# Transfer Learning Based GAN For Augmenting Covid-19 Chest X-Ray Dataset

Giacomo Fiorentini
g.fiorentini@students.uu.nl
4861310

Maria Galanty
m.galanty@students.uu.nl
8111875

Luca Lin
l.lin1@students.uu.nl
5522146

Jakub Myśliwiec
j.mysliwiec@students.uu.nl
0632759

***Abstract*—The Covid-19 pandemic is causing worldwide repercussions, with early and rapid detection of the disease being crucial in reducing its spread and mortality rate. Radiological findings can be used to detect the disease, and multiple machine learning models so far have learnt to recognize Covid-19 from chest x-rays with good results. One major obstacle in improving the accuracy and generality of these models is the small size of the available datasets. In this paper we employ StyleGan2 to generate synthetic images for the Covid-19 label of the Covidx5 dataset and then fine-tune an AlexNet model to classify images into three labels: Covid-19, Pneumonia and Normal. Three models are built for comparison: one on the base dataset, one on the augmented dataset and one the dataset with GAN-generated images. Results show that with a Frechet Inception Distance score of 73.471, the images generated by the GAN perform worse than the base and augmented images in multilabel classification. When comparing performance for the Covid-19 label we report ROC-AUC scores of 89.95, 87.06 and 84.26 for the Base, Augmented and GAN-based model respectively, indicating the generated images were not representative of the Covid-19 label. We hope future research into different GAN models can improve on our results.**

## I. Introduction

Coronavirus disease (Covid-19) is an infectious illness caused by a newly discovered coronavirus (SARS-CoV-2). It was initially detected in December 2019, in Wuhan, China, and has spread since then leading to the ongoing worldwide pandemic. This virus spreads mainly through droplets of saliva or discharge from the nose when an infected person coughs or sneezes. The incubation period of the virus, which is a time between being exposure to the virus and developing symptoms is usually 5-6 days, but can take up to 14 days. Due to long incubation period, easy transmission of the virus and the possibility of asymptomatic course of the disease, Covid-19 spreads very quickly from human to human. Around 80% of people who develop symptoms will experience mild to moderate symptoms and recover without requiring hospitalization. Around 15% become seriously ill and require specialized medical treatment and around 5% become critically ill and need intensive care. Older people and those with underlying medical conditions like high blood pressure, heart and lung problems, obesity are more likely to develop serious illness. However, anyone can become seriously ill regardless of age and health condition [1].

Radiological findings are deemed to be useful in aiding the diagnosis of pneumonia and Covid-19. Currently, a common diagnostic tool is laboratory testing, which is associated with high costs and long waiting period for the results. While diagnostic using the CXR is quick to capture, cheap and available, even in poorer world's areas [2]. Creation of a tool that supports the detection of Covid-19 on the basis of only X-rays could reduce the burden on medical personnel and improve risk prioritization [3]. This attracted the attention of machine learning a deep learning community and led to the construction of models enabling disease detection based on the X-ray images.

Due to the novelty of the Covid-19 virus, CXR images from Covid-19 patients are scarce. As a result, datasets made available such as Covidx5 [4] are either imbalanced in the class labels or small in overall size. Classic methods to combat class imbalance include resampling, and over- and under sampling, and data augmentations as a method for regularization when training the model. A more novel method includes the use of Generative Adversarial Networks (GANs) to generate synthetic images to populate the dataset.

GANs, originally described by Goodfellow et al. [5] are well known for excelling at generating realistic looking images within a given domain. The basic principle is that there are 2 networks: a generator and a discriminator which learn simultaneously during training. The generator tries to come up with some image that looks like it is from the training data while the discriminator tries to tell the real images apart from the generated ones.

This project explores transfer learning using a state of the art GAN, StyleGan2[6] to augment Covid-19 chest X-ray images. One key research question we aim to answer is: Does data augmentation of the Covid-19 label using a state of the art GAN trained with transfer learning improve image classification performance in the multi-label task by generating representative Covid-19 CXR images?

