the nonlinear function such as neural networks. With this mapping, GAN has an ability to generate new plausible images from existing dataset.

Evaluation metric for GAN generator

The objective measurement of GAN performance remains an open problem. There are several metrics for evaluating the generator performance of GAN. The most intuitive method would be a visual examination by human experts. Some studies[15,16] had performed a visual Turing test – a test that how well domain experts classify real and synthetic images in a blind manner. Authors of these studies asserted that synthetic images generated by GAN is so realistic that even domain experts could confuse. Although visual inspection is a simple and natural way to evaluate, it has drawbacks. Domain experts may evaluate with biases and may

not notice subtle artifacts on pixel level. To overcome these subjectivities in measurement, quantitative evaluation metrics were devised. To begin with, Inception Score (IS) was proposed. The score seeks to incorporate the image quality and image diversity by using a well-known classification network, Inception[17]. Given a set of synthetic images, IS is calculated from how likely an image belongs to a pre-defined category and how diverse the categories that the images fall into. Fréchet Inception distance (FID), a kind of advanced version of IS that utilizes the real images as well as synthetic ones, was proposed and used in measuring performance of StyleGAN[18]. Other methods such as precision and recall (P&R)[19] and perceptual path length (PPL) were used in measuring performance of StyleGAN2[20]. These methods take advantage of a pre-trained classification network such as Inception v3[21] and the feature vectors extracted by those networks. However, most pretrained networks were trained on ImageNet[22], whose feature extraction could be quite different in medical images. Therefore, it is inadequate for medical domain to see generated synthetic images – this method is inefficient and not quantitative – nor using ImageNet-pretrained network.

High-resolution Image synthesis

Image synthesis is the core capability of GAN showing compelling results in various computer vision tasks[24-26] without the need for label responses. Despite impressive results, conventional GANs have inherent difficulty in training. This is partly explained by the fact that optimizing the two components of GANs, generator and discriminator, have to be optimized in parallel and are dependent on each other. While these issues are severe in the generation of high resolution images. This appears to be intuitively possible since starting with a high-resolution image makes the classification task of discriminator easier compared when the generator generates a nearly accurate image from scratch. Thus, the task of the discriminator is easy to be optimized, and it tends to dominate in the early training process, therefore preventing successful training. The novel approach[23] was to start with a low-resolution GAN and increasing image size step by step during training (hence the name progressively growing GAN, PGGAN), thereby assisting the generator and stabilizing the model. It was demonstrated the utility of this approach in the image synthesis of human faces by generating a large number of high resolution (1024×1024 pixel). After the release of

PGGAN, significant improvements have been developed towards high-quality synthetic images such as StyleGAN[18], and recently released StyleGAN2[20].

Disentanglement of latent space

Well-trained generator of GAN can generate synthetic images with the encoded representations (latent vectors). The latent space of the learned distribution can be compose of multiple subspaces of semantic representations. The semantic representations of a training dataset could be a structural features of pulmonary disease patterns such as texture, shape, or the size of in CXR images. By using the features, we could build an understanding of how the disease pattern initiates and how its size changes. However, the latent space of GAN is likely to be entangled when learning a distributed representation of training datasets. Thus, a disentangling the latent space is required to generate or manipulate synthetic images with semantic control. By discovering the disentangled the semantic feature representations, manipulation of synthetic images could be possible. Previous researches on disentangling semantic feature representations in the latent space of GAN achieved a conditional manipulation of a given synthetic image[25,27-32].

Unsupervised Anomaly Detection

Anomaly detection refers to the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data. Unsupervised anomaly detection is another research area that has been actively explored with the rise of generative models such as GAN and variational auto-encoder (VAE). The main concept is that anomalous samples can be distinguishable in the latent space of GAN which trained with normal samples only. By establishing a decision boundary that deviate from the normal samples, the model can detect other possible anomalous samples has never seen before. However, there exists an ill-posed problem in a mapping of GAN. To generate a synthetic image closest to the given unseen sample, it is required to find a corresponding latent vector. Unfortunately, GAN does not yield this reverse mapping automatically. One of the tractable solutions to detour these inversion problem is to approximate the latent vector by an iterative optimization algorithm. With this approach, unsupervised anomaly detection could provide

complementary information to radiologists in diagnostic procedures.

Objectives

This study was conducted toward GAN-based applications in medical imaging. Each task of high-fidelity image synthesis, controllable image synthesis (image manipulation), and anomaly detection on CXR images can be exploited to address existing problems in supervised learning such as patient privacy, unbalanced dataset, and insufficient annotation in medical datasets.

Methods

1. Datasets

We constructed three CXR datasets according to the following purposes. Two datasets were collected to train GAN generators for the generation of synthetic CXR images. The other one was collected to train a CNN classifier for feature learning of pulmonary disease patterns. The Institutional review board for human investigations at Asan Medical Center (AMC) and Seoul National University Bundang Hospital (SNUBH) approved the retrospective study with a waiver of informed consent. The imaging data were de-identified in accordance with the Health Insurance Portability and Accountability Act privacy rule.

1.1. GAN training dataset

We retrospectively collected 217,924 CXR scans from AMC between Jan 1, 2011 and Jun 30, 2016. The CXR scans had a size of about 2000 x 2000 pixels and were stored in 12-bit digital imaging and communications in medicine (DICOM) format. Each of the scans had a radiological report associated with it. First, diagnostic codes were used to identify whether the case has any diseases. Then, each scan was divided into two image-level classes: normal and abnormal. Of the 217,924 scans, 109,201 and 108,723 cases were labeled as normal and abnormal, respectively. After the image-level labeling with normal/abnormal, i.e. weak labeling, each of the CXR scans were then screened for the exclusion criteria: subjects under the age of 19, scans not from GE manufacturer, i.e. the majority of X-ray equipment in AMC, and scans not taken in a posteroanterior (PA) view. Additionally, normal scans with presence

of devices (e.g. catheter, pace maker, wire, etc.) were excluded using a simple CNN classifier. In total, the dataset is consisted of 164,101 CXR scans, of which 72,938 are normal cases, of which 91,163 are abnormal cases. Each cases were used for generating synthetic CXR images with and without pulmonary abnormalities, respectively.

For the experiments of high-fidelity image synthesis and controllable image synthesis, a total of 111,163 CXR images were used, of which 20,000 were normal cases subsampled from 72,938, of which 91,163 were abnormal cases. For the experiments of anomaly detection, 72,938 CXR images of normal cases were used. The CXR scans were converted into 8-bit portable network graphics (PNG) format and 99th percentile normalization was conducted for all converted images.

1.2. CNN classifier training dataset

The CXR images were collected from two medical centers, AMC, Seoul, South Korea, and SNUBH, Bundang, South Korea. The dataset is consisted of 6,069 normal CXR scans and 3,417 CXR scans of the patients at AMC including 944, 550, 280, 1364, and 331 cases with nodule[s], consolidation, interstitial opacity, pleural effusion, and pneumothorax, respectively. 1,035 normal CXR scans and 4,404 cases at SNUBH[33] including 1189, 853, 1009, 998, and 944 cases with nodule[s], consolidation, interstitial opacity, pleural effusion, and pneumothorax, respectively. One subject may have multiple abnormalities in a given CXR. All CXR scans have mask annotations and confirmed by corresponding chest CT images. In addition, abnormal cases with pleural effusion and pneumothorax were determined by consensus of two thoracic radiologists with the corresponding chest CT images.

For the experiment of high-fidelity image synthesis and its evaluation, we divided all the data into two categories of normal and abnormal. We considered five pulmonary abnormality cases as abnormal in terms of binary classification. Therefore, there were 7,104 normal CXR images from 7,104 healthy subjects and 10,234 abnormal CXR images from 7,821 patients when counting multi-labeled cases. For the experiment of controllable image synthesis, on the other hand, normal and the aforementioned five pulmonary disease abnormality cases were used to deal with 6-class classification.

