Better Pneumonia Diagnosis with GAN-enabling Data Augmentation

Nguyen Thanh Tung, Nguyen Tien Dat, Lai Dang Quoc Vinh, Dinh Khac Tuyen
Korea Advanced Institute of Science and Technology (KAIST)
{tungnt, tiendat, ldqvinh, tuyen.dk}@kaist.ac.kr

Abstract

In recent years, Deep Learning has proven its huge potential in terms of medical diagnosis. Pneumonia, a condition that can be confirmed based on X-ray images by a doctor, now can be confirmed by a Convolutional Neural Network as well and with high precision. However, as medical datasets are scarce, they can also be naturally imbalanced. In this paper, we explore the use of data augmentation in training a classification network, CheXNet. We first acquire the generated data using two generative adversarial networks (GANs), called DCGAN and CGAN, which then is combined with the original dataset to train the classifier under different settings. Our experimental results show that with the new augmented dataset, the performance of the classifier increases by 1 - 6% in terms of accuracy and precision. We also provide qualitative and quantitative analysis regarding the augmented dataset. The implementations of this work are avaiable at: https://github.com/tungnt101294/CheXNet-for-Augmented-Pneumonia

1. Introduction

Pneumonia is a lung condition caused by respiratory illnesses, which was ranked eighth among the top ten causes of death in the United States in 2016 [1]. To diagnose pneumonia, chest X-ray is one of the most used methods [2] because it gives doctors visual confirmation as well as a means to locate the infected area. However, the scarcity of qualified doctors and radiologists, for example, during difficult times like the pandemic we are experiencing, raises a demand computer-aided diagnosis methods, especially artificial intelligence-based solutions with high performance. The application of deep learning, particularly Convolutional Neural Networks (CNNs), in biomedical image diagnosis has been proven to be highly useful in providing a quick and accurate diagnosis of health conditions including skin lesions [3], colon cancer [4], blood cancer, breast cancer [5], heart anomalies [6]. Li *et al.* [7] presented a customized CNN with shallow ConvLayer to classify image patches of lung disease. The authors also found that the system can be generalized to other medical image datasets. Another widely used CNN architecture for Chest X-ray classification is CheXNet, which was proposed in [8]. The authors implemented a 121-layer Dense Convolutional Network trained on ChestXray-14, which achieved a better performance than the average performance of four radiologists.

One of the major challenges in this area is that high-quality medical image data is scarce as it is considered sensitive and confidential information. Take pneumonia X-ray datasets as an example. Currently, there are only a few publicly available. Moreover, just as they are scarce, medical datasets can also be severely imbalanced, especially for rare diseases. To overcome the challenge of limited datasets, data augmentation is usually employed by researchers. Amongst data augmentation methods, Generative Adversarial Network (GAN) has been a top candidate in recent years. It was introduced by Goodfellow *et al.* [9] as a form of synthetic data augmentation which can overcome the limitations of classical data augmentation because GANs are able to generate outputs with far more complex features. It is a powerful method to generate unseen samples by using a generator model for generating new examples to "trick" a discriminator whose job is to classify the generated sample. This process can be done completely without supervision [10]. Various methods for using GANs to expand datasets have been proposed. One variant of GAN, conditional GAN [11] (CGAN), involves the conditional generation of images by a generator model. Image generation can be conditional on a class label, if available, allowing targeting generated images of a given type. Progressive Growing GAN (PCGAN) [12] is an extension to the GAN training process that allows for the stable training of generator models that capable of producing large high-quality images. Another widely used GAN network in generating image dataset is Deep convolutional GAN (DCGAN) [13]. It mainly composes of convolution layers without max pooling layers. It uses strided convolution and transposed convolution for downsampling and upsampling image samples.

In this paper, we tackle the issue of imbalanced datasets used for pneumonia diagnosis with generated datasets from

GANs. First, we use DCGAN and CGAN to generate more data based on the original dataset provided in [14]. Next, we train a pneumonia-classifying model, called CheXNet [8], on different combinations of the generated and original datasets. This aims to evaluate how the new datasets can improve the classification performance as well as explore the severity of dataset imbalance. In short, our main contributions are: (1) our generated datasets improve the classification performance by 1 - 6% in terms of accuracy and precision, compared to training on the original dataset and (2) we offer qualitative and quantitative analysis regarding generated datasets by DCGAN and CGAN, as well as different combinations of datasets.

The rest of this paper is organized as follows. Section 2 reviews the existing literature in the field. Next, in Section 3 we introduce the architectures of generative and classifier networks used in this research. Then, Section 4 provide details on experimental setups, results, and analysis. Finally, Section 5 concludes the paper and proposes future directions.

