Additional Look into GAN-based Augmentation for Deep Learning COVID-19 Image Classification

Oleksandr Fedoruk[1], Konrad Klimaszewski[1],
Aleksander Ogonowski[1] and Michał Kruk[2]
[1] *Department of Complex Systems, National Centre for Nuclear Research, Otwock-Świerk, Poland*
[2] *Institute of Information Technology, Warsaw University of Life Sciences – SGGW, Warsaw, Poland*

Abstract Data augmentation is a popular approach to overcome the insufficiency of training data for medical imaging. Classical augmentation is based on modification (rotations, shears, brightness changes, etc.) of the images from the original dataset. Another possible approach is the usage of Generative Adversarial Networks (GAN). This work is a continuation of the previous research where we trained StyleGAN2-ADA by Nvidia on the limited COVID-19 chest X-ray image dataset. In this paper, we study the dependence of the GAN-based augmentation performance on dataset size with a focus on small samples. Two datasets are considered, one with 1000 images per class (4000 images in total) and the second with 500 images per class (2000 images in total). We train StyleGAN2-ADA with both sets and then, after validating the quality of generated images, we use trained GANs as one of the augmentations approaches in multi-class classification problems. We compare the quality of the GAN-based augmentation approach to two different approaches (classical augmentation and no augmentation at all) by employing transfer learning-based classification of COVID-19 chest X-ray images. The results are quantified using different classification quality metrics and compared to the results from the previous article and literature. The GAN-based augmentation approach is found to be comparable with classical augmentation in the case of medium and large datasets but underperforms in the case of smaller datasets. The correlation between the size of the original dataset and the quality of classification is visible independently from the augmentation approach.

Keywords: computer vision, deep learning, image classification, generative adversarial networks, medical imaging.

1. Introduction

Computer vision techniques are used in different medical applications for various purposes. They accelerate decision-making while diagnosing patients and support medical personnel on a daily basis. As available algorithms and solutions advance, the problem of medical data accessibility is becoming a bottleneck for new researchers and breakthroughs [25]. Almost all modern algorithms are data-driven and require a lot of data samples to perform well [41]. However, the process of gathering medical data is not easy and often blocked by the high costs of procedures required to obtain the data, patients' personal data access limitations (as GDPR in the European Union or CCPA in the United States of America), rare diseases for which there is just not much data at all.

To overcome that problem, researchers use data augmentation techniques. The main idea is to train models on several modified copies of original data. This reduces overfitting and allows to achieve better training results with less data [36]. In the graphical

Machine GRAPHICS & VISION 32(3/4):107–124, 2023. DOI: 10.22630/MGV.2023.32.3.6 .

data domain, the classical augmentation pipeline includes transformations such as rotating, scaling, changing the brightness of the image, etc. With the rapid development of Generative Adversarial Networks (GAN) [12] it is possible to generate images similar to ones from the given dataset so it is possible to apply GANs as an augmentation pipeline.

In our original experiment, we showed, on a dataset that contains 2000 images per class, that the GAN-based augmentation approach is comparable to but not outperforming classical augmentation [10]. In this paper, we want to move forward and verify if the same is true for even smaller datasets.

The manuscript is organized as follows: in this Section the research problem is described, followed by a literature review in the *Related works* section (Sect. 2). The section on the *Methods* (Sect. 3) describes the *Motivation and methodology* (Sect. 3.1) and the *Dataset* (Sect. 3.2) including its preprocessing. The *GAN-based augmentation* as well as the *Classical augmentation* are both described in Sections 3.3 and 3.4, respectively, followed by Sect. 3.5 on *Image comparison metrics* used to assess their quality. This section ends with a description of the approach used for the *Classification evaluation* for the images, in Section 3.6. In Section 4 on the *Results* the dependence of the classifier output on the sample size is described. The paper ends with a *Conclusion* in Section 5.

Our main contributions include: 1) a study of the multi-class classification performance dependence on the dataset size, from small to moderate samples, of X-ray chest scans; 2) a study of the GAN-based augmentation performance for small datasets using a StyleGAN2-ADA [19] architecture, including Adaptive Discrimination Augmentation, designed to improve GAN training on limited datasets; 3) validation of the StyleGAN2-ADA multi-class training mode to obtain a single generative model for multiple classes; 4) comparison of the GAN-based augmentation to classical augmentation techniques.

This paper presents improved and updated materials originally presented at the 9th Conference on Symbiosis of Technology and IT (SIT) in Kiry, June 2023. These materials have not been previously published.

2. Related works

While chest radiography and computed tomography (CT) scans are not recommended as primary diagnostic tools, they are reported to be highly sensitive in detecting COVID-19 [2]. This led to numerous studies [13] on the applicability of X-ray chest scans in COVID-19 diagnosing. The first to our knowledge ML-based lung imaging classification methods for COVID-19 were works by Xu et al. [42] and by Gozes et al. [14]. Both focus on binary classification of COVID-19 and healthy images with many works that followed [9, 16, 28, 33, 34]. While the binary classification problem for COVID-19 X-ray images is already extensively studied, there were only several attempts to distinguish COVID-19 from other diseases affecting the respiratory system. In particular, the four-class problem is underrepresented in the literature [13]. The lung patterns seen in COVID-19 are unique but bear a resemblance to those found in pneumonia from other causes [32]. This is concerning, as challenges in distinguishing viral pneumonia from bacterial and fungal pathogen-induced cases have been reported [22]. For instance, A. I. Khan et al. [23] found that a limited dataset leads to problems in the classification of patients with COVID-19 when considering also patients with viral pneumonia and lung opacity. To our knowledge, the CoroNet [23] was the first proposed model that includes four classes with a recent result by E. Khan et al. [24] that achieved 96.13% accuracy utilising a modified EfficientNet-B1 model.