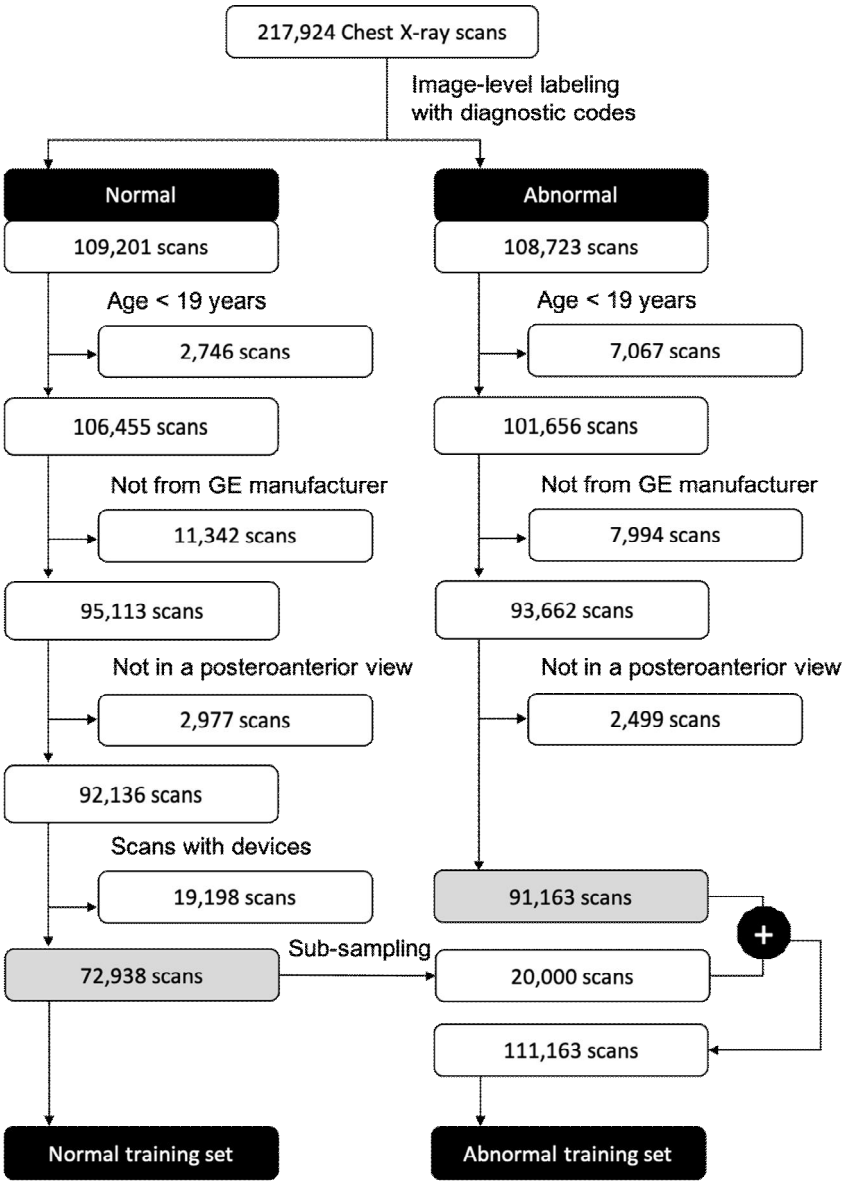

Figure 1. A flowchart of dataset selection for GAN training

Table 1. Dataset compositions for GAN training for normal and abnormal CXR image synthesis.

Diagnosis	Normal CXR dataset	Abnormal CXR dataset
Normal	72,938	20,000
Abnormal	-	91,163
Total	72,938	111,163

Table 2. Dataset description of the AMC and SNUBH datasets.

Diagnosis		Number of images			
		AMC*	SNUBH**	Total	
Normal (NM)		6,069	1,035	**7,104**	
Abnormal	Nodule (ND)	1,012	1,516	**10,234**	2,528
	Consolidation (CS)	653	1,114		1,767
	Interstitial Opacity (IO)	312	1,222		1,534
	Pleural Effusion (PE)	1,599	1,302		2,901
	Pneumothorax (PT)	421	1,083		1,504

2. Evaluating GAN generator

Our goal is to devise a method that can be used for quantitative comparison of GAN generator performance and to validate its possible application in medical imaging. To this end, our approach is to utilize CNN-based models to measure prediction performances. In the evaluation, we compared the performances of each of two CNN classifiers which have been trained on real and synthetic CXR dataset, respectively.

2.1. PGGAN training

To generate synthetic images, we utilized PGGAN[23] due to its superior performance of realistic image generation with a high-resolution quality. Training scheme of PGGAN is shown in Figure 2. PGGAN learns to generate synthetic images starting from a low resolution of 4x4 pixel to a high resolution of 1024x1024 pixels by growing image sizes progressively. For PGGAN training, 20,000 normal CXRs and 91,163 abnormal CXRs were used. The PGGAN training was done for 130 epochs and took around 12.2 days with two Titan RTX GPUs.

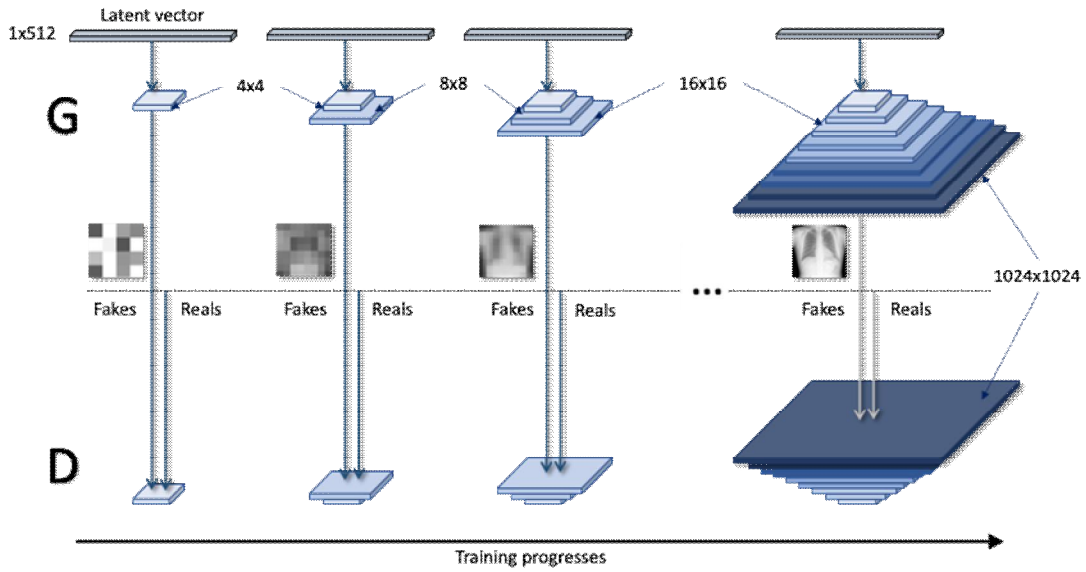

Figure 2. Generation of high-quality CXRs using a progressively growing training scheme of PGGAN.

2.2. CNN classifier training.

For the training of a CNN classifier, we chose a network architecture of ResNet-50[33] which based on deep residual network (ResNet)[34] since it has a high performance, which has been frequently used to solve a real-world classification problem. We modified the softmax layer to a sigmoid layer in Resnet-50 to deal with binary classification; normal or abnormal. The classifiers were trained using real and synthetic images, respectively, with the same strategy including the identical architecture[33] and hyperparameters. For fair comparison, the identical number of images and ratio of normal and abnormal cases were used. We divided the CXR dataset for CNN classifier training into three groups – train set (60% of the total dataset, 8,981 out of 14,925 subjects), validation set (20% of the total dataset, 2,972 out of 14,925 patients), test set (20% of the total dataset, 2,972 out of 14,925 patients). The detailed number of images are stated in Figure 3. Finally, the evaluation was performed using the real CXR images for both classifiers trained on real and synthetic dataset.

2.3. 3-step classification scheme.

We suggest the 3-step concept in the utilization of classification network. To begin with, a classifier was trained using the real dataset with normal and abnormal CXR images. Then, we created pseudo-labels for the PGGAN-generated images according to the classification results of a classifier trained on real dataset, i.e. classifier (real). We set a likelihood threshold to be above 0.7 and below 0.3 for labeling normal and abnormal samples, respectively. Also, a classifier was trained using the synthetic dataset, i.e. classifier (synthetic), containing normal and abnormal CXR images. Finally, a real test set was evaluated and compared in the prediction using both classifiers (real and synthetic). For all these three steps, sensitivity, specificity, and AUROC were calculated.

2.4. Evaluation of the fidelity of GAN generator

To compare performances of (1) classifier trained on real dataset on real test set, (2) classifier trained on synthetic dataset on synthetic test set, (3) classifier trained on synthetic dataset, we evaluated their area under the receiver operating characteristic curves (AUROCs). We used AUROC comparison method[35] for statistical comparison, which is to evaluate how

significantly different two AUROCs are. Overall classification scheme is shown in Figure 3.

