# Improving COVID-19 CXR Detection with Synthetic Data Augmentation

Daniel Schaudt[1], Christopher Kloth[2], Christian Späte[1], Andreas Hinteregger[2], Meinrad Beer[2], and Reinhold von Schwerin[1]

[1] Technische Hochschule Ulm - Ulm University of Applied Sciences
`daniel.schaudt@thu.de`, `spaete@mail.hs-ulm.de`, `reinhold.vonschwerin@thu.de`
[2] Universitätsklinikum Ulm - Ulm University Medical Center
`christopher.kloth@uniklinik-ulm.de`, `andreas.hinteregger@uni-ulm.de`,
`meinrad.beer@uniklinik-ulm.de`

**Abstract.** Since the beginning of the COVID-19 pandemic, researchers have developed deep learning models to classify COVID-19 induced pneumonia. As with many medical imaging tasks, the quality and quantity of the available data is often limited. In this work we train a deep learning model on publicly available COVID-19 image data and evaluate the model on local hospital chest X-ray data. The data has been reviewed and labeled by two radiologists to ensure a high-quality estimation of the generalization capabilities of the model. Furthermore, we are using a Generative Adversarial Network to generate synthetic X-ray images based on this data. Our results show that using those synthetic images for data augmentation can improve the model's performance significantly. This can be a promising approach for many sparse data domains.

**Keywords:** Deep Learning, Medical Imaging, GANs, Data Augmentation

## 1 Introduction

The ongoing COVID-19 pandemic brings many challenges for societies all around the globe. For the healthcare sector, it is important to screen infected patients in an effective and reliable manner. This is especially true in an emergency setting, where patients already experience advanced symptoms. The prevalent test used for COVID-19 detection is the reverse transcription polymerase chain reaction (RT-PCR) [1–3]. This method has a high false negative rate and the processing requires dedicated personnel and can take hours to days [4].

Since chest X-ray (CXR) images of COVID-19 patients show typical findings including peripheral opacities and ground class patterns in the absence of pleural effusion [4,5], they can be used as a first-line triage tool [6]. This could speed up the identification process, as CXR images are easy to obtain and rather inexpensive with a lower radiation dose than computed tomography (CT) images. Using deep learning models for detection of COVID-19 prevalence in CXR images is promising, because it eliminates the need for specialized medical staff in an emergency setting. This can further help to alleviate the challenges to the healthcare systems around the world and has the potential to save lives.

In this retrospective study, we are training a deep convolutional neural network (CNN) on the openly available COVIDx V8b dataset [7] and evaluate the model on local hospital CXR data. We specifically choose this learning framework to assess the generalization abilities of a CNN in the medical imaging context. Since high quality CXR image data is sparse, we see this as the most common use case for models in production.

Furthermore, we are using a modified version of the StyleGAN architecture [8] to generate synthetic COVID-19 positive and COVID-19 negative CXR images for data augmentation. This is done to offset some negative side effects encountered by a distributional shift between the training and testing data.

## 2 Related Work

There has been a lot of previous work on applying deep learning to CXR images to detect a COVID-19 pulmonary disease [7,9–12]. However, most of the existing work is using publicly available CXR and COVID-19 image data. Most of those images are collected from heterogeneous sources with varying image and label quality, which raises concerns about the quality and valid evaluation of deep learning models [13, 14].

Generative Adversarial Networks (GANs) [15] have been used for many applications in the medical imaging domain [16–18]. Some studies show promising results specifically for the CXR and COVID-19 domain [19, 20]. In contrast to existing work, we integrate differentiable augmentation [21] into our GAN architecture. This enables us to train on a very small dataset and still get meaningful results.

## 3 Materials and Methods

Our goal for this work is to correctly detect a COVID-19 pulmonary disease in chest X-ray images on local university hospital study data. Therefore, we train a deep learning model on publicly available COVID-19 image data and evaluate the model based on our study data. We further enhance the amount of available training data by generating synthetic X-ray images. In this section we explain the origin and distribution of the data, as well as the deep learning model and training process.

### 3.1 Data

In this work we analyze chest X-ray images in posteroanterior (PA) and anteroposterior (AP) front view. Typically the AP view is encountered for cases where the patient is bedridden. Figure 1 shows two male patient example CXR images from our study data.