## II. Related Work

Lately, Generative Adversarial Networks (GANs) are considered to be one of the most effective data augmentation methods [7]. More recently they have been gaining traction within the domain of medical imaging. The literature review by Yi et al. [8] demonstrates a significant upward trend in the number of papers that appear each year in this field. After the Covid-19 world pandemic a need to diagnose patients on the basis of their X-rays arose. The lack of sick patient data due to privacy concerns led to the use of data augmentation techniques including GANs to improve performance of the Covid-19 detection task.

Waheed and al. [9] created a Auxiliary Classier Generative Adversarial Network (ACGAN) based model called CovidGAN to generate chest X-rays. Including synthetic images in the training set improved Covid-19 detection accuracy by CNN from 85% to 95%. ACGAN is an extension of the conditional GAN where discriminator apart from predicting whether an image is real also predicts which class it belongs to [10]. The dataset consisted of 403 images of Covid-19 patients and 721 healthy X-rays. CovidGAN was trained to synthesize X-rays for both Covid-19 and Normal class. It produced 1669 synthetic Covid-19 images and 1399 synthetic Normal class images.

Loey and al. [11] used GANs to produce more synthetic X-ray images, which were used later in three experiments involving classification problems with different number of classes and different transfer learning models: Alexnet, Googlenet and Restnet18. The initial dataset consisted of 69 Covid-19, 79 Normal, 79, Pneumonia bacteria, 79 pneumonia virus images, which gives a total of 306. The dataset increased to 8100 images after using GANs for all 4 classes. The test set consisted around 10% of the original dataset. The paper obtained the following results. Test accuracy for the best transfer learning model: four different diseases - 80.56%, three classes (excluding pneumonia virus) - 85.19%, two classes (Covid-19 and Normal) - 100%.

Khalifa and al. [12] were trying to distinguish pneumonia from healthy lungs by using GANs to enlarge a dataset and applying transfer learning using models such as AlexNet, GoogLeNet, Squeeznet, and Resnet18. Zulkifley and al. [2] used a conditional generative adversarial network (conditional GAN) to reduce imbalance between classes in the classification task for three different cases: Covid-19, viral pneumonia and normal. All three classes were used during training conditional GAN, but only the Covid-19 class was augmented with the synthetic data. The study proposes its own lightweight deep learning model with only 841,771 parameters. However, a potential limitation of this research is that they include the GAN generated images in the test set, potentially inducing bias in their model.

Comparing our work with the literature, for our research, we make use of the largest dataset available at the time, Covidx5. This provides us with a more reliable estimation of our models performance due to an increase in test data. Another difference is the GAN architecture we are using - unconditional GAN with transfer learning. Loey and al. also make use of unconditional GANs, but they are augmenting all classes with synthetic images, while we are using it only to expand the Covid-19 class. Previous work obtained the best accuracy for binary classification for normal and Covid-19 classes, therefore we will focus our efforts on multi label classification.

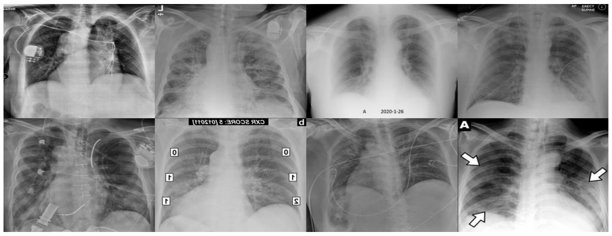

Fig. 1. example images of Covid-19 label in Covidx5 dataset

| | Normal | Pneumonia | Covid-19 |
|---|---|---|---|
| Train set | 6835 | 4673 | 450 |
| Development set | 1157 | 760 | 83 |
| Test set | 885 | 594 | 100 |

TABLE I
SPLIT OF THE DATASET.

## III. METHODOLOGY

### A. Dataset

Wang et al. [4] released Covidx5 in October 2020, a dataset containing 8851 images labeled as normal, 5778 images as pneumonia and 617 images as COVID-19. The images vary in positioning of the torso, contrast and contain artefacts such as arrows and textual information. Moreover, medical devices can be present in the images. Figure 1 shows some example images in dataset labeled Covid-19. To the best of our knowledge this is the largest Covid-19-related dataset available to the public at the moment of writing. Nevertheless, the number of relevant CXRs remains extremely limited, especially in the context of deep learning.