Data augmentation with GANs is not a new idea – multiple papers [36, 37, 43] are showing that this approach is worth deeper investigation, yet the majority of them consider problems with large datasets available [18]. For instance, Bowles et al. [5] study the performance of classical and GAN-based augmentation on data samples ranging from 8 K to 80 K images. Also, GAN-based augmentation for chest X-ray medical imaging (with or without COVID-19 included) has been studied in the literature mostly on moderate-sized datasets. For example, in the proposed IAGAN architecture [26], the models (IAGAN and DCGAN) were trained on two datasets: the "chest X-ray" [21] dataset with two categories – Normal (1575 images) and Pneumonia (4265 images), and the "covid-chestxray" [8] dataset with 3 classes – Normal (8066 images), Pneumonia (5999 images) and COVID-19 (589 images). That results in generative networks trained on more than 4000 images in both cases. In addition, the study focused on synthetic image generation for only two classes while disregarding the possibility of augmentation for the COVID-19 data sample. The very interesting RANDGAN model, proposed by Motamed et al. [27], was also trained on large samples with only two classes — 7493 Normal and 4986 Pneumonia images. In another study [35], a DCGAN architecture tailored for chest X-rays image generation was trained with a dataset of 2000 chest X-rays per 5 classes – Normal, Cardiomegaly, Pleural Effusions, Pulmonary Edema and Pneumothorax. A more recent study by Bali and Mahara used a similar methodology [3] but with DCGAN architecture and a bigger training dataset that contained only two classes – Normal (1314 images) and Pneumonia (3875 images). Finally, Albahli proposed [1] a different GAN architecture – a combination of variational auto-encoder and GAN which was trained on 16 classes with 5000 images per class.

There are only a few studies that consider GAN applications for very small datasets. A very interesting research was performed on a very small dataset of liver lesion images [11]. The dataset contained 3 classes – cysts (53 images), metastases (64 images) and hemangiomas (65 images). Two approaches were tried – DCGAN trained individually per class and Auxiliary Classifier GAN (ACGAN) [29] as a single network that is able to generate images of different classes. The dedicated DCGAN per class approach performed better and resulted in ≈7% improvement over the classic augmentation approach while multi-class ACGAN was not able to improve the classification over DCGANs. In the study, both generative networks were trained on classically augmented data samples and then used as an additional source of synthetic images to further increase of the training dataset. It is worthwhile to investigate whether models with built-in augmentation like StyleGAN2-ADA will behave similarly.

Regretfully, a number of studies on GAN-based augmentation available in the literature [27, 35, 40] omit a comparison with classical augmentation approach. With the training of a GAN being both time and computationally expensive, this limits the proper evaluation of such results. The expected improvement over the simpler method is key information for an informed selection of the most effective augmentation approach. It is also apparent that only several of studies on the topic focus on augmentation for classification problems with more than three classes. Based on our review of the available literature, we find that there is a need for a dedicated study of the multi-class GAN-based augmentation performance in comparison with classical methods, in particular for small and very small datasets. In addition to our knowledge GAN models with built-in augmentation were not evaluated for applicability for small medical data samples.

3. Methods

3.1. Motivation and methodology

In this paper, we test the hypothesis that data augmentation plays a more crucial role in the deep learning process with small datasets than it does with large ones. Also, we try to verify that, in the case of medical imaging, data augmentation based on GAN-generated images could result in bigger data diversity and thus improve deep learning results in comparison to the classical augmentation approach.

We compare 3 data-augmentation approaches with two datasets – the first dataset contains 1000 images per 4 classes and the second contains 500 images per 4 classes respectively. Further in the paper we call the dataset with 1000 images per class a *small* dataset. The dataset with 500 images per class is called *micro*, respectively. To create the small dataset 1000 images of each class were randomly picked from the original dataset after the preprocessing. The micro dataset is a subset of the small dataset and was also created by picking 500 random images of each class. The datasets used are described in detail in the following section.

Data augmentation approaches being compared were: no augmentation at all, classical augmentation, and GAN-based augmentation. To estimate which approach is better we trained a convolutional neural network on data augmented by each approach and evaluate based on different classification quality metrics. The scheme of operations performed is shown in Fig. 1.

3.2. Dataset

The dataset used in the research is the “COVID-19 Radiography Database” developed by a team of researchers from Qatar University, Doha, Qatar, and the University of Dhaka,

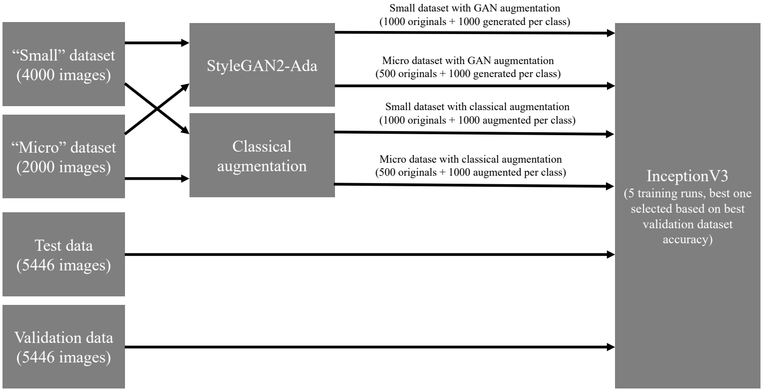
"Small" dataset (4000 images)
"Micro" dataset (2000 images)
Test data (5446 images)
Validation data (5446 images)
StyleGAN2-Ada
Classical augmentation
Small dataset with GAN augmentation (1000 originals + 1000 generated per class)
Micro dataset with GAN augmentation (500 originals + 1000 generated per class)
Small dataset with classical augmentation (1000 originals + 1000 augmented per class)
Micro datase with classical augmentation (500 originals + 1000 augmented per class)
InceptionV3 (5 training runs, best one selected based on best validation dataset accuracy)

Fig. 1. Visualization of experiment steps and data flow.

