Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Informatics in Medicine Unlocked 27 (2021) 100779

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Data augmentation using Generative Adversarial Networks (GANs) for GAN-based detection of Pneumonia and COVID-19 in chest X-ray images

Saman Motamed [a,b,*], Patrik Rogalla [d], Farzad Khalvati [a,b,c]

[a] Institute of Medical Science, University of Toronto, Canada
[b] Department of Diagnostic Imaging, Neurosciences and Mental Health, The Hospital for Sick Children, Canada
[c] Department of Mechanical and Industrial Engineering, University of Toronto, Canada
[d] University Health Network, Toronto, Ontario, Canada

ARTICLE INFO

Keywords:
Data augmentation
Semi-supervised learning
Generative adversarial networks
Disease detection
Medical Imaging

ABSTRACT

Successful training of convolutional neural networks (CNNs) requires a substantial amount of data. With small datasets, networks generalize poorly. Data Augmentation techniques improve the generalizability of neural networks by using existing training data more effectively. Standard data augmentation methods, however, produce limited plausible alternative data. Generative Adversarial Networks (GANs) have been utilized to generate new data and improve the performance of CNNs. Nevertheless, data augmentation techniques for training GANs are underexplored compared to CNNs. In this work, we propose a new GAN architecture for augmentation of chest X-rays for semi-supervised detection of pneumonia and COVID-19 using generative models. We show that the proposed GAN can be used to effectively augment data and improve classification accuracy of disease in chest X-rays for pneumonia and COVID-19. We compare our augmentation GAN model with Deep Convolutional GAN and traditional augmentation methods (rotate, zoom, etc.) on two different X-ray datasets and show our GAN-based augmentation method surpasses other augmentation methods for training a GAN in detecting anomalies in X-ray images.

1. Introduction

In recent years, Convolutional Neural Networks (CNNs) have shown excellent results on several tasks using sufficient training data [1–3]. One of the main reasons for poor CNN performance and overfitting on training data remains limited-sized datasets in many domains such as medical imaging. Improving the performance of CNNs can be achieved by using the existing data more effectively. Augmentation methods such as random rotations, flips, and adding various noise profiles have been proposed [4,5] as some methods of augmentation. Typical data augmentation techniques use a limited series of invariances that are easy to compute however (rotation, flips, etc.), limited in the amount of new data they can generate.

Generative Adversarial Networks (GANs) [6] have been used for data augmentation to improve the training of CNNs by generating new data without any pre-determined augmentation method. Cycle-GAN was used to generate synthetic non-contrast CT images by learning the transformation of contrast to non-contrast CT images [7]. This improved the segmentation of abdominal organs in CT images using a U-Net model [8]. Using Deep Convolutional-GAN (DCGAN) [9] and Conditional-GAN [10] to augment medical CT images of liver lesion and mammograms showed improved results in classification of lesions using CNNs [11,12]. Data Augmentation GAN (DAGAN) [13] was able to improve the performance of basic CNN classifiers on EMNIST (images of handwritten digits), VGG-Face (images of human faces) and Omniglot (images of handwritten characters from 50 different alphabets) datasets by training DAGAN in a source domain and generating new data for the target domain. There has not been any study on data augmentation using GANs for training other GANs. The challenge with using a GAN to augment data for another GAN is that newly generated images with the trained generator of the GAN follow the same distribution as the training images, and hence there is no new information to be learned by another GAN that is trained on the original images combined with the newly generated (augmented) images.

In this paper, we propose Inception-Augmentation GAN (IAGAN) model inspired by DAGAN [13] for the task of data augmentation that specifically improves the performance of another GAN architecture. While a growing number of supervised Deep Learning models have achieved promising results in the diagnostic medical imaging domain, they require large amounts of labeled data to learn and generalize to classify diseases, such as Pneumonia and COVID-19, accurately. Using

* Corresponding author at: Institute of Medical Science, University of Toronto, Canada.
E-mail address: smotamed@andrew.cmu.edu (S. Motamed).

https://doi.org/10.1016/j.imu.2021.100779
Received 20 August 2021; Received in revised form 24 October 2021; Accepted 30 October 2021
Available online 22 November 2021

