Generating Realistic COVID-19 x-rays with a Mean Teacher + Transfer Learning GAN

Sumeet Menon[1], Joshua Galita[1], David Chapman[1], Aryya Gangopadhyay[1], Jayalakshmi Mangalagiri[1], Phuong Nguyen[1]

Yaacov Yesha[1], Yelena Yesha[1], Babak Saboury[1,2], Michael Morris[1,2,3]

[1]University of Maryland, Baltimore County
1000 Hilltop Circle, Baltimore, MD, 21250
sumeet1@umbc.edu

[2]National Institutes of Health Clinical Center
9000 Rockville Pike, Building 10, Room 1C455,
Bethesda, MD 21201

[3]Networking Health
331 Oak Manor Drive, Suite 201, Glen Burnie MD, 21061

2020 IEEE International Conference on Big Data (Big Data) | 978-1-7281-6251-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/BigData50022.2020.9377878

***Abstract*—COVID-19 is a novel infectious disease responsible for over 1.2 million deaths worldwide as of November 2020. The need for rapid testing is a high priority and alternative testing strategies including x-ray image classification are a promising area of research. However, at present, public datasets for COVID-19 x-ray images have low data volumes, making it challenging to develop accurate image classifiers. Several recent papers have made use of Generative Adversarial Networks (GANs) in order to increase the training data volumes. But realistic synthetic COVID-19 x-rays remain challenging to generate. We present a novel Mean Teacher + Transfer GAN (MTT-GAN) that generates COVID-19 chest x-ray images of high quality. In order to create a more accurate GAN, we employ transfer learning from the Kaggle pneumonia x-ray dataset, a highly relevant data source orders of magnitude larger than public COVID-19 datasets. Furthermore, we employ the Mean Teacher algorithm as a constraint to improve stability of training. Our qualitative analysis shows that the MTT-GAN generates x-ray images that are greatly superior to a baseline GAN and visually comparable to real x-rays. Although board-certified radiologists can distinguish MTT-GAN fakes from real COVID-19 x-rays, quantitative analysis shows that MTT-GAN greatly improves the accuracy of both a binary COVID-19 classifier as well as a multi-class pneumonia classifier as compared to a baseline GAN. Our classification accuracy is favorable as compared to recently reported results in the literature for similar binary and multi-class COVID-19 screening tasks.**

***Index Terms*—Coronavirus, deep transfer learning, mean teacher, artificial intelligence, diagnostic radiology, x-ray.**

I. Introduction

The SARS-CoV-2 novel coronavirus was reported to originate from Wuhan, Hubei province, China in 2019. COVID-19, the disease caused by this virus, is transmitted by inhalation or contact with infected droplets, and the incubation period ranges from 2 to 14 days [1]. In the study of a patient who was a worker at the market and was admitted to the Central Hospital of Wuhan on 26 December 2019, the patient was reported to be experiencing a severe respiratory syndrome that included fever, dizziness and a cough which proved to be one of the major symptoms of the virus. The complete 16 biological analysis stated that the virus showed similarities to a group of SARS-like coronaviruses which was previously found in bats in China [2], [3]. Across 150 states, over 750,000 individuals were reported to be infected by SARS-CoV-2 with a death rate of 4% [4], [5].

Rapid testing is a major need across the world. The primary testing modality is molecular testing [13], of which nucleic acid testing for discriminating genes is the dominant approach. However, a challenge is that nucleic acid testing requires culturing a sample, which can take several days. An alternative is *rapid* serological testing [12] which detects COVID-19 antigens. However, rapid serological testing is not intended, and may be less effective, for detecting the currently infected individuals. It is nevertheless widely used in hospitals even for this purpose due to its rapid turnaround time despite the high potential for false negatives [14]. An alternative modality that is sometimes employed to make use of x-ray or CT to detect the presence of traces of pulmonary infectious or inflammatory processes [15], [16]. For diagnostic radiologists, CT scans are a modality that may offer discriminating power of the disease at early stages [15]. Chest x-rays are also used, and much more widely available than CT. But x-rays are often difficult to interpret, as several indicators of COVID-19 infection, such as ground glass opacities (GGO), are not practical to discern by human eye in x-rays versus CT scans by radiologists [16].

Many recent papers have attempted to develop a deep learning classifier for COVID-19 using x-ray imagery [17]–[20]. Image classification has the potential to provide immediate testing results by identifying distinguishing imaging biomarkers. Image classification algorithms for COVID-19 have relied heavily on public datasets; in particular the covid-chestxray-dataset in conjunction with the Kaggle pneumonia competition dataset [21]–[23]. Due to the availability of these datasets, it is not uncommon for investigators to construct a multi-class classifier to distinguish Normal x-rays, Bacterial pneumonia, Viral pneumonia and COVID-19. However, public availability of COVID-19 x-ray datasets have limited data

This research was supported by NSF award titled *RAPID: Deep Learning Models for Early Screening of COVID-19 using CT Images*, award # 2027628.

volumes numbering in low hundreds of images. As such, several recent papers have investigated ways of increasing data volumes by making use of Generative Adversarial Networks (GANs) for deep augmentation [5], [32]. To the best of our knowledge, no GAN algorithms for COVID-19 chest x-rays, including ours, have achieved clinical quality for use by human radiologists. Yet, improvements to image quality translate to improved classification accuracy for automated screening algorithms. A difficulty with GAN-generated COVID-19 x-rays is the presence of fuzzy boundaries over major anatomical features such as heart, liver, and ribcage [5], [32]. Fuzzy image quality for generated x-rays is attributable to insufficient data volumes of COVID-19 images. Nevertheless, these deep generated x-rays still yield discriminating features and improve classification accuracy [5], [32]. One of the most discriminating features of COVID-19 versus other causes of pneumonia is the presence of ground glass opacities on CT that appear less electron-dense relative to consolidation and more electron-dense relative to normal healthy lungs [15], [16]. Subtle differences in opacities, though difficult to discriminate by qualitative visual assessment, could potentially be detectable through deep learning even in images that cannot reliably generate crisp boundaries around organs and the rib-cage.

Ideally, however, the generated COVID-19 x-rays should achieve the highest quality possible such as to approximate the real COVID-19 images. For example, if the generated COVID-19 images are overly fuzzy, then a classifier might mistakenly learn that the fuzzy images are indicative of COVID-19, but the sharply defined images are non-COVID-19.