2. Related Work

Pneumonia Detection using Deep Learning has been proposed in a number of previous works. Stephen *et al.* [15] proposed a CNN model trained from scratch to classify and detect the presence of pneumonia from a collection of chest X-ray image samples. The authors also deployed several data augmentation algorithms to improve the validation and classification accuracy of the CNN model, but the method simply using manual rescale and shift operations. A similar approach was proposed by Hashmi *et al.* [16], which combines the weighted predictions from the state-of-the-art deep learning models such as ResNet18 [17], Xception [18], InceptionV3 [19], DenseNet121 [20], and MobileNetV2 [21] in an optimal way. In the another research, Ayan *et al.* [22] compared two well-known CNN models Xception and VGG16 [23] for diagnosing of pneumonia. They used transfer learning and fine-tuning of the model in the training stage. The test results showed that VGG16 network exceed Xception network at the accuracy, but the Xception network achieved a more successful result in detecting pneuamonia cases. Rajpurkar *et al.* [8] introduced a novel network called CheXNet, which is a 121-layer Dense Convolutional neural network (DenseNet) trained on ChestX-ray14, the largest publicly available chest X-ray dataset, containing over 100,000 frontal-view X-ray images with 14 diseases, one of which is pneumonia. The detection performance of CheXNet was proved to have surpassed that of radiologists.

Despite the many advancements in classification models, high-quality data still plays a vital role in the diagnosis performance. The lack of data in medical imaging proposed the need for data augmentation. Nevertheless, the common data augmentation method that uses the same picture for scaling and rotation make the diversity of the dataset is not enough so that it leaves their applications ad-hoc and empirical. Recently, the GAN framework has been used by many medical imaging techniques. Salehinejad *et al.* [24] used Deep Convolutional Generative Adversarial Network (DCGAN) method to detect pathology across five classes of chest X-rays. With the balanced dataset containing synthetic images made by DCGAN, they trained a modified version of AlexNet and showed significant improvements in the chest pathology classification task, comparedonly training with the original dataset. Beers *et al.* [12] trained PGGAN, a model that can produce realistic medical images in two different domains: fundus photographs exhibiting vascular pathology associated with retinopathy of prematurity (ROP) and multi-modal magnetic resonance images of glioma. Waheed *et al.*[25] presented the Auxiliary Classifier Generative Adversarial Network (ACGAN) to generate synthetic chest X-ray images that can be utilized to enhance the performance of CNN for COVID-19 detection. Furthermore, one factor that could heavily influence the quality of a dataset is the balance between its classes. Models trained with imbalanced dataset will have inherent bias towards one or a number of classes while neglecting the others, which decreases its generalization ability at inference time. While much effort has been put in handling imbalance medical datasets, using GANs for such purpose is still uncharted territory. In this work, with the aim of tackling dataset imbalance, we use two different GAN models, namely DCGAN and CGAN to generate more data for the mini classes. The quality of the augmented data would be confirmed using CheXNet.

3. Model Architectures

In this section, we introduce the two generative networks, DCGAN and CGAN, employed to generate synthetic X-ray images of a Pneumonia dataset. The expected output for these two networks are 256x256, which is the size of input for the classification network, CheXNet.

3.1. DCGAN

The first architecture that we use is DCGAN [13], which is a direct extension of the original GAN, except that the discriminator and generator explicitly use convolutional and transpose-convolutional layers, respectively. In this architecture, as describe by Fig. 1, a 100x1 noise vector is fed as an input to the generator. There are then five Convolutional layers with 2D-upsampling layers applied with Leaky ReLu activation function interlaced in between to scale to the appropriate 256x256 image size. The discriminator network is a similar network with five convolutional layers with a stride of 2, using Leaky ReLU as the activation function except for the final node which is a sigmoid activation function