the covid-chestxray [14] and COVIDx [15] datasets, multiple studies have built supervised models to detect COVID-19 markers using X-ray images of the chest [15–20]. Wang et al.'s CNN-based COVID-NET [15] achieved a 93.3% test accuracy for multi-class classification on a test cohort of 100 Normal, 100 Pneumonia and 100 COVID-19 from the COVIDx dataset with the rest of the images of each class being used to train their model. Ozturk et al.'s DarkNet [16] carried out both multi-class classification (Pneumonia vs. COVID-19 vs. No Findings) and binary classification (COVID-19 vs. No Findings). They reported a multi-class classification with accuracy of 0.87% on 25 COVID-19, 100 Normal, and 100 Pneumonia images and binary classification accuracy of 98.08%. Hemdan et al.'s COVIDX-Net [18], comprised of multiple architectures such as DenseNet121, VGG19, and InceptionV3, was tested on 50 X-ray images from the covid-chest X-ray dataset. 25 COVID-19 negative and 25 COVID-19 positive. Their reported accuracy is anywhere between 50% (InceptionV3) to 90% (VGG19 and DenseNet201), for each investigated architecture. Afshar et al. used capsule networks to detect COVID-19 positive cases using COVIDx dataset. Their model was pre-trained using non-COVID chest X-ray images from other datasets. They report an Accuracy of 95.7%, Sensitivity of 90%, Specificity of 95.8%, and the area under the ROC curve (AUC) of 0.97. The number of test images from each class is not disclosed in their paper. A recent study by DeGrave et al. [21] showed the effects of supervised models, trained on imbalanced covid-chestxray [14] and COVIDx [15] datasets, demonstrating the overfitting of these models and failure to generalize to ocher datasets. With recent success of GANs in detecting anomalies in medical images [22,23] For these reasons, we explored data augmentation methods to improve the performance of GAN based networks.

We trained our proposed IAGAN on two chest X-rays datasets, one containing normal and pneumonia images and the other dataset containing normal, pneumonia and COVID-19 images. We showed that a trained IAGAN model can generate new X-ray images, independent of image labels, and improve the accuracy of generative models. We evaluated the performance of IAGAN model by training a DCGAN for anomaly detection (AnoGAN) [22] and showed improved results in classifying pneumonia and COVID-19 positive cases with improved area under the receiver operating characteristic (ROC) curve (AUC), sensitivity, and specificity. We showed our trained IAGAN is able to generate new domain specific data regardless of the class of its input images. This allowed for an unsupervised data augmentation, in the case of not having labels for a subset of the images in the dataset. By training the same DCGAN model on the augmented data using traditional augmentation methods and generating new data using another DCGAN for the task of augmentation, we showed the ineffectiveness of these methods in successful augmentation of data for training a generative model compared to our IAGAN for detecting pneumonia and COVID-19 images.

2. IAGAN architecture

Fig. 1 shows the architecture of the proposed IAGAN's Generator. At each iteration i, as input, the generator (**G**) takes a Gaussian noise vector z_i and a batch of real training image x_i. By encoding the input images x_i using convolution and attention layers to a lower-dimensional representation, before concatenating this representation of the image with the projected noise vector z_i (concatenation happens after z_i goes through a dense layer and non-linearity), we aim to not only use the full image representation using the discriminator, but also get a lower representation of images fed through the generator for better generalizability of **G** in generating images. The dual input to the generator also allows the trained generator to use images from different classes and generate a broader range of images to augment our specific training data class. The use of attention layers in GANs (Fig. 2) has shown to capture long-range dependencies in the image [24] where simple convolution layers focus on local features restricted by their receptive field, self-attention layers capture a broader range of features within the image. The attention layer uses three 1×1 convolutions. 1×1 convolution helps to reduce the number of channels in the network. Two of the convolution outputs, as suggested by Fig. 2, are multiplied (matrix multiplication) and fed to a *softmax* activation, which results in producing the attention map. The attention map acts as the probability of each pixel affecting the output of the third convolution layer. Feeding a lower-dimensional representation of an input image x allows for the trained generator to use images from different classes to produce similar never-before-seen images of the class it was trained on.

Using inception and residual architectures [25] increase GAN's ability to capture more details from training image-space without losing spatial information after each convolution and pooling layer. Making **G**'s network deeper is theoretically a compelling way to capture more details in the image, however deep GANs are unstable and hard to train [9,26]. A trained generator learns the mapping $G(z) : z \longmapsto x$ from latent space representations z to realistic, $2D$, chest X-ray images.

The discriminator (**D**) (Fig. 3) is a 4-layer CNN that maps a 2D image to a scalar output that can be interpreted as the probability of the given input being a real chest X-ray image sampled from training data or image G(z) generated by the generator G. Optimization of D and G can be thought of as the following game of minimax [6] with the value function $V(G, D)$:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim P\,data(x)}[\log D(x)] + \mathbb{E}_{z \sim P\,z(z)}[\log(1 - D(G(z)))] \quad (1)$$

During training, generator G is trained to minimize the accuracy of discriminator D's ability in distinguishing between real and generated images while the discriminator is trying to maximize the probability of assigning real training images the "real" and generated images from G, "fake" labels. During the training, G improves at generating more realistic images while D gets better at correctly identifying between real and generated images.

3. Datasets

3.1. Dataset I

We used the publicly available chest X-ray dataset [27] with two categories of Normal (1575 images) and Pneumonia (4265 images). The images were in JPEG format and varied in size with pixel values in $[0, 255]$ range. We resized all images to 128×128 pixels. Images were normalized to have $[-1, 1]$ range for $\tanh$ non-linearity activation in the IAGAN architecture. We use our bigger cohort (pneumonia) as the training class. 500 images from each class were randomly selected to evaluate the models' performance while the rest of the images were used for augmentation and training different models.