The proposed MTT-GAN architecture greatly improves image quality through transfer learning from the Kaggle pneumonia dataset. In order to further improve the accuracy for screening, we propose to combine this transfer learning with an exponential moving average approach based on the mean teacher algorithm. Although transfer learning from ImageNet is widely adopted and employed for COVID-19 classification, ImageNet is not an x-ray dataset. MTT-GAN is unique in that it employs transfer learning to both the generator and discriminator from the Kaggle pneumonia x-ray dataset not ImageNet. The Kaggle pneumonia x-ray dataset is close to the target domain and is orders of magnitude larger than the covid-chestxray-dataset, thereby making an ideal data source for transfer learning to COVID-19.

MTT-GAN is also unique because it employs the exponential moving average component of the mean teacher algorithm [7]. Mean Teacher combines two models: a student and teacher in which the teacher performs exponential moving average of student weights in order to estimate an improved gradient direction. The mean teacher algorithm makes the gradient descent converge more consistently and to a better global optimum than Adam optimization alone for both fully supervised and semi-supervised models [7].

II. Related Work

Perhaps the most similar recent work to ours is that of Loey et al. (2020) [5] which employs conditional GAN augmentation to improve the accuracy of multiclass classification to identify COVID-19 versus normal, bacterial pneumonia and viral pneumonia as well as binary classification between COVID-19 and normal x-rays [5]. The testing accuracy on 4 classes (covid, normal, bacterial pneumonia and viral pneumonia) was 66.7%, 80.56% and 69.46% using AlexNet, GoogleNet and ResNet18 respectively [5]. We have trained and tested our model on a comparably large amount of data with 4 classes and we achieve a test accuracy of 83.45% and 84.91% on VGG-19 and AlexNet respectively.

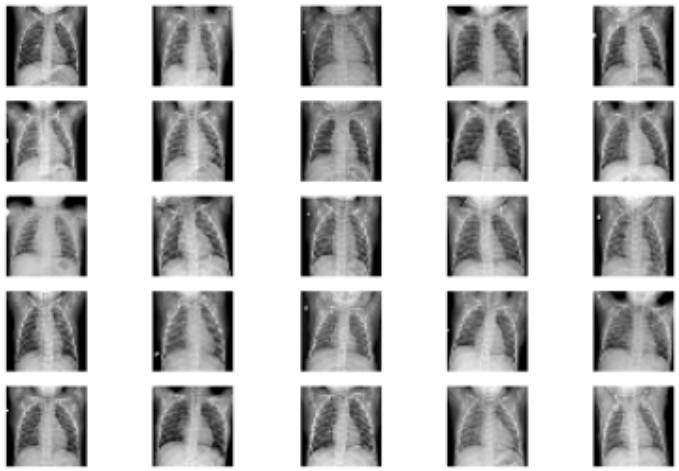

Fig. 1. Kaggle pneumonia/Normal Chest x-rays

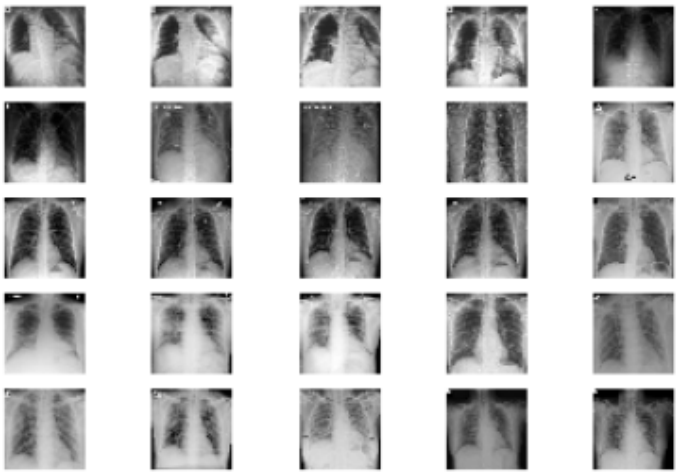

Fig. 2. COVID-19 Chest x-rays

A notable distinction between our approach and Loey et al. (2020), is that we propose the use of transfer learning from Kaggle pneumonia to train the GAN generator and discriminator models, rather than only the final classification step [5]. As such, MTT-GAN is capable of generating higher quality images with more anatomical detail.

Narin et al. (2020) also generate synthetic COVID-19 x-rays using GANs [32]. The authors compare 3 different generator architectures viz., ResNet50, InceptionV3 and InceptionResNetV2, of which ResNet50 provides the best accuracy.

Binary classification is performed between the COVID-19 and the normal x-rays by training and testing (10 COVID-19 + 10 normals x-rays) on a relatively small data-set. [32]. Similarly to Narin et al. (2020), we make use of a ResNet-like architecture for the generator. We extend this approach by incorporating Mean Teacher and transfer learning from the Kaggle pneumonia dataset. Our classification results are comparable but evaluated with a larger testing dataset [32].

The use of GANs to improve pulmonary disease classification with x-rays was first performed in 2018 simultaneously by Salehinejad et al. (2018) and Madani et al. (2018) [35], [36]. Salehinejad et al. (2018) generate chest x-rays to improve the performance of a classifier model for five categories of lung diseases [36]. Qualitatively, a board certified radiologist was able to identify the features pertaining to the five categories in the generated images. Quantitatively, the use of the GAN to augment their dataset improved the performance of the classifier. In the same year, Madani et al. (2018) trained two deep convolutional GANs to generate normal x-rays and x-rays with cardiovascular abnormalities [35]. The authors compared the accuracy of a classifier using unaugmented training data, data augmented using traditional methods, such as shifts and cropping, and data augmented using a GAN and traditional methods.

In addition to works that have incorporated the use of GANs a variety of papers have worked on classifiers between COVID-19, bacterial pneumonia, viral pneumonia, and normal healthy lungs [18], [19]. Sethy, et al. (2020) have created a dataset of 381 x-rays amongst 3 classes: COVID-19, pneumonia and normal [18]. The authors compare several models and achieve the highest accuracy of 98.6% with a ResNet50 plus SVM architecture as compared to the 93.4% by the traditional approach [18]. Wang, L. et al. (2020) have investigated a similar classification problem with a novel COVID-NET architecture featuring a lightweight residual projection-expansion projection-extension (PEPX) design pattern [19].

III. Data Preparation

The datasets used for our study are the Kaggle pneumonia chest x-rays dataset [21] and the COVID-19 open source chest x-rays [22], [23]. The Kaggle pneumonia dataset consists of 5,856 x-ray images (JPEG) with 3 categories (normal, bacterial pneumonia, viral pneumonia). This dataset (anterior and posterior) was taken from a collection of pediatric patients. Some sample images from the Kaggle dataset are shown in Fig. 1. The COVID-19 dataset, after removing images from patients without COVID-19 and those from a lateral view, consists of 227 x-ray images of patients with the coronavirus. Some sample images from the COVID-19 dataset are shown in Fig. 2. For our study, we downsampled all of the images to 128x128 resolution.