The dataset combines 5 different sources: COVID-19 Image Data Collection [13], Figure 1 COVID-19 Chest X-ray Dataset Initiative [14], Actualmed COVID-19 Chest X-ray Dataset Initiative [14], COVID-19 Radiography Database [15] and the ChestX-ray8 dataset [16]. All the datasets are open-source allowing for reproducibility and further experimentation.

We make use of the training and test split proposed by Wang et al. The final split can be seen in table I. The training set was further split into training set and development set using the random split function and a fixed seed of 0 in PyTorch. The size for the development set is set to 2000 for the base case and data augmentation and 4000 for the GAN augmented dataset. During each epoch, the model is evaluated against the development set. Finally, the best performing model is tested on the test set and results are evaluated.

To significantly reduce training time and due to the scarcity

| | Normal | Pneumonia | Covid-19 |
|---|---|---|---|
| Train set | 6468 | 4449 | 6042 |
| Development set | 1531 | 1073 | 1396 |
| Test set | 885 | 594 | 100 |

TABLE II
SPLIT OF THE GAN AUGMENTED DATASET. THIS DATASET HAS BEEN AUGMENTED WITH 7001 IMAGES OF COVID-19.

| D | Params |
|---|---|
| images_in | - |
| labels_in | - |
| 256x256/FromRGB | 256 |
| 256x256/Conv0 | 36928 |
| 256x256/Conv1_down | 73856 |
| 256x256/Skip | 8192 |
| 128x128/Conv0 | 147584 |
| 128x128/Conv1_down | 295168 |
| 128x128/Skip | 32768 |
| 64x64/Conv0 | 590080 |
| 64x64/Conv1_down | 1180160 |
| 64x64/Skip | 131072 |
| 32x32/Conv0 | 2359808 |
| 32x32/Conv1_down | 2359808 |
| 32x32/Skip | 262144 |
| 16x16/Conv0 | 2359808 |
| 16x16/Conv1_down | 2359808 |
| 16x16/Skip | 262144 |
| 8x8/Conv0 | 2359808 |
| 8x8/Conv1_down | 2359808 |
| 8x8/Skip | 262144 |
| 4x4/MinibatchStddev | - |
| 4x4/Conv | 2364416 |
| 4x4/Dense0 | 4194816 |
| Output | 513 |
| Total | 24001089 |

TABLE III
DISCRIMINATOR NETWORK

| G | Params |
|---|---|
| latents_in | - |
| labels_in | - |
| G_mapping/Normalize | - |
| G_mapping/Dense0 | 262656 |
| G_mapping/Dense1 | 262656 |
| G_mapping/Dense2 | 262656 |
| G_mapping/Dense3 | 262656 |
| G_mapping/Dense4 | 262656 |
| G_mapping/Dense5 | 262656 |
| G_mapping/Dense6 | 262656 |
| G_mapping/Dense7 | 262656 |
| G_mapping/Broadcast | - |
| dlatent_avg | - |
| Truncation/Lerp | - |
| G_synthesis/4x4/Const | 8192 |
| G_synthesis/4x4/Conv | 2622465 |
| G_synthesis/4x4/ToRGB | 264195 |
| G_synthesis/8x8/Conv0_up | 2622465 |
| G_synthesis/8x8/Conv1 | 2622465 |
| G_synthesis/8x8/Upsample | - |
| G_synthesis/8x8/ToRGB | 264195 |
| G_synthesis/16x16/Conv0_up | 2622465 |
| G_synthesis/16x16/Conv1 | 2622465 |
| G_synthesis/16x16/Upsample | - |
| G_synthesis/16x16/ToRGB | 264195 |
| G_synthesis/32x32/Conv0_up | 2622465 |
| G_synthesis/32x32/Conv1 | 2622465 |
| G_synthesis/32x32/Upsample | - |
| G_synthesis/32x32/ToRGB | 264195 |
| G_synthesis/64x64/Conv0_up | 1442561 |
| G_synthesis/64x64/Conv1 | 721409 |
| G_synthesis/64x64/Upsample | - |
| G_synthesis/64x64/ToRGB | 132099 |
| G_synthesis/128x128/Conv0_up | 426369 |
| G_synthesis/128x128/Conv1 | 213249 |
| G_synthesis/128x128/Upsample | - |
| G_synthesis/128x128/ToRGB | 66051 |
| G_synthesis/256x256/Conv0_up | 139457 |
| G_synthesis/256x256/Conv1 | 69761 |
| G_synthesis/256x256/Upsample | - |
| G_synthesis/256x256/ToRGB | 33027 |
| Total | 24767458 |