3.2. Dataset II

Covid-chestxray dataset [14] is an ongoing effort by Cohen et al. to make a public COVID-19 dataset of chest X-ray images with COVID-19 radiological readings. Wang et al. used covid-chestxray dataset, along with four other publicly available datasets and compiled the COVIDx [15] dataset. With the number of images growing, many deep learning models are trained and tested on this public dataset [15,16,18]. At the time of this study, the COVIDx dataset is comprised of 8066 normal, 5559 pneumonia, and 589 COVID-19 images. The images are in RGB format with pixel range of $[0, 255]$ and have various sizes. To train the generative models in this study, all images were converted to gray scale, resized to 128×128 pixels and normalized to have pixel intensities in the $[-1, 1]$ range. 589 images from normal and pneumonia classes were randomly selected along with 589 COVID-19 images to test the models while the rest of the images were used for augmentation and training different models.

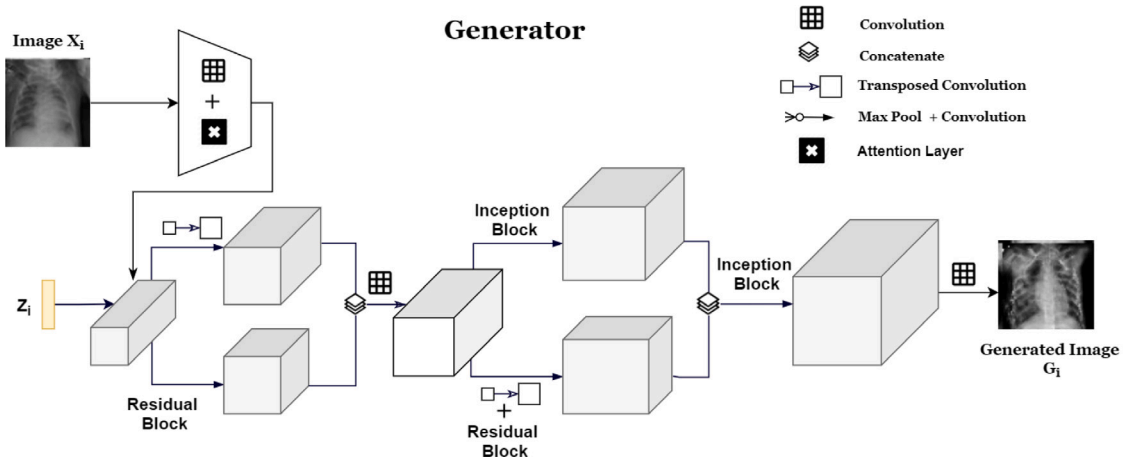

Fig. 1. IAGAN's generator architecture.

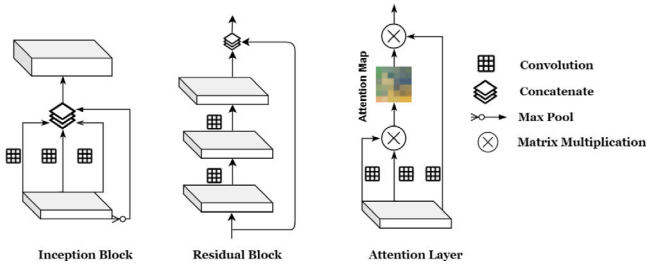

Fig. 2. IAGAN's generator specific architecture breakdown.

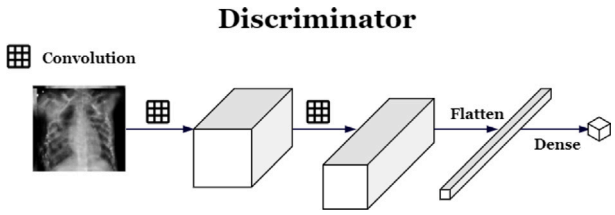

Fig. 3. Discriminator architecture.

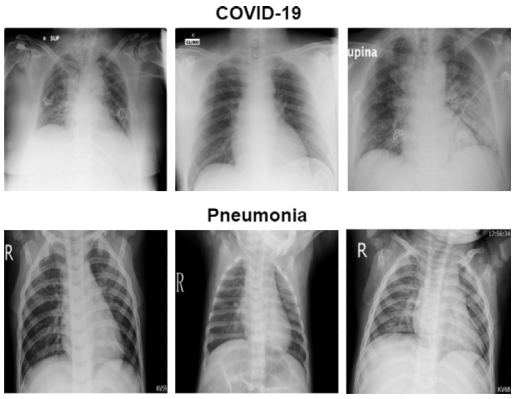

Fig. 4. Pneumonia and COVID-19 sample images from COVIDx dataset with class consistent annotations.