Bangladesh, along with their cooperators from Pakistan and Malaysia in collaboration with medical doctors. It is worth noticing that the dataset is being updated and the latest version of it contains way more images than it was when the original experiment was done. As this paper's goal is an additional investigation of the GAN augmentation technique, we continue to use the same version of the dataset that we used in the original experiment.

The dataset used in this paper contains 3616 images of COVID-19-positive cases, 6012 images of lung opacity (non-COVID lung infection), 1345 images of viral pneumonia, and 10192 images of healthy lungs. Each image is represented in PNG format withdimensions 256×256. For each image, the dataset authors provided a corresponding lung segmentation mask obtained using a dedicated U-Net model [30]. The examples of the images and segmentations masks are displayed in Fig. 2 and Fig. 3. The dataset was split into 3 subsets: train, validation, and test. The small train subset contains 1000 (500 in the case of the micro dataset) images of each class and is used as a source for classical augmentation, training of the GAN, and for no augmentation approach. The rest of the images were split in half to form validation and test subsets. Also, we use the same test and validation subsets for both experiments with small and micro datasets. The validation subset is used as validation data in target classification CNN training. The test subset was used only as the final trained CNN benchmark which simulates new data from new patients to show real world usage of the trained classification network.

We have preprocessed all images from the original dataset with the following 3-step procedure:

1. All images were cropped to the lung region according to the provided masks.
2. Cropped images were manually reviewed and all images containing any text or graphical annotations/marks were removed.

3. Remaining images were resized into dimensions 128 × 128 and converted to 1-channel (grayscale) to reduce the amount of data processed while training. The conversion was done by leaving only the first channel of the original images.

After all the described steps, the final version of the dataset used in the experiment contained 3242 images of COVID-19 cases, 2982 images of lung opacity (non-COVID lung infection), 1264 images of viral pneumonia, and 7404 images of healthy lungs. As mentioned earlier 2 experimental cases have been considered in the research. The first experiment was conducted with the small dataset (1000 images per class in the training subset) and the second one with the micro dataset (500 images per class in the training subset). As the original dataset contained 4 classes, there are 4000 images in total in the train subset of the small dataset and 2000 images in the train subset of the micro dataset. Test and validation subsets remained the same for both cases and were prepared during small dataset preparation (see Tab. 1). This made it possible to compare the final results for both experiments as the data those results were calculated on remained the same.

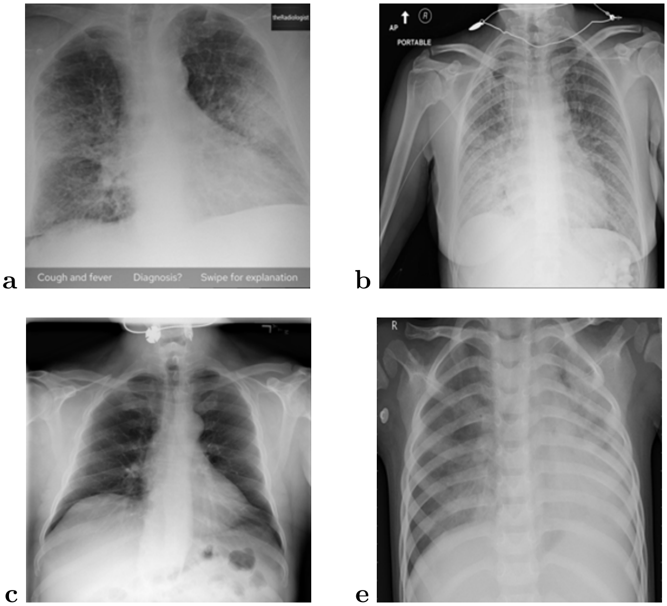

Fig. 2. Examples of unprocessed images from the original dataset. (**a**) COVID-19; (**b**) Normal; (**c**) Lung opacity; (**e**) Viral pneumonia.

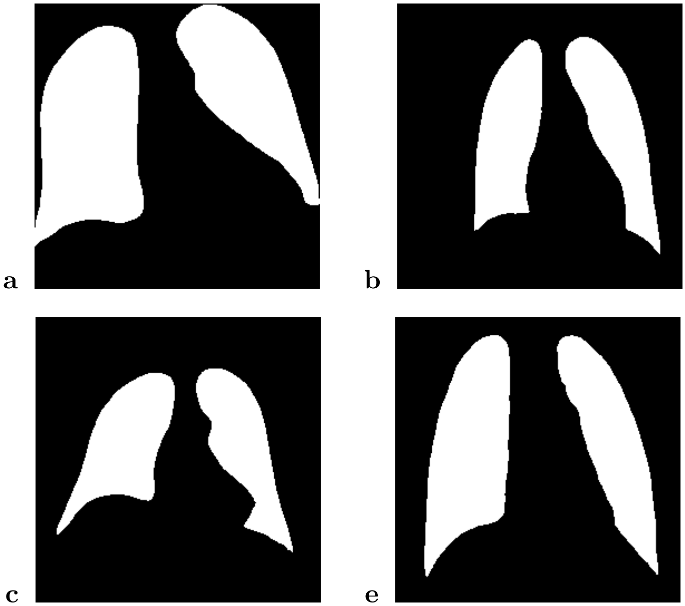

Fig. 3. Example binary masks from the original dataset for the images shown in Fig. 2.

3.3. GAN-based augmentation

In this research we continued to use StyleGAN2 with adaptive discriminator augmentation (ADA) mechanism by NVIDIA [19] as one of its features is the ability to be trained on relatively small datasets and support of class-conditional image generation. For each of the experiments, StyleGAN2-ADA was trained on the corresponding train subset. The training process was monitored, using the validation dataset, to prevent network overfitting. After the training, the epoch with the best Kernel Inception Distance (KID) score was picked as a source of future data generation. The target for r_t ADA heuristic is set to 0.6, both generator and discriminator learning rates were set to 0.0025 while

Tab. 1. Number of images per class in each subset for both small and micro datasets.