TABLE IV
GENERATOR NETWORK

of data, we make use of transfer learning for both GAN and the proposed classification models. Tan et al. [17] suggest that transfer learning can alleviate the high requirement for training data by deep learning models in highly specialized domains. We therefore apply transfer learning for both GAN image generation and CNN image classification.

### *B. StyleGan2*

In order to generate high quality images and due to the scarcity of the training set, we make use of a state-of-the-art Generative Adversarial Network made available by Karras et al. [6] and apply transfer learning. The discriminator network has a total of 24.001.089 parameters, whereas the generator has 24.767.458 parameters. The network architecture for the discriminator can be found in table III and for the generator in table IV.

The generator of StyleGan2 is trained on the 517 COVID-19 images from the training set for 2000kimg by using a pretrained network called ”lsundog256”, a network trained on 256x256 images of dog pictures, while keeping the discriminator frozen. Then, both generator and discriminator are trained for 2000 more kimg, for a total of 3650 kimg. We apply early stopping and use the network with the best Frechet Inception Distance (FID) score to generate 7001 images of Covid-19. The images are added to the training set and split into training and development set using the same procedure as the original dataset. The resulting split can be found in table II.

### *C. Frechet Inception Distance*

Due to the lack of ground truth when training GANs, the classic metrics used in supervised learning are not suitable for evaluating GAN generated images. Borji [18] lists pros and cons of metrics used for GANs, one of which is the Frechet Inception Distance, introduced by Heusel et al.[19].

This metric captures how similar the generated images are to the real images by comparing the means and covariance matrices of the two datasets while penalizing the GAN for noise and homogeneity. The FID score is reportedly robust to noise and computationally efficent, making it a good choice for our research. It was measured after every 24kimg while training StyleGan2.

$$FID(r,g) = ||\mu_r - \mu_g||^2 + Tr(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r\Sigma_g}) \qquad (1)$$

The FID is computed as a function of r and g, the real and generated images, where $\mu$ and $\Sigma$ represent respectively the means and covariances of the distributions of the real and generated data. In StyleGan2, the FID score is measured by generating 50k images by the generator and comparing them to the real dataset.

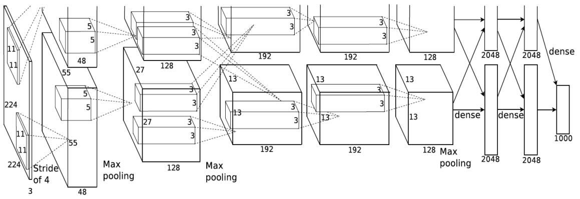

Fig. 2. Alexnet architecture from the original paper

```
Resize((256,256))
RandomHorizontalFlip(p=0.5)
RandomAffine((0,0), translate=(0.1,0.1))
RandomRotation((-5,5))
RandomCrop(224),
ToTensor(),
Normalize([0.485, 0.456, 0.406], [0.229, 0.224, 0.225]),
```

TABLE V
DATA AUGMENTATIONS APPLIED IN PYTORCH ON THE TRAINING SET

### D. Alexnet

To assess the performance of the GAN augmentation, a pre-trained Alexnet [20] model was fine-tuned with the given dataset under three different conditions. Firstly, a model was trained on the preprocessed data, serving as a base case. Secondly, a model trained on simple data augmentation and lastly, a model trained on a dataset augmented with synthetic images generated by a GAN trained on the training set of 517 images. Figure 2 illustrates the Alexnet architecture from the original paper.