3.2.1. Segmentation of COVIDx dataset

A recent study [21] using the COVIDx dataset showed that existing markers such as annotations and arrows outside of the lung on the X-ray images can act as shortcuts [28] in detecting COVID-19 using those shortcuts instead of actual COVID-19 disease markers. Fig. 4 shows annotations on the top left of COVID-19 images which are consistent with the rest of the COVID-19 images and the *R* symbol positioned on the left of pneumonia images consistent with images from the pneumonia class in COVIDx dataset.

To mitigate the effect of non-disease markers on our model, we segmented the lungs for the COVIDx dataset images. 900 randomly selected images (300 from each class) were manually segmented by an expert radiologist. A modified U-NET model [29], pre-trained on the Montgomery chest X-ray dataset [30] was fine-tuned using the 800 COVIDx segmentations. The segmentation model was tested on the 100 remaining ground truth images and achieved a Sørensen–Dice coefficient of 0.835.

4. Data augmentation

4.1. IAGAN

We trained multiple instances of IAGAN outlined below. The architecture of IAGAN was kept unchanged for each instance and learning rates of 0.0004 and 0.0001 were used for the discriminator and generator, respectively. Experimenting with the size of the Gaussian noise vector z showed 120 to be the optimal size. We trained our IAGAN for 250 epochs on an Nvidia GeForce RTX 2080 Ti - 11 GB with a batch size of 32. For dataset I, IAGAN was trained on 3765 pneumonia images and tested on 500 pneumonia vs. 500 normal cases. For dataset II, one IAGAN was trained on 4700 Pneumonia images and one IAGAN was trained on 7477 Normal images. After successful training of the IAGAN, the generator has learned the distribution of the images of the training class.

To generate new data, for each input image to IAGAN, 3 random noise vectors were initiated and 3 new images were generated from the generator. For dataset I, 3765 pneumonia training images were put through G and for each image, three new images were generated (11,295). For each normal image that was not used for testing the model's performance, we did the same and generated 3225 images from 1075 normal images. Similarly, for dataset II, normal and pneumonia training images were put through the two trained generators, one

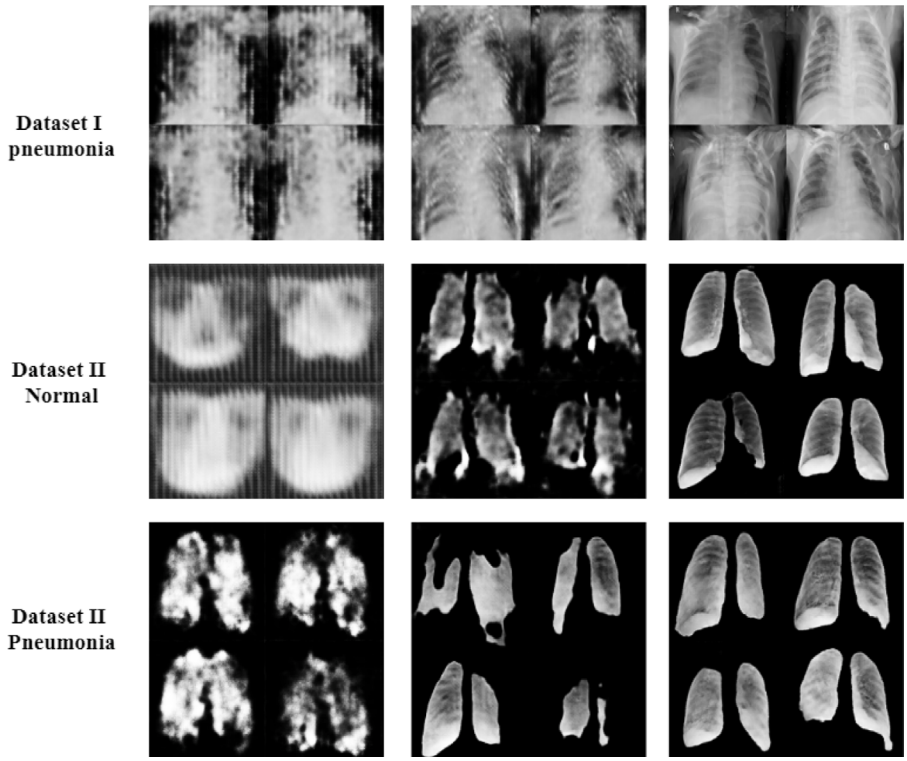

Fig. 5. Generator's output during training.

Table 1
IAGAN augmentation.

	Normal (Training/Test)	Pneumonia (Training/Test)	COVID-19 (Training/Test)
Dataset I	0/500	3,765/500	N/A
Augmented Dataset I	0/500	19,360/500	N/A
Dataset II	7,477/589	4,700/589	0/589
Augmented Dataset II	48,708/589	48,708/589	0/589

Table 2
DCGAN augmentation.