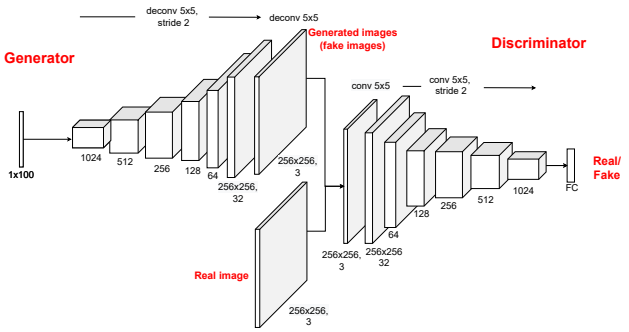

Figure 1: The architecture of DCGAN

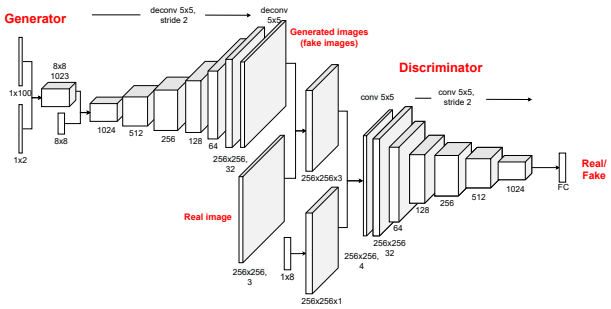

Figure 2: The architecture of CGAN

to output if the image is real (original data) or fake (generated data). Despite suggested by the authors, we did not use batch normalization layers as we found them make the output unstable. We use Adam optimizer [26] with $\beta_1 = 0.5$.

3.2. CGAN

Looking at Fig. 2, proposed by [11], has a quite similar architecture to that of DCGAN, but the difference is that in CGAN, the class label vector is fed into both generator and discriminator as additional input layer. Through this addition, our model would be able to generate images that are conditioned on class labels, therefore making it possible to direct the data generation process. The remaining part is quite the same as DCGAN: The generator network has 5 convolutional layers with 2D-upsampling layers to scale to the appropriate 256x256 image size. The discriminator network has 5 convolutional layers and a stride of 2 with Leaky RELU as the activation function except for the final node which is a sigmoid activation function. Similar to CGAN, an Adam optimizer with $\beta_1 = 0.5$ is used. We expect that this additional information would help CGAN to generate images of high quality, since it can exploit additional information (class label).

3.3. CheXNet for Pneumonia Diagnosis

Our classification network, CheXNet is a 121-layer Dense Convolutional Network (DenseNet) that inputs a chest X-ray image and outputs the probability of pneumonia along with a heatmap localizing the areas of the image most indicative of pneumonia. DenseNets improve flow of information and gradients through the network, making the optimization of such deep network tractable. In this architecture, the final fully connected layer is replaced with one that has a single output, and then followed by a sigmoid activation layer. The weights of the network are initialized with weights from a model pretrained on ImageNet [27]. The network is trained end-to-end using Adam with a standard parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.99$) [26]. We train the model using minibatches of size 16. We use an initial learning rate of 0.001 that is decayed by a factor of 10 each time the validation loss plateaus after every 5 epoch, and pick the model with the lowest validation loss.

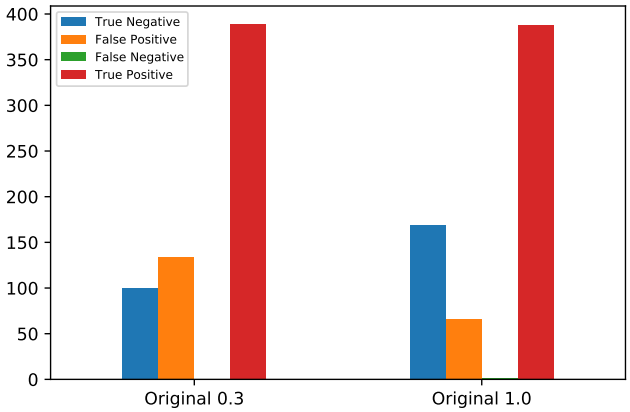

Figure 3: Classification result from the original dataset

4. Experiments and Result Analysis

In this section, firstly, the dataset used in this research is analysed in detail to prove the need for data augmentation. Secondly, we provide the detailed setups used in the experiments. Lastly, we show and discuss the experimental results.

4.1. Dataset

In this paper, we use the Pneumonia X-ray dataset provided in [14]. The dataset contains of nearly 6000 frontal chest X-ray images, about 5100 of which are used for training and the rest are used as the test set. Approximately three fourths of the training images contain positive results, which makes the dataset significantly imbalanced. We trained the network 3 times on 30% and 100% of the original dataset and the results are shown in Figure 3. Since the train set is heavily skewed towards the positive class, the classifier made a significant number of False Positives in both cases. This degenerates the performance of the classifier and can potentially cause unwanted psychological to patients. Therefore, we try to tackle this issue by using