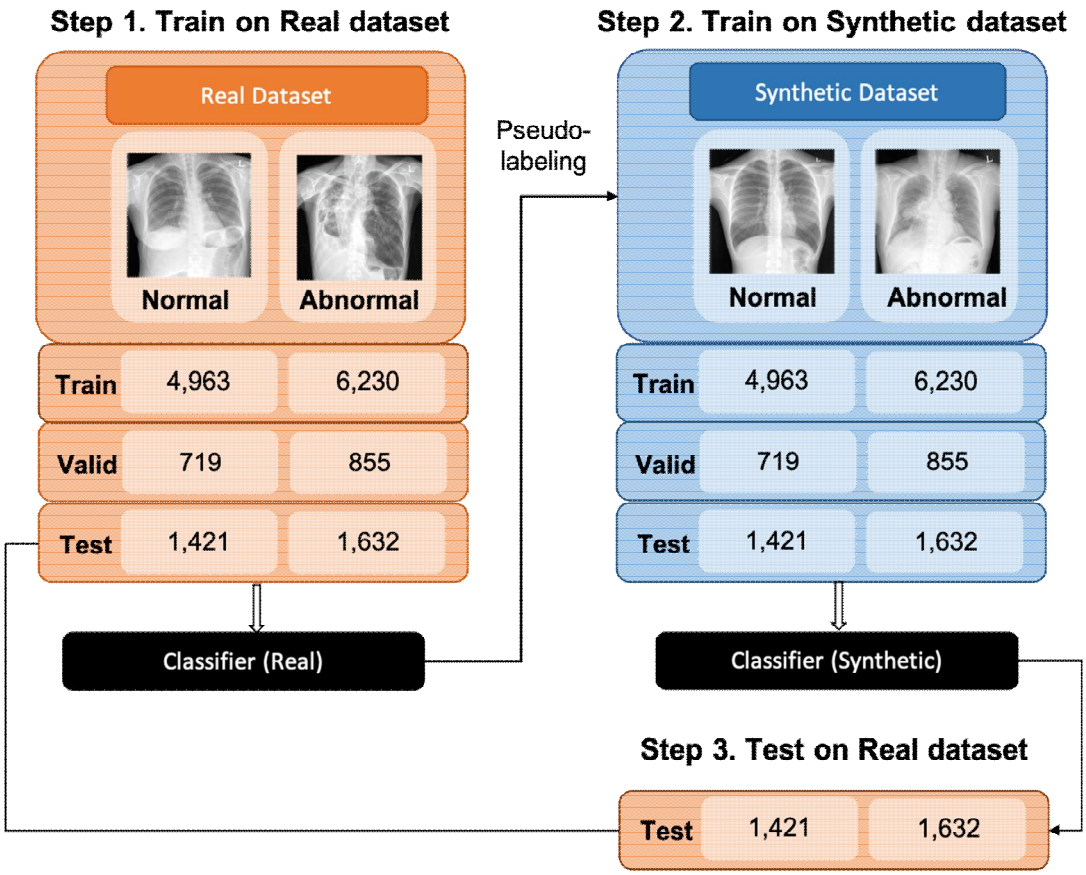

Figure 3. An overall classification scheme for evaluating performance of the GAN generator.

3. Disentanglement of the latent space in GAN

Our goal is to generate synthetic CXR images with desired pulmonary disease patterns. Inspired by transparent latent space GAN (TL-GAN)[28], we chose an approach of a direct mapping between the latent vector z from GAN[1] and the numeric prediction output y from a CNN pulmonary disease classifier[4]. An overview of the experimental setting is shown in Figure 4. To discover the axes of the feature representations of pulmonary diseases in the latent space of GAN[1]. First, we trained PGGAN on the weakly labeled dataset of 111,163 CXR images. We generated 30,000 synthetic CXR images from this network, then classified by pulmonary disease using a multi-label CNN classifier[4]. Finally, we performed a linear regression on the latent vector z of the synthetic CXR images and the output y from the CNN

classifier. The regression slopes for each disease class are considered as the axes of disease patterns.

The evaluation was conducted qualitatively and quantitatively. To validate whether the discovered axes are plausible, a visual scoring test was performed by an expert thoracic radiologist. We additionally evaluated the likelihood of the synthetic CXR images with the CNN classifier.

3.1. PGGAN training

The training procedure is identical to 2.1. PGGAN model was selected to implement to generate synthetic CXR images since this model performed better in reconstructing both of a global structure and fine details with a high-resolution quality among other GAN variant models [31,36,37]. PGGAN was intended to learn meaningful feature representations of pulmonary disease patterns in each real CXR image during training. After training, a generator network of PGGAN is capable of generating synthetic CXR images with random disease patterns from a 512-dimensional variables of random noises, i.e. latent vector.

3.2. CNN classifier training

We utilized a multi-label CNN classifier with a high performance to obtain the pulmonary disease classification outputs of the synthetic CXR images. The base model architecture is Resnet-50 replaced the last softmax layer with six sigmoid layers to deal with CXR images with multiple disease patterns to deal with a multi-label classification problem. The particulars of altered architecture can be found in[4]. The classifier has an ability to detect and classify five disease patterns—nodule[s], consolidation, interstitial opacity, pleural effusion, and pneumothorax— simultaneously on CXR images. The generated 30,000 synthetic CXR images were given to the classifier and classified with six classes (five disease patterns and normal), whose outputs are numeric numbers between 0 and 5. A CNN classifier can be used to predict pulmonary abnormalities in the randomly generated CXR images and classify into six classes including normal and five classes of abnormal patterns.

3.3. Exclusion of synthetic images with multiple disease patterns

Before performing a linear regression, we excluded synthetic CXR images having multiple disease patterns. We set a criteria; the highest likelihood output of disease pattern to be 0.9 or above and the rest of other likelihood outputs to be 0.1 or below. In this case that we can assume that there is a single disease pattern in the image. We selected the CXR images with a single disease pattern using the 6-class CNN classifier.

3.4. Linear transformation

The feature representations of disease patterns could be encoded in latent vector of the PGGAN. We performed a linear regression to discover a correlation among the samples by disease pattern class. With the regression of the latent vectors and the classification outputs of synthetic CXR images, disentangled feature axis of disease pattern could be discovered. Also, if PGGAN is well trained, it generates synthetic images of a various diseases, reflecting the incidence distribution of each disease pattern in the training dataset. To reduce the natural bias on the regression result of derived from the data imbalance, we added weights on the class with rare disease patterns in a proportion of the number of images each class. After the linear regression, we explored the disentanglement by manipulating the given image toward the regression slope.

3.5. Evaluation of manipulated synthetic images

Qualitative and quantitative evaluations were performed to evaluate the disentanglement. With synthetic images that has been manipulated with an axis control, a visual scoring with an expert radiologist was used. Additionally, the classification-based evaluation metric was used suggested with a 3-step scheme.

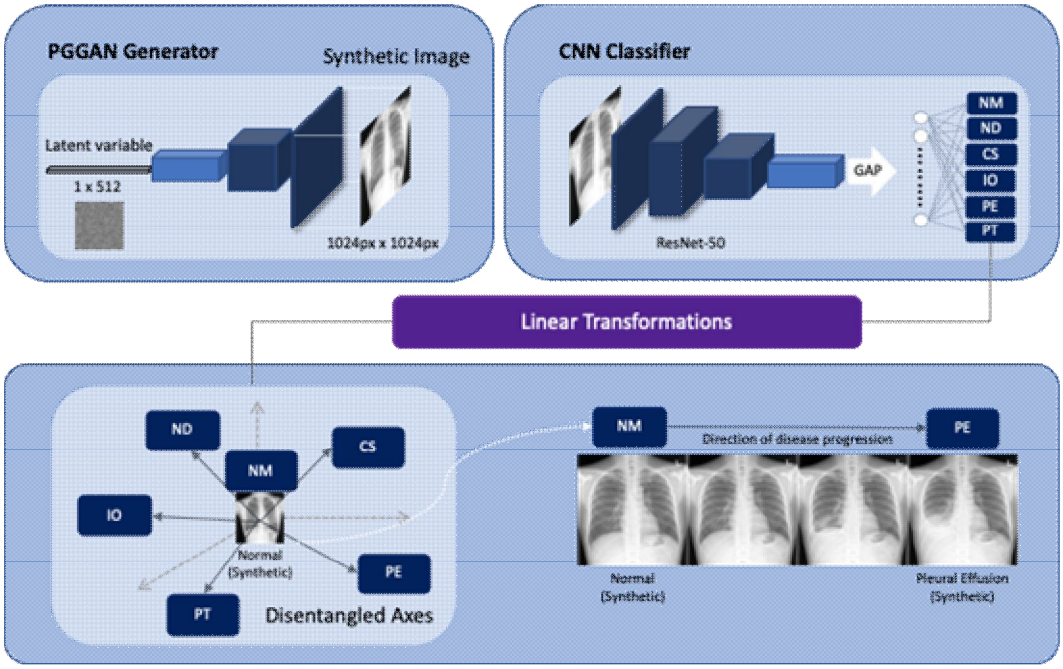

Figure 4. An overview of the experimental setting for disentangling the latent space of PGGAN and manipulating the synthetic images.

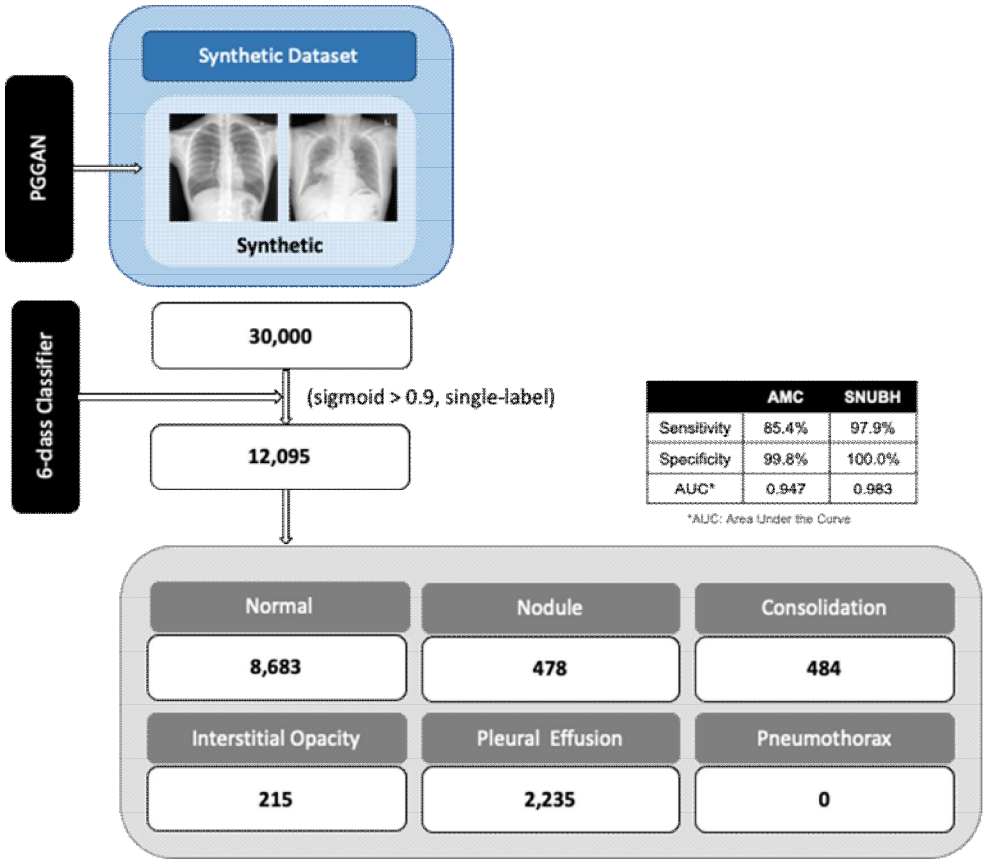

Figure 5. Dataset tree for classifying the generated synthetic images.

