scientific reports

Check for updates

OPEN

RANDGAN: Randomized generative adversarial network for detection of COVID-19 in chest X-ray

Saman Motamed[1,2]✉, Patrik Rogalla[4] & Farzad Khalvati[1,2,3]

COVID-19 spread across the globe at an immense rate and has left healthcare systems incapacitated to diagnose and test patients at the needed rate. Studies have shown promising results for detection of COVID-19 from viral bacterial pneumonia in chest X-rays. Automation of COVID-19 testing using medical images can speed up the testing process of patients where health care systems lack sufficient numbers of the reverse-transcription polymerase chain reaction tests. Supervised deep learning models such as convolutional neural networks need enough labeled data for all classes to correctly learn the task of detection. Gathering labeled data is a cumbersome task and requires time and resources which could further strain health care systems and radiologists at the early stages of a pandemic such as COVID-19. In this study, we propose a randomized generative adversarial network (RANDGAN) that detects images of an unknown class (COVID-19) from known and labelled classes (Normal and Viral Pneumonia) without the need for labels and training data from the unknown class of images (COVID-19). We used the largest publicly available COVID-19 chest X-ray dataset, COVIDx, which is comprised of Normal, Pneumonia, and COVID-19 images from multiple public databases. In this work, we use transfer learning to segment the lungs in the COVIDx dataset. Next, we show why segmentation of the region of interest (lungs) is vital to correctly learn the task of classification, specifically in datasets that contain images from different resources as it is the case for the COVIDx dataset. Finally, we show improved results in detection of COVID-19 cases using our generative model (RANDGAN) compared to conventional generative adversarial networks for anomaly detection in medical images, improving the area under the ROC curve from 0.71 to 0.77.

COVID-19 spread globally over a short period of time and became a deadly pandemic[1]. Early diagnosis and detection of pneumonia can minimize the risk factors of the illness[2] and help break the transmission chain. The standard test for diagnosis of COVID-19 is reverse transcriptase polymerase chain reaction (RT-PCR)[3]. The lack of accessibility and slowness of RT-PCR, along with its high false negative rate (39–61%), drew attention to diagnosis of COVID-19 using chest radiographs[4,5]. Automation of COVID-19 diagnosis using chest X-rays can help healthcare systems keep up with demands for patients testing as X-rays are more readily available than RT-PCR and reduce strain from radiologists and healthcare systems. Medical imaging based diagnosis can also help control the high false negative rate of RT-PCR tests by acting as a secondary control. Computer-aided disease diagnosis using medical imaging techniques have accelerated over the past decade due to the breakthroughs in the field of machine learning and the development of detection and classification models that are based on convolutional neural networks (CNNs)[6–8]. CNNs, which are mainly used in supervised frameworks, require large amounts of labeled data to learn the task of anomaly detection, such as detecting COVID-19 in chest X-rays. Supervised architectures require training data with complete labels for all image classes (e.g., normal and COVID-19). Nevertheless, this requires accurate labeling of the data for all cases and the cumbersome annotation effort, and the diagnosis variation amongst expert radiologists limits the performance of these supervised models on new data. Specially, in pandemics such as COVID-19, at the beginning, there is limited COVID-19 data (if any data at all) available for training a supervised classification model. In contrast, solutions based on

[1]Institute of Medical Science, University of Toronto, Toronto, ON, Canada. [2]Department of Diagnostic Imaging, Neurosciences and Mental Health, The Hospital for Sick Children, University of Toronto, Toronto, ON, Canada. [3]Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada. [4]University Health Network, Toronto, ON, Canada. ✉email: sam.motamed@mail.utoronto.ca

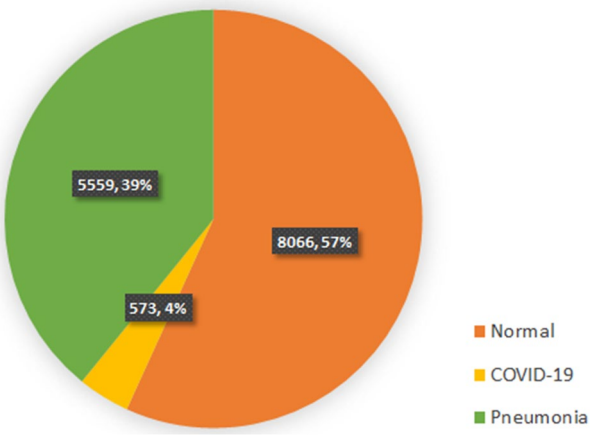

Figure 1. Class distribution of COVIDx dataset.

semi-supervised learning only require partial labels for the training data[9]. Semi-supervised learning significantly reduces the cost of creating training data and, thus, opens new opportunities for automated disease detection using training data with only single class labels.

In this study, we propose a semi-supervised generative model (randomized generative adversarial network-RANDGAN) for detection of COVID-19 positive chest X-ray images. The idea behind anomaly detection using generative adversarial networks (GANs) comes from the great ability of generative models in learning the image-space manifold where training images lie on, and being able to generate never-before-seen images that lie on the learned image-space[10]. Anomaly detection may be seen as only detecting abnormality in medical images such as a tumour or pneumonia. We extend the definition of anomaly in medical images as the deviation from the image-space manifold of training data. In other words, if the training data only includes COVID-19 negative cases (i.e., healthy or viral pneumonia), the anomaly detected in test cases is indeed an abnormality such as COVID-19. On the other hand, if the training data only includes COVID-19 positive cases, the "anomaly" detected in the test cases are the deviation from COVID-19 cases, meaning that the test case does not contain the abnormality in the training class (i.e., healthy or viral pneumonia). We show our proposed RANDGAN model is able to differentiate between COVID-19 positive and negative images. To the best of our knowledge, this study is the first of its kind, using semi-supervised learning for detection of COVID-19 in medical images and reporting performance accuracy on the entire cohort of COVID-19 positive images without the need to use any of the COVID-19 positive images to train our model. The code for our RANDGAN along with instructions to create the datasets used in this study can be found here; https://github.com/samxmot/RANDGAN.