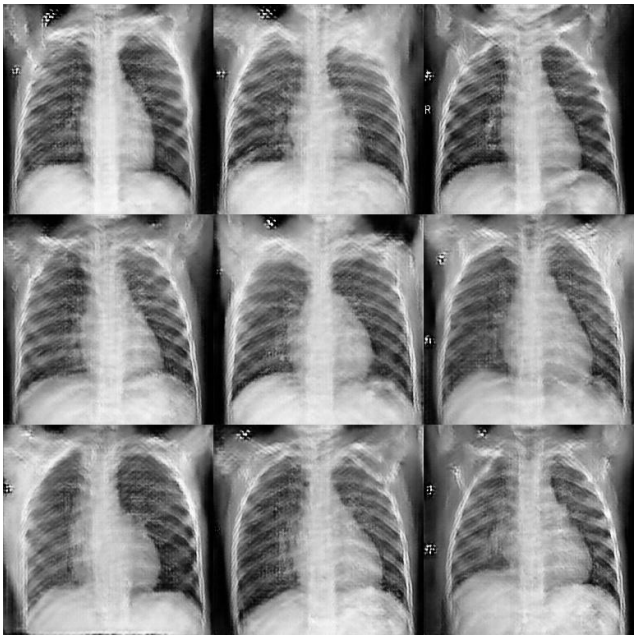

Figure 4: Samples of X-ray images generated by DCGAN

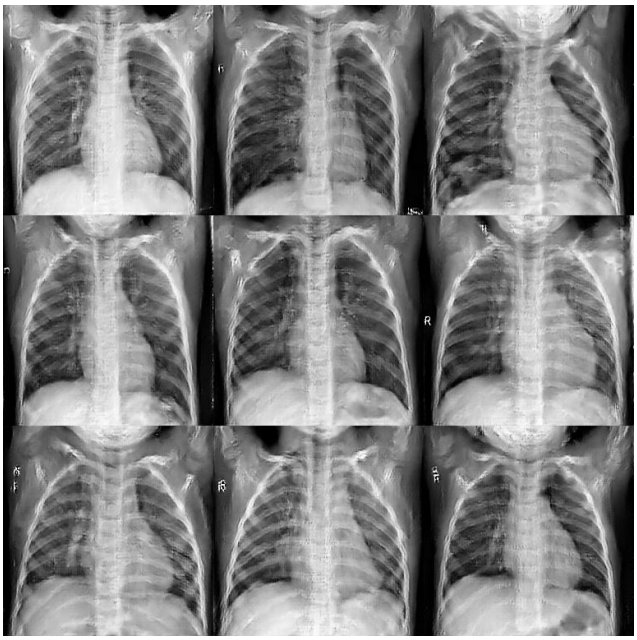

Figure 5: Samples of X-ray images generated by CGAN

CGAN and DCGAN to generate complementary images for the negative class. With each generative network, we generated 4000 frontal X-ray images. Figure 4 and 5 show the sample images generated by DCGAN [13] and CGAN [11] respectively. The detailed setups to generate the above-mentioned images, their quality, and how they will be used as training data to CheXNet [8] will be discuss later this section.

4.2. Experimental setups

This research was performed on a computer running Ubuntu 18.04.2 LTS with Intel Xeon Processor (19.25M Cache, 2.60 GHz x 48), 110GB RAM, and NVIDIA Tesla P40 GPU (24GB). The versions of driver and CUDA toolkit are 11.3 and 9.1.8. We used Tensorflow-gpu v2.4.1 and PyTorch 1.4.0 along with TorchVision 0.5.0 on the base of Python 3.6.9.

To generate the images shown in Figure 4 and 5, we used the architectures of DCGAN [13] and CGAN [11] described in Section 3. Each network was trained for 500 epochs with learning rate of 0.005 and minibatches of 128, which roughly took 15 hours on average. After the training ended, the generators were fed random vectors to generate 4000 images each.

The evaluate the effect of different magnitudes of imbalance, we take 30% and 100% of the original dataset, which are then mixed with 30%, 60% and 100% of the generated one, respectively. This mixing creates 5 different settings for each augmented dataset. From now on, CGAN-0.3-1.0 would refer to the augmented dataset between the 30% of the original and 100% of the GAN-generated dataset. Similarly, Original-1.0-0.0 would refer to 100% of the original dataset. Each setting is used to train the CheXNet model [8] for 3 times, 100 epoch each, which takes roughly 7 hours. The Tensorflow and Pytorch implementations of this research are available at: https://github.com/tungnt101294/CheXNet-for-Augmented-Pneumonia

4.3. Qualitative Analysis

Looking at Figre 4 and 5, we can see that both GAN models can generate decent images of the torso, rib cage, heart and first stomach. We believe these are the important features that help the classifier locate important areas, which is the lung. However, DCGAN seems to model the shape of the torso better than CGAN. Both generative model could not completely suppress the noisy details in the lung which can confuse the classifier. This is because in both bacterial [28] and [29] pneumonia cases, X-ray images show low opacity around the lung area as a result of inflammation and the thickening of interlobular septal.