4. Anomaly detection system with GAN

Our goal is to explore the unsupervised anomaly detection scheme in CXR images. Our method is inspired by anomaly detection with generative adversarial networks (AnoGAN)[2], which trains normal samples to identify anomalous samples. Given a real CXR image, possibly with an anomalous patterns, an iterative algorithm with a residual loss was adapted to obtain the closest normal image and its corresponding latent vector. In the evaluation, we compared the performance of two systems trained with supervised and unsupervised manner. Specifically, we compared the sensitivities of the classifier trained with a pre-defined disease patterns and the proposed anomaly detection system. The classifier is identical to the one in used 3.2.

4.1. PGGAN training

The training procedure is identical to 2.1 and 3.1 but with normal CXR images to generate synthetic CXR images without any abnormality. A total of 72,938 weakly labeled normal cases were used. The data selection and curation criteria is identical to the PGGAN training for abnormal cases.

4.2. Approximating the latent vector

To approximate the most similar synthetic image that can be generated by a generator of PGGAN, an optimization technique using a residual loss was used. With a certain iteration number, two images of unseen and approximated images were acquired. By subtracting two images, the difference map in a pixel-level can be calculated. By setting the residual difference map as a loss function to be minimized, the generator is able to discover the matching latent vector in the latent space. As a residual difference score, we utilized mean pairwise squared error.

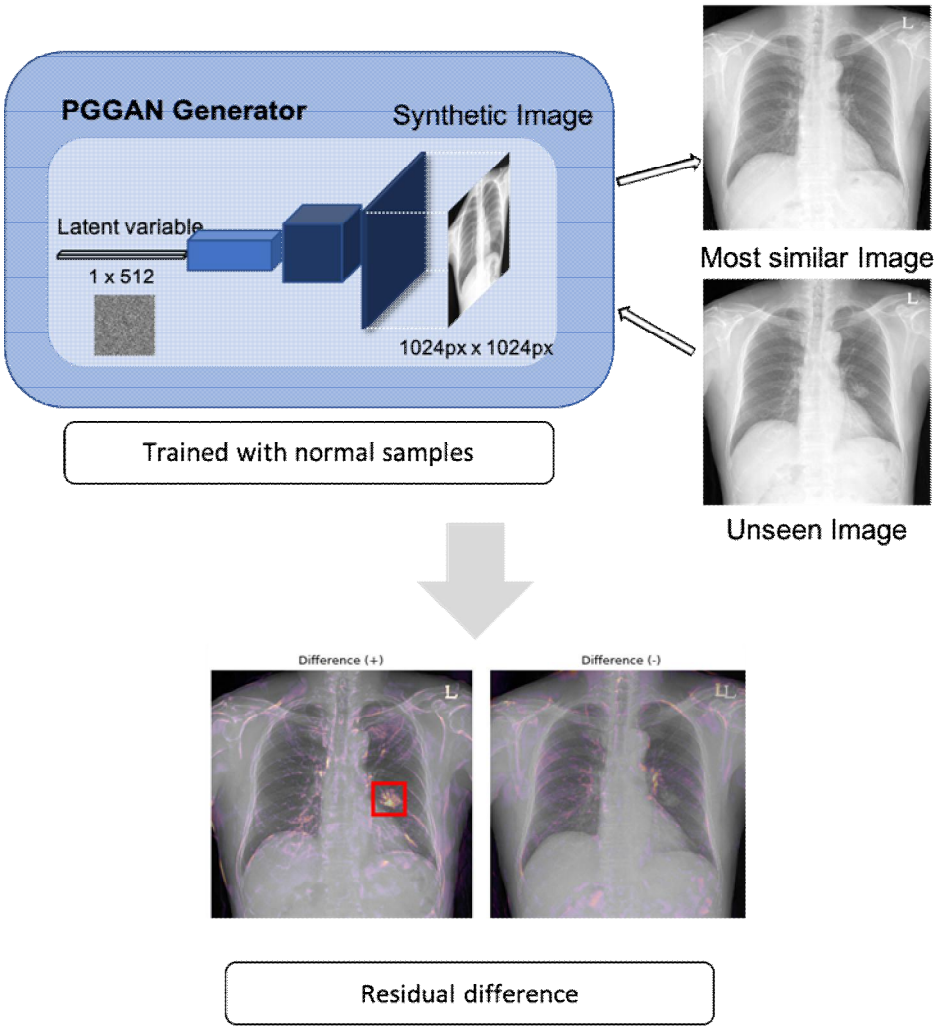

Figure 6. An overview of the experimental setting for detecting anomalous patterns in synthetic CXR images.

Results

1. Evaluation of image fidelity of PGGAN-generated data

Performance of classifier (real) on real test set had AUROC 0.9830, sensitivity 93.4%, specificity 93.3%. Also, performance of classifier (synthetic) on synthetic test set was AUROC 0.9673, sensitivity 90.8%, and specificity 90.5%. Finally, the performance of the classifier (synthetic) on the real test set was AUROC 0.8810, sensitivity 85.1%, and specificity 81.6%. AUROC comparison between two classifiers (real and synthetic) on the real test set showed there was no statistical difference between them ($p<0.0001$). Results have demonstrated that the performance gap between classifiers trained on PGGAN-generated data and real data was negligible in the abnormality classification of CXR images. Confusion matrices of three experiments are shown in Table 3, Table 4, Table 5, respectively.

Table 3. A confusion matric of a classifier (real) on real test set.

Classifier (real) on real test set		Ground truth		Total
		Normal	Abnormal	
Predicted	Normal	1,312	92	1,404
	Abnormal	109	1,540	1,549
Total		1,321	1,632	3,053

Table 4. A confusion matrix of a classifier (synthetic) on synthetic test set.

Classifier (synthetic) on synthetic test set		Ground truth		Total
		Normal	Abnormal	
Predicted	Normal	1,264	128	1,392
	Abnormal	157	1,504	1,661
Total		1,421	1,632	3,053

Table 5. A confusion matrix of a classifier (synthetic) on real test set.

Classifier (synthetic) on real test set		Ground truth		Total
		Normal	Abnormal	
Predicted	Normal	1,095	191	1,286
	Abnormal	326	1,441	1,767
Total		1,421	1,632	3,053

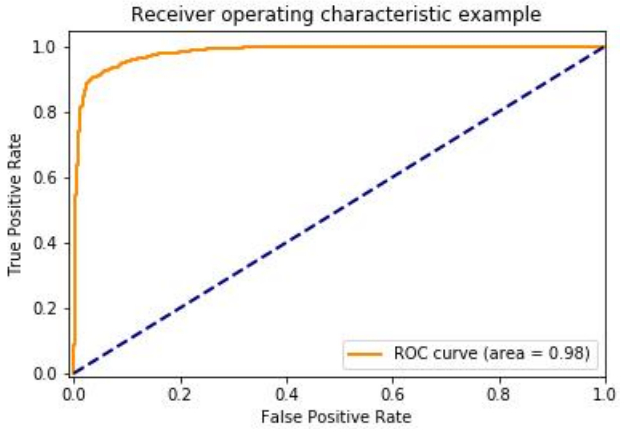

Figure 7. A result of the ROC curve with AUROC score using a classifier (real) on real test set.

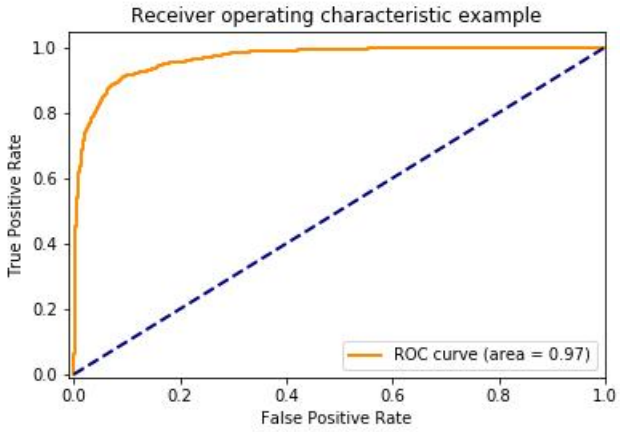

Figure 8. A result of the ROC curve with AUROC score using a classifier (synthetic) on synthetic test set.