Both the Kaggle pneumonia and COVID-19 datasets are augmented using soft augmentation through cropping. For each original image, we generate a certain number of augmented images, which we call the augmentation factor. For each augmented image, the image is cropped on all four sides by a percentage randomly chosen between zero and five percent. Thus, at a minimum, the middle 90% of the original image, measured in both the horizontal and vertical directions, remains. For the Kaggle pneumonia dataset, we choose an augmentation factor of 5; for the COVID-19 dataset, we choose an augmentation factor of 50.

We were careful not to employ horizontal flipping. Imposing a horizontal flip would cause the cardiac silhouette to appear on the opposite side of the body, which would be clinically incorrect for the majority of patients.

IV. Methodology

A. Transfer Learning

Transfer learning is a method for addressing the problem of insufficient training data by using additional data from another larger source [8], [11], [25], [26]. Due to the low volume of COVID-19 images available, we wish to employ transfer learning from the Kaggle pneumonia dataset. Intuitively, we assume that the ideal weights for the model for generating COVID-19 x-rays are closer to the weights of the model after training on pneumonia and normal x-rays (the Kaggle dataset) than they are to the weights at initialization. For the most part, COVID-19 x-rays are more similar to Kaggle pneumonia x-rays than they are to white noise. Thus, we expect that the difference between the ideal COVID-19 weights of the model and the weights after transfer learning to be substantially smaller than the difference between the COVID-19 weights and the randomly initialized weights.

The MTT-GAN is first trained to convergence on the Kaggle dataset and is subsequently fine tuned on the COVID-19 dataset. To improve convergence, the fine tuning makes use of an exponential moving average learning algorithm based on Mean Teacher [7]. Care was taken to ensure that the training and testing splits were completely separated, both for pre-training and fine tuning. 30% of the real covid x-rays (i.e. 68 x-rays) were removed from the COVID-19 dataset and 68 images of each class (normal, bacterial pneumonia, and viral pneumonia) were removed from the Kaggle dataset. For each training epoch, the training was split into mini-batches of size 32. For each mini-batch, the discriminator model was trained using 16 real images (labeled 1) and 16 fake images (labeled 0), with the fake images generated by passing Gaussian noise on the interval [0, 1] through the generator model. Then, the combined model, consisting of both the discriminator and the generator, was trained using 32 noise vectors. A crossentropy loss was employed with the Adam optimizer using a learning rate of 10^{-5} and a β_1 of 0.5. The training was run for 100 epochs using the Kaggle dataset followed by 100 epochs with the exponential moving average algorithm using the COVID-19 dataset.

B. Exponential Moving Average Training

MTT-GAN employs a supervised version of the Mean Teacher algorithm featuring exponential moving average of model weights in order to improve training convergence of the generator and discriminator as seen in figure 3. Mean

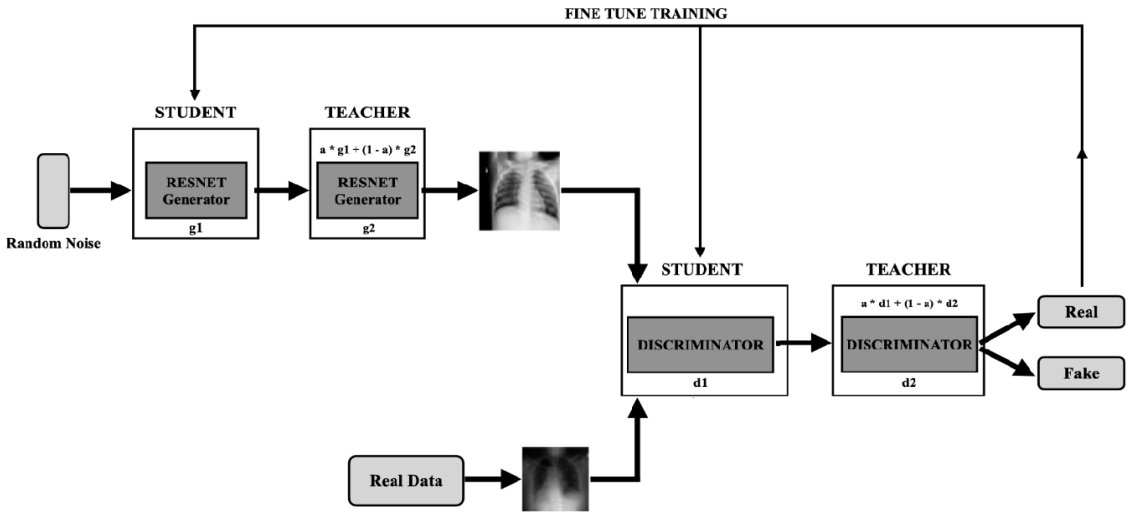

Fig. 3. MTT-GAN fine-tuning algorithm showing student and teacher for generator and discriminator models.

Teacher is equally applicable to both semi-supervised as well as supervised learning, and it is widely regarded as a state-of-the-art semi-supervised learning approach [7]. Mean teacher employs two models in parallel, a student and a teacher model. After each gradient step, the student model performs gradient descent, but the teacher model is updated to use an exponential moving average of the student weights. Due to the use of this exponential moving average, the teacher model is expected to converge faster than the student model. As such, the teacher model is ultimately used for classification [7].

The training loss of MTT-GAN differs from the original Mean Teacher, because MTT-GAN does not include a consistency loss between pseudo labels of the teacher models and the predicted labels of the student models. The original Mean Teacher incorporates this consistency loss, in which the teacher model predicts the labels of unlabeled samples, and the student model performs an additional gradient descent toward the teacher's predictions. However, this step is not necessary for MTT-GAN, because the minimax loss function is completely supervised [7].

The student discriminator loss is defined as follows,

$$L_D = -\mathbf{E}_{x_r \in S} \log(D(x_r)) - \mathbf{E}_{z \in Z} \log(1 - D(G(z))) \quad (1)$$

where S is the set of real images and Z is the latent space from which noise, z, taken. On the other hand, the student generator, G, tries to optimize its weights to maximize that same equation, or equivalently, minimize the equation,

$$L_G = -\mathbf{E}_{z \in Z} \log(D(G(z))) - \mathbf{E}_{x_r \in S} \log(1 - D(x_r)) \quad (2)$$

After each gradient step of the student generator and student discriminator models as seen in figure 3, the respective teacher's weights are updated using an exponential moving average of the student weights as follows,

$$\theta_t^{'} = \alpha\, \theta_{t-1}^{'} + (1 - \alpha)\, \theta_t \quad (3)$$

Where $\theta_t^{'}$ represents the weights of the respective teacher model at time-step t, whereas θ_t represents the weights of the respective student model at time-step t.