Dataset

Covid-chestxray dataset[11] is an 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[12] dataset. With the number of images growing, many deep learning models are trained and tested on this public dataset[12–14]. Figure 1 shows the class distribution of the COVIDx dataset. 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.

Related work

Using the covid-chestxray and COVIDx datasets, multiple studies have utilized supervised deep learning models to detect COVID-19 in chest X-rays[12–17]. Wang et al.'s CNN based COVID-NET[12] achieved a 93.3% test accuracy for multi-class classification on a test cohort of 100 Normal, 100 Pneumonia and 100 COVID-19 images from the COVIDx dataset with the rest of the images of each class being used to train their model. Hemdan et al.'s COVIDX-Net[16], comprised of multiple architectures such as VGG19, DenseNet121 and InceptionV3, was tested on a small set of 50 X-ray images from the covid-chestxray dataset; 25 COVID-19 positive and 25 COVID-19 negative. They reported accuracies of anywhere between 50% (InceptionV3) to 90% (VGG19 and DenseNet201) for each investigated architecture. Ozturk et al.'s DarkNet[13] experimented with both binary classification (COVID-19 vs. No Findings) and multi-class classification (Pneumonia vs. COVID-19 vs. No Findings). They reported a binary classification accuracy of 98.08% and multi-class classification with accuracy of 87% on 25 COVID-19, 100 Normal and 100 Pneumonia images. Afshar et al. proposed using capsule networks for binary classification of COVID-19 positive and negative cases using COVIDx dataset, pre-trained on non-COVID chest X-ray images from other datasets. They reported 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.

The high accuracy achieved in these models, despite the imbalanced dataset with only 4% of the images belonging to COVID-19 and the multi-centric nature of the dataset which could cause images from different

scanners and health centres to have inherent characteristic differences, put the robustness of these models under question.

DeGrave et al.[18] conducted a few experiments to test the generalizability and robustness of these models trained on the COVIDx dataset. Replicating the same underlying supervised model structure such as COVID-NET[12] and training the model on COVIDx dataset, they achieved high test accuracy when tested on COVIDx data. Their predictive performance, however, dropped by 50% when they validated their model on an external COVID and Non-COVID dataset[19] where the images were from a single institution. Furthermore, using saliency maps, which highlight the region of each X-ray image that contributed most to the classification decision of the CNNs, they found non-COVID markers such as image edges, diaphragm and cardiac silhouette have contributed to the classification of COVID-19; markers that do not have a predictive value for detection of COVID-19[20]. This confirms that using the full images from a dataset that comes from different scanners can be problematic where non-disease specific markers could act as a shortcut[21] and help CNNs achieve high accuracy on a particular dataset yet fail to generalize to any other dataset. To minimize the effect of shortcuts, we created a segmented COVIDx dataset that includes only the lungs where the true markers of COVID-19 and Pneumonia appear.

Segmentation of lung in COVIDx images

To mitigate the issue of deep learning models picking non-disease related markers from the images, we created a new dataset by segmenting lungs in COVIDx dataset. For the training set, we used the Montgomery County chest X-ray set[22], which contains 138 frontal chest X-rays from Montgomery County's Tuberculosis screening program with corresponding masks manually annotated by radiologists. We resized the images to 256×256 pixels and normalized them to have pixel intensities between 0 and 1. We trained a U-NET[23] based model, that has been augmented with inception and residual architectures, with these normalized images[24,25]. Transfer learning[26] has shown promise in adapting tasks from one domain (source) to another (target). For the task of lung segmentation for the COVIDx dataset, the Montgomery dataset was used as the source and COVIDx dataset was used as the target domain. For the task of transfer learning, Sefexa[27], an image segmentation tool, was used for manual segmentation of 900 randomly selected images (300 from each class) of the COVIDx dataset. All segmentation masks were corrected by an experienced radiologist and intentionally over-segmented to ensure no region of lung is excluded. Thus, these masks are best to be used for classification algorithms for detection of COVID-19 and pneumonia, and not for precise segmentation of lung boundaries. 850 segmented X-ray images from COVIDx were used for performing transfer learning[24] from the Montgomery dataset to COVIDx. We kept 50 manual segmentations to evaluate our model's accuracy. Since source domain is smaller than our target domain, we fine-tuned 75% of the pre-trained model's layers (encoder part), and trained on the Montgomery dataset images. We froze the first 25% layers of the pre-trained U-NET and fine-tuned the rest of the encoder and decoder components based on our manual segmentation for COVIDX images. We used open and close operations as a post-processing step to fill any holes in the masks and reduce noise in the predicted masks. We tested the accuracy of our model using Sørensen–Dice coefficient (DSC). We achieved a DSC of 0.83 on our test set of 50 images.

Figure 2 shows the output of our segmentation model. We include some of the failed segmentation attempts of the model as well. Accurate segmentation of lung images are a limitation of using automated segmentation models.