	Normal (Train/Test)	Pneumonia (Train/Test)	COVID-19 (Train/Test)
Augmented Dataset I	0/500	15,060/500	N/A
Augmented Dataset II	29,908/589	18,800/589	0/589

generator from the IAGAN trained on normal images and one trained on pneumonia images. Similar to dataset I, each generator generated 3 new images using pneumonia and normal images that are not used in testing the model. Fig. 5 shows the generator's output at early, mid and later stages (from left to right respectively) of the training on datasets I and II.

Table 1 shows the number of images for each class, before and after data augmentation using IAGAN. Dataset I does not have any COVID-19 images and does not use any normal images for training. Dataset II uses all COVID-19 images (589) for testing the model and hence, no augmentation is done using this class. Both normal and pneumonia class images are used for training the model and therefore, 589 randomly selected images are fixed to test the model from each class, the rest of the images are augmented using two separately trained IAGANs. One IAGAN trained on normal images, uses normal and pneumonia images to generated more normal images. The other IAGAN, uses normal and pneumonia images to generate more pneumonia images.

4.2. DCGAN

To understand the effect of our input image to IAGAN's generator, which allows using images from all classes to be fed into a trained generator for augmentation, we trained a DCGAN [9] that uses only the traditional Gaussian noise vector input to the generator. We used the same hyper-parameters and number of epochs as IAGAN. The only difference in the number of generated images is that images from classes other than what the DCGAN's Generator was trained on cannot be fed to the trained **G** for generating new images. For this reason, we generate 3 images for each image the DCGAN was trained on; for dataset I, 3 images were generated for each pneumonia training image (3 similar images were generated using the anomaly score defined by Schlegl et al. [22] and for dataset II, two DCGANs were trained similar to IAGAN, 3 images were generated for each normal training image with the **G** trained on normal images and 3 images were generated for each pneumonia training images with the **G** trained on pneumonia images. Table 2 shows the number of images for each class, before and after data augmentation using DCGAN .

4.3. Traditional augmentation

Based on recent literature on data augmentation for chest X-ray pathology classification using CNNs [31], we used Keras' data generator function for data augmentation by using random rotations in the range of 20 degrees, width and height shift in the range of 0.2 and zoom in the range of 0.2. For each training image, 8 new images were randomly generated using the aforementioned augmentation methods. Fig. 6 shows the sample output of this function. Table 3 shows the number of images for each class, before and after data augmentation using traditional augmentation methods.

5. Experiments

Schlegl et al. [22] proposed AnoGAN for detecting anomalies in optical coherence tomography images of the retina. The AnoGAN architecture follows DCGAN [9] in terms of overall generator and discriminator

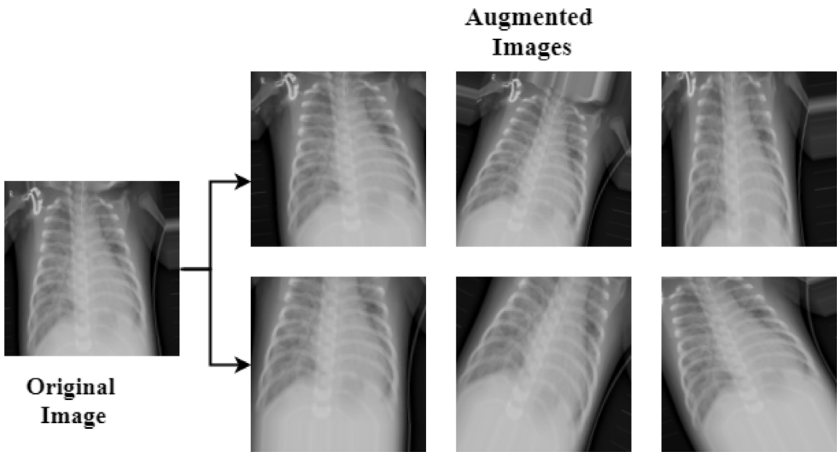

Fig. 6. Traditional augmentation output sample.

Table 3
Traditional augmentation.

	Normal (Train/Test)	Pneumonia (Train/Test)	COVID-19 (Train/Test)
Augmented Dataset I	0/500	33,885/500	N/A
Augmented Dataset II	67,293/589	42,300/589	0/589

architecture. They trained the AnoGAN model on one class of images. With the trained generator **G** at hand, in order to find anomalies in test image x, back-propagation (using Eq. (4) with $\lambda = 0.2$) was used to find a point z_i that generates an image that looks similar to x. Upon finding a point z after a set number of iterations (800 iterations in our experiments), the anomaly score $A(x)$ (Eq. (5)) is defined using residual and discrimination losses as shown below, calculated at point z. L_R and L_D are the residual and discriminator loss that enforce visual and image characteristic similarity between real image x and generated image $G(z_i)$. The discriminator loss captures image characteristics using the output of an intermediate layer of the discriminator, $f(.)$, making the discriminator act as an image encoder.

$$\mathcal{L}_R(z_i) = \sum |x - G(z_i)| \tag{2}$$

$$\mathcal{L}_D(z_i) = \sum |f(x) - f(G(z_i))| \tag{3}$$

$$\mathcal{L}(z_i) = (1-\lambda) \times \mathcal{L}_R(z_i) + \lambda \times \mathcal{L}_D(z_i) \tag{4}$$

$$A(x) = (1-\lambda) \times \mathcal{L}_R(z) + \lambda \times \mathcal{L}_D(z) \tag{5}$$

5.1. Identifying information statement

Datasets used in this study are publicly available and have been anonymized to protect any identifying patient information.