4.4. Quantitative Analysis

On the training process of GANs shown in Figure 6, we can notice on important detail that is DCGAN's generator loss starts slowly going up, while discriminator loss decreases gradually, after about 270 epochs until the end of the training. This shows that DCGAN generator cannot "catch up" with its discriminator, which has gotten better and now been able to pick up the difference in distributions of real and fake images. On the other hand, CGAN generator, ex-

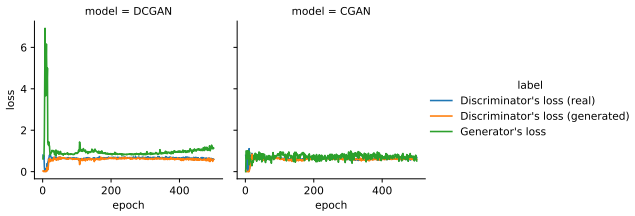

Figure 6: The loss values of DCGAN and CGAN

Setting	Accuracy	Recall	Precision
Original-0.3-0.0	78.365	0.997	0.744
Original-1.0-0.0	89.423	0.997	0.857
CGAN-0.3-0.3	80.929	0.997	0.767
CGAN-0.3-0.6	74.359	0.997	0.710
CGAN-0.3-1.0	84.295	0.997	0.800
CGAN-1.0-0.3	90.385	0.995	0.870
CGAN-1.0-0.6	90.545	0.992	0.874
DCGAN-0.3-0.3	80.449	0.997	0.763
DCGAN-0.3-0.6	75.962	0.997	0.723
DCGAN-0.3-1.0	79.968	0.995	0.759
DCGAN-1.0-0.3	**91.879**	0.997	**0.886**
DCGAN-1.0-0.6	**92.468**	0.995	**0.896**

Table 1: Comparisons in terms of accuracy, precision and recall from different settings

ploiting the extra label information, could keep up with the generator.

Table 1 shows the comparisons in terms of accuracy, precision and recall of CheXNet trained with different combinations of datasets. It is obvious that most cases have very high recall resulted from extremely low number of False Negatives. This may seem to be a good sign, but in this case, it could be due to the skewness of the dataset. Accuracies and precisions, on the other hand, see improvements of 1 - 6% with the use of augmented datasets. High precisions, resulted from lower numbers of False Positives, prove that the generated images have been able to pull the data distribution closer to the balance.

5. Conclusion and Future Work

In this paper, we have explore the use of GANs to generate X-rays images of a heavily imbalanced pneumonia dataset. Our results show certains improvement in terms of accuracy and precision. However, these improvements are quite small due to the output of the GAN models' not being able to match the original images. We believe that in the future, by applying newer models of GANs, we can significantly improve the results. Moreover, CheXNet [8] is able to classify different diseases and conditions besides pneumonia. Therefore, we plan to apply new GAN models to these diseases and conditions as well, because each of them would show up at a different location and in a different manner in X-ray images. So, their classification could provide crucial information regarding the characteristics of each GAN.

References

[1] University of Utah Healthcare. Pneumonia makes list for top 10 causes of death. https://healthcare.utah.edu/the-scope/shows.php, accessed on 06/05/2021. 1

[2] World Health Organization. Standardization of interpretation of chest radiographs for the diagnosis of pneumonia in children / World Health Organization Pneumonia Vaccine Trial Investigators' Group. *Technical Report*, 2001. 1

[3] E. Nasr-Esfahani, S. Samavi, N. Karimi, S. M.R. Soroushmehr, M. H. Jafari, K. Ward, and K. Najarian. Melanoma detection by analysis of clinical images using convolutional neural network. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, volume 2016-October, pages 1373–1376, 2016. 1

[4] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David R. J. Snead, Ian A. Cree, and Nasir M. Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1196–1206, 2016. 1

[5] Michiel Kallenberg, Kersten Petersen, Mads Nielsen, Andrew Y. Ng, Pengfei Diao, Christian Igel, Celine M. Vachon, Katharina Holland, Rikke Rass Winkel, Nico Karssemeijer, and Martin Lillholm. Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring. *IEEE Transactions on Medical Imaging*, 35(5):1322–1331, 2016. 1

[6] Serkan Kiranyaz, Turker Ince, and Moncef Gabbouj. Real-time patient-specific ecg classification by 1-d convolutional neural networks. *IEEE Transactions on Biomedical Engineering*, 63(3):664–675, 2016. 1