Subset	COVID-19	Healthy	Lung opacity	Viral pneumonia
Train subset 1000	1000	1000	1000	1000
Train subset 500	500	500	500	500
Validation subset	1121	3202	991	132
Test subset	1121	3202	991	132

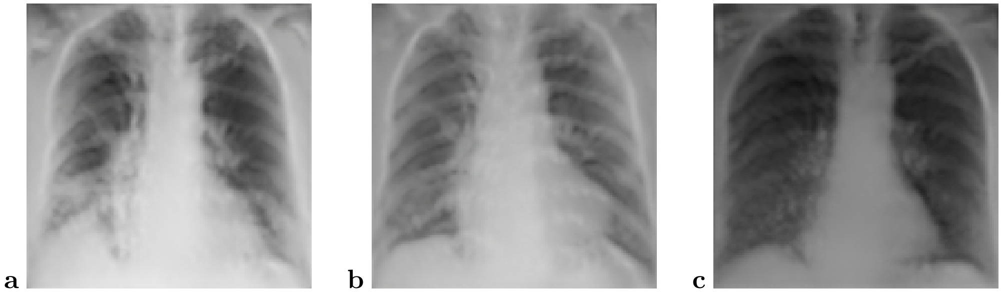

Fig. 4. Kernel Inception Distance values with correlated example image generated by GAN trained on the small dataset. (**a**) KID ≈ 19.46; (**c**) KID ≈ 13.29; (**b**) KID ≈ 12.26.

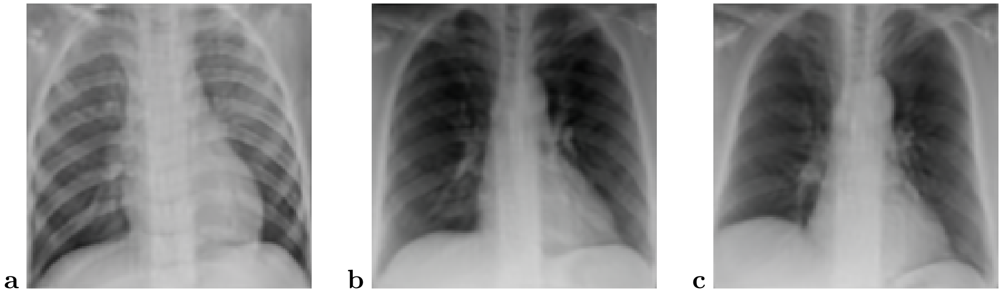

Fig. 5. Kernel Inception Distance values with correlated example image generated by GAN trained on the micro dataset. (**a**) KID ≈ 18.82; (**b**) KID ≈ 13.30; (**c**) KID ≈ 12.89.

the batch size was set to 32. Following StyleGAN2 authors [20], we use non-saturating logistic loss with R_1 regularization with the regularization term weight γ set to 1.024. Multiclass training was enabled so one network was able to generate images for all target classes after the training was done. All other parameters are set to default values provided by the NVIDIA implementation [19]. The network was trained with a single NVIDIA Tesla K80 GPU and for each case, training took around 7 days.

After the training had been finished we generated 1000 artificial images per class for both experiments, with example images presented in Fig. 4 and 5. The GAN-augmented training dataset for CNN contained 2000 images per class (8000 images in total) for the small dataset and 1500 images per class (6000 in total) for the micro dataset.

3.4. Classical augmentation

Similarly to the GAN-based augmentation described earlier, we have generated 1000 additional images (Fig. 6) by applying classical augmentation with parameters as follows:

- rotation – randomly rotate the image by an angle of up to 5 degrees clockwise or counterclockwise;
- shift – randomly shift the image along cardinal axes within the range of 5% of the specific image size, the empty field is filled with the trace of the last shifted pixels;
- stretch – randomly stretch the image between opposite vertices by up to 5 %;
- zoom – randomly zoom in pictures up to 15% of the specific value of the image size;
- brightness change – randomly brighten or darken the image by up to 40%.

The values of the parameters remain the same as in the previous work and were picked to maximize the accuracy score on the validation subset [10].

3.5. Image comparison metrics

In the previous work, we have used the Fréchet Inception Distance (FID) [15] metric to select the best-performing state of the StyleGAN2-ADA network [10] throughout the training process as it is commonly used to evaluate the quality of images generated by GANs. We calculated a mean FID value, for each training epoch, across all classes between the train subset and generated images. In this research, we added the Kernel Inception Distance (KID) [4] metric as FID is biased for smaller datasets [6]. KID is very similar to FID in that it measures the difference between two sets of samples by calculating the square of the maximum mean discrepancy between vectors of vision-relevant features as extracted by the Inception-v3 [38] classifier network. KID compared to FID has several advantages and performs better with smaller sets and more consistently matches human perception. Similarly to FID, a smaller value of KID means that compared images are more similar to each other, and comparing the same set of images will result in a value equal to 0. Looking at the graphs of both KID and FID values per epoch (Figures 7–10) it is visible that the overall training trend is the same – the value drops with each epoch of training until around epoch 250 and then it starts to grow slowly. But at the same time, KID values change more drastically per each epoch

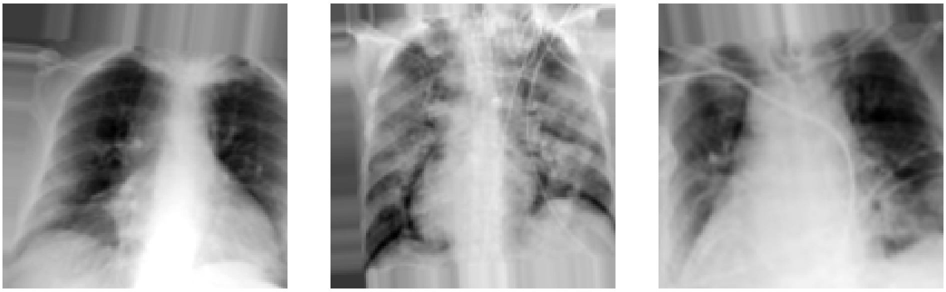

Fig. 6. Examples of images with applied classical augmentation pipeline.

