International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Improving AI Model Performance by Augmenting Synthetic Data

Monojit Banerjee

Lead, AI Platform, Independent Researcher

Abstract

In recent years, supervised learning has improved many computer vision problems. However, data scarcity, lack of labeled data, and imbalanced datasets have created issues in adopting this improvement in the medical imaging domain. With the recent advancement in other large language and vision language models(eg: chatgpt, DALL-E) generating synthetic data has become easier. However, this is still cost-prohibitive for large-scale datasets specifically image dataset generation. This approach can also may not be suitable for privacy-first datasets. In this work, the proposed methodology is to generate synthetic images based on available labeled images and then use these generated images along with the existing data to solve above mentioned issues. Chest X-ray datasets are one of the complex datasets that suffer from label imbalance problems and strict data privacy is required for handling any such kind of data. In this work, a simplified generative adversarial network-based solution is used which is cost-effective and provides better results than only using available datasets. This proposed method is especially useful for privacy-first, imbalanced datasets. Finally, this solution was compared with some existing proposals. The promising result obtained using this methodology shows that this proposed solution can be expanded to other domains.

Keywords: Artificial Intelligence, Machine Learning, Synthetic Data, Model training, GAN, MLOps

Introduction

Recent improvement in deep learning provides state-of-the-art results in image classifications. However, not all of these recent improvements can be applied to medical imaging due to various reasons such as lack of data, insufficient expert labeled annotations, and other privacy concerns. However, a relatively new unsupervised classification learning technique called Generative Adversarial Networks(GAN)[1] shows promising results in synthetic image generation.

As mentioned earlier, data is scarce in medical imaging research and most of the time it can't be shared due to privacy reasons and limited to only selected research group[2]. Apart from that, due to its default nature, the medical image system captures mostly non-positive disease data, and hence available data sets are often imbalanced[3].

In this project, I have used a modified GAN-based architecture named Deep Convolutional Generative Adversarial Networks(DCGAN)[4] to generate synthetic images. After that, those images along with available labeled images are used to train a deep-learning image classification model. For the classification part, a state-of-the-art image classification model for image recognition ResNet (and its various enhancements such as ResNet34, ResNet50, and ResNet101)[5] was used.

As I have used GAN to generate images, a single class 'Pneumonia' with only 1431 labeled images was

chosen to experiment with my approach can realistically solve the data scarcity and imbalanced dataset problems. Whereas, compared to this class, more than 60000 images were present in the chest x-ray dataset for the 'normal' or 'non-pneumonia' class.

Related Works

X-rays are the most frequently used form of medical imaging. This is one of the oldest methods of medical imaging. However, the scarcity of annotated X-ray data prevented any meaningful advancement in medical imaging with deep learning. However with the introduction of the Chest X-ray dataset (Wang et al[6] introduced the Chest X-ray dataset in 2017), this limitation is slightly reduced. This data set contains over 110K images from more than 30,000 patients which can be used for deep learning techniques.

Wang et al[6] propose a deep convolutional network model using this data set. Various convolutional network models are also used on this data set to train the models. These models are then used to detect pathologies. The authors also provide a bounding box for a subset of these data(200 instances for each pathology) which can be used as ground truth data. Moreover, weakly supervised localization is also proposed using weighted maps based on the weight of the prediction layers which localizes the active X-ray areas for various pathologies.

Yao et al.[7] use an LSTM(Long short-term memory) model to leverage statistical interdependence on the target labels. The proposed method uses a densely connected convolutional network(DenseNet) for the encoder and an LSTM decoder which exploits the interdependence on the target labels and produces an average AUC score better than the method mentioned by Wang et al.

Rajpurkar et al.[8] use a convolutional neural network on this data set and achieve better results. This model is a 121-layer Dense Convolutional Network(DenseNet). This network was pre-trained on the ImageNet[9] data set and then it was trained end to end on the chest X-ray data set using Adam. This model achieves an F1 score of 0.435 whereas the average radiologist F1 score is 0.387. Moreover, this model is then used to detect 14 pathologies. The AUROC score for all these 14 pathologies was in the 0.73 to 0.94 range which was more than 0.05 over previous results.