Random input generative adversarial networks

Generative adversarial networks (GANs)[10] revolutionized the field of deep learning by allowing generation of never-before-seen data that follows the distribution of real data. Applications of GANs have expanded from generating human-like faces, to image style transfer and detection of anomalies in images[28]. In the following, we describe the components of our proposed random input generative adversarial network (RANDGAN).

Generator network. The Generator (**G**) (Fig. 3) learns a distribution P_g over the input data x via mapping of input noise z, to $2D$ images by function $G(z)$. The trained Generator learns the mapping $G(z) : z \longmapsto x$ from latent space representations z to realistic, $2D$, X-ray images. Our Generator model follows DCGAN's architecture (named AnoGAN for anomaly detection GAN in the study)[28] (used for anomaly detection for retina) with three main modifications; the use of randomized 2D image inputs to the generator, inception layers, and residual connections as shown in Fig. 4.

Feeding real training images as an input to the generator has shown improvement in using GANs for augmenting images[29,30]. Real images are encoded into a lower dimensional space before being concatenated with the noise input vector z. To improve generalizability of our generator, specially due to the multi-centric nature of COVIDx data, we randomly select batches of 32 images from the cohort of our training class and encode them to a lower-representation space using inception layers. This helps in adding variability to each iteration of the generator's training by not only using a random noise vector, but also real, random image representations of the training class. Doing so shows improved results when using the trained GAN to classify images of an unknown class from other known classes. The idea behind the inception and residual architecture[31] is being able to increase GAN's ability to capture more details from training image-space without losing spatial information after each convolution and pooling layer. Although making the Generator deeper is theoretically a valid way to capture more long-range details in the image, deep GANs are unstable and hard to train[28,32].

Discriminator. The discriminator (**D**) (Fig. 5) 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 images sampled from training data or generated image G(z) by the Generator G.

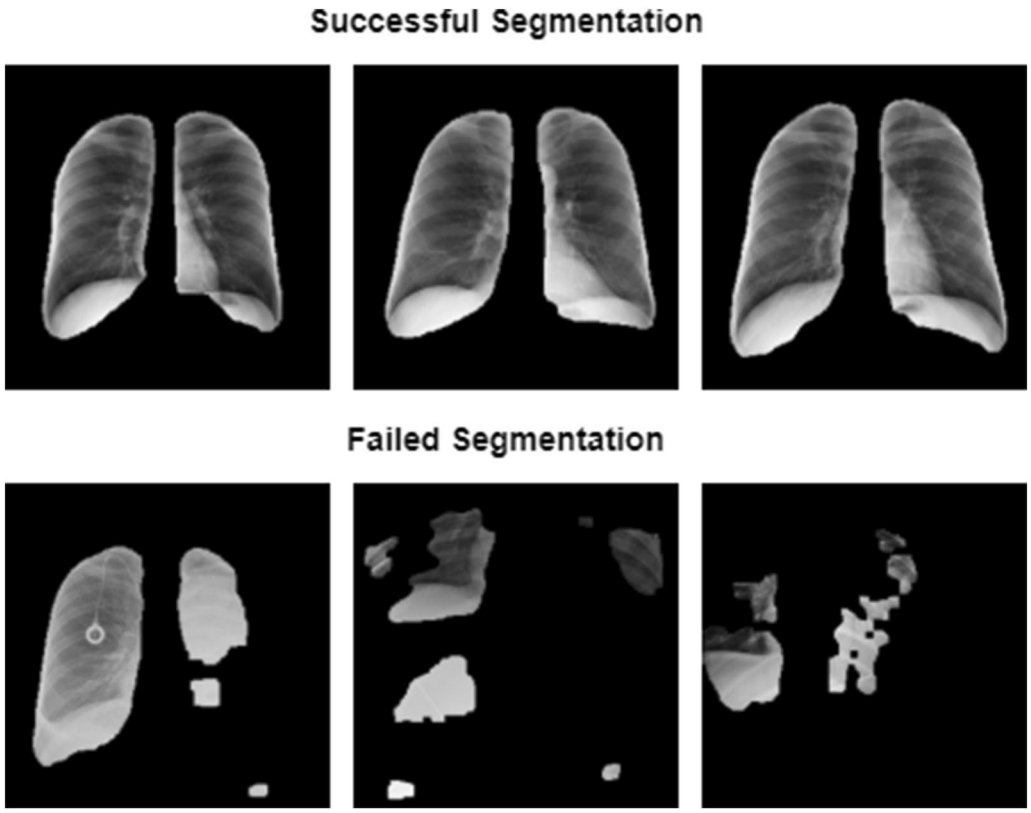

Figure 2. Output samples of our segmentation model on COVIDx images.