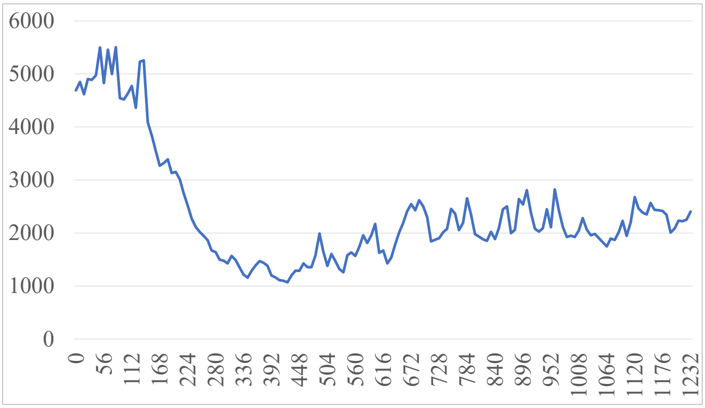

Fig. 7. Fréchet Inception Distance graph for the small dataset (1000 images per class in the train subset) GAN training.

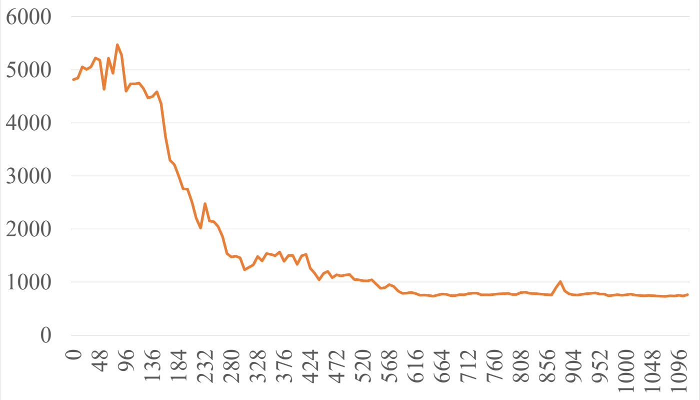

Fig. 8. Fréchet Inception Distance graph for the micro dataset (500 images per class in the train subset) GAN training.

which may indicate, that the KID metric is more sensitive to differences between real and generated images. In addition, as in the original paper, we used RMSE, SRE, and SSIM metrics to verify the quality of generated images [10].

3.6. Classification evaluation

To evaluate and compare the augmentation techniques described above, we trained a convolutional neural network using each of them. The network was trained to classify

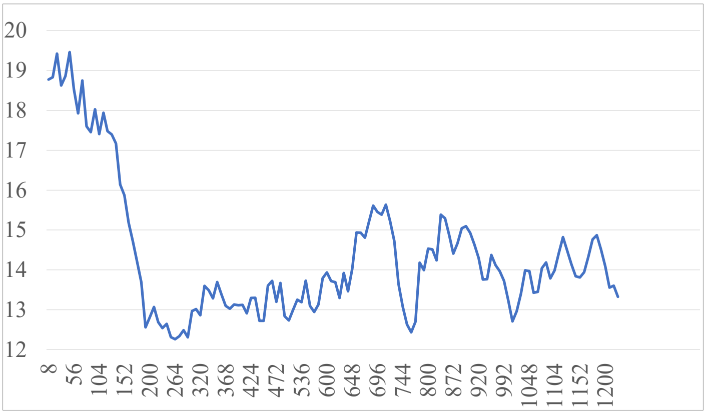

Fig. 9. Kernel Inception Distance graph for case the small dataset (1000 images per class in the train subset) GAN training.

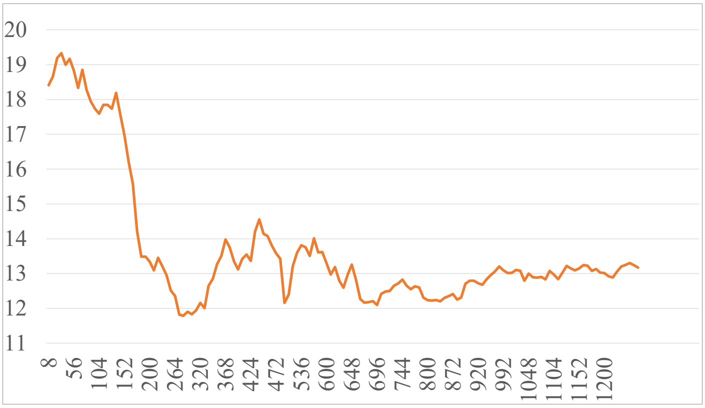

Fig. 10. Kernel Inception Distance graph for the micro dataset (500 images per class in the train subset) GAN training.

4 classes present in the original dataset. We used Inception-v3 with a transfer learning technique as an augmentation benchmark network [38]. We used Keras default implementation with library-provided ImageNet weights [17]. The head of the network was replaced with the following output layers for the transfer learning:

- flatten layer;
- dense layer with 10 neurons and an Exponential Linear Unit (ELU) activation function [7];
- dense layer with 20 neurons and an ELU activation function;
- dense layer with 4 neurons and a softmax activation function.