[7] Qing Li, Weidong Cai, Xiaogang Wang, Yun Zhou, David Dagan Feng, and Mei Chen. Medical image classification with convolutional neural network. *Proceedings of International Conference on Control Automation Robotics and Vision (ICARCV)*, pages 844–848, 2014. 1

[8] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017. 1, 2, 4, 5

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Y. Bengio. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 3, 06 2014. 1

[10] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *ICML*, volume 2019-June, pages 3218–3238, 2019. 1

[11] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. pages 1–7, 2014. 1, 3, 4

[12] Andrew Beers, James Brown, Ken Chang, J Peter Campbell, Susan Ostmo, Michael F Chiang, and Jayashree Kalpathy-Cramer. High-resolution medical image synthesis using progressively grown generative adversarial networks. *arXiv preprint arXiv:1805.03144*, 2018. 1, 2

[13] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, pages 1–16, 2016. 1, 2, 4

[14] Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, Carolina C.S. Valentim, Huiying Liang, Sally L. Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, Justin Dong, Made K. Prasadha, Jacqueline Pei, Magdalena Ting, Jie Zhu, Christina Li, Sierra Hewett, Jason Dong, Ian Ziyar, Alexander Shi, Runze Zhang, Lianghong Zheng, Rui Hou, William Shi, Xin Fu, Yaou Duan, Viet A.N. Huu, Cindy Wen, Edward D. Zhang, Charlotte L. Zhang, Oulan Li, Xiaobo Wang, Michael A. Singer, Xiaodong Sun, Jie Xu, Ali Tafreshi, M. Anthony Lewis, Huimin Xia, and Kang Zhang. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 172(5):1122–1131.e9, 2018. 2, 3

[15] Okeke Stephen, Mangal Sain, Uchenna Joseph Maduh, and Do-Un Jeong. An efficient deep learning approach to pneumonia classification in healthcare. *Journal of healthcare engineering*, 2019, 2019. 2

[16] Mohammad Farukh Hashmi, Satyarth Katiyar, Avinash G Keskar, Neeraj Dhanraj Bokde, and Zong Woo Geem. Efficient pneumonia detection in chest xray images using deep transfer learning. *Diagnostics*, 10(6):417, 2020. 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[18] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pages 1251–1258, 2017. 2

[19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 2

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 2

[21] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018. 2

[22] Enes Ayan and Halil Murat Ünver. Diagnosis of pneumonia from chest x-ray images using deep learning. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pages 1–5. Ieee, 2019. 2

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[24] Hojjat Salehinejad, Shahrokh Valaee, Tim Dowdell, Errol Colak, and Joseph Barfett. Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 990–994. IEEE, 2018. 2

[25] Abdul Waheed, Muskan Goyal, Deepak Gupta, Ashish Khanna, Fadi Al-Turjman, and Placido Rogerio Pinheiro. CovidGAN: Data Augmentation Using Auxiliary Classifier GAN for Improved Covid-19 Detection. *IEEE Access*, 8:91916–91923, 2020. 2

[26] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, pages 1–15, 2015. 3

[27] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009. 3

[28] José Vilar, Maria Luisa Domingo, Cristina Soto, and Jonathan Cogollos. Radiology of Bacterial Pneumonia. *European Journal of Radiology*, 51(2):102–113, 2004. 4

[29] Hyun Jung Koo, Soyeoun Lim, Jooae Choe, Sang Ho Choi, Heungsup Sung, and Kyung Hyun Do. Radiographic and CT features of viral pneumonia. *Radiographics*, 38(3):719–739, 2018. 4