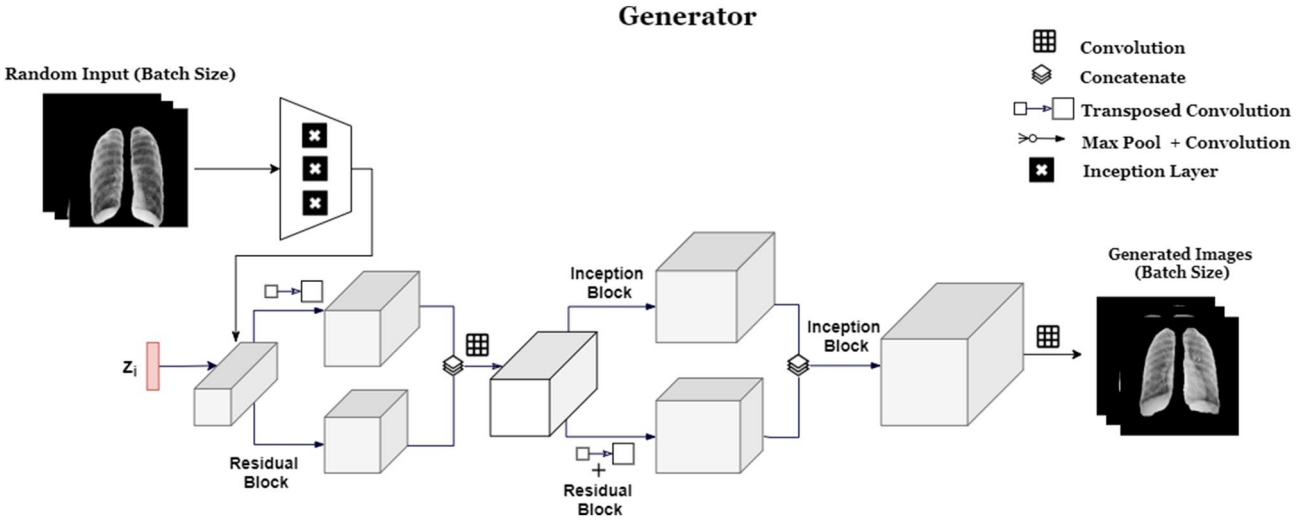

Figure 3. RANDGAN's generator architecture.

Optimization of D and G can be thought of as the following game of minimax[10] 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. The Generator improves at generating more realistic images while Discriminator gets better at correctly identifying between real and generated images.

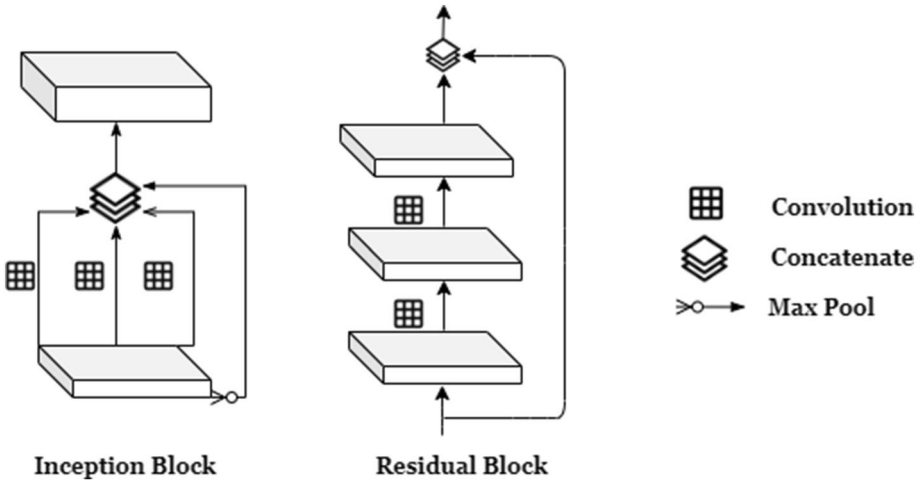

Figure 4. Inception and residual block architecture.

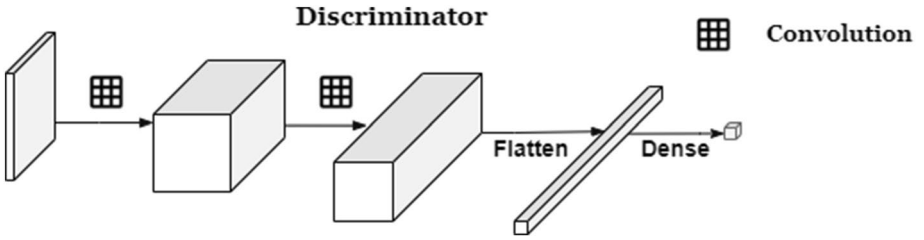

Figure 5. RANDGAN's discriminator architecture.

Label	Train	Test
Normal	7493	573
Pneumonia	4986	573
COVID-19	N/A	573

Table 1. Train and test class distribution of COVIDx and COVIDx segmentation dataset.

Experiments

Data and pre-processing. We used both full images from COVIDx dataset and our segmentation of the COVIDx data to train separate models and compare the results. One of the advantages of our semi-supervised model compared to supervised models is the ability to test our model on not only a subset, but all of COVID-19 positive images as we do not use any of these images to train our model. While studies such as Wang et. al's COVID-NET use 100 images of COVID-19, Hedman et al.[16] and Ozturk et al.[13] using 25 COVID-19 positive images to test their models, we used 573 images of each class; the entire dataset for COVID-19 was used and for normal and pneumonia classes, 573 images were randomly selected for each class. All images, converted from RGB to grayscale, were resized to 128×128 pixels, with pixel intensities normalized to have values between -1 and 1. The models were trained using an NVIDIA GeForce RTX 2080 Ti with 11 GB of memory.

Table 1 shows the Train and Test split of our COVIDx and Segmented COVIDx images.