**Training data** We use two different training datasets, see Table 1. As a first step, we use the COVIDx V8b dataset [7] to train our model. This dataset is one of the biggest curated and publicly available COVID-19 CXR datasets. We use the training split of the dataset, which contains 13.794 COVID-19 *negative* and 2.158 COVID-19 *positive* frontal view X-ray images of 14.978 unique patients.

In a second step we enhance this training data by using 20.000 additional synthetic CXR images that we generated based on our study data. With that, we can add 10.000 COVID-19 *positive* and 10.000 COVID-19 *negative* images to our existing COVIDx V8b training data. This synthetic data is used to further augment the training of the classification model and increase image diversity. A sample of the generated images has been reviewed by a radiologist to ensure that the model produces meaningful data.

**Validation data** We validate the model by calculating loss metrics on the so called *test* split of the COVIDx V8B dataset. This dataset contains 200 COVID-19 positive and 200 COVID-19 negative images. We used this dataset to tune model parameters. This is to avoid overfitting our model to the testing data.
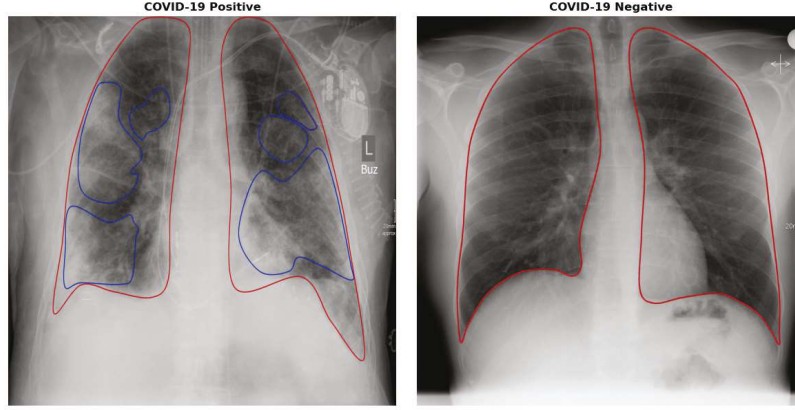
**Fig. 1.** Chest X-ray images with lungs marked in red. (Left) COVID-19 positive image, typical Ground-glass opacification marked in blue. (Right) COVID-19 negative image.

**Testing data** The central data in this work comes from a single center retrospective study of the Universitätsklinikum Ulm. For this study 566 patients (average age 51.12y +/- 18.73y; range 23-82y, 315 women) of a single institution (11/2019-05/2020) were included. The data has been carefully reviewed and labeled by two radiologists after dedicated training into *COVID-19 positive* and *COVID-19 negative*. The senior radiologist (CK) has 8 years of experience in thoracic imaging. This resulted in 110 positive images and 223 negative images, as seen in Table 1.

This testing data is used as a holdout set for final model evaluation. With this method we make sure to avoid any patient overlap between the training and testing data. Furthermore, we get a high-quality estimation of the generalization capabilities of the model when evaluating on the testing data. This is because the testing images come from a different data source, which leads to a *distributional shift*.

**Table 1.** Distribution of images for all datasets

| Dataset | Split | COVID-19 positive | COVID-19 negative |
|---|---|---|---|
| COVIDx V8B | Training | 2.158 | 13.794 |
| COVIDx V8B + Synthetic | Training | 12.158 | 23.794 |
| COVIDx V8B | Validation | 200 | 200 |
| Uniklinik Ulm Study | Test | 110 | 223 |

### 3.2 Network Architecture

For classification we use the ResNet50 architecture [22]. The network has been pretrained on the ImageNet [23] database. We replace the final fully connected layer with a linear

layer of two outputs, one for each class. To get the predictions we apply a softmax activation function. Since training was very stable, we did not use any additional dropout layers or regularization methods.