C. GAN Architecture

The MTT-GAN architecture consists of two separate models, a generator network and a discriminator network. Our discriminator network, shown in Fig. 4, consists of a total of nine convolutional layers, each with a kernel size of 3x3, and each followed by a LeakyRelu activation function with an alpha value of 0.2. The first three layers have 64 filters, the next three have 128 filters, and the last three have 256. The first, fourth, and seventh layers have a stride of 2x2, the other six layers have a 1x1 stride. Additionally, a dropout layer with a dropout rate of 0.4 is included after every third layer. The last convolutional layer is fully connected to a layer with one node, and the sigmoid activation function is used.

The generator network, shown in Fig. 5, takes a vector in 100-dimensional latent space. This is followed by a fully connected layer that is shaped into 128 feature maps of size 16x16. The generator contains three blocks, each of which increases the size of the feature maps. Each of these blocks starts with a deconvolutional layer with 128 filters, a 4x4 kernel, and a 2x2 stride, doubling both dimensions of the feature map. Each of these upscaling deconvolutional layers is followed by two residual blocks. Each residual block consists of a deconvolutional layer (128 filters, 4x4 kernel), a batch normalization layer with a momentum of 0.8, a leakyRelu activation function, another deconvolutional layer (same specifications), and another leakyRelu activation. After all three different-sized blocks, a convolutional layer with a 3x3 kernel is used with a sigmoid activation. This generates a 128x128 image with the pixel intensities normalized to the interval [0, 1]. The loss function used was binary cross-entropy and we used the adam optimizer with a learning rate of 0.00001 and the size of each mini-batch was 32.

Residual networks, or ResNets, allow for a model to be deeper without the gradients vanishing or exploding by creating "skip connections," where a portion of the model learns the difference between the output and the input rather than learning the output from scratch [27]. ResNets have been used extensively, including with similar datasets as in Wang et al. (2020), where a residual architecture is used in a classification model for COVID-19 x-ray images [31]. While first used in CNNs, Gulrajani et al. (2017) demonstrated that residual architectures can be used to improve the performance of a GAN [30]. Since then, ResNets have become commonly used in the generator portion of a GAN, such as in Ledig et al. (2017), where a ResNet is used in the generator of a super-resolution GAN. As such, we employ a residual network in our generator architecture [28].

V. Experimental Design

The MTT-GAN architecture was evaluated by observing the effect on a classifier model when the training dataset is augmented using images generated by the GAN. Two classifier models were used for our evaluation, VGG-19 [33] and AlexNet [34]. The architectures of the models were modified only to change the number of outputs to two and four for the binary and multi-class classification problems respectively. For all experiments, the classifiers were trained from scratch for 50 epochs. The adam optimizer was used with a learning rate of 10^{-5}. For classifier training, 30% of the training data was reserved for validation.

Further, the MTT-GAN was compared against a Transfer-GAN and a baseline GAN. The baseline GAN has the same generator and discriminator architectures, but does not use the mean-teacher exponential moving average or the transfer learning. The transfer-GAN is similar but does not incorporate the exponential moving average training. All three GANs were trained using a crossentropy loss and the adam optimizer with a learning rate of 10^{-5}. Both GANs were trained for 100 epochs on the specified COVID-19 dataset, and the MTT-GAN is trained on the transfer dataset prior to the COVID-19 training.

For all experiments, 68 images of each class are withheld for testing. For the binary classifier, the classes are COVID-19 and normal, and for the multi-class classifier, the classes are COVID-19, normal, bacterial pneumonia, and viral pneumonia. The images withheld for testing are not used in any way for training the GAN models or the classifier models.

The training dataset consists of 1400 images of each class prior to the validation split (30% of the training data is used for validation). Thus, a total of 2800 images are used for training the binary classifier, and 5600 images are used for the multi-class classifier. For the normal, bacterial pneumonia, and viral pneumonia classes, the images used for training are all real images taken from the Kaggle dataset. For the COVID-19 class, the composition of the images varies between experiments. Six experiments are run on each of the classifiers. In the first two experiments, the COVID-19 images are images generated by the baseline GAN model. In the first, data augmentation is not used for training the GAN, while in the second, data augmentation is used, using an augmentation factor of 50 (the same as in training the MTT-GAN on COVID-19 images). In the third experiment, the COVID-19 images are images generated by the Transfer-GAN model. In the fourth experiment, the 1400 images consist of the 159 real COVID-19 images (the remaining images after 68 were reserved for testing) and 1241 images generated using the Transfer-GAN. The fifth and sixth experiments are similar but employ the full MTT-GAN with exponential moving average training and transfer learning. For each experiment, we report the testing accuracy, and for the experiments with only GAN-generated images, we also report a confusion matrix. Furthermore, confidence intervals for the reported accuracy numbers are included in Tables I and II using the Clopper-Pearson method.

VI. Quantitative Results

TABLE I
Binary Classifier Experiments

Training Dataset (2800 x-rays = 1400 covid x-rays+1400 normal x-rays)	*VGG-19 Test Accuracy with confidence intervals (136 x-rays = 68 covid x-rays+68 normal x-rays)*	*AlexNet Test Accuracy with confidence intervals (136 x-rays = 68 covid x-rays+68 normal x-rays)*
Baseline Method (without augmentation) with only Generated Images	**74.26%** (0.66,0.81)	**77.20%** (0.69,0.83)
Baseline Method (with augmentation) with only Generated Images	**91.91%** (0.85,0.95)	**82.35%** (0.74,0.88)
Transfer GAN with only generated x-rays	**95.58%** (0.90,0.98)	**93.38%** (0.87,0.96)
Transfer GAN with Real and Generated covid x-rays (1400 = 159 real + 1241 generated)	**99.26%** (0.95,0.99)	**99.26%** (0.95,0.99)
MTT-GAN with only generated x-rays	**96.32%** (0.91,0.98)	**94.11%** (0.88,0.97)
MTT-GAN with Real and Generated covid x-rays (1400 = 159 real + 1241 generated)	**99.26%** (0.95,0.99)	**100%** (1,1)