Evaluation. We trained two instances of our RANDGAN. For comparison, we repeated the same training using the GAN model (AnoGAN) used in Radford et al.'s[28] anomaly detection study. One RANDGAN/AnoGAN was trained using Normal images and the other RANDGAN/AnoGAN was trained using Pneumonia images. When the model's training is done, the generator has learned the mapping $G(z) : z \longmapsto x$ from latent space representation z to realistic images. Given a query image x in test, we want to find a point z from the latent space such that, given the Generator's output on that point ($G(z)$), that is most similar to the query image x. The expected behaviour after successful training is that the query image x, if affected by pneumonia, will result in finding an image $G(z)$, which is visually closer to image x than if the query image was a normal case (given the GAN was trained on Pneumonia images).

Model	Dataset	AUC
AnoGAN	COVIDx (balanced test set)	0.54
AnoGAN	Segmented COVIDx (balanced test set)	0.71
RANDGAN	Segmented COVIDx (balanced test set)	**0.77**
RANDGAN	Segmented COVIDx (imbalanced test set)	0.76

Table 2. Performance comparison of RANDGAN and AnoGAN.

To find latent variable z that generates the most similar image $G(z)$ to the query image x, we used back propagation with a predefined number of steps. The loss function defined to find such z through back-propagation is comprised of two components; *residual loss* and *discrimination loss*. Residual loss ($\mathcal{L}_{\mathcal{R}}$) calculates the L1 distance between $G(z)$ and the query image x and enforces visual similarity between the query image and generated image.

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

Schlegl et al.[33] proposed a discrimination loss ($\mathcal{L}_{\mathcal{D}}$) inspired by the concept of feature matching[34] that enforces generated images $G(z_i)$ to follow the statistical characteristics of the training images. $\mathcal{L}_{\mathcal{R}}$ is defined below where the output of an intermediate layer of the discriminator, $f(.)$, is used to represent the statistical characteristics of the input image.

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

The overall loss used to back-propagate and find the best z is a weighted sum of residual and discrimination loss;

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

The Anomaly score $A(x)$ for the query image x is defined as;

$$A(x) = (1 - \lambda) \times R(x) + \lambda \times D(x) \tag{5}$$

where R(x) and D(x) are respectively the residual and discrimination loss of the best z_i found through back-propagation. λ adjusts the weighted sum of the overall loss and anomaly score. We used $\lambda = 0.2$ to train our proposed RANDGAN and AnoGAN[33]. Both architectures were trained with the same initial conditions for performance comparison.

With two trained models, one on Normal and one on Pneumonia images, we calculate two anomaly scores for each test image. One anomaly score from inputting the test image into Normal trained GAN and one from Pneumonia trained GAN. The anomaly score generated from the Normal trained GAN will be lower for Normal test images compared to Pneumonia and COVID-19 images. Respectively, the anomaly score generated from Pneumonia trained GAN will be lower for Pneumonia test images compared to Normal and COVID-19 images. For each test image and the corresponding two anomaly scores, we generate a single anomaly score by summing the two scores together. The idea is that COVID-19 (unknown) images would score high anomalies from both networks while Normal and Pneumonia images score low in one model and high in the other. This should lead to the COVID-19 (unknown) images to score higher overall than the two other (known) classes.

Results

We generated a single anomaly score, comprised of two anomaly scores from the two trained models (Normal, Pneumonia), for the images in our test set. 573 anomaly scores were computed for each class (Normal, Pneumonia and COVID-19) of our COVIDx and segmented COVIDx dataset. To evaluate the performance of our COVID-19 positive detection model on a balanced test set, we randomly selected 286 Normal labeled and 287 Pneumonia labeled images and combined them into a COVID-19 negative test set with corresponding anomaly scores. We repeated the random selection of images from Normal and Pneumonia test cohorts 5 times in order to achieve an average performance metric of our models. The experiments were performed using AnoGAN trained on full COVIDx images, AnoGAN trained on segmented COVIDx images and RANDGAN trained on segmented COVIDx images. Table 2 shows the average AUC of our models for the 5 calculations. We also report the AUC on the unbalanced test set, using 573 COVID-19 positive and 1146 COVID-19 negative (573 normal and 573 Pneumonia) images.

Figure 6 shows the ROC curve of the 3 trained models. With an AUC of 0.54, the AnoGAN model fails to classify COVID-19 positive and negative cases in full images. The same model performs significantly better when trained on lung segmented COVIDx dataset and achieves an AUC of 0.71. This shows the markers outside of the lung that were irrelevant to the disease, hindered the performance of the GAN. The same markers were shown by DeGrave et al.[18] to act as shortcuts in wrongfully helping CNNs classify the classes of COVID-19, Normal and Pneumonia images. Our RANDGAN model achieved an AUC of 0.77, a 6% improvement compared to that of the AnoGAN model on the segmented dataset, showing the effectiveness of our down-sampling and feeding randomly selected images to the Generator during the training.