Categorical cross-entropy was picked as the loss function and the learning rate was set to Keras default of 0.001 [31]. We continue using RMSprop as an optimizer as it was selected on validation accuracy in the original article [10]. The network was trained for 11 epochs with a batch size equal to 32. The total amount of epochs was reduced in comparison with previous work as the best results were achieved in the first 11 epochs in almost all training runs. For the clarity of the experiment, the network was trained 5 times for each experimental case and each augmentation pipeline (no augmentation, classical augmentation, GAN-augmentation). The network frozen weights with the highest validation accuracy score were picked as a final version that was later evaluated on the test subset. To examine and compare the quality of the trained network, several model evaluation metrics were calculated: accuracy, precision, recall, F1, specificity, and Matthew's correlation coefficient (MCC). Metrics were calculated on the test dataset. Values of the calculated metrics are presented in the Results section of the paper.

4. Results

After the classification metrics are calculated on the test dataset, it is visible, that any augmentation approach is better than no augmentation at all. At the same time, GAN-based augmentation and classical augmentation perform comparably for the dataset with at least 1000 images per class in the training dataset, as shown in Table 2 and Fig. 11. Classical augmentation outperforms GAN-based augmentation with datasets containing 500 images per class, as presented in Table 3 and Fig. 12. The overall networks' results are slightly worse independently of the augmentation approach in comparison with the ones achieved in our previous work [10]. Finally, the classical augmentation shows itself as the best augmentation approach while GAN-based augmentation can achieve comparable results but requires significantly more time and hardware to be performed.

Tab. 2. Classification metrics values for the small dataset (1000 images in the train subset)

Augmentation pipeline	Accuracy	Precision	Recall	F1	Specificity	MCC
No augmentation	0.85	0.845	0.769	0.805	0.931	0.74
Classical augmentation	0.87	0.861	0.815	0.837	0.934	0.78
GAN-augmentation	0.862	0.822	0.815	0.819	0.936	0.76

Tab. 3. Classification metrics values for the micro dataset (500 images in the train subset)

Augmentation pipeline	Accuracy	Precision	Recall	F1	Specificity	MCC
No augmentation	0.783	0.659	0.68	0.669	0.903	0.573
Classical augmentation	0.842	0.85	0.789	0.818	0.93	0.749
GAN-augmentation	0.81	0.83	0.685	0.751	0.9	0.665

Tab. 4. Accuracy metric value from the original article (2000 images per dataset)

Augmentation pipeline	Accuracy
No augmentation	0.855
Classic augmentation	0.891
GAN-augmentation	0.871

We can take a look (Table 4) at the accuracy value obtained in our previous work where 2000 images per class were used [10], there is a visible correlation between classification accuracy and the size of the training dataset independently from data augmentation applied.

5. Conclusion

We have studied the performance of GAN-based augmentation for the classification of lung X-ray medical images as a function of dataset size. The obtained results show that GAN-based augmentation is comparable with classical augmentation for medium and large datasets. Unfortunately, the time and hardware requirements make it unreasonable to use such an approach as the main augmentation technique. In the case of small datasets, the GAN model wasn't able to train well enough to be a source of valuable training data. At the same time, the fact of GAN being able to compete with classical augmentation for larger datasets, potentially allows researchers and medical institutions to solve the problem of medical data availability by sharing synthetically generated images instead of real ones [39]. Therefore, the topic of GAN-based augmentation should be investigated further.

Acknowledgments

This work was completed in part with resources provided by the Świerk Computing Centre at the National Centre for Nuclear Research. This work benefited from the software tools developed in the frame of the EuroHPC PL Project, Smart Growth Operational

Programme 4.2. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centres: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2022/015617.

References

[1] S. Albahli. Efficient GAN-based chest radiographs (CXR) augmentation to diagnose coronavirus disease pneumonia. *International Journal of Medical Sciences*, 17(10):1439–1448, 2020. doi:10.7150/ijms.46684.

[2] H. X. Bai, B. Hsieh, Z. Xiong, K. Halsey, J. W. Choi, et al. Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT. *Radiology*, 296(2):E46–E54, Aug 2020. doi:10.1148/radiol.2020200823.

[3] M. Bali and T. Mahara. Comparison of affine and DCGAN-based data augmentation techniques for chest X-ray classification. *Procedia Computer Science*, 218:283–290, 2023. International Conference on Machine Learning and Data Engineering. doi:10.1016/j.procs.2023.01.010.

[4] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In: *Proc. Int. Conf. Learning Representations (ICLR 2018)*, 2018. https://openreview.net/forum?id=r1lUOzWCW.

[5] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, et al. GAN augmentation: Augmenting

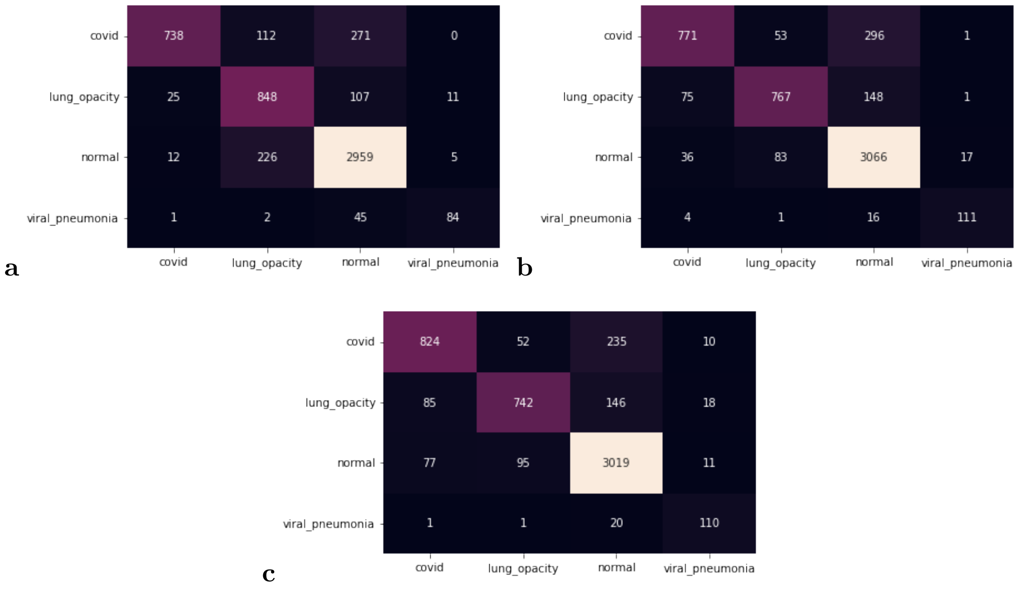

Fig. 11. Confusion matrices for different augmentations of the small dataset. The vertical axis represents predicted values, horizontal axis represents real values. (**a**) No augmentations; (**b**) Classical augmentations; (**c**) GAN-based augmentations.