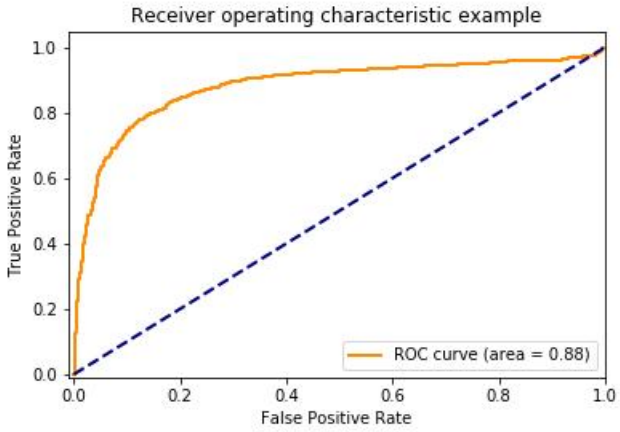

Figure 9. A result of the ROC curve with AUROC score using a classifier (synthetic) on real test set.

2. Disentangled pulmonary representations in PGGAN

Linear regression was performed on 8,683, 478, 484, 215, 2235 synthetic CXR images classified as normal, nodule[s], consolidation, interstitial opacity, and pleural effusion using the 6-class classifier. The synthetic CXR images classified as pneumothorax were found to be none. We observed that three feature representations of pulmonary disease patterns of consolidation, interstitial opacity, and pleural effusion were discovered among the aforementioned five disease classes in latent space of PGGAN. We then manipulated CXR images starting from normal synthetic image to abnormal with desired pulmonary disease patterns. Figure 10. shows the continuous changes on pulmonary regions with abnormalities when manipulating images by moving along the discovered axes of (a) consolidation, (b) interstitial opacity, and (c) pleural effusion. In Figure 10, each of the first column of (a), (b), and (c) represents the conditional image synthesis and the second column represents the corresponding activation maps when predicting using a 6-class CNN classifier. For the quality evaluation, we performed a visual Turing test with a board-certified radiologist with 20+ years of experience. The generated images on the axis of consolidation, interstitial opacity, and pleural effusion were scored from normal to severe, moderate and moderate stages, respectively. For the quantity evaluation, the likelihood of each disease evaluated by the CNN classifier increased from 0 to 0.97, 0.89, 0.99, respectively. We observed that both of the visual scoring of the severity of disease patterns and the likelihood of each pulmonary disease pattern class increased according to the intensity of manipulation.

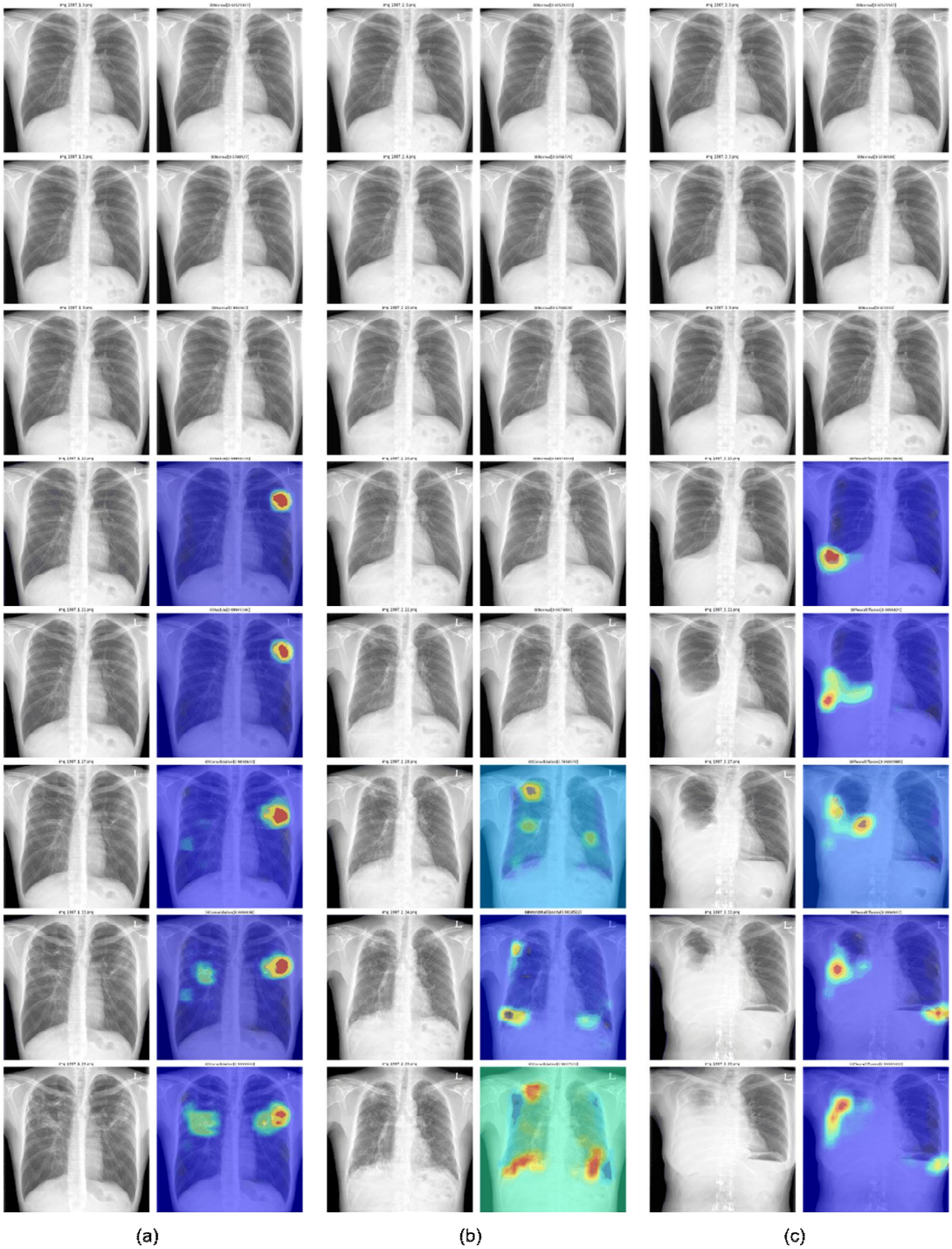

Figure 10. A visualization of CAM results on CXR image manipulation by moving along the pulmonary disease axis of consolidation, interstitial opacity, and pleural effusion.

3. Anomaly detection with PGGAN

To compare the performance of disease detection on CXR scans with that of AnoGAN model and that of preexisting supervised anomaly detection with convolutional neural net (CNN) model. Our approach to validate the detection performance is by a visual scoring by an expert radiologist. Given an unseen CXR image, AnoGAN will provide the most similar case with some differences. Ideally, the difference map with the highest gap with be localized as representing anomaly patterns, since the anomalous region could not possibly be learned in the training of normal samples. Material and methods. Using a modified AnoGAN model, PGGAN-trained generator yields corresponding synthetic normal images for given query images by minimizing mean squared error between the query and the synthetic images. Test set consisted of 100 CXR scans to validate and compare the models. The 100 examinations were composed of 12 classes as follows: nodule (n=8), calcification (n=9), consolidation (n=6), interstitial opacity (n=9), atelectasis (n=9), mediastinal widening (n=6), pleural effusion (n=8), pneumothorax (n=9), rib fracture (n=10), pneumomediastinum (n=10), subcutaneous emphysema (n=6), and pneumoperitoneum (n=10). AnoGAN model and 6-class CNN model (6-class: normal, nodule, consolidation, interstitial opacity, pleural effusion, and pneumothorax) were used for disease detection of the images. For these 100 query images, AnoGAN model generated normal fake image most similar to that image and then detected disease by subtraction of two images. After the test set, one board certified cardiothoracic radiologist reviewed the images to evaluate the model performance. Results of 100 CXR scans, 90 diseases were detected using AnoGAN model as follows: Nodule 100.0% (8/8), calcification 100.0% (9/9), consolidation 100.0% (6/6), interstitial opacity 100.0% (9/9), atelectasis 88.9% (8/9), mediastinal widening 83.3% (5/6), pleural effusion 87.5% (7/8), pneumothorax 100.0% (8/9), rib fracture 80.0% (8/10), pneumomediastinum 60.0% (6/10), subcutaneous emphysema 83.3% (5/6), and pneumoperitoneum 100.0% (10/10). The 6-class CNN model detected 37 of 40 diseases for trained classes (92.5%) whereas that detected only 14 of 60 disease for untrained classes (23.3%). Overall detection rate was higher in AnoGAN model than that of 6-class CAD model (90.0% vs. 51.0%, $p < 0.001$).