The false negative rate (FNR) of the RT-PCR test[5] varies depending on the time of test in comparison with the time of contracting the COVID-19 virus. It has been shown that at specificity of 90%, on the day of symptom onset, the median FNR for RT-PCR test was 38% with Confidence Interval (CI) 18–65%. On day 8, the

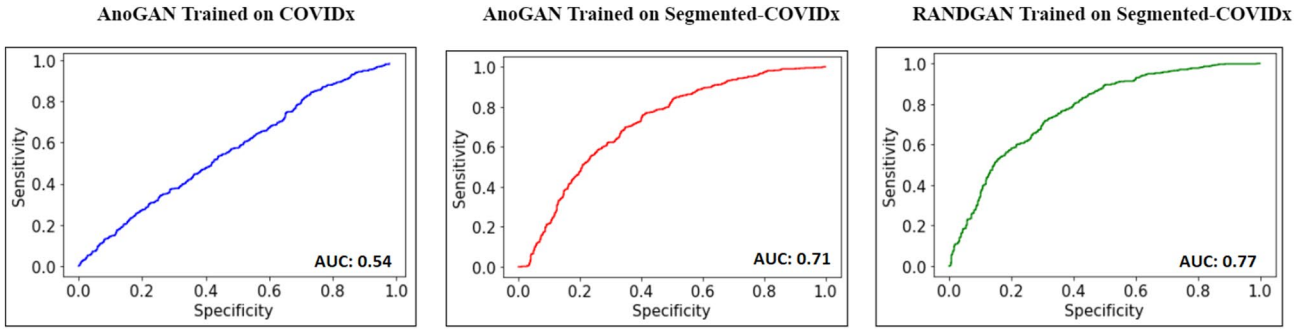

Figure 6. ROC curve of the trained generative models.

Model	Specificity (%)	Sensitivity (%)	False negative rate (%)
RANDGAN	90	34	65
AnoGAN	90	30	69
RANDGAN	85	49	50
AnoGAN	85	48	51
RANDGAN	80	57	42
AnoGAN	80	57	42

Table 3. Sensitivity, specificity and false negative rate for AnoGAN and RANDGAN model.

median FNR for RT-PCR test is 20% with CI 12–30% and on day 21 from symptom onset, the FNR increases to median of 66% with CI 54–77%. Although we do not have the information for the date of symptom onset and the date of X-ray acquisition from symptom onset in the COVIDx dataset, Table 3 shows the sensitivity and FNR of our proposed RANDGAN model and AnoGAN at specificity of 90%, 85% and 80%. With the wide confidence interval for RT-PCR test's FNR at the early (18–65%) and later (54–77%) stages of contracting the virus, our model matches the upper bound of the RT-PCR's FNR early on (65%) and outperforms the median FNR of RT-PCR at later stages of the disease (65% vs. 66%). The joint use of both tests (RT-PCR and imaging) could lower the overall FNR given that while one test's result is False Negative result for a patient, the other test may call the patient positive.

Figure 7 shows the normalized average anomaly score of the 5 runs of each of our three models; RANDGAN trained on segmented X-ray images, AnoGAN trained on segmented images and AnoGAN trained on full images. The highest score of each trained GAN, among the 3 classes of Normal, Pneumonia and COVID-19 is normalized to anomaly score of 10. Other anomalies are normalized accordingly by dividing the score by highest anomaly score and multiplying by 10. Despite the ROC curve that combines a balanced number of Normal and Pneumonia images in comparison to COVID-19 images, we present the anomaly scores in their entirety (573 Normal, 573 Pneumonia and 573 COVID-19 images). The desired anomaly score for the purpose of detecting COVID-19 positive and negative images is achieving higher anomaly score for COVID-19 positive images and lower scores for COVID-19 negative (Normal and Pneumonia) images. Figure 7 shows the AnoGAN and RANDGAN trained on both the full and segmented COVIDx datasets satisfy this characteristic. However, the gap between COVID-19 positive and negative scores defines the accuracy of each model. The bigger the anomaly score gap is between the two classes, the higher our classification confidence becomes. RANDGAN shows the biggest gap of normalized mean anomaly score (MAS) between COVID-19 (MAS = 4.48) and Pneumonia (3.01) and COVID-19 and Normal (3.56) images which are 1.47 and 0.92 respectively. AnoGAN trained on segmented COVIDx dataset shows 1.36 as the gap between COVID-19 (MAS = 4.42) and 0.91 between COVID-19 and Normal (3.51). AnoGAN trained on full COVIDx images shows a small gap between COVID-19, Pneumonia and Normal images (0.36 between COVID-19 and Pneumonia and 0.12 between COVID-19 and Normal).

Discussion

In this study, we introduced RANDGAN, a novel generative adversarial network for semi-supervised detection of an unknown (COVID-19) class in chest X-ray images from a pool of known (Normal and Pneumonia) and unknown classes (COVID-19) by only using the known classes for training. With this model, unknown cases can be screened and flagged for further investigations by radiologists increasing the probability of catching such cases early on. Using semi-supervised approaches for a problem such as detection of COVID-19, specially at the beginning of a pandemic are preferred over supervised approaches for they allow faster training of models without the need for gathering and annotation of data from the spreading disease. The result of semi-supervised models are more reliable where number of images are limited for the unknown (COVID-19) class. Where our semi-supervised model uses all COVID-19 images to test the model's performance, supervised models have to use majority of the images ($\sim$ 90%) for training the model and test the model on a small subset of the images.