5.2. Dataset I

We used the AnoGAN architecture to evaluate the effects of different approaches to data augmentation. We trained 4 AnoGAN models; one trained on pneumonia images from dataset I and the other 3 were trained on augmented pneumonia images with IAGAN, DCGAN and traditional augmentation methods.

5.3. Dataset II

To detect COVID-19 positive from COVID-19 negative images, one AnoGAN was trained on normal images and another identical network was trained on pneumonia images. After calculating two anomaly scores for each test image, one calculated by each AnoGAN, the sum of two anomaly scores was assigned as the final anomaly score for the test image. The idea is that the AnoGAN trained on normal images will result in lower anomaly score for normal images during test while AnoGAN trained on pneumonia images results in lower scores for pneumonia images. In both networks, the COVID-19 images produce higher anomaly scores hence the COVID-19 final anomaly score will be higher than the normal and pneumonia classes.

The AnoGAN pair model were trained similar to AnoGAN on dataset I; trained on normal and pneumonia training images without augmentation, normal and pneumonia images augmented using IAGAN, DCGAN and traditional augmentation methods.

6. Results

We calculated the area under the ROC curve (AUC) for each model trained on datasets I and II, before and after data augmentation. For dataset I, AUC represents the classification capability of detecting pneumonia vs. normal chest X-rays. For dataset II, we classify COVID-19 positive from COVID-19 negative images. With 589 test images from each class (normal, pneumonia and COVID-19) in dataset II, we calculated the AUC for the balanced COVID-19 negative class vs. COVID-19 positive test images. The balanced COVID-19 negative class was created by randomly sampling 294 normal and 295 pneumonia images from 589 normal and 589 pneumonia test images.

Table 4 shows the calculated AUC for datasets I and II. It can be seen that our proposed IAGAN augmentation method outperforms all other three models for both Dataset I and II: no augmentation, DCGAN, and traditional augmentation methods. DeLong test [32] was used to compare the AUC of the models by calculating the *p-value* for significance difference. The *p-values* are added next to the AUC of each augmentation method and measures the significance of the model compared to the model trained with no augmentation.

We calculated the accuracy of each model at the highest sensitivity/specificity pair points (with minimum 0.80 sensitivity and specificity) for each model trained on datasets I and II. Table 5 shows the sensitivity, specificity and accuracy of different trained models on both datasets where it can be seen that our proposed IAGAN outperforms all other models in both sensitivity and specificity.

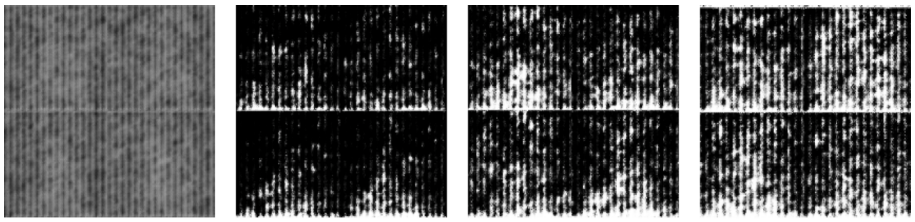

Fig. 7. IAGAN's generator output at different epochs of the model training with random generated input images.

Table 4
AUC and *p*-value for datasets I and II.

	No augmentation	IAGAN	DCGAN	Traditional augmentation
Dataset I	0.87	**0.90** ($p = 3.17 \times 10^{-7}$)	0.87 ($p = 0.5$)	0.88 ($p = 0.08$)
Dataset II	0.74	**0.76** ($p = 0.01$)	0.75 ($p = 0.43$)	0.75 ($p = 0.57$)

Table 5
Sensitivity, Specificity and Accuracy for datasets I and II, respectively.

Model (Datasets I/II)	Sensitivity	Specificity	Accuracy
No augmentation	0.80/0.67	0.81/0.68	0.80/0.67
IAGAN	**0.82/0.69**	**0.84/0.69**	0.80/**0.69**
DCGAN	0.80/0.67	0.81/0.67	0.80/0.67
Traditional augmentation	0.80/0.68	0.81/0.68	0.80/0.68

7. Discussion

Harnessing GANs' ability to generate never-before-seen data, by learning the distribution of images, allows for augmentation of data that is not limited to applying different transformations to existing images. By using the proposed IAGAN, not only are we able to generate new images for the same class used to augment data (e.g., using normal images to augment normal dataset), but also generate new images of any class withing that domain of images using one class of images (e.g., generating chest X-rays with pneumonia, COVID-19 or healthy cases using normal images).