Each Alexnet model is trained for 20 epochs, 64 size mini-batches, Soft-max loss function and the Stochastic Gradient Descent optimizer with 0.9 momentum and a learning rate of 0.0001 overall and 0.0002 for the last layer for fine-tuning the network. The SGD optimizer in PyTorch still makes use of the given mini-batch sizes and we are thus performing mini-batch GD. Lastly, we take the best performing model on the development set for each condition and evaluate it on the test set.

### E. Preprocessing

As a preprocessing step, each image is resized to 360x360 and the top 50 rows of pixels are removed. This is done in line with the work of Wang et al., where they remove the top 8% of each image to remove common textual information in the CXR images. Following the work of Maguolo et al. [21], more of the images was removed due to 8% not being sufficient in removing dataset related information.

For the data augmentation case, we make use of simple data augmentation methods due to the nature of the problem. Because CXR images are taken only from specific angles, the applied augmentations on the CXR images must be limited. We follow the work of Wang et al. and apply the augmentations in table V for the training set. The development and test sets are only resized to 256x256 and centercropped to 224x224 in addition to being normalized for Alexnet.

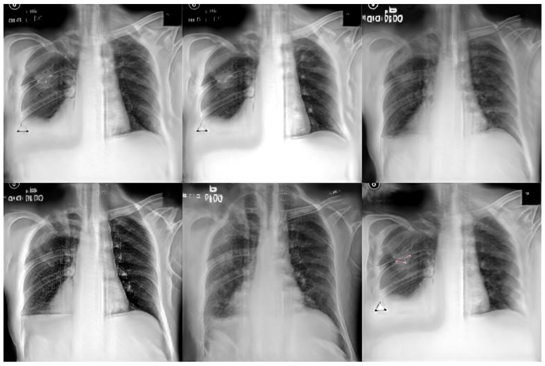

Fig. 3. Sample images generated by StyleGan2 after training for 5416kimg.

StyleGan2's 256x256 generated images are topcropped to 224x256 after generation, similarly to the rest of the images and added to the Covid-19 training data. Lastly, each image in this condition is resized to 224x224 and normalized.

For the base case and data augmentation, oversampling and undersampling is applied to the training set with PyTorch's WeightedRandomSampler due to the scarcity of the COVID-19 label. Each sample corresponding to the COVID-19 label is assigned a higher chance to be sampled during the training process according to how many samples there are w.r.t. the majority label. For the Covid-19 label, we set this to $\frac{7966}{517}$ and for pneumonia, we set this to $\frac{7966}{5475}$.

## IV. RESULTS

### A. Training Time of StyleGan2

Although we made use of transfer learning for the initial 1983 kimg, total training time lasted for several days. We made use of Google Colab's Pro feature, which provided us with P100-GPU's and TPU's, as well as up to 24 hours of runtime. However, the GPU upgrade is not guaranteed and Google still imposes limitations on the use of their GPU's even for Pro members. For the training of the first 1983 kimg using transfer learning, we estimate a training time of about 20 hours. Between 1983 and 3360 kimg, we estimate a training time of 43 hours, excluding downtime due to Google Colab limitations. In their documentation, StyleGan2 reports training times of several days.

### B. Quality of Synthesized X-Rays

The Generative Adversarial Network with the best FID score was used to generate synthetic X-rays images, which were later used in the classification task. The lowest FID score was at 3360 kimg and equalled to 73.471, where we early stopped but let the GAN model keep training. However, along with further training the score started to reach very high values, images were becoming more homogenous and were inheriting artefacts from the original dataset, as can be seen in figure 3. The trajectory of the FID score is shown in the figure 4. Synthetic X-rays generated by the chosen network can be seen in the figure 5. When evaluating the synthesized X-rays, it should be mentioned that around 20% of them didn't