This project defines the problem differently than the other related projects such as ChexNet[8] where the problem was treated as a multi-class classification with 14 different pathogens.

All of these methods exploit the available data set and rely on the curated labels. Since this type of curated data is not readily available for most medical imaging tasks, other methods are also proposed.

Salehinejad et al.[10] propose a deep convolutional generative adversarial network (DCGAN) to overcome these problems. Synthetic images are used to counter the imbalanced data. A GAN network is used to generate the synthetic images which are then fed to a deep convolutional neural network(DCNN) to obtain the classification result.

Madani et al.[11] utilize this concept and use GAN to create a semi-supervised learning architecture. This paper claims that using GAN to generate data has two benefits. First, it avoids the problem of data scarcity and second, it avoids data domain over-fitting.

Data

I have used the Chest X-ray data provided by NIH[6]. This dataset has 112110 images with 14 pathological labels and one "no finding" label. Each image has a patient ID and follow-up number. The 14 diseases in this dataset are shown in the Figure 2.

Descriptive statistics for the data set are shown in the below table.

Table 1: Descriptive statistics for the data

Statistics	Follow-up	Patient ID	Patient Age
Count	112110	112110	112110
Mean	8.573751	14346.381743	46.901463
Std. Dev.	15.406320	8403.876972	16.839923
Min	0.000000	1.000000	1.000000
Max	183.000000	30805.000000	414.000000
25 percentile	0.000000	7310.750000	35.000000
50 percentile	3.000000	13993.000000	49.000000

Exploratory data analysis shows that the majority of the labels are of ”no-findings”. Also, the majority of the data is for Male patients except for Hernia.

Experimental Setup

For this experiment, Apache Spark(version 2.3.0) for the ETL pipeline is used and Pyspark is used to load the data and perform exploratory data analysis. For the GAN and DCGAN part, pytorch library is used and various other python libraries (such as matplotlib, torchvision, etc.) were also used. Fast Ai library was used to create the classification portion of the experiment.

For Hardware, a machine with 16G RAM, a single 8G Nvidia 3060 GPU, and 8 core CPU was used. In the experiment, image batch size was varied, and for the majority of the experiments, batch size = 96 is selected so that optimal speed is achieved and memory of the GPU can be utilized to its full extent.

Method

A two-step approach is used here - first, use DCGAN to generate the artificial images, and second, use the generated images as training data. Initially, the data is cleaned and split as training validation and test sets(80:10:10 ratio). The test data was never used in any steps-for example, neither in image generation nor in model selection during

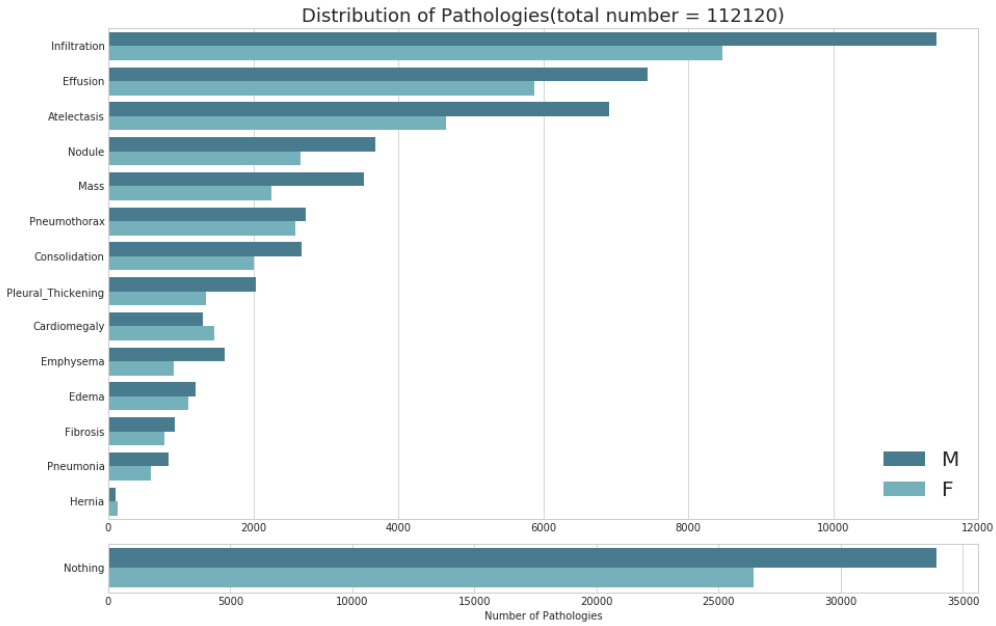

Figure 1: Explanatory data analysis.

second phase. This test data was used at the final step to measure the ROC score after the best classification model was selected based on the validation accuracy.