We showed that a traditional DCGAN with a single random noise vector input to the generator fails to effectively augment data for a GAN. Traditional augmentation methods showed improved prediction in a subset of the tasks (AUC of 0.75 vs 0.74 for dataset II), yet failed to effectively improve the accuracy of the overall models with statistical significance. Our proposed IAGAN architecture, however, improves the models' accuracy when used for augmentation of the training cohort, with statistical significance. We used the AnoGAN [22] architecture to show when the training data is augmented using our proposed IAGAN method, the AUC improves by 3% and 2%, compared to no augmentation, for dataset I and II, respectively. IAGAN also showed improved sensitivity/specificity for the AnoGAN model (2%–3% for dataset I and 2%–1% for dataset II in sensitivity and specificity respectively).

IAGAN architecture allows for semi-supervised augmentation of data for a specific class of labels. We showed that by training IAGAN on a specific class, we were able to use all classes to generate new data for that specific class. Effective training of generative models for medical imaging can be specially helpful to detect anomalies in classes where we do not have enough data/labels for effectively training CNN models. The COVID-19 pandemic is a great example for the importance of generative models, where no images are required for this class of images in order to detect images of this class [23]. Advances in generative models for detection of anomalies can allow for fast deployment of such models at a time where adequate number of labeled images for the new disease are not available for the effective training of CNNs. It is worth mentioning that while an architecture like CycleGAN [33] uses images as input to its generator, to train a CycleGAN, images from two different domain (i.e normal and pneumonia) are used to learn the transition of one image domain to the other. While this could allow for augmenting data from one class to the other, it would require having enough labeled data for all classes and does not allow for single class data augmentation (i.e augmenting normal dataset using partially labeled chest X-rays with only available label being normal) as is enabled by IAGAN.

Early on in this study, it was not immediately clear whether the effects of feeding real images to GAN's generator (**G**) was due to image specific information, or providing the model with a larger vector size in the generator's up-sampling path. Since the down-sampled image is concatenated with **G**'s other input early on in the network, the effects of the input image might be associated with the added vector size, having the same effect as adding the same image with randomly sampled pixel valued. We trained the IAGAN but this time, the input images were randomly generated. The IAGAN failed to generate realistic images using random input images. This confirms that our proposed IAGAN architecture that encodes the input images using convolution and attention layers to a lower-dimensional representation, before concatenating with the projected noise is an effective way to generate meaningful images and augment data. Fig. 7 shows **G**'s output in epochs 5–150.

One of the disadvantages of using a dataset such as COVIDx, compared to dataset I, is the multicentric nature of the images. Since images have been collected from multiple sources and health centers with possibly different acquisition parameters and different scanner models, we observed that our GAN for anomaly detection does not perform as well as dataset I, with or without augmentation. With a more consistent dataset, we hope to achieve improved results on dataset II, compared to dataset I.

8. Conclusion

In this paper, we presented IAGAN; a semi-supervised GAN-based augmentation method to improve training GANs for detection of anomalies (pneumonia and COVID-19) in chest X-rays. IAGAN showed to be statistically significant in augmenting data, improving the AUC, sensitivity and specificity of GAN for detection of anomalies.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by Chair in Medical Imaging and Artificial Intelligence, a joint Hospital-University Chair between the University of Toronto, The Hospital for Sick Children, and the SickKids Foundation.

References

[1] Krizhevsky Alex, Sutskever Ilya, Hinton Geoffrey E. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. 2012. p. 1097–105.

[2] He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. 2015. p. 1026–34.

[3] He Kaiming, Zhang Xiangyu, Ren Shaoqing, Sun Jian. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 770–8.

[4] Zhang Yu-Dong, Dong Zhengchao, Chen Xianqing, Jia Wenjuan, Du Sidan, Muhammad Khan, et al. Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation. Multimedia Tools Appl 2019;78(3):3613–32.

[5] Hao Ruqian, Namdar Khashayar, Liu Lin, Haider Masoom A, Khalvati Farzad. A comprehensive study of data augmentation strategies for prostate cancer detection in diffusion-weighted mri using convolutional neural networks. 2020, arXiv preprint arXiv:2006.01693.

[6] Goodfellow Ian. Nips 2016 tutorial: Generative adversarial networks. 2016, arXiv preprint arXiv:1701.00160.

[7] Sandfort Veit, Yan Ke, Pickhardt Perry J, Summers Ronald M. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. Sci Rep 2019;9(1):1–9.

[8] Ronneberger Olaf, Fischer Philipp, Brox Thomas. U-net: Convolutional networks for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention. Springer; 2015, p. 234–41.

[9] Radford Alec, Metz Luke, Chintala Soumith. Unsupervised representation learning with deep convolutional generative adversarial networks. 2015, arXiv preprint arXiv:1511.06434.