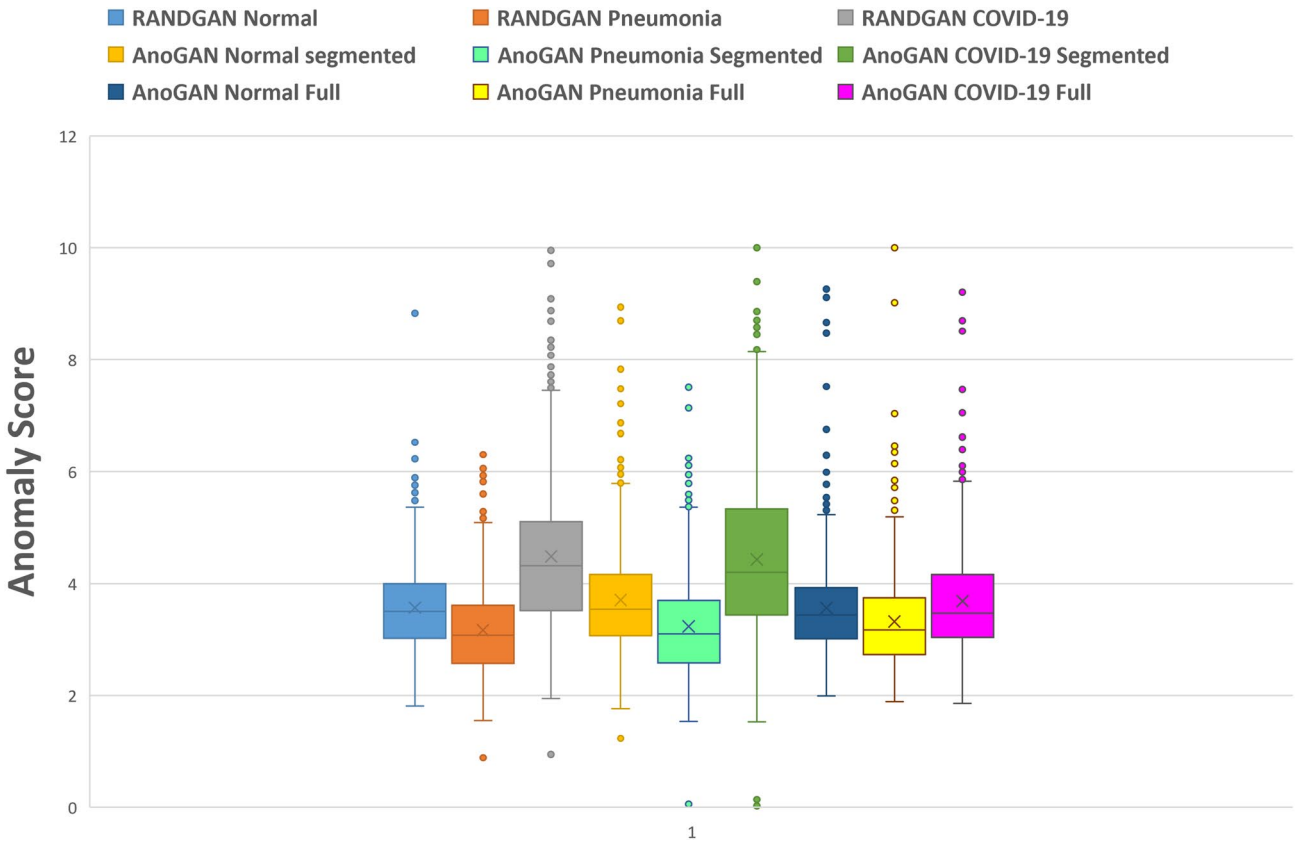

Figure 7. Normalized average anomaly score of the trained generative models.

We demonstrated the importance of segmentation of lungs for the COVIDx dataset. DeGrave et al.[18] showed non-disease markers outside the lung act as shortcuts[21] in helping CNNs performance on specific datasets on which the model is trained. By using transfer learning and segmenting the lung, we observed that using lung only images boosts the performance of generative models in detecting COVID-19 from Pneumonia and Normal images. AnoGAN[28] achieved an average AUC of 0.54 when using full images from COVIDx images while using segmented COVIDx images achieved an average AUC of 0.71. While the semi-supervised RANDGAN approach for detection of COVID-19 in chest X-ray results in overall AUC of 77%, which is lower compared to supervised counterpart models, the main advantages of our proposed RANDGAN model is that it requires no COVID-19 X-ray images for training. This is crucial in dealing with a pandemic such as COVID-19 when there is virtually no data or very little data at the onset. This is in contrast to supervised detection models that require a large COVID-19 dataset for training. Our model can be trained and used as soon as a new disease emerges without the need for the cumbersome process of acquiring enough images and annotating the images by radiologists. This could take months to compete as it is seen with the COVIDx dataset, in which after months of emergence of COVID-19 and becoming a pandemic, only 4% of the dataset is made up of COVID-19 cases. Another advantage of our model is that since it does not require COVID-19 data for training, we are able to test our model on all available COVID-19 X-ray images and report the AUC on the complete dataset (573 images) making it immediately more reliable although reporting a lower AUC. In contrast, supervised detection models report their results on around 25–100 COVID-19 X-ray images[14,16,35] (10–20% of the available COVID-19 images). Future directions will focus on improving the performance of our proposed RANDGAN (AUC of 0.77) model by performing data augmentation, and as more data is collected, it is important to validate the model on external data sources (different scanners/health care systems).

Limitations

One limitation of working with data of relatively early stages of a disease such as COVID-19 is dataset size. Even though our semi-supervised model is able to use all COVID-19 images to evaluate the performance of the model, while supervised models have to use majority of the already small COVID-19 cohort to train their images, more images would allow for a better understanding of the true performance of both supervised and semi-supervised models. Segmentation accuracy of the lungs is another limiting factor. Although the performance of the base model greatly improves (AUC of 0.54–0.71), segmentation model fails in some cases (Fig. 2). As more data gets collected and becomes available from different health care systems, any model trained for detection of COVID-19 needs validation from external sources. Without validation, these models need to be used as a secondary measure for detection of COVID-19.