In the first step, DCGAN was used to generate artificial images from the labeled pneumonia images. Since there are 5X more numbers of non-pneumonia images than pneumonia images, we only generate synthetic images for pneumonia. Once DCGAN training was completed and synthetic images were produced, these were also labeled as pneumonia class. In the second step, the combination of original pneumonia and generated images labeled along with normal images was used to train the classification model. Various ResNet models such as ResNet34, ResNet50, and ResNet101 model were used to experiment and it has been found that ResNet34 provides the best performance and accuracy among these 3 models. Finally, to recap, 2000 images were generated from the initial labeled training pneumonia images, and then these were added back into the pneumonia training class.

For the DCGAN part, the network architecture is relatively simple. It consists of a generator and a discriminator network. For the generator network, A 100-dimensional vector is projected onto the 1024 feature maps. After that 4 fractional-strided(stride=2) convolutional layer was used. ReLU activation was used in all layers in the generator except for the output where the tanh activation function is used. For the discriminator network, the network layer is the same but is in the opposite order from the generator. Also, instead of using ReLU, leaky ReLU activation is used and Batchnorm is used in both networks to minimize the over-fitting. With any type of GAN, we generally face a problem called mode collapse(generator collapses and in turn produces limited varieties of objects) which is minimized here by using the above

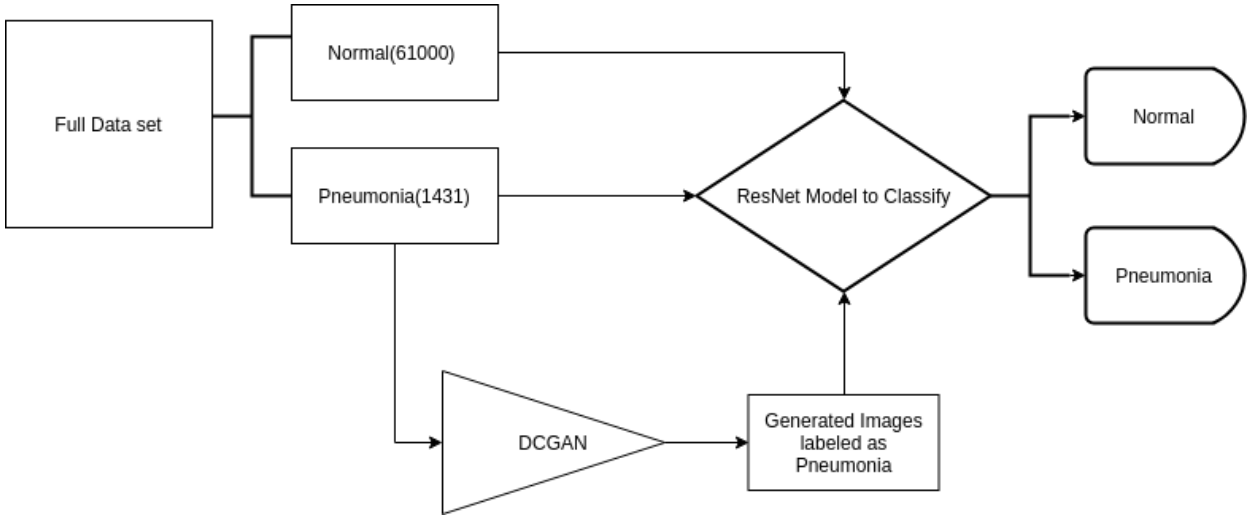

Figure 2: Proposed methodology

techniques. I have used 1000 epochs in DCGAN to train the network with a learning rate=0.0002 and adam optimizer with beta=0.5. The original images were down-sampled to 64x64 and a batch size= 400 was used. This model outputs 64x64 synthetic images.