6. Supplementary Materials

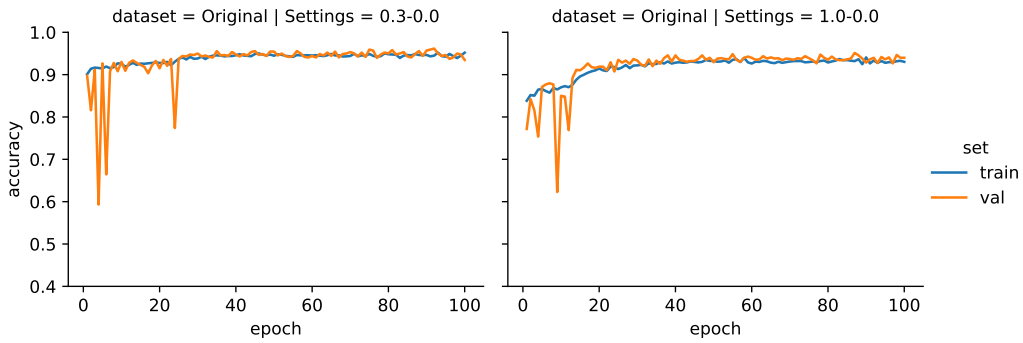

Figure 7: Accuracies of original datasets during training

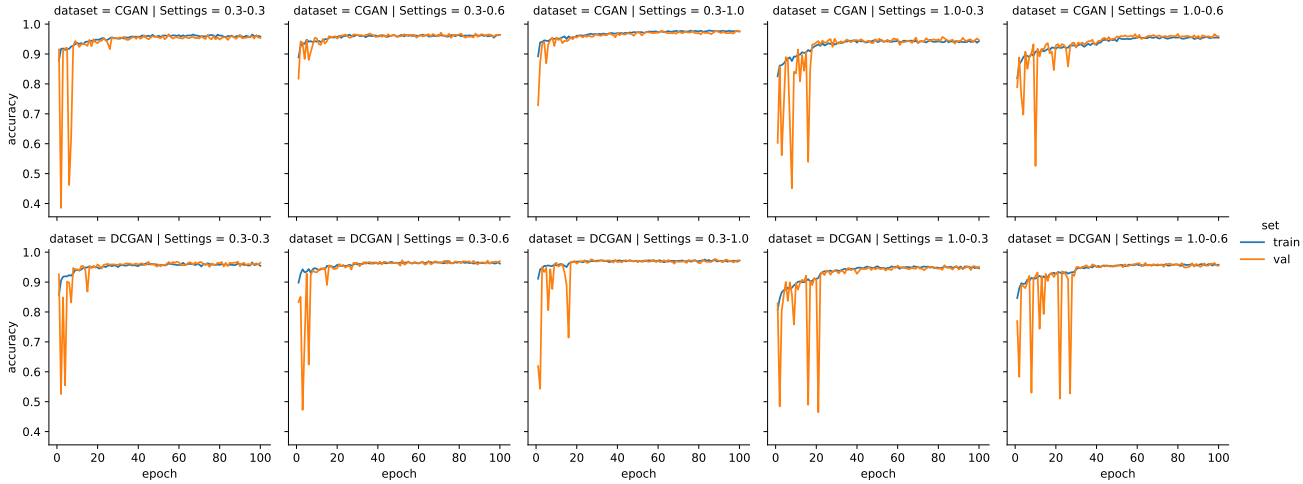

Figure 8: Accuracies of augmented datasets during training

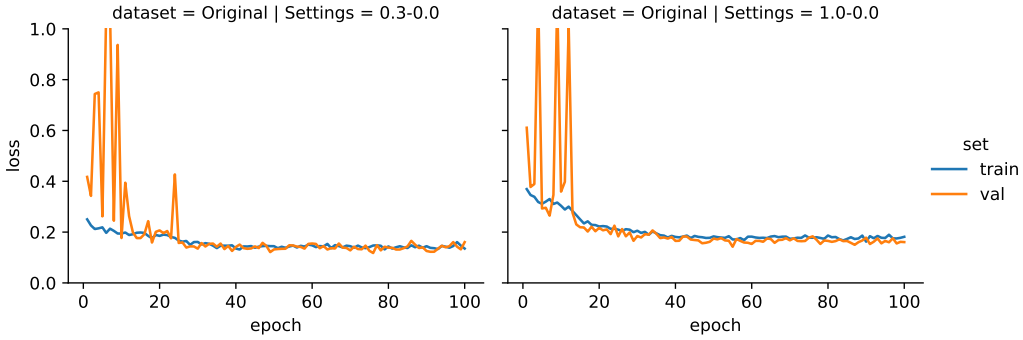

Figure 9: Loss values of original datasets during training