Table 6. Results of AnoGAN with a various disease samples with the metric of sensitivity

Index	Diagnosis	Number of Subjects	Num of Subjects Detected by AnoGAN	Sensitivity
1	Nodule*	8	8	100.0%
2	Calcification	9	9	100.0%
3	Consolidation*	6	6	100.0%
4	Interstitial opacity*	9	9	100.0%
5	Atelectasis	9	8	88.9%
6	Mediastinal widening	6	5	83.3%
7	Pleural effusion*	8	7	87.5%
8	Pneumothorax*	9	8	88.9%
9	Rib fracture	10	8	80.0%
10	Pneumomediastinum	10	6	60.0%
11	Subcutaneous emphysema	6	5	83.3%
12	Pneumoperitoneum	10	10	100.0%
Total		100	90	90.0%

Discussion

Medical image synthesis

Questions have been raised as to whether the utilization of GAN-generated synthetic images in medical imaging is acceptable. Although many researches have been conducted based on GAN-based augmentation in medical domain[16,38-42], the metric of validating whether GAN-generated synthetic images have enough information was less explored. Our study focused on measuring how well the GAN generator generates synthetic CXR images towards the use of those in downstream tasks, especially classification. The experimental results have shown that a CNN classifier trained on PGGAN-derived synthetic dataset has only slight superiority in performance compared to the one trained on real CXR dataset. One possible interpretation of the results could be that GAN-generated CXR images are so realistic that even a classifier trained on synthetic dataset can learn anatomical variations with pulmonary abnormalities comparable to the one with real dataset. For better explanations, we additionally analyzed what regions of a given CXR image influenced the CNN classifier the most when predicting. Figure 11. represents Gradient-weighted Class Activation Mapping (Grad-CAM)[43] results of

the classifiers trained on real and synthetic datasets on testing. The first row represents a sample of normal case (left) and the second row represents a sample of pneumothorax case with ground truth annotation (left). The second and third columns represent the respective Grad-CAM results (center and right) when tested on the classifiers trained on real and synthetic datasets. In both cases, the activated regions from each classifier have anatomical consistency. These visual explanations can be a strong indication of radiological information preserved in GAN-generated CXR images.

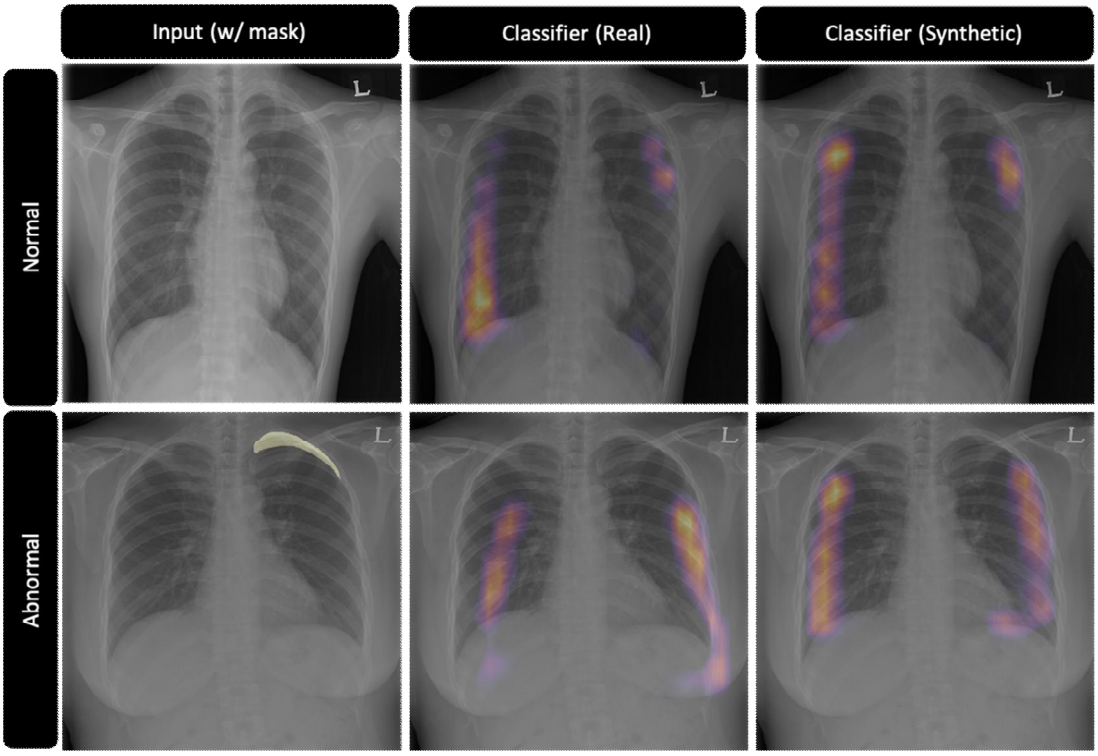

Figure 11. A Visualization of Grad-CAM results on false negative and false positive cases.

We derived a quantitative measure of image fidelity by exploiting a CNN-based classification network trained on PGGAN-generated synthetic images. By using AUROC comparison, we implemented a metric that has statistical validations. This has benefits compared to visual scoring test, IS, FID, PPL which does not have any statistical validations. Additionally, unlike FID, which uses an ImageNet-pretrained network, we can utilize the semantic features trained on the medical domain; these will measure the anatomical fidelity instead of general features of natural images.

Leveraging the high-fidelity image synthesis, we moved forward the controllable image synthesis in the medical domain. Our study aimed at CXR image synthesis with desired pulmonary disease patterns by disentangling the latent space of PGGAN. Visual scoring of our results has demonstrated that the control of CXR images with a specific pulmonary abnormality was plausible as well as moving towards the axis represent the severity of the disease patterns. In addition, we adapted our suggested evaluation metric for validating the manipulated CXR images with pulmonary abnormalities. The results have shown that visual scoring and the confidence score of a softmax output have an agreement in terms of disease severity.

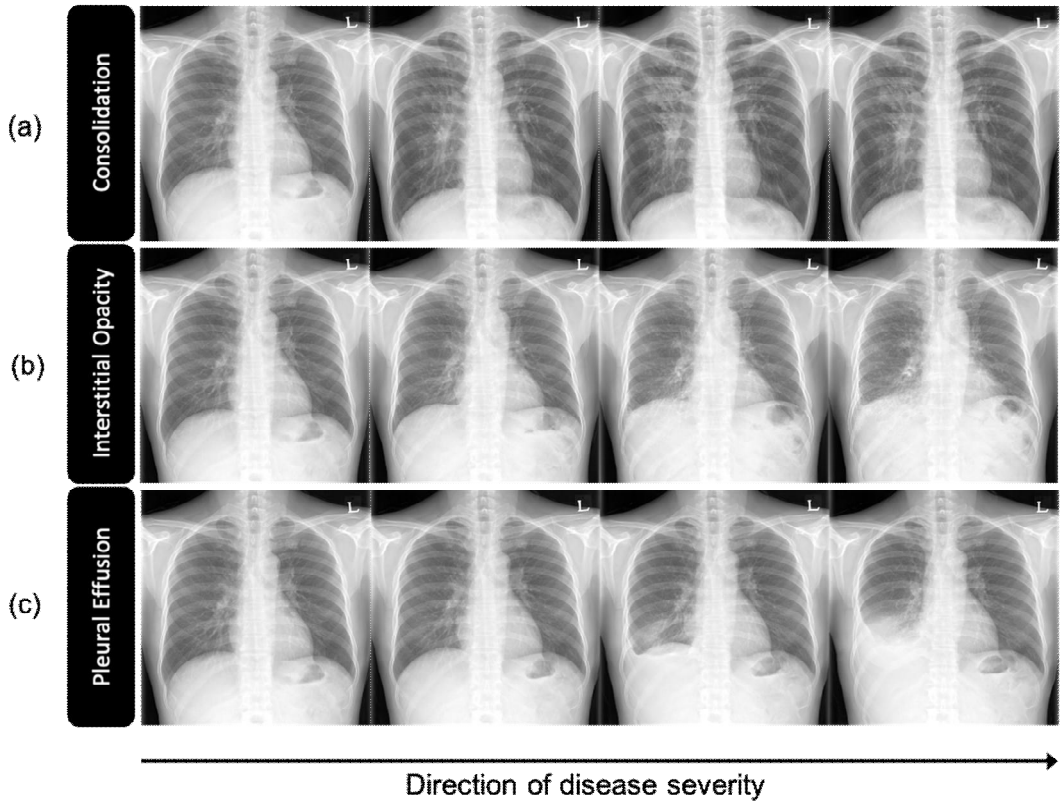

Figure 12. Examples of synthetic image manipulation using disentangled feature axes.

Despite the controversy, GAN-generated synthetic images are still beneficial in downstream tasks in medical imaging in the following three viewpoints. First, patient privacy can be resolved when constructing datasets. Second, GAN-generated synthetic images can be useful to improve the generalizability of deep learning models, as the training of deep learning

models heavily relies on the quality and quantity of dataset. However, existing data augmentation techniques are manually designed, e.g. rotation, flipping, color jittering, and can not cover the whole variation of the data. GANs can allow us to sample the training data distribution which offers more flexibility in augmenting the training data as shown in Figure 13. PGGAN is able to generate high resolution realistic images with unprecedented level of details. This could be readily applied to CXR dataset to generate images with pulmonary abnormalities that has insufficient number of cases. Third, another potential of GANs will be in synthesizing uncommon cases. Through conditional image synthesis with disentangling the latent space, many issues driven from imbalanced dataset or rare cases can be detoured.