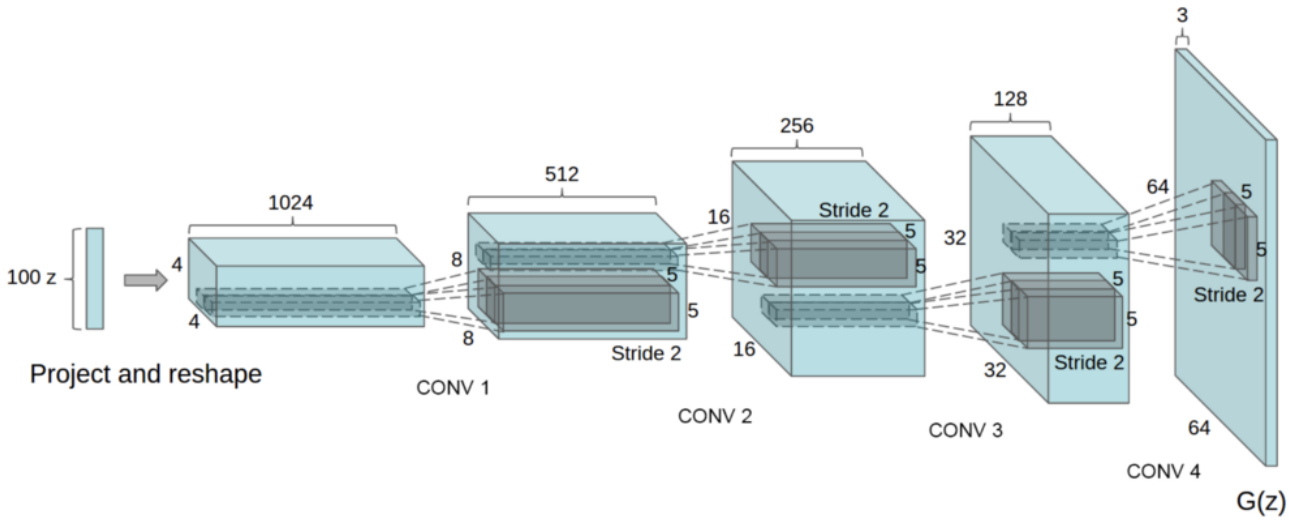

Figure 3: DCGAN Generator Architecture-figure is from the original DCGAN paper

For the classification model part, a standard ResNet architecture is used. As mentioned earlier, after using ResNet34, ResNet50, and ResNet101 architecture, I have selected ResNet34 as my final model due to its speed and performance compared to the other two. Different learning rates(ranges from 0.00001 to 0.001) were also used to fine-tune the model. A method called one-cycle policy is used here to fit the trained model which takes care of both regularization and fast training[12].

Experimental Results and Discussion

During DCGAN training, one needs to be careful to avoid mode collapse. From the generator and discriminator loss graph shown here, we don't see a large difference between these two losses, indicating a well-trained model.

From the loss curve, we see the model is slightly over-fitting. This is somewhat expected as the images were not normalized before it was used to generate images.