Quantitative analysis over the binary and multi-class classifiers demonstrate that MTT-GAN achieves the highest classification accuracy and that both Transfer GAN and MTT-GAN achieve superior accuracy over the baseline GAN. Furthermore, the highest accuracy is achieved when a small amount of real imagery is included along with the generated COVID-19 imagery. Table I shows the quantitative test accuracy for binary COVID-19 vs. normal classification. The first experiment, where the models were trained on the combination of baseline GAN images and the normal x-rays, gave an accuracy of 74.26% with a confidence interval of (0.66068,0.81374) for VGG-19 and 77.20% (0.69233,0.83957) for AlexNet. In the second experiment, incorporating 159 real x-ray images increases this accuracy to 91.91% (0.85989,0.95893) and 82.35% (0.74890,0.88354) respectively. For the Transfer GAN

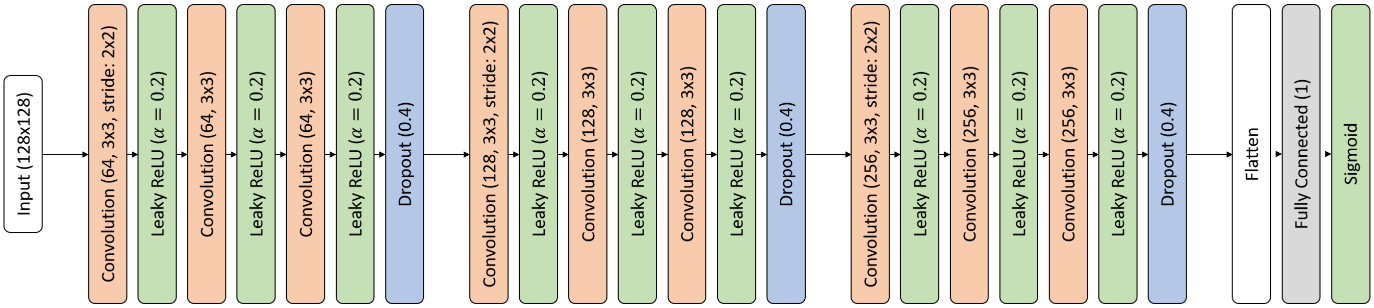

Fig. 4. Discriminator Architecture

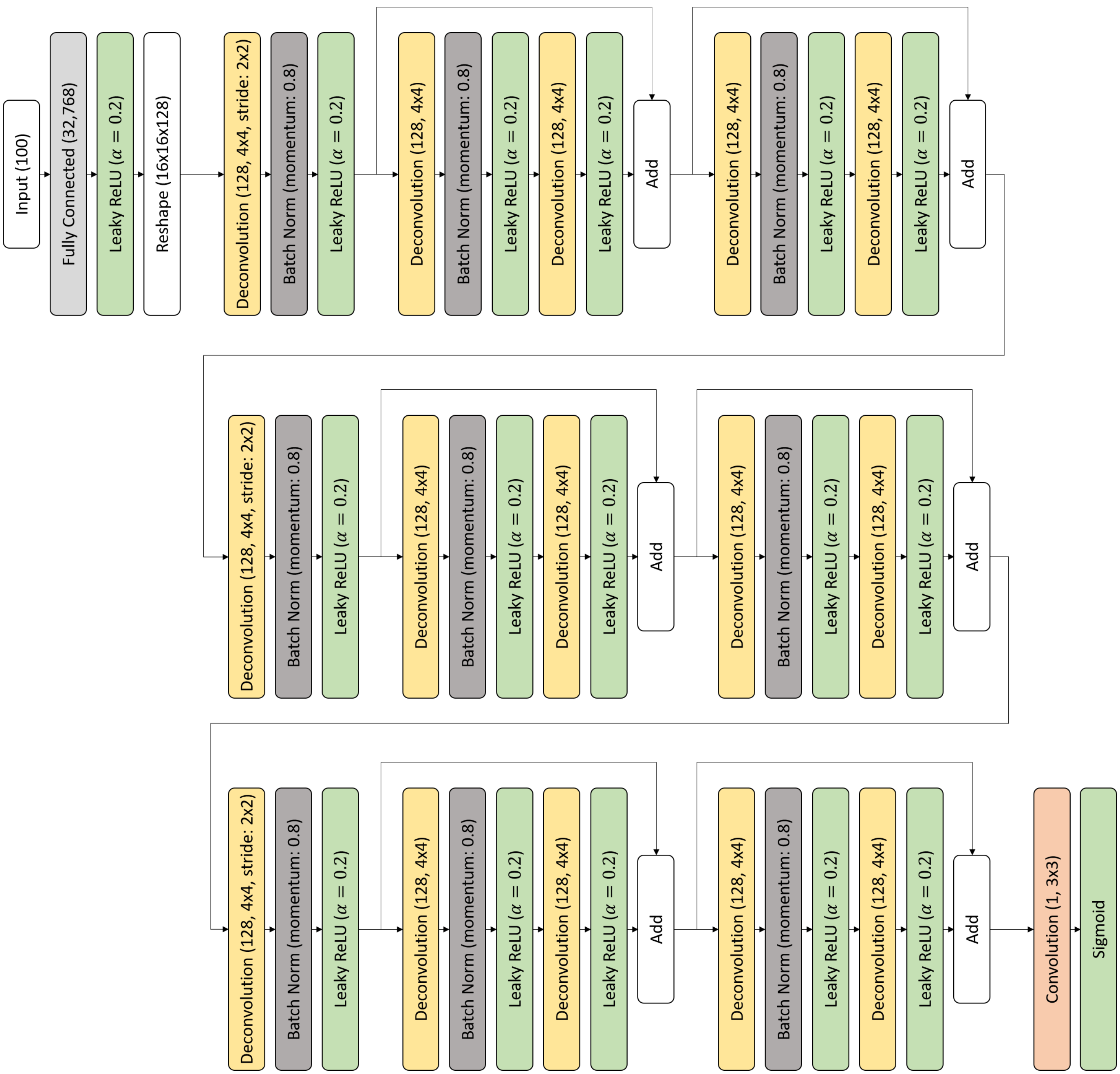

Fig. 5. Generator Architecture

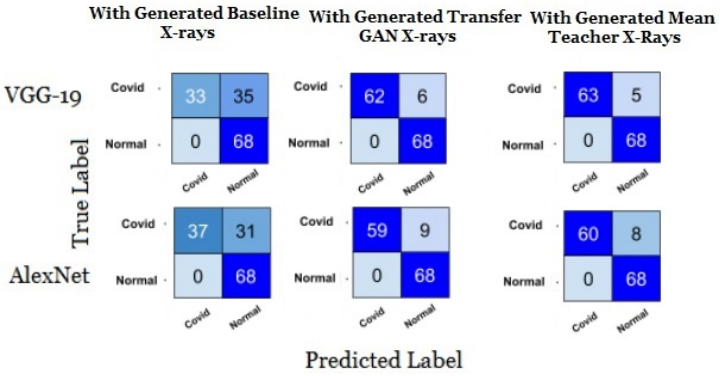

Fig. 6. Binary Classification Confusion Matrix

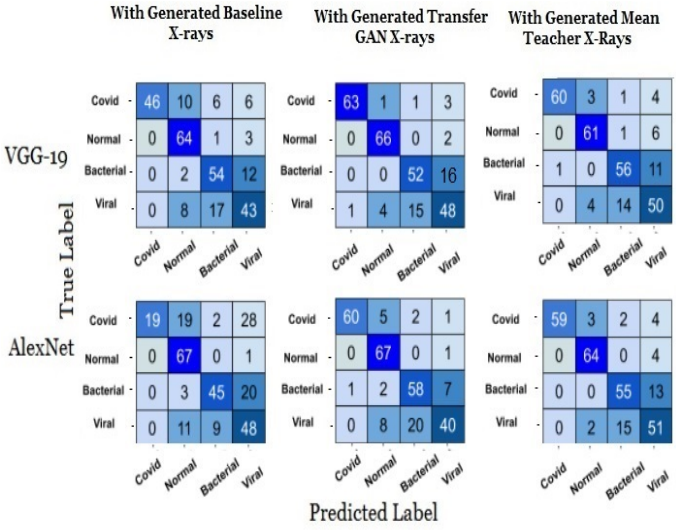

Fig. 7. MultiClass Classification Confusion Matrix

experiments, the VGG-19 model gives an accuracy of 95.58% (0.90,0.98) and the AlexNet gives us an accuracy of 93.38% (0.87,0.96). with GAN images only. The fourth experiment on binary classification consisted of 159 real covid x-rays, 1241 generated covid x-rays and 1400 real normal x-rays, and the VGG-19 model yielded a test accuracy of 99.26% (0.95,0.99) and the AlexNet yielded an accuracy of 99.26% (0.95,0.99). Experiments five and six show that MTT-GAN further improves accuracy. Using only MTT-GAN augmentations gives us an accuracy of 96.32% (0.91,0.98) and 94.11% (0.88,0.97) respectively. Combining MTT-GAN agumentations with 159 real x-rays, VGG-19 and AlexNet produce their highest accuracy of 99.26% (0.95,0.99) and 100% (1,1) respectively. Fisher tests show that the improvement in accuracy of MTT-GAN versus Baseline is significant with p-values < 0.0001 for all rows that compare MTT-GAN with Baseline Method without augmentation, and $p < 0.05$ for all rows that compare MTT-GAN with Baseline with augmentation, except for VGG with only generated versus Baseline, which is near significant with augmentation where $p = 0.064$. Furthermore, the confusion matrices in figure 6 demonstrate that MTT-GAN greatly increases the sensitivity relative to baseline GAN.