The generative model is based on a modified version of StyleGAN [8]. We specifically integrate differentiable augmentation [21] into our StyleGAN architecture. This is to prevent memorization of the training data and helps to stabilize the training process. This combined architecture enables us to generate meaningful synthetic images based on our small study dataset [24]. Figure 2 shows one example image along a classification by the COVIDx+Synth model.
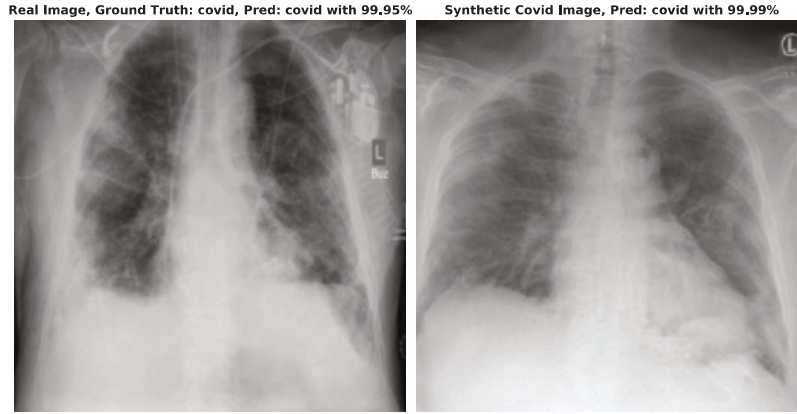


**Fig. 2.** Example images for classification by the COVIDx+Synth model. (Left) Real COVID-19 positive image, correctly predicted as *positive*. (Right) Synthetic COVID-19 positive image, correctly predicted as *positive*.

### 3.3 Training details

To train the ResNet classifier we use the Adam solver [25] with default parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$) and a cross-entropy loss. We train the model using minibatches of size 16. We use an initial learning rate of 0.001 and apply the One-cycle learning rate scheduler [26] with a maximum learning rate of 0.006. We initially freeze all but the new last network layer for 5 epochs of training. After those 5 epochs all network parameters are trained for 30 additional epochs. The One-cycle learning rate scheduler is only applied after the initial freeze period. Increasing the amount of training showed no further improvement empirically.

All images are being scaled down to $224 \times 224$ and normalized with the mean and standard deviation of images in the COVIDx V8B dataset [7] before feeding them into the network. During training, we augment the images with random horizontal flipping and random rotation ($\pm 5°$).

Since we use two different training datasets (see Section 3.1) we get two different models: *COVIDx* and *COVIDx+Synth*. Both classification models use the exact same

hyperparameters and training procedures as described. This is to evaluate the effect of using the synthetic data and make the results comparable.

For the StyleGAN generator, we train two different models: one for each class of COVID-19 positive and negative images. This is a simple method to ensure that we can generate a specific image class. For further details regarding the training process of the StyleGAN generator see Späte 2021 [24].

## 4   Results

To investigate our models in a quantitative manner, we computed the accuracy, as well as F1-score, precision and recall for each class on the validation and testing data. The metrics for the validation data are shown in Table 2. Both models perform quite well on the validation data with an accuracy of 96 % and 95.5 % respectively. The results are in line with Wang et al. 2020 [7] and their COVIDNet-CXR-2 model. Interestingly, the Covidx+Synth model falls behind the other models, despite having a lot more training data. This could be another indication of a distributional shift between the COVIDx dataset and the study data of the Universitätsklinikum Ulm.

The results for the testing data are also shown in Table 2. The table shows that a model trained on the COVIDx dataset can adapt quite well to the testing data, with an accuracy of 89.49 %. The model achieves a decent precision for COVID-19 cases (90.32 %), which is good since too many false positives would increase the burden for the healthcare system due to the need for additional PCR testing. With a rather low recall of 76.36 % the model does miss quite a lot of COVID-19 cases. This can be especially problematic in this sensitive medical setting, since false negatives lead to undetected cases of COVID-19.

This drawback can be controlled by using additional synthetic data to train the model. Table 2 shows an increase in accuracy (92.49 %) and most evaluation metrics. Especially the improved recall of 95.45 % is very desirable. This comes with the cost of a slight reduction in precision (-6.32 %). Based on those results, it can be seen that our models perform quite well, especially when incorporating the synthetic data, but there are still several areas for improvement.