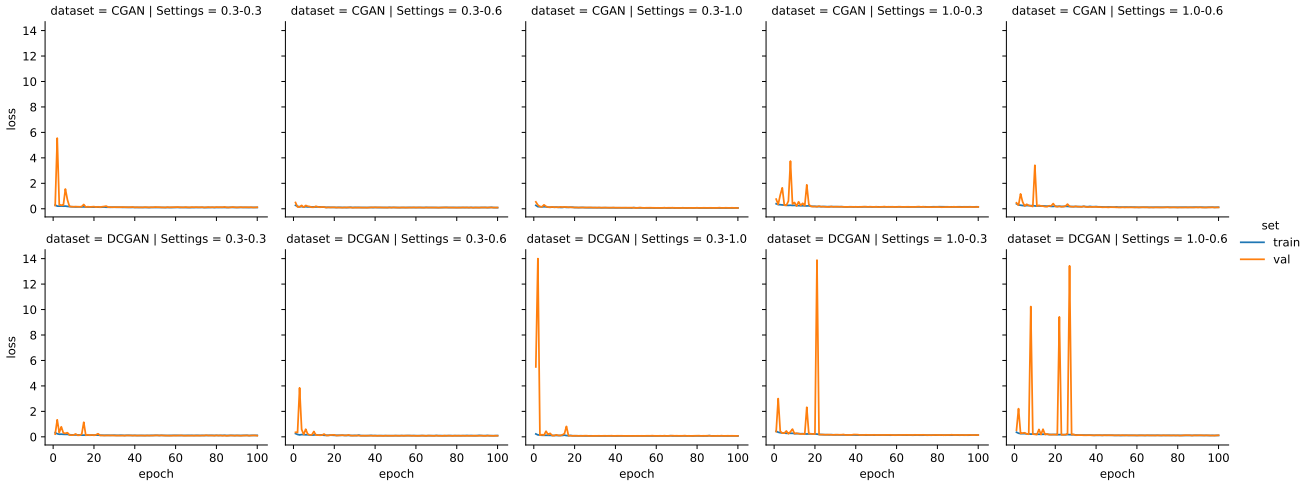

Figure 10: Loss values of augmented datasets during training

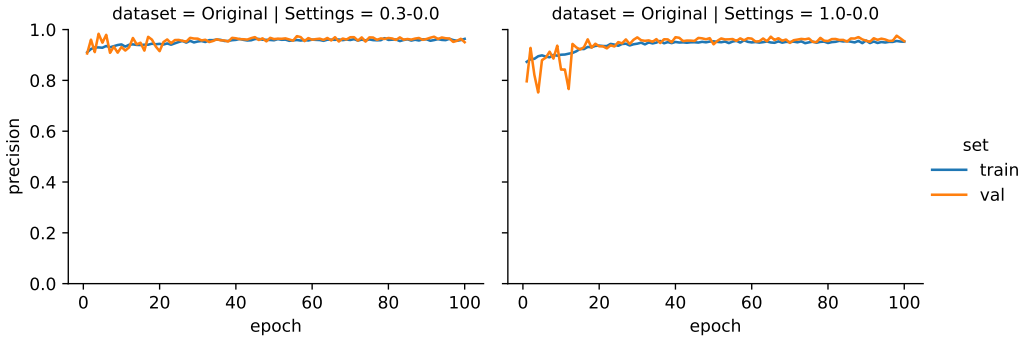

Figure 11: Precision rates of original datasets during training

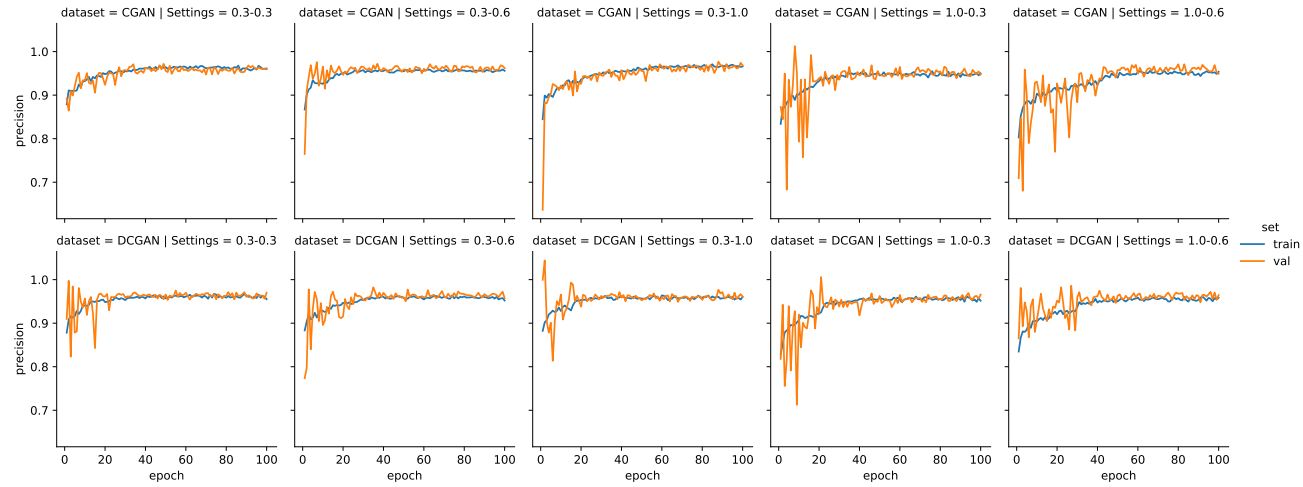

Figure 12: Precision rates of augmented datasets during training