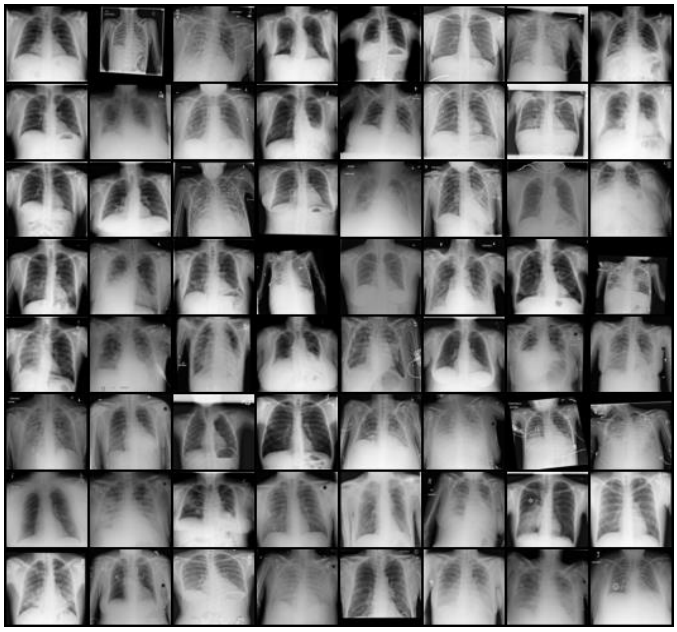

Figure 4: Real Samples

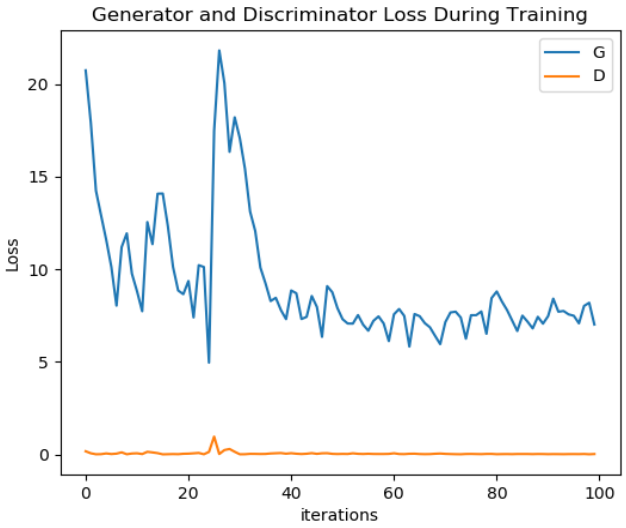

Figure 5: Generator and Discriminator loss

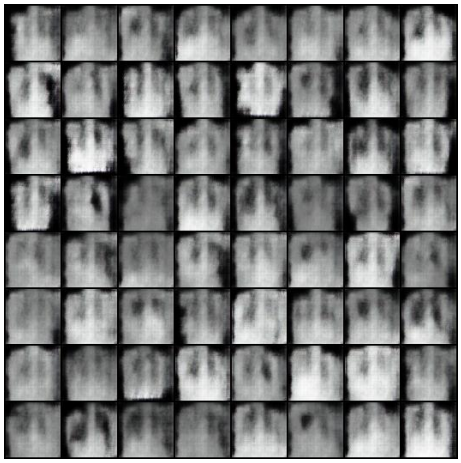

Figure 6: Generated -30 Epochs

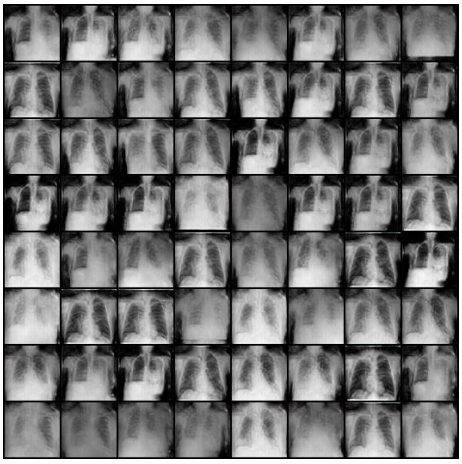

Figure 7: Generated 1000 Epochs

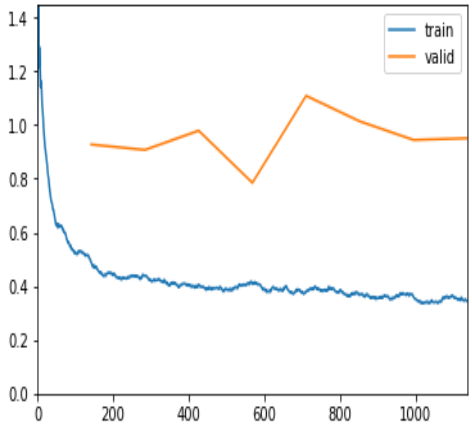

Figure 8: Proposed model Loss curve

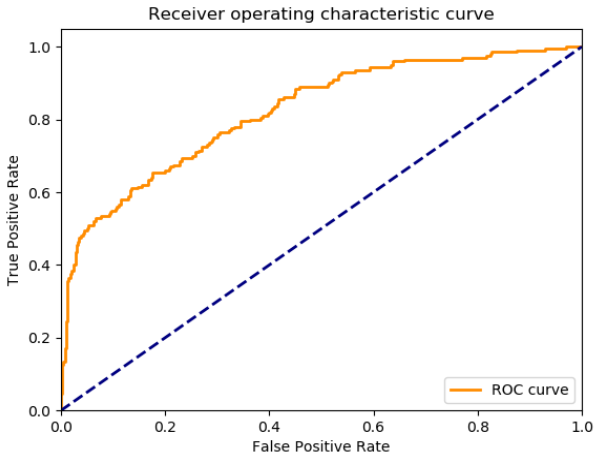

Figure 9: Proposed model ROC Curve

Moreover, the intensity of the raw images(both normal and pneumonia) is different than the generated images. This might contribute to the model over-fit as the learner might classify both raw pneumonia and normal images in the same class as the intensity and size are similar for these than the generated images. Also, the size of generated images is only 64x64 and hence it might be possible that down-sampling of

the raw input causes some feature data loss, and as a result, the learner is not able to fit the model as well as it could have if full size generated images were used. To compare the effectiveness of this proposed method, a separate model was also created with original images, and no generated images were used. This model(will be referred to as the Non-GAN or NG model) is trained similarly with a ResNet34 classifier and hyperparameters were also tuned similarly to what was done for the proposed model as explained earlier. From the validation curve of the NG model, it can be seen that this model also suffers from overfitting more so than the proposed model. Finally, from the ROC curve for the NG model, we can observe that the proposed model does a better job of classifying the pathologies.

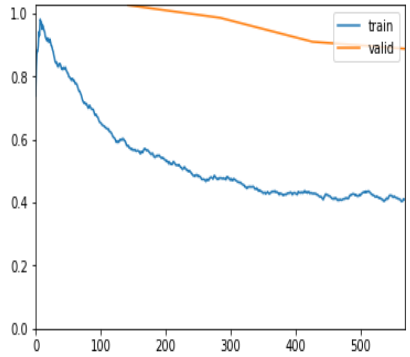

Figure 10: NG Model loss plot