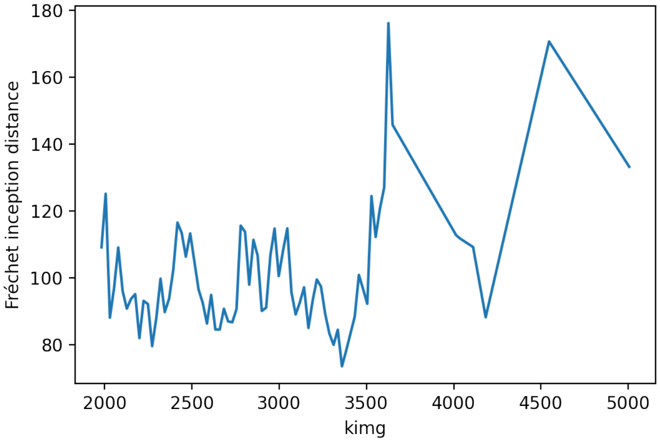

Fig. 4. Progress of FID score between 1983 kimg and 3650 kimg.

look like chest's X-rays. This is an approximate number after viewing the received dataset, a more accurate approximation would require consultation with a medical expert.

### C. Classification results

In the table VI we report a confusion matrix for base case on test data, where no data augmentation methods were used and classes remain highly imbalanced during the training phase but oversampling was applied. AlexNet had the greatest difficulty in distinguishing pneumonia and normal X-rays. Specificity and sensitivity are two measures, which are widely report in medical domain. They are both measures for binary classification task, while applying them to multiclassification *one against all* approach might be used. Sensitivity (True Positive Rate) measures the proportion of positives that are correctly classified relative to all positives cases. In our study it gives the proportion of cases with Covid-19 which are correctly classified as ill patients. This measure is especially valuable during diseases classification, it is important to not overlook any case. Specificity (True Negative Rate) measures the proportion of negatives that are correctly classified. Here, the proportion of cases without Covid-19 which are correctly classified as patients without illness. Sensitivity for Covid-19 class in a base case is 81%, while specificity equals to 98.12%.

In the table VII a confusion matrix for test set with applied standard data augmentation methods is shown. AlexNet had troubles to distinguish Normal and Pneumonia images. Sensitivity for Covid-19 class is 75%, while specificity equals to 99.12%.

In the table VIII confusion matrix for GAN augmented test dataset is presented. AlexNet struggled to differentiate Normal from Pneumonia images, also Covid-19 X-rays had a tendency to be classified as Pneumonia. Sensitivity for Covid-19 class is 69%, while specificity equals to 99.53%.

In the table IX Receiver Operator Characteristic Area Under the Curve (ROC AUC) score for all three different conditions can be seen. A receiver operating characteristic curve is a plot that shows the performance ability of a binary classifier. It plots true positive rate against false positive rate at various threshold values. An area under the curve (AUC) is a measure that indicates how well the classifier distinguish between classes, and it is use as a ROC summary, which allows to compare different curves between each other. ROC AUC is designed for binary classification problems. As in the case of sensitivity and specificity, when using this measure for multiclassification, the *one against all* approach can be used. It can be seen that normal class was the easiest to distinguished for all three cases, while Covid-19 was the hardest. Macro ROC AUC is an average results of ROC AUC score for three classes, which obtained the highest result for the base case. Overall accuracy has the best result for the GAN augmented case, but the difference between results is small and might be not significant.

TABLE VI
BASE NO AUGMENT

| | | Predicted | | |
|---|---|---|---|---|
| | | Covid | Normal | Pneumonia |
| | Covid | 81 | 5 | 14 |
| True label | Normal | 7 | 808 | 70 |
| | Pneumonia | 9 | 37 | 548 |

TABLE VII
STANDARD AUGMENT

| | | Predicted | | |
|---|---|---|---|---|
| | | Covid | Normal | Pneumonia |
| | Covid | 75 | 6 | 19 |
| True label | Normal | 5 | 821 | 59 |
| | Pneumonia | 8 | 43 | 543 |

## V. CONCLUSION

### A. Summary of findings

While the accuracy of the model with basic augmentations is marginally higher, and the accuracy of the GAN-augmented model is higher yet, the ROC_AUC score for Covid-19 classification did not improve with the more elaborate methods of data augmentation. The improvement in the accuracy can likely be explained by the imbalance of classes in the test data. If the classes were balanced we would likely see a decrease. This leads us to believe that our method of generating synthetic images when using StyleGan2 with FID=73.471 does not produce representative images of Covid-19.