Table II shows similar results for the multi-class classifiers. The baseline GAN with only generated images achieves the lowest results, where the VGG-19 model yields an accuracy of 76.10% with a confidence interval of (0.70,0.81), and the AlexNet yields an accuracy of 65.80% (0.59,0.71). In the second experiment for multi-class classification, using the baseline GAN with augmentation, the VGG-19 model yields an accuracy of 79.41% (0.74,0.84), and the AlexNet yields an accuracy of 76.47% (0.70,0.81). When using images generated with the Transfer GAN, the accuracy is 84.19% (0.79,0.88) using VGG-19, whereas the AlexNet yields an accuracy of 82.72% (0.77,0.87). Combining the images generated by the Transfer GAN with the real images gives an accuracy of 84.92% (0.80,0.88) using VGG-19 and 83.89% (0.78,0.87) using AlexNet. MTT-GAN with only generated x-rays achieved a slightly lower accuracy for VGG-19 of 83.45% (0.78,0.87), but an improved accuracy of 84.19% (0.79,0.88) for AlexNet. The highest accuracy was achieved by combining MTT-GAN with 159 real COVID-19 x-rays, yielding 84.93% (0.80,0.88) accuracy for VGG-19 and 85.61% (0.80,0.88) accuracy for AlexNet.

TABLE II
MULTICLASS CLASSIFIER EXPERIMENTS

Training Dataset (5600 x-rays = 1400 covid x-rays+1400 normal x-rays+1400 bacterial pneumonia+1400 viral pneumonia)	*VGG-19 Test Accuracy with confidence intervals (272 x-rays = 68 covid x-rays+68 normal x-rays + 68 bacterial + 68 viral)*	*AlexNet Test Accuracy with confidence intervals (272 x-rays = 68 covid x-rays+68 normal x-rays + 68 bacterial + 68 viral)*
Baseline Method (without augmentation) with only Generated Images	**76.10%** (0.70,0.81)	**65.80%** (0.59,0.71)
Baseline Method (with augmentation) with only Generated Images	**79.41%** (0.74,0.84)	**76.47%** (0.70,0.81)
Transfer-GAN method with only generated x-rays	**84.19%** (0.79,0.88)	**82.72%** (0.77,0.87)
Transfer-GAN method with Real and Generated covid x-rays (1400 = 159 real + 1241 generated)	**84.92%** (0.80,0.88)	**83.89%** (0.78,0.87)
MTT-GAN with only generated x-rays	**83.45%** (0.78,0.87)	**84.19%** (0.79,0.88)
MTT-GAN with Real and Generated covid x-rays (1400 = 159 real + 1241 generated)	**84.93%** (0.80,0.88)	**85.61%** (0.80,0.88)

The confusion matrix in Fig. 7 shows that the VGG-19 classifier is able to predict 46 out of 68 covid images accurately with the baseline generated x-rays, whereas it predicts 63 of the 68 x-rays accurately using the x-rays generated by our MTT-GAN algorithm and architecture, which is detecting 92.6% of the covid cases accurately. Using the AlexNet, the baseline predicts only 19 out of the 68 positive covid cases,