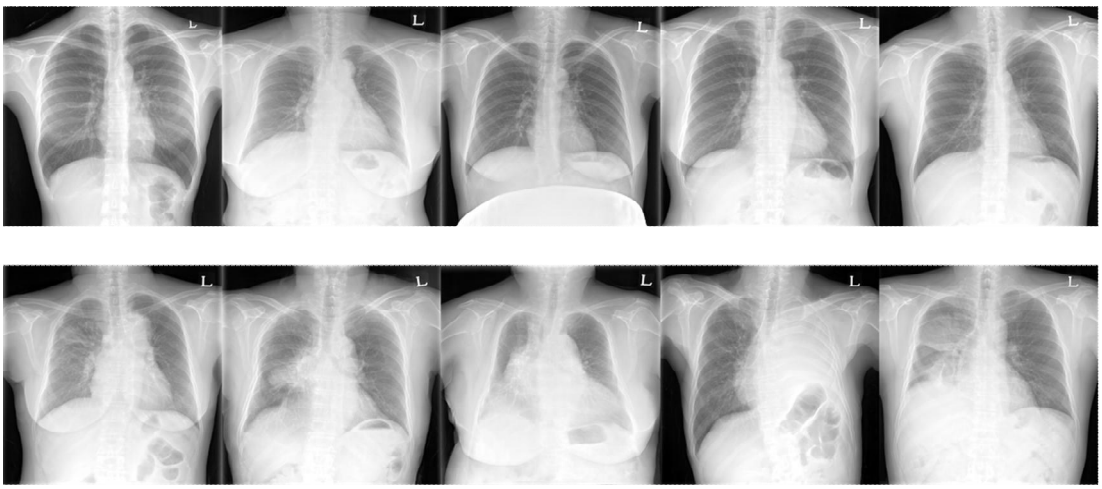

Figure 13. Examples of normal (top) and abnormal (bottom) cases of PGGAN-generated CXR images.

In short, our study concerning medical image synthesis suggested two points: (1) PGGAN can generate high-fidelity CXR images preserving radiologic information so that can be utilized for possible applications in medical imaging and (2) the disentanglement of PGGAN enabled CXR image synthesis with control of pulmonary disease features. Not only our method and result imply GAN can generate realistic images on image information level which can be used in data augmentation, anomaly detection, and other various fields, but also show performance of GAN can be measured with classifier with statistical manner.

Alongside many promising results of GANs, there are limitations in our study. First, although our evaluation metric for GAN generator is practically useful, there is still in need

of confirmation of medical experts to be applicable to diagnostic assistance in clinical settings. Second, the feature axes of nodule and pneumothorax were not discovered, which needs more exploration in the latent space of PGGAN.

Unsupervised Anomaly Detection

Unsupervised anomaly detection has drawn many attentions with the rise of GAN-based applications. Our study is to explore the anomaly detection system in CXR images with high-quality image synthesis using PGGAN. Figure 14. and Figure 15. show each result of unsupervised and supervised methods in detecting pulmonary abnormal patterns. In Figure 14. when unseen data with abnormality is given, the unsupervised anomaly detection system can detect and localize the region of interest in positive residual difference map. Our study suggested that the unsupervised scheme of anomaly detection has potentials with unseen data with rare disease patterns, which could not sensitively be detected with a supervised scheme. By highlighting possible anomalous regions in CXR images, physicians or radiologists could use this system as a complementary or a second opinion in diagnosis. However, there are several weaknesses of this system. The optimization process is relatively time-consuming compared to the supervised system since it requires an iterative algorithm. Also, false positive maps should be handled. In Figure 15., false positive maps are found when the unseen data without abnormality is given. The false positive region makes difficult to recognize anomalous patterns in lung regions and set a threshold to differentiate between normal and abnormal. Although PGGAN can generate high-quality data and optimization method can successfully discover the matching latent vector, additional post processing is required to obtain accurate locations or boundaries.

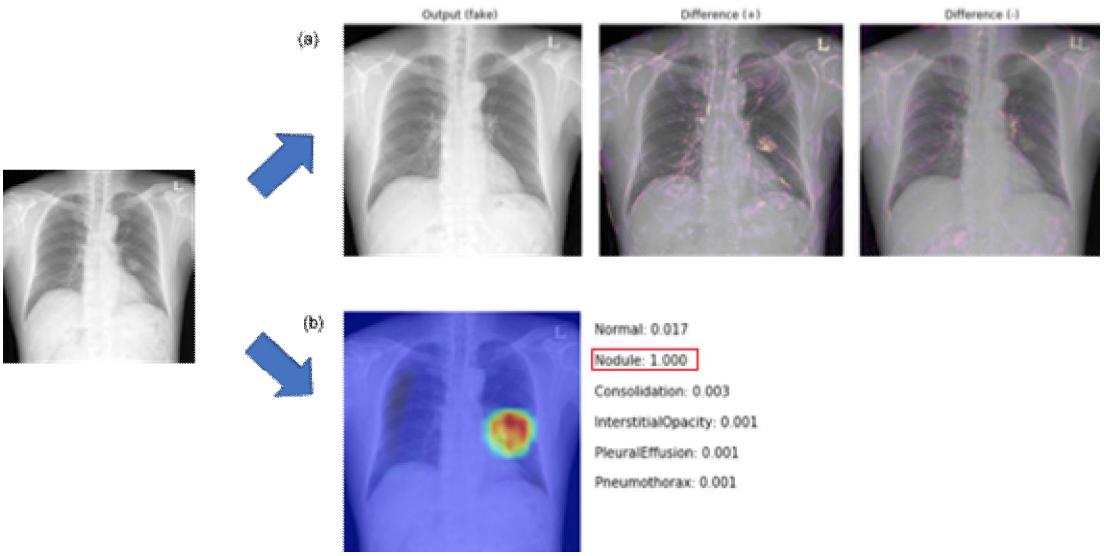

Figure 14. A visual comparison of the prediction results of abnormal case using unsupervised and supervised methods in anomaly detection.

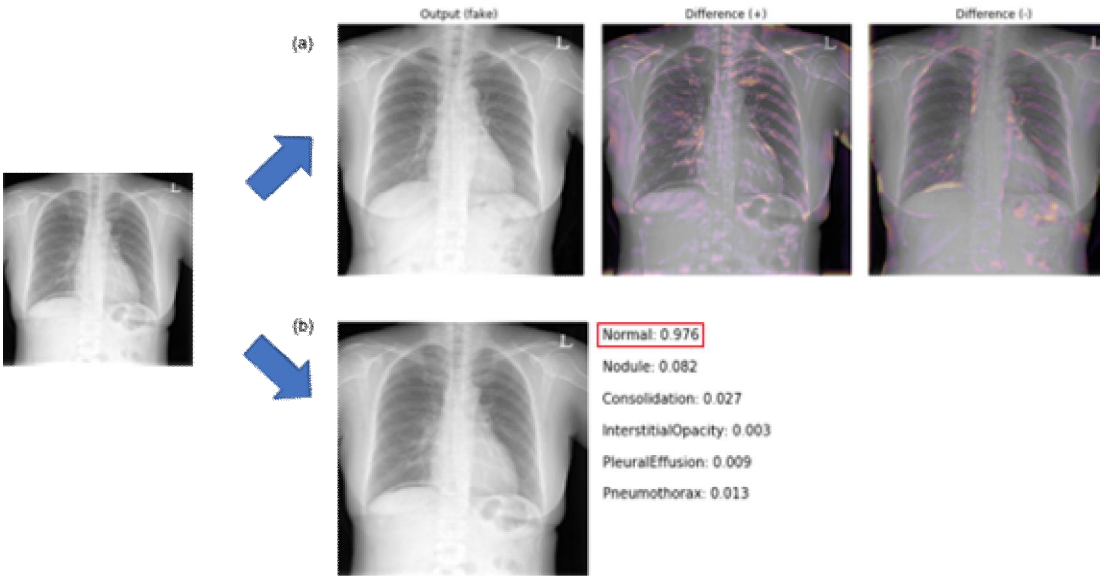

Figure 15. A visual comparison of the prediction results of normal case using unsupervised and supervised methods in anomaly detection.

Conclusion

In this study, we have explored GANs, especially PGGAN, in the task of the evaluation of the GAN generator, control of image synthesis, and anomaly detection on CXR images. We suggested an unsupervised method that uses GANs and CNN classifiers for CXR image generation and its quantitative evaluation. Additionally, we suggested an unsupervised

anomaly detection in CXR images using GANs with an iterative optimization. Results have demonstrated that PGGAN-generated synthetic CXR images can be utilized for downstream tasks such as data augmentation. Furthermore, disentanglement of the latent space of PGGAN can be exploited to generate insufficient data with rare disease patterns. Lastly, detection of pulmonary abnormalities in CXR images was feasible in an unsupervised manner without the need for data unusual or rare patterns. Our method has shown the potential of utilizing GANs in various applications in medical imaging, bypassing patient privacy, data imbalance, and data deficiency issues.

References

1 Goodfellow, I. *et al.* Generative Adversarial Networks. *Advances in Neural Information Processing Systems* **3** (2014).

2 Schlegl, T., Seeböck, P., Waldstein, S., Schmidt-Erfurth, U. & Langs, G. *Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery*. (2017).

3 Karras, T., Aila, T., Laine, S. & Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. (2017).

4 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. *Scientific Reports* **9**, 15352, doi:10.1038/s41598-019-51832-3 (2019).

5 Rajpurkar, P. *et al.* CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. (2017).

6 Wang, X. *et al.* ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. *arXiv:1705.02315* (2017).

7 Sayyouh, M., Hadjiiyski, L., Chan, H.-P. & Agarwal, P. Lung Nodule: Imaging Features and Evaluation in the Age of Machine Learning. *Current Pulmonology Reports* **8**, doi:10.1007/s13665-019-00229-8 (2019).