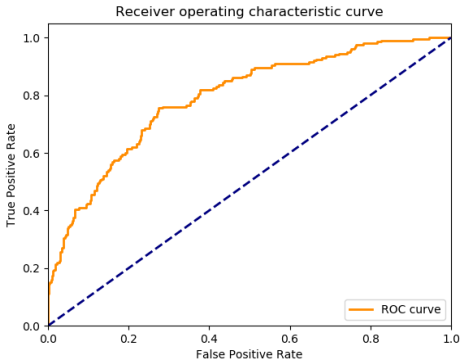

Figure 11: NG Model ROC curve

The proposed model was also compared with some of the existing models in the literature and it's observed that the model mentioned here has a significantly higher AUC score than others(Table 2). Accuracy for the proposed model is 0.79 whereas the NG model is 0.62.

Table 2: ROC comparison of the proposed method with other methods

Pathology	Wang et al. (2017)	Yao et al. (2017)	CheXNet (2017)	Proposed
Pneumonia	0.633	0.713	0.7680	0.8599

Table 3: ROC score-Model with generated and non-generated data

Pathology	NG Model	Proposed
Pneumonia	0.7925	0.8599

Conclusion

In this paper, a two-stage process to classify pneumonia with a limited number of available labeled images is proposed. This proposed model performs relatively well and achieves a better ROC score than other methods in the literature. Recent breakthroughs in deep learning techniques in various sectors especially in image applications provide us a unique opportunity to use these techniques in different domains such as medicine and health care. Using Generative Adversarial Networks to produce artificial images can solve many problems that arise due to the lack of labeled data in the healthcare domain. However, before using this technique on a large scale, we need to perform a sanity check with the help of a trained medical practitioner to confirm that the generated images are capturing the data and not adding random noise. Also, in this project, I have hypothesized that the result can be improved by changing the output probability threshold. Further work is needed to explore these topics.

References

1. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.
2. Elisa Bertino, Beng Chin Ooi, Yanjiang Yang