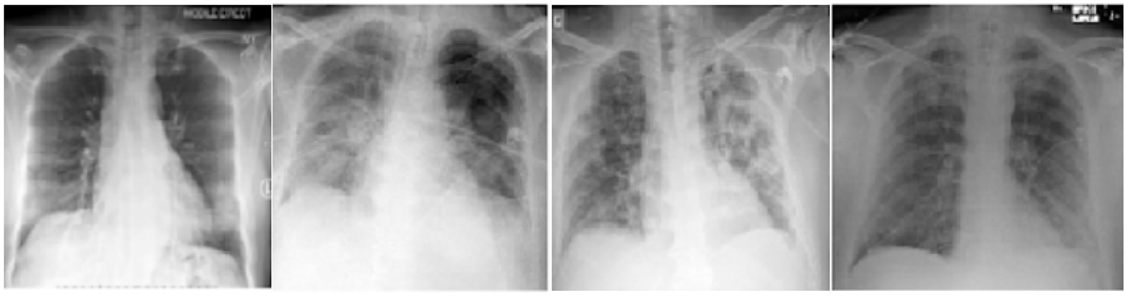

Fig. 8. Real COVID-19 x-rays

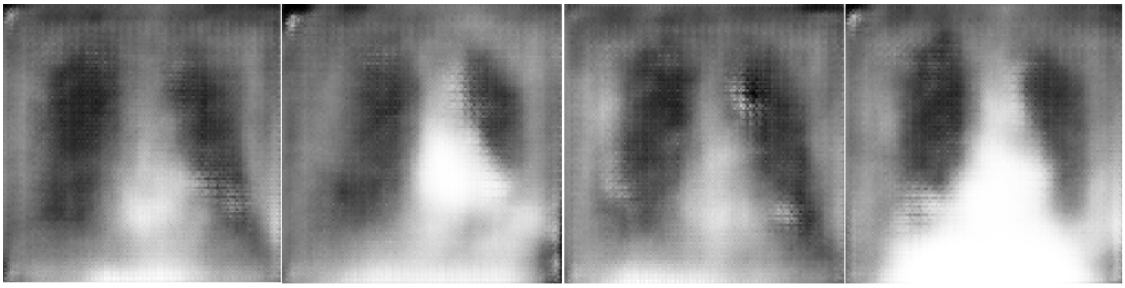

Fig. 9. Baseline Model Generated x-rays

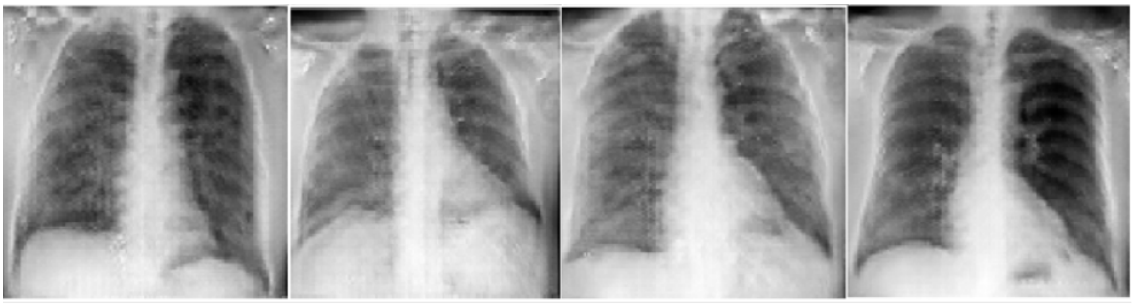

Fig. 10. X-rays Generated using MTT-GAN

whereas with the MTT-GAN, 60 out of the 68 positive images are correctly identified. Fisher tests show that the MTT-GAN improves accuracy relative to the baseline. We achieve $p < 0.001$ for VGG comparisons of MTT-GAN versus Baseline without augmentation, and $p < 0.0001$ for equivalent Alex-net comparison. Furthermore, $p < 0.05$ for VGG comparison of MTT-GAN versus Baseline with augmentation, and $p < 0.01$ for equivalent Alex-net comparison.

VII. Qualitative Results

Our qualitative results show that MTT-GAN greatly outperforms the baseline GAN and generates images that approximate many anatomical features of the real images. The real images are shown in Fig. 8, while the baseline generated x-rays and the x-rays generated with transfer learning and mean teacher look are shown in Fig. 9 and Fig. 10 respectively.

Although the MTT-GAN generated x-rays have many well defined anatomical features, they are nonetheless distinguishable from real x-rays by board-certified radiologists. We conducted a survey with two board-certified diagnostic radiologists, each with over 6 years of clinical experience interpreting chest radiographs, where we displayed 25 pairs of real covid x-rays versus generated covid x-rays. We asked the radiologists to classify between the real and the generated x-rays and asked them to provide comments on the features that the generated x-rays have in comparison to the real x-rays. Both radiologists were able to correctly identify which image was real and which was fake in all 25 cases.

The radiologists mentioned that the x-rays have greatly improved quality relative to the baseline, but fall short of diagnostic quality due to the following limitations: 1. low resolution (128x128) and the methodology would further need to be improved, namely with increased available memory capacity, a more memory efficient process, or increased available processing time, in order to produce diagnostic quality images. 2. systematic errors in the scapula and the clavicle bones as

highlighted in Fig 11.

The radiologists suggested that a potential area of future work would be to incorporate skeletal background removal and/or style transfer methods to ensure that background features such as the scapula and clavicle bone structures are consistent between generated and real images. Additional more subtle differences were perceptible by the board-certified radiologists in the extracorporeal space and along the cardiomediastinal sillouette.

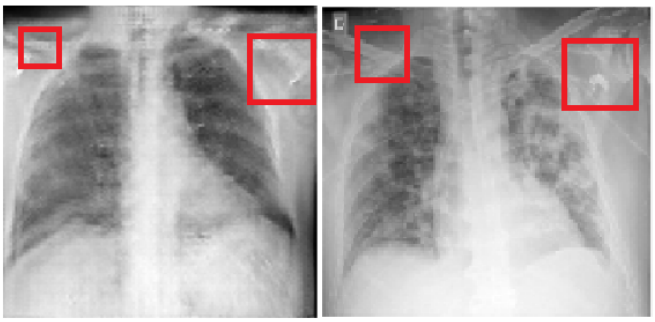

Fig. 11. Survey Results by Radiologists

VIII. Conclusion

We present a novel MTT-GAN architecture for generating high quality synthetic chest x-ray images for patients with COVID-19, and we demonstrate improved accuracy of binary and multi-class classifiers for automated COVID-19 x-ray screening. MTT-GAN addresses a notable challenge in that public datasets for COVID-19 x-rays have highly limited data volumes. To the best of our knowledge MTT-GAN is the first architecture to employ transfer learning from Kaggle pneumonia to COVID-19 for both the generator and discriminator models, thereby greatly improving image quality. This improved image quality translates to highly competitive COVID-19 classification accuracy. To the untrained eye, MTT-GAN images appear similar to real COVID-19 x-rays, although board certified radiologists can distinguish these images and suggest that more research is necessary to achieve diagnostic quality for human performance tasks. Nevertheless, quality improvements to deep fakes are invaluable to improve classification accuracy for computer aided diagnosis. In conclusion, MTT-GAN is a novel approach that provides a notable improvement in the realism of generated deep fake COVID-19 x-rays images.