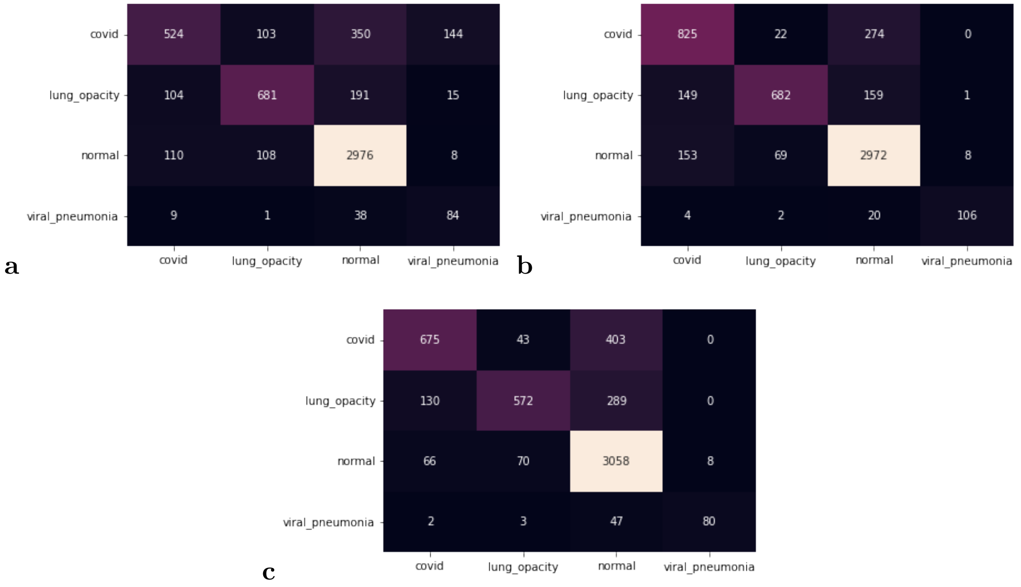

Fig. 12. Confusion matrices for different augmentations of the micro dataset. The vertical axis represents predicted values, horizontal axis represents real values. (**a**) No augmentations; (**b**) Classical augmentations; (**c**) GAN-based augmentations.

training data using Generative Adversarial Networks. *arXiv*, 2018. ArXiv.1810.10863. https://arxiv.org/abs/1810.10863.

[6] M. J. Chong and D. Forsyth. Effectively unbiased FID and inception score and where to find them. In: *Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6069–6078, 2020. doi:10.1109/CVPR42600.2020.00611.

[7] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). In: *Proc. Int. Conf. Learning Representations (ICLR 2016)*, 2016. https://arxiv.org/abs/1511.07289.

[8] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Duong, et al. COVID-19 image data collection: Prospective predictions are the future. *Machine Learning for Biomedical Imaging*, 1:1–38, 2020. doi:10.59275/j.melba.2020-48g7.

[9] D. Ezzat, A. E. Hassanien, and H. A. Ella. An optimized deep learning architecture for the diagnosis of COVID-19 disease based on gravitational search optimization. *Applied Soft Computing*, 98:106742, Jan 2021. doi:10.1016/j.asoc.2020.106742.

[10] O. Fedoruk, K. Klimaszewski, A. Ogonowski, and R. Możdżonek. Performance of GAN-based augmentation for deep learning COVID-19 image classification. In: *Proc. Int. Workshop on Machine Learning and Quantum Computing Applications in Medicine and Physics*. Warsaw, Poland, 13-16 Sep 2022. Accepted for publication in AIP Conference Proceedings. https://events.ncbj.gov.pl/event/141/page/65-home.

[11] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, et al. GAN-based synthetic medical

image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321:321–331, Dec 2018. doi:10.1016/j.neucom.2018.09.013.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, et al. Generative Adversarial Networks. *Advances in Neural Information Processing Systems*, 3, 06 2014. doi:10.1145/3422622.

[13] W. Gouda, M. Almurafeh, M. Humayun, and N. Z. Jhanjhi. Detection of COVID-19 based on chest X-rays using deep learning. *Healthcare*, 10(2):343, 2022. doi:10.3390/healthcare10020343.

[14] O. Gozes, M. Frid-Adar, H. Greenspan, P. D. Browning, H. Zhang, et al. Rapid AI development cycle for the coronavirus (COVID-19) pandemic: Initial results for automated detection & patient monitoring using Deep Learning CT image analysis. *arXiv*, 2020. ArXiv.2003.05037. doi:10.48550/arXiv.2003.05037.

[15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, et al., eds., *Advances in Neural Information Processing Systems: Proc. NIPS 2017*, vol. 30. Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

[16] Md. I. Zabirul, Md. I. Milon, and A. Asraf. A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images. *Informatics in Medicine Unlocked*, 20:100412, 2020. doi:10.1016/j.imu.2020.100412.

[17] InceptionV3 – Keras Applications API Reference. https://keras.io/api/applications/inceptionv3/, [Accessed Dec 2023].

[18] J. Jeong, A. Tariq, T. Adejumo, H. Trivedi, J. Gichoya, et al. Systematic review of generative adversarial networks (GANs) for medical image classification and segmentation. *Journal of Digital Imaging*, 35, 01 2022. doi:10.1007/s10278-021-00556-w.