TABLE VIII
GAN AUGMENT

| | | Predicted | | |
|---|---|---|---|---|
| | | Covid | Normal | Pneumonia |
| | Covid | 69 | 8 | 23 |
| True label | Normal | 2 | 841 | 42 |
| | Pneumonia | 5 | 57 | 532 |

TABLE IX
1-VS-REST ROC_AUC SCORE

| | Covid-19 | Normal | Pneumonia | Macro ROC_AUC | Acc |
|---|---|---|---|---|---|
| Base | 89.95 | 92.62 | 91.86 | 91.48 | 91.01 |
| Augment | 87.06 | 92.85 | 91.75 | 90.55 | 91.13 |
| GAN | 84.26 | 92.83 | 91.48 | 89.53 | 91.32 |

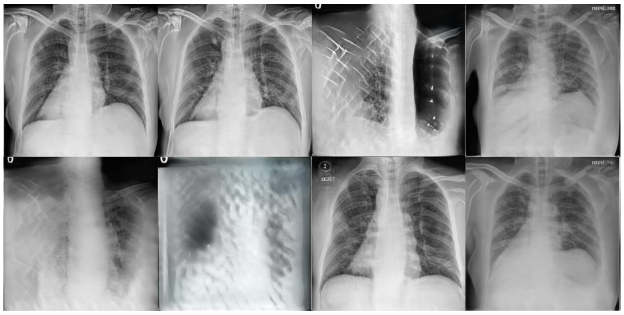

Fig. 5. Example results of StyleGan2 generated images with FID score of 73.471

When it comes to the difference in performance when considering the model trained with the use of simpler augmentations the cause of the difference is harder to identify. One explanation may be that the non-augmented model was overfitting on features that are obfuscated by the augmentations. Features that are unlikely to be caused by Covid-19 and thus are unwanted sources of potential bias, such as textual information present in the image. An alternative explanation is that by oversampling the Covid label, as well as applying random augmentations, the training set distribution becomes less similar to the test set. Moreover, the number of cases of Covid-19 in the test set is small, leading to high variance in the performance estimation of the trained model. When looking at the results for the other two labels, they are similar to the base case.

Further, when we compare our results with other studies that attempt to tackle the problem of Covid-19 classification, notably the work of Waheed et al. [9], it seems that using a form of conditional GANs such as ACGANs has an advantage of seeing more examples and implicitly learning the differences between the classes. However, the documentation of the GAN training process and the results leaves more precision, such as reporting an FID score, to be desired.

### B. Improvements

One possible improvement includes the development of a new GAN architecture which takes as input high resolution images. More specifically, conditional GANs as the one used by Khalifa et al. would be able to discriminate between each label and thus generate more representative images. However, transfer learning from domains with large datasets to specialized domains lacking in data can be invaluable. Thus, another simple improvement would be to explore multiple training settings of StyleGan2 and try out different hyperparameters during training to try to improve the FID score. Unfortunately, due to time limitations, we were only able to train the network using the default hyperparameters, which may not have been optimal for our dataset.

An alternative improvement to our work could be to have an expert in this field, filter the output of the GAN and discard the images that do not resemble lungs. This suggestion comes from qualitative inspection of the GAN-augmented images and the observation that roughly a fifth of them do not look like lungs to our untrained eyes. Due to time limitations and lack of medical expertise, we were unable to explore this option ourselves.

It has been pointed out [21, 22] that due to the lack of data and the composition of the available Covid-19 x-ray datasets models trained to predict Covid-19 may not necessarily be looking at the correct features to be helpful in real-world settings. Some of the models may be looking at various annotations, of medical devices present on the scan. Other unwanted biases may include focus on the characteristics of the imaging device used as Covid-19 samples tend to come from few sources. In future work, if a clear improvement in performance can be established using GANs, tests such as the ones proposed by Maguolo et al. should be conducted as a follow up test to determine the source bias in the results. When analysing the images generated after training for 5416 kimg, we noticed that the images were inhereting artefacts from the original dataset, as well as becoming more homogenous, as explained in the results section. While an improvement in performance is welcome, it is thus also likely that the GAN generated images inherit dataset source information and that the model may use this to classify the images.