References

[1] Singhal, T. A Review of Coronavirus Disease-2019 (COVID-19). Indian J. Pediatrics 2020, 87, 281–286.

[2] Wu, F., Zhao, S., Yu, B. et al. A new coronavirus associated with human respiratory disease in China. Nature 579, 265–269 (2020).

[3] Hu, D. et al. Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bat. Emerg. Microbes Infect. 7, 1–10 (2018).

[4] Shereen, M.A.; Khan, S.; Kazmi, A.; Bashir, N.; Siddique, R. COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. J. Adv. Res. 2020, 24, 91–98

[5] Mohamed Loey, Florentin Smarandache, Nour Eldeen M. Khalifa. Within the Lack of Chest COVID-19 x-ray Dataset: A Novel Detection Model Based on GAN and Deep Transfer Learning (2020)

[6] Goodfellow, Ian J., Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron C., and Bengio, Yoshua. Generative adversarial nets. NIPS, 2014.

[7] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in Neural Information Processing Systems, 2017.

[8] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," 2018, arXiv:1808.01974. [Online]. Available: https://arxiv.org/abs/1808.01974

[9] Dai, W., Yang, Q., Xue, G.R., Yu, Y.: Boosting for transfer learning. In: Proceedings of the 24th international conference on Machine learning. pp. 193–200. ACM (2007)

[10] Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q.: Domain adaptation via transfer component analysis. IEEE Transactions on Neural Networks 22(2), 199–210 (2011)

[11] Huang, J.T., Li, J., Yu, D., Deng, L., Gong, Y.: Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. pp. 7304–7308. IEEE (2013)

[12] Bastos, M. L., Tavaziva, G., Abidi, S. K., Campbell, J. R. and Haraoui, L.P., and Johnston, J.C. and Lan, Z., Law, S., MacLean, E., Trajman, A., et. al, "Diagnostic accuracy of serological tests for covid-19: systematic review and meta-analysis," British Medical Journal (BMJ) vol. 370 (2020)

[13] Kubina, R., and Dziedzic, A., "Molecular and Serological Tests for COVID-19 a Comparative Review of SARS-CoV-2 Coronavirus Laboratory and Point-of-Care Diagnostics," Diagnostics, 10(6), 434. (2020)

[14] Cassaniti, I., Novazzi, F., Giardina, F., Salinaro, F., Sachs, M., Perlini, S., ... & Baldanti, F., "Performance of VivaDiag COVID-19 IgM/IgG Rapid Test is inadequate for diagnosis of COVID-19 in acute patients referring to emergency room department." Journal of medical virology. (2020)

[15] Zhou, S., Wang, Y., Zhu, T., & Xia, L., "CT features of coronavirus disease 2019 (COVID-19) pneumonia in 62 patients in Wuhan, China," American Journal of Roentgenology, 214(6), 1287-1294 (2020).

[16] Shi, H. Han, X. Jiang, N. et al., "Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study." Lancet Infect. Dis. (2020)

[17] Hemdan E.E.D., Shouman M.A., Karar M.E. (2020)., "COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in x-ray Images.," arXiv preprint arXiv 2003.11055. (2020)

[18] Sethy, P. K., Behera, S. K., Ratha, P. K., & Biswas, P., "Detection of coronavirus Disease (COVID-19) based on Deep Features and Support Vector Machine." International Journal of Mathematical, Engineering and Management Sciences, 5(4), 643–651. (2020)

[19] Wang, L., & Wong, A., "COVID-Net: A Tailored Deep Convolutional Neural Network" arxiv preprint 2003.09871v1 (2020)

[20] Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., & Rajendra Acharya, U. (2020). Automated detection of COVID-19 cases using deep neural networks with x-ray images. Computers in Biology and Medicine, 103792. Advance online publication.

[21] Mooney, P. (2020). Chest x-ray Images (pneumonia). Kaggle. https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia

[22] Cohen, J.P., Morrison, P., Dao, L., "COVID-19 image data collection, arXiv:2003.11597," arXiv:2003.11597, (2020) https://github.com/ieee8023/covid-chestxray-dataset

[23] Cohen, J.P., Morrison, P., Dao, L., Roth, K., Duong, T.Q., Ghassemi, M., "COVID-19 Image Data Collection: Prospective Predictions Are the Future," arXiv:2006.11988 (2020) https://github.com/ieee8023/covid-chestxray-dataset,

[24] J. Lemley, S. Bazrafkan and P. Corcoran, "Smart Augmentation Learning an Optimal Data Augmentation Strategy," in IEEE Access, vol. 5, pp. 5858-5869, 2017, doi: 10.1109/ACCESS.2017.2696121.

[25] Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and transferring mid-level image representations using convolutional neural networks. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. pp. 1717–1724. IEEE (2014)

[26] Tang, Y., Cai, J., Lu, L., Harrison, A. P., Yan, K., Xiao, J., ... & Summers, R. M. (2018, September). CT image enhancement using stacked generative adversarial networks and transfer learning for lesion segmentation improvement. In International Workshop on Machine Learning in Medical Imaging (pp. 46-54). Springer, Cham.

[27] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[28] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., ... & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4681-4690).

[29] Laine, Samuli and Aila, Timo. Temporal Ensembling for Semi-Supervised Learning. arXiv:1610.02242 [cs], October 2016. arXiv: 1610.02242.

[30] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. In Advances in neural information processing systems (pp. 5767-5777).

[31] Wang, L., & Wong, A. (2020). 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.

[32] Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., & Pinheiro, P. R. (2020). Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection. IEEE Access, 8, 91916-91923.

[33] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[34] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[35] Madani, A., Moradi, M., Karargyris, A., & Syeda-Mahmood, T. (2018, March). Chest x-ray generation and data augmentation for cardiovascular abnormality classification. In Medical Imaging 2018: Image Processing (Vol. 10574, p. 105741M). International Society for Optics and Photonics.

[36] Salehinejad, H., Valaee, S., Dowdell, T., Colak, E., & Barfett, J. (2018, April). Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 990-994). IEEE.