Received: 27 October 2020; Accepted: 6 April 2021
Published online: 21 April 2021

References

1. Du, R.-H. *et al.* Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: a prospective cohort study. *Eur. Respir. J.* **55**(5), 2000524 (2020).
2. Pereira, J. M., Paiva, J. A. & Rello, J. Severe sepsis in community-acquired pneumonia-early recognition and treatment. *Eur. J. Intern. Med.* **23**(5), 412–419 (2012).
3. Wang, W. *et al.* Detection of SARS-CoV-2 in different types of clinical specimens. *JAMA* **323**(18), 1843–1844 (2020).
4. Kundu, S., Elhalawani, H., Gichoya, J. W. & Kahn, C. E. Jr. How might AI and chest imaging help unravel COVID-19's mysteries?. *Radiol. Artif. Intell.* **2**, e200053 (2020).
5. Kucirka, L. M., Lauer, S. A., Laeyendecker, O., Boon, D. & Lessler, J. Variation in false-negative rate of reverse transcriptase polymerase chain reaction-based SARS-CoV-2 tests by time since exposure. *Ann. Intern. Med.* **173**(4), 262–267 (2020).
6. Long, J., Shelhamer, E. & Darrell, T. Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015).
7. Girshick, R. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448 (2015).
8. Krizhevsky, A., Sutskever, I., & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105 (2012).
9. Zhu, X. & Goldberg, A. B. Introduction to semi-supervised learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **3**(1), 1–130 (2009).
10. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680 (2014).
11. Cohen, J. P., Morrison, P. & Dao, L. Covid-19 image data collection. https://arxiv.org/abs/2003.11597. URL https://github.com/ieee8023/covid-chestxray-dataset (2020).
12. Wang, L. & Wong, A. Covid-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. arXiv preprint arXiv:2003.09871 (2020).
13. Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O. & Rajendra A. Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Comput. Biol. Med.* pp. 103792 (2020).
14. Afshar, P., Heidarian, S., Naderkhani, F., Oikonomou, A., Plataniotis, K. N. & Mohammadi, A. Covid-caps: A capsule network-based framework for identification of COVID-19 cases from X-ray images. arXiv preprint arXiv:2004.02696 (2020).
15. Karim, M., Döhmen, T., Rebholz-Schuhmann, D., Decker, S., Cochez, M., Beyan, O., *et al.* Deepcovidexplainer: Explainable COVID-19 predictions based on chest X-ray images. arXiv preprint arXiv:2004.04582 (2020).
16. Hemdan, E. E.-D., Shouman, M. A. & Karar, M. E. Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in X-ray images. arXiv preprint arXiv:2003.11055 (2020).
17. Ghoshal, B. & Tucker, A. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. arXiv preprint arXiv:2003.10769 (2020).
18. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *medRxiv* (2020).
19. de la Iglesia Vayá, M., Saborit, J. M., Montell, J. A., Pertusa, A., Bustos, A., Cazorla, M., Galant, J., Barber, X., Orozco-Beltrán, D., Garcia, F., *et al.* Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. arXiv preprint arXiv:2006.01174 (2020).
20. Ng, M.-Y. *et al.* Imaging profile of the COVID-19 infection: radiologic findings and literature review. *Radiol. Cardiothor. Imaging* **2**(1), e200034 (2020).
21. Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M. & Wichmann, F. A. Shortcut learning in deep neural networks. arXiv preprint arXiv:2004.07780 (2020).
22. Jaeger, S. *et al.* Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant. Imaging Med. Surg.* **4**(6), 475 (2014).
23. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, pp. 234–241 (2015).
24. Motamed, S., Gujrathi, I., Deniffel, D., Oentoro, A., Haider, M. A. & Khalvati, F. A transfer learning approach for automated segmentation of prostate whole gland and transition zone in diffusion weighted MRI. arXiv preprint arXiv:1909.09541 (2019).
25. Clark, T., Wong, A., Haider, M. A. & Khalvati, F. Fully deep convolutional neural networks for segmentation of the prostate gland in diffusion-weighted MR images. In *International Conference Image Analysis and Recognition*, pp. 97–104. Springer (2017).
26. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C. & Liu, C. A survey on deep transfer learning. In *International Conference on Artificial Neural Networks*, pp. 270–279. Springer (2018).
27. Fexa, A. *Sefexa Image Segmentation Tool.*
28. Radford, A., Metz, L. & Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015).
29. Antoniou, A., Storkey, A. & Edwards, H. Data augmentation generative adversarial networks. arXiv preprint arXiv:1711.04340 (2017).
30. Motamed, S. & Khalvati, F. *Inception Augmentation Generative Adversarial Network* (2020).
31. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826 (2016).
32. Kodali, N., Abernethy, J., Hays, J. & Kira, Z. On convergence and stability of gans. arXiv preprint arXiv:1705.07215 (2017).
33. Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U. & Langs, G. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International Conference on Information Processing in Medical Imaging*, pp. 146–157. Springer (2017).
34. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. & Chen, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242 (2016).
35. Wang, L. & Wong, A. Covid-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images (2020).

Acknowledgements

This research received funding support from 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.

Author contributions

S.M. and F.K. contributed to the design and implementation of the concept. P.R. contributed to collecting and reviewing the data and provided lung segmentation for the data. All authors contributed to the writing and reviewing of the paper. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.M.

Reprints and permissions information is available at www.nature.com/reprints.

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