[19] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, et al. Training generative adversarial networks with limited data. In: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds., *Advances in Neural Information Processing Systems: Proc. NeurIps 2020*, vol. 33, pp. 12104–12114. Curran Associates, Inc., 2020. https://proceedings.neurips.cc/paper_files/paper/2020/hash/8d30aa96e72440759f74bd2306c1fa3d-Abstract.html.

[20] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, et al. Analyzing and improving the image quality of StyleGAN. In: *Proc. CVPR*, 2020. 1912.04958, https://arxiv.org/abs/1912.04958.

[21] D. Kermany. Labeled optical coherence tomography (OCT) and chest X-ray images for classification. *Mendeley Data*, V2, 2018. doi:10.17632/rscbjbr9sj.2.

[22] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131.e9, Feb 2018. doi:10.1016/j.cell.2018.02.010.

[23] A. I. Khan, J. L. Shah, and M. M. Bhat. Coronet: A deep neural network for detection and diagnosis of COVID-19 from chest X-ray images. *Computer Methods and Programs in Biomedicine*, 196:105581, Nov 2020. doi:10.1016/j.cmpb.2020.105581.

[24] E. Khan, M. Z. U. Rehman, F. Ahmed, F. A. Alfouzan, N. M. Alzahrani, et al. Chest X-ray classification for the detection of COVID-19 using deep learning techniques. *Sensors*, 22(3):1211, 2022. doi:10.3390/s22031211.

[25] J. Li, G. Zhu, C. Hua, M. Feng, B. Bennamoun, et al. A systematic collection of medical image datasets for deep learning. *ACM Computing Surveys*, 56(5), nov 2023. doi:10.1145/3615862.

[26] S. Motamed, P. Rogalla, and F. Khalvati. Data augmentation using Generative Adversarial Networks (gans) for gan-based detection of pneumonia and COVID-19 in chest X-ray images. *Informatics in Medicine Unlocked*, 27:100779, 2021. doi:10.1016/j.imu.2021.100779.

[27] S. Motamed, P. Rogalla, and F. Khalvati. RANDGAN: Randomized generative adversarial network for detection of COVID-19 in chest X-ray. *Scientific Reports*, 11(1):8602, Apr 2021. doi:10.1038/s41598-021-87994-2.

[28] A. Narin, C. Kaya, and Z. Pamuk. Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. *Pattern Analysis and Applications*, 24(3):1207–1220, May 2021. doi:10.1007/s10044-021-00984-y.

[29] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. In: D. Precup and Y. W. Teh, eds., *Proc. 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, pp. 2642–2651. PMLR, 06-11 Aug 2017. https://proceedings.mlr.press/v70/odena17a.html.

[30] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, et al. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Computers in Biology and Medicine*, 132:104319, 2021. doi:10.1016/j.compbiomed.2021.104319.

[31] RMSprop – Keras Optimizers API Reference. https://keras.io/api/optimizers/rmsprop/, [Accessed Dec 2023].

[32] G. D. Rubin, C. J. Ryerson, L. B. Haramati, N. Sverzellati, J. P. Kanne, et al. The role of chest imaging in patient management during the COVID-19 pandemic: A multinational consensus statement from the fleischner society. *Chest*, 158(1):106–116, Jul 2020. doi:10.1016/j.chest.2020.04.003.

[33] W. Saad, W. A. Shalaby, M. Shokair, F. A. El-Samie, M. Dessouky, et al. COVID-19 classification using deep feature concatenation technique. *Jornal of Ambient Intelligence and Humanized Computing*, 13(4):2025–2043, 2022. doi:10.1007/s12652-021-02967-7.

[34] K. Sahinbas and F. O. Catak. Transfer learning-based convolutional neural network for COVID-19 detection with X-ray images. In: U. Kose, D. Gupta, V. H. C. de Albuquerque, and A. Khanna, eds., *Data Science for COVID-19 – Computational Prespective*, chap. 24, pp. 451–466. Academic Press, 2021. doi:10.1016/B978-0-12-824536-1.00003-4.

[35] H. Salehinejad, S. Valaee, T. Dowdell, E. Colak, and J. Barfett. 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, 2018. doi:10.1109/ICASSP.2018.8461430.

[36] C. Shorten and T. M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019. doi:10.1186/s40537-019-0197-0.

[37] N. K. Singh and K. Raza. *Medical Image Generation Using Generative Adversarial Networks: A Review*, pp. 77–96. Springer Singapore, 2021. doi:10.1007/978-981-15-9735-0_5.

[38] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception architecture for computer vision. In: *2016 IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016. doi:10.1109/CVPR.2016.308.

[39] R. Venugopal, N. Shafqat, I. Venugopal, B. M. J. Tillbury, H. D. Stafford, et al. Privacy preserving generative adversarial networks to model electronic health records. *Neural Networks*, 153:339–348, 2022. doi:10.1016/j.neunet.2022.06.022.

[40] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, et al. CovidGAN: Data augmentation using auxiliary classifier GAN for improved Covid-19 detection. *IEEE Access*, 8:91916–91923, 2020. doi:10.1109/ACCESS.2020.2994762.

[41] S. E. Whang, Y. Roh, H. Song, and J.-G. Lee. Data collection and quality challenges in deep learning: A data-centric AI perspective. *The VLDB Journal*, 32(4):791–813, 2023. doi:10.1007/s00778-022-00775-9.

[42] X. Xu, X. Jiang, C. Ma, P. Du, X. Li, et al. A deep learning system to screen novel coronavirus disease 2019 pneumonia. *Engineering*, 6(10):1122–1129, 2020. doi:10.1016/j.eng.2020.04.010.

[43] X. Yi, E. Walia, and P. Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58:101552, 2019. doi:https://doi.org/10.1016/j.media.2019.101552.