8 Almotairi, S., Kareem, G., Aouf, M., Almutairi, B. & Salem, M. A.-M. M. Liver Tumor Segmentation in CT Scans Using Modified SegNet. *Sensors* **20**, 1516,

doi:10.3390/s20051516 (2020).

9 Sun, Y. & Shi, C. Liver Tumor Segmentation and Subsequent Risk Prediction Based on Deeplabv3+. *IOP Conference Series: Materials Science and Engineering* **612**, 022051, doi:10.1088/1757-899X/612/2/022051 (2019).

10 Hauptmann, A., Arridge, S., Lucka, F., Muthurangu, V. & Steeden, J. Real-time Cardiovascular MR with Spatio-temporal Artifact Suppression using Deep Learning - Proof of Concept in Congenital Heart Disease. *Magnetic Resonance in Medicine*, doi:10.1002/mrm.27480 (2018).

11 Lee, D., Yoo, J., Tak, S. & Ye, J. C. Deep Residual Learning for Accelerated MRI Using Magnitude and Phase Networks. *IEEE Transactions on Biomedical Engineering* **PP**, 1-1, doi:10.1109/TBME.2018.2821699 (2018).

12 Schlemper, J., Caballero, J., Hajnal, J., Price, A. & Rueckert, D. A Deep Cascade of Convolutional Neural Networks for Dynamic MR Image Reconstruction. *IEEE Transactions on Medical Imaging* **PP**, doi:10.1109/TMI.2017.2760978 (2017).

13 Demner-Fushman, D. *et al.* Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association : JAMIA* **23**, doi:10.1093/jamia/ocv080 (2015).

14 Jaeger, S. *et al.* Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery* **4**, 475-477, doi:10.3978/j.issn.2223-4292.2014.11.20 (2014).

15 Chuquicusma, M. J., Hussein, S., Burt, J. & Bagci, U. in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. 240-244 (IEEE).

16 Han, C. *et al.* in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 734-738 (IEEE).

17 Szegedy, C. *et al.* in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 1-9.

18 Karras, T., Laine, S. & Aila, T. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 4401-4410.

19 Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O. & Gelly, S. in *Advances in Neural Information Processing Systems.* 5228-5237.

20 Karras, T. *et al.* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 8110-8119.

21 Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. in *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2818-2826.

22 Deng, J. *et al.* in *2009 IEEE conference on computer vision and pattern recognition.* 248-255 (Ieee).

23 Karras, T., Aila, T., Laine, S. & Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).

24 Chen, Y., Lai, Y.-K. & Liu, Y.-J. *CartoonGAN: Generative Adversarial Networks for Photo Cartoonization*. (2018).

25 Karras, T., Laine, S. & Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PP**, 1-1, doi:10.1109/TPAMI.2020.2970919 (2020).

26 Shaham, T., Dekel, T. & Michaeli, T. *SinGAN: Learning a Generative Model From a Single Natural Image*. (2019).

27 Chen, X., Houthooft, R., Schulman, J., Sutskever, I. & Abbeel, P. *InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets*. (2016).

28 Guan, S. *TL-GAN: transparent latent-space GAN*, <https://github.com/SummitKwan/transparent_latent_gan> (2018).

29 Mirza, M. & Osindero, S. Conditional Generative Adversarial Nets. (2014).

30 Nie, W. *et al.* Semi-Supervised StyleGAN for Disentanglement Learning. (2020).

31 Odena, A., Olah, C. & Shlens, J. Conditional Image Synthesis With Auxiliary Classifier GANs. (2016).

32 Shen, Y., Yang, C., Tang, X. & Zhou, B. *InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs*. (2020).

33 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. *Scientific reports* **9**, 1-9 (2019).

34 He, K., Zhang, X., Ren, S. & Sun, J. in *Proceedings of the IEEE conference on*

computer vision and pattern recognition. 770-778.

35 Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29-36 (1982).

36 Arjovsky, M., Chintala, S. & Bottou, L. Wasserstein GAN. (2017).

37 Radford, A., Metz, L. & Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. (2015).

38 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).

39 Frid-Adar, M. *et al.* GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing* **321**, 321-331 (2018).

40 Mirsky, Y., Mahler, T., Shelef, I. & Elovici, Y. in *28th {USENIX} Security Symposium ({USENIX} Security 19)*. 461-478.

41 Lee, D., Kim, J., Moon, W.-J. & Ye, J. C. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2487-2496.

42 Bi, L., Kim, J., Kumar, A., Feng, D. & Fulham, M. in *molecular imaging, reconstruction and analysis of moving body organs, and stroke imaging and treatment* 43-51 (Springer, 2017).

43 Rs, R., Cogswell, M., Vedantam, R., Parikh, D. & Batra, D. *Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization*. (2017).

Abstract (in Korean)

최근 주목 받고 있는 인공지능 기술 중 하나인 딥러닝은 수많은 컴퓨터 비전 작업에서 유망한 결과를 보여주었다. 컨볼루션 신경망(CNN)이 발전함에 따라 병변 패턴의 분류 및 감지, 장기의 자동 분할, 의료 이미지 재구성 등과 같은 의료 이미징에 딥러닝이 빠르게 채택되었다. 전통적인 방법인 지도 학습을 위해서는 고품질 이미지 및 정확한 주석이 필요하다. 그러나 고품질 데이터 세트를 구성하는 것은 의학 분야에서 어렵다. 제한된 데이터 액세스, 불균형 데이터 세트 및 값 비싼 주석 과정은 딥러닝의 예측 능력을 제한하고 편향된 결과를 유도 할 수 있다.

고품질 의료 이미지 데이터 세트에 대한 요구가 증가하는 가운데 생성적 적대 신경망(GAN)의 출현은 새로운 돌파구가 되었다. GAN 은 적대적인 학습 과정을 통해 기존 데이터 세트에서 그럴듯한 새로운 샘플을 생성하는 능력을 배운다. GAN 은 도메인 적응, 초해상화, 이미지 대 이미지 변환, 이미지 스타일 전송, 이상 탐지 등 다양한 작업에서 잠재력을 입증했다.

GAN 의 유망한 결과에도 불구하고, 의료 분야에서는 적은 수의 연구만이 진행되었다. 이 연구에서는 점진적으로 성장하는 생성적 적대 신경망 (PGGAN)을 의료 영상, 특히 흉부 X 선 (CXR) 영상에 응용하기 위한 몇 가지 비지도적 학습 방법을 제안한다. 다루는 주제는 다음과 같다. (a) PGGAN 에서 생성된 합성 CXR 영상의 충실도 평가하고, (b) (a)에서 학습한 PGGAN 의 잠복 공간에서 의미론적 표현을 풀어내어 원하는 폐질환 패턴을 가진 합성 CXR 영상을 생성하며, (c) PGGAN 을 사용하여 비지도적 방식으로 CXR 영상의 이상 패턴을 식별하는 이상 감지 시스템을 개발한다.

(a)의 첫 번째 주제에서는 딥러닝 기반 분류 네트워크 (분류기)를 활용한 3 단계 방법을 제안한다. 단계 1 및 2 에서는 실제 CXR 이미지와 합성 CXR 이미지에 대해 별도로 학습한 두 분류기의 성능을 비교한다. 단계 3 에서는 실제 CXR 영상으로 구성된 동일한 테스트 데이터 세트에서 이진 분류(정상 또는 비정상) 성능을 평가한다. PGGAN 에서 생성된 합성 CXR 영상이 실제 영상에 준하는 방사선 정보를 보존한다는 것을 발견하였다. (b)의 두 번째 주제에서는 PGGAN 의 잠복 공간에서 미리 정의된 폐질환 패턴의 의미론적 표현을 탐색하고 발견한다. 간단한 선형 회귀를 통해 원하는 질병 패턴을 가진 CXR 영상의 제어 가능한 생성이 가능함을 입증하였다. 평가는 20 년 이상의 경험을 가진 전문 방사선 전문의의 시각적 채점과 (a)에서 제안된 분류기를 사용하는 지표에 의해 정성적 및 정량적으로 각각 수행되었다. (c)의 세 번째 주제에서는 PGGAN 을 이

용하여 비지도 방법으로 CXR 영상의 이상 탐지 시스템을 제안한다. 비정상 CXR 샘플을 식별하기 위해 정상 CXR 데이터 세트로 PGGAN 을 훈련하였다. 이상이 있는 실제 CXR 영상이 주어지면 반복 알고리즘을 통해 잠재 벡터를 최적화하여 주어진 영상과 가장 유사한 정상 CXR 이미지를 근사 하였다. 평가에서 질병 주석이 필요없이 CXR 의 비정상적인 패턴을 민감하게 감지할 수 있음을 입증하였다.

이 연구는 의료 영상에서 비지도 딥러닝 기반 애플리케이션을 개발하는 데 있어 GAN 의 잠재력을 보여준다. CXR 영상에 대한 고 충실도 영상 생성, 제어 가능한 영상 생성 및 이상 감지의 각 결과를 활용하여, 환자의 개인 정보 보호, 불균형 데이터 세트 및 의료 영상의 값 비싼 주석과 같은 감독 학습의 기존 문제를 해결할 수 있다.