**Table 2.** Evaluation metrics for models COVIDx and COVIDx+Synth on validation data (with reported metrics from Wang et al. [7] for comparison) and on testing data.

|  | Accuracy | F1-Score | | Precision | | Recall | |
|---|---|---|---|---|---|---|---|
| Model | | C19 pos. | C19 neg. | C19 pos. | C19 neg. | C19 pos. | C19 neg. |
| *Validation Data* | | | | | | | |
| COVIDx | **0.9600** | **0.9583** | **0.9615** | **1.0000** | 0.9259 | 0.9200 | **1.0000** |
| COVIDx+Synth | 0.9550 | 0.9548 | 0.9552 | 0.9596 | 0.9505 | 0.9500 | 0.9600 |
| COVIDNet-CXR-2 [7] | - | - | - | 0.9700 | **0.9560** | **0.9550** | 0.9700 |
| *Testing Data* | | | | | | | |
| COVIDx | 0.8949 | 0.8276 | 0.9244 | **0.9032** | 0.8917 | 0.7636 | **0.9596** |
| COVIDx+Synth | **0.9249** | **0.8936** | **0.9420** | 0.8400 | **0.9760** | **0.9545** | 0.9103 |

## 5 Limitations and Discussion

In this work we showed that a deep learning model trained with a comparatively large volume of publicly available data for COVID-19 detection is able to generalize well to single source, local hospital data with patient demographics and technical parameters independent of the training data. This is not without limitations, since the distributional shift between the training and testing data can lead to some undesirable results, especially for important metrics like low recall values.

We show that this can be improved by using synthetically generated data to augment the training data. Although this works quite well, one of the reasons could be a rebalancing effect, that could have been achieved with various resampling methods as well. Another reason could be a light form of *data leakage*, since the synthetic data was generated based on the testing data. This is not fully clear, since the StyleGAN generator has no direct access to the ground truth data and just learns based on the feedback of a discriminator. Despite these concerns, the model shows promising first results and using such a model in an emergency setting could give a fast estimation for the prevalence of pulmonary infiltrates and therefore improve clinical decision-making and resource allocation.

## References

1. Vogels, C.B., Brito, A.F., Wyllie, A.L., Fauver, J.R., Ott, I.M., Kalinich, C.C., Petrone, M.E., Casanovas-Massana, A., Muenker, M.C., Moore, A.J., Klein, J., Lu, P., Lu-Culligan, A., Jiang, X., Kim, D.J., Kudo, E., Mao, T., Moriyama, M., Oh, J.E., Park, A., Silva, J., Song, E., Takahashi, T., Taura, M., Tokuyama, M., Venkataraman, A., Weizman, O.E., Wong, P., Yang, Y., Cheemarla, N.R., White, E.B., Lapidus, S., Earnest, R., Geng, B., Vijayakumar, P., Odio, C., Fournier, J., Bermejo, S., Farhadian, S., Cruz, C.S.D., Iwasaki, A., Ko, A.I., Landry, M.L., Foxman, E.F., Grubaugh, N.D.: Analytical sensitivity and efficiency comparisons of SARS-COV-2 qRT-PCR primer-probe sets. (apr 2020)
2. Udugama, B., Kadhiresan, P., Kozlowski, H.N., Malekjahani, A., Osborne, M., Li, V.Y.C., Chen, H., Mubareka, S., Gubbay, J.B., Chan, W.C.W.: Diagnosing COVID-19: The disease and tools for detection. ACS Nano **14**(4) (mar 2020) 3822–3835
3. Yang, Y., Yang, M., Shen, C., Wang, F., Yuan, J., Li, J., Zhang, M., Zhang, Z., Xing, L., Wei, J., Peng, L., Wong, G., Zheng, H., Wu, W., Liao, M., Feng, K., Li, J., Yang, Q., Zhao, J., Zhang, Z., Liu, L., Liu, Y.: Evaluating the accuracy of different respiratory specimens in the laboratory diagnosis and monitoring the viral shedding of 2019-nCoV infections. (feb 2020)
4. Arevalo-Rodriguez, I., Buitrago-Garcia, D., Simancas-Racines, D., Zambrano-Achig, P., Campo, R.D., Ciapponi, A., Sued, O., Martínez-García, L., Rutjes, A., Low, N., Bossuyt, P.M., Perez-Molina, J.A., Zamora, J.: FALSE-NEGATIVE RESULTS OF INITIAL RT-PCR ASSAYS FOR COVID-19: A SYSTEMATIC REVIEW. (apr 2020)
5. Kong, W., Agarwal, P.P.: Chest imaging appearance of COVID-19 infection. Radiology: Cardiothoracic Imaging **2**(1) (feb 2020) e200028
6. Rubin, G.D., Ryerson, C.J., Haramati, L.B., Sverzellati, N., Kanne, J.P., Raoof, S., Schluger, N.W., Volpi, A., Yim, J.J., Martin, I.B.K., Anderson, D.J., Kong, C., Altes, T., Bush, A., Desai, S.R., onathan Goldin, Goo, J.M., Humbert, M., Inoue, Y., Kauczor, H.U., Luo, F., Mazzone, P.J., Prokop, M., Remy-Jardin, M., Richeldi, L., Schaefer-Prokop, C.M., Tomiyama, N., Wells, A.U., Leung, A.N.: The role of chest imaging in patient management during the COVID-19 pandemic: A multinational consensus statement from the fleischner society. Radiology **296**(1) (July 2020) 172–180
7. Wang, L., Lin, Z.Q., Wong, A.: COVID-net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest x-ray images. Scientific Reports **10**(1) (November 2020)

8. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (June 2019)

9. Khan, A.I., Shah, J.L., Bhat, M.M.: CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. Computer Methods and Programs in Biomedicine **196** (November 2020) 105581

10. Ucar, F., Korkmaz, D.: COVIDiagnosis-net: Deep bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from x-ray images. Medical Hypotheses **140** (July 2020) 109761

11. Keidar, D., Yaron, D., Goldstein, E., Shachar, Y., Blass, A., Charbinsky, L., Aharony, I., Lifshitz, L., Lumelsky, D., Neeman, Z., Mizrachi, M., Hajouj, M., Eizenbach, N., Sela, E., Weiss, C.S., Levin, P., Benjaminov, O., Bachar, G.N., Tamir, S., Rapson, Y., Suhami, D., Atar, E., Dror, A.A., Bogot, N.R., Grubstein, A., Shabshin, N., Elyada, Y.M., Eldar, Y.C.: COVID-19 classification of x-ray images using deep neural networks. European Radiology (may 2021)

12. Shamout, F.E., Shen, Y., Wu, N., Kaku, A., Park, J., Makino, T., Jastrzębski, S., Witowski, J., Wang, D., Zhang, B., Dogra, S., Cao, M., Razavian, N., Kudlowitz, D., Azour, L., Moore, W., Lui, Y.W., Aphinyanaphongs, Y., Fernandez-Granda, C., Geras, K.J.: An artificial intelligence system for predicting the deterioration of COVID-19 patients in the emergency department. npj Digital Medicine 4(1) (may 2021)

13. Tartaglione, E., Barbano, C.A., Berzovini, C., Calandri, M., Grangetto, M.: Unveiling COVID-19 from CHEST x-ray with deep learning: A hurdles race with small data. International Journal of Environmental Research and Public Health **17**(18) (September 2020) 6933

14. Oakden-Rayner, L.: Exploring the chestxray14 dataset: problems (Dec 2017)

15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q., eds.: Advances in Neural Information Processing Systems. Volume 27., Curran Associates, Inc. (2014)

16. Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing **321** (dec 2018) 321–331

17. Yi, X., Walia, E., Babyn, P.: Generative adversarial network in medical imaging: A review. Medical Image Analysis **58** (dec 2019) 101552

18. Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., Mukhopadhyay, A.: GANs for medical image analysis. Artificial Intelligence in Medicine **109** (sep 2020) 101938

19. Karbhari, Y., Basu, A., Geem, Z.W., Han, G.T., Sarkar, R.: Generation of synthetic chest x-ray images and detection of COVID-19: A deep learning based approach. Diagnostics **11**(5) (may 2021) 895

20. Motamed, S., Rogalla, P., Khalvati, F.: RANDGAN: Randomized generative adversarial network for detection of COVID-19 in chest x-ray. Scientific Reports **11**(1) (apr 2021)

21. Zhao, S., Liu, Z., Lin, J., Zhu, J.Y., Han, S.: Differentiable augmentation for data-efficient gan training. In: Conference on Neural Information Processing Systems (NeurIPS). (2020)

22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (June 2016)

23. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (jun 2009)

24. Spaete, C.: Synthetic generation of medical images (unpublished master's thesis). Master's thesis, Technische Hochschule Ulm, Ulm (2021)

25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. (2014)

26. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. (2017)