[10] Mirza Mehdi, Osindero Simon. Conditional generative adversarial nets. 2014, arXiv preprint arXiv:1411.1784.

[11] Frid-Adar Maayan, Diamant Idit, Klang Eyal, Amitai Michal, Goldberger Jacob, Greenspan Hayit. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. Neurocomputing 2018;321:321–31.

[12] Wu Eric, Wu Kevin, Cox David, Lotter William. Conditional infilling gans for data augmentation in mammogram classification. In: Image analysis for moving organ, breast, and thoracic images. Springer; 2018, p. 98–106.

[13] Antoniou Antreas, Storkey Amos, Edwards Harrison. Data augmentation generative adversarial networks. 2017, arXiv preprint arXiv:1711.04340.

[14] Cohen Joseph Paul, Morrison Paul, Dao Lan. Covid-19 Image data collection. 2020, arxiv:2003.11597, URL https://github.com/ieee8023/covid-chestxray-dataset.

[15] Wang Linda, Wong Alexander. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. 2020, arXiv preprint arXiv:2003.09871.

[16] Ozturk Tulin, Talo Muhammed, Yildirim Eylul Azra, Baloglu Ulas Baran, Yildirim Ozal, U. Rajendra Acharya. Automated detection of covid-19 cases using deep neural networks with x-ray images. Comput Biol Med 2020;103792.

[17] Karim Md, Döhmen Till, Rebholz-Schuhmann Dietrich, Decker Stefan, Cochez Michael, Beyan Oya, et al. Deepcovidexplainer: Explainable covid-19 predictions based on chest x-ray images. 2020, arXiv preprint arXiv:2004.04582.

[18] Hemdan Ezz El-Din, Shouman Marwa A, Karar Mohamed Esmail. Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images. 2020, arXiv preprint arXiv:2003.11055.

[19] Ghoshal Biraja, Tucker Allan. Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection. 2020, arXiv preprint arXiv:2003.10769.

[20] Afshar Parnian, Heidarian Shahin, Naderkhani Farnoosh, Oikonomou Anastasia, Plataniotis Konstantinos N, Mohammadi Arash. Covid-caps: A capsule network-based framework for identification of covid-19 cases from x-ray images. 2020, arXiv preprint arXiv:2004.02696.

[21] DeGrave Alex J, Janizek Joseph D, Lee Su-In. Ai for radiographic covid-19 detection selects shortcuts over signal. medRxiv. 2020.

[22] Schlegl Thomas, Seeböck Philipp, Waldstein Sebastian M, Schmidt-Erfurth Ursula, Langs Georg. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International conference on information processing in medical imaging. Springer; 2017, p. 146–57.

[23] Motamed Saman, Rogalla Patrik, Khalvati Farzad. Randgan: Randomized generative adversarial network for detection of covid-19 in chest x-ray. Sci Rep 2021;11(1):1–10.

[24] Zhang Han, Goodfellow Ian, Metaxas Dimitris, Odena Augustus. Self-attention generative adversarial networks. 2018, arXiv preprint arXiv:1805.08318.

[25] Szegedy Christian, Vanhoucke Vincent, Ioffe Sergey, Shlens Jon, Wojna Zbigniew. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. p. 2818–26.

[26] Kodali Naveen, Abernethy Jacob, Hays James, Kira Zsolt. On convergence and stability of gans. 2017, arXiv preprint arXiv:1705.07215.

[27] Kermany Daniel, Zhang Kang, Goldbaum Michael. Labeled optical coherence tomography (oct) and chest x-ray images for classification. Mendeley Data 2018;2.

[28] Geirhos Robert, Jacobsen Jörn-Henrik, Michaelis Claudio, Zemel Richard, Wiel Brendel, Bethge Matthias, et al. Shortcut learning in deep neural networks. 2020, arXiv preprint arXiv:2004.07780.

[29] Motamed Saman, Gujrathi Isha, Deniffel Dominik, Oentoro Anton, Haider Masoom A, Khalvati Farzad. A transfer learning approach for automated segmentation of prostate whole gland and transition zone in diffusion weighted mri. 2019, arXiv preprint arXiv:1909.09541.

[30] Jaeger Stefan, Candemir Sema, Antani Sameer, Wáng Yì-Xiáng J, Lu Pu-Xuan, Thoma George. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quant Imaging Med Surg 2014;4(6):475.

[31] Stirenko Sergii, Kochura Yuriy, Alienin Oleg, Rokovyi Oleksandr, Gordienko Yuri, Gang Peng, et al. Chest x-ray analysis of tuberculosis by deep learning with segmentation and augmentation. In: 2018 IEEE 38th international conference on electronics and nanotechnology. IEEE; 2018, p. 422–8.

[32] DeLong Elizabeth R, DeLong David M, Clarke-Pearson Daniel L. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 1988;837–45.

[33] Zhu Jun-Yan, Park Taesung, Isola Phillip, Efros Alexei A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. 2017. p. 2223–32.