We hope that our work will help guide future research. The use of publicly available models and dataset should make our findings easily replicable. In addition to the paper we make the fine-tuned model weights of StyleGan2-ada as well as the generated images accessible. [1]

## REFERENCES

[1] World Health Organisation, https://www.who.int/emergencies/diseases/novel-coronavirus-2019,.

[2] M. A. Zulkifley, S. R. Abdani, and N. H. Zulkifley, "Covid-19 screening using a lightweight convolutional neural network with generative adversarial network data augmentation," *Symmetry*, vol. 12, no. 9, p. 1530, 2020.

[3] H. Kim, "Outbreak of novel coronavirus (covid-19): What is the role of radiologists?" *European Radiology*, vol. 30, no. 6, p. 3266–3267, 2020.

[4] L. Wang, Z. Q. Lin, and A. Wong, "Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, p. 19549, Nov 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-76550-z

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville,

[1] https://github.com/lucalin17081994/PR_project

and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014, pp. 2672–2680. [Online]. Available: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[6] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. CVPR*, 2020.

[7] M. Shams, O. Elzeki, M. Abd Elfattah, T. Medhat, and A. E. Hassanien, "Why are generative adversarial networks vital for deep neural networks? a case study on covid-19 chest x-ray images," in *Big Data Analytics and Artificial Intelligence Against COVID-19: Innovation Vision and Approach*. Springer, 2020, pp. 147–162.

[8] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical image analysis*, vol. 58, p. 101552, 2019.

[9] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro, "Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection," *IEEE Access*, vol. 8, pp. 91 916–91 923, 2020.

[10] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," 2017.

[11] M. Loey, F. Smarandache, and N. E. M Khalifa, "Within the lack of chest covid-19 x-ray dataset: A novel detection model based on gan and deep transfer learning," *Symmetry*, vol. 12, no. 4, p. 651, 2020.

[12] N. E. M. Khalifa, M. H. N. Taha, A. E. Hassanien, and S. Elghamrawy, "Detection of coronavirus (covid-19) associated pneumonia based on generative adversarial networks and a fine-tuned deep transfer learning model using chest x-ray dataset," *arXiv preprint arXiv:2004.01184*, 2020.

[13] J. P. Cohen, P. Morrison, and L. Dao, "Covid-19 image data collection," *arXiv 2003.11597*, 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset

[14] L. Wang, Z. Q. Lin, and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, pp. 1–12, 2020.

[15] A. K. R. M. M. K. Z. M. K. I. M. K. A. I. N. A.-E. M. R. M. T. I. M.E.H. Chowdhury, T. Rahman, "Can ai help in screening viral and covid-19 pneumonia?" *IEEE Access*, vol. 8, pp. pp. 132 665 – 132 676, 2020.

[16] L. L. L. Z. B. M. S. R. Wang X, Peng Y, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." *IEEE CVPR*, pp. pp. 132 665 – 132 676, 2017.

[17] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," *CoRR*, vol. abs/1808.01974, 2018. [Online]. Available: http://arxiv.org/abs/1808.01974

[18] A. Borji, "Pros and cons of GAN evaluation measures," *CoRR*, vol. abs/1802.03446, 2018. [Online]. Available: http://arxiv.org/abs/1802.03446

[19] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a nash equilibrium," *CoRR*, vol. abs/1706.08500, 2017. [Online]. Available: http://arxiv.org/abs/1706.08500

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[21] G. Maguolo and L. Nanni, "A critic evaluation of methods for covid-19 automatic detection from x-ray images," 2020.

[22] J. P. Cohen, M. Hashir, R. Brooks, and H. Bertrand, "On the limits of cross-domain generalization in automated x-ray prediction," *arXiv preprint arXiv:2002.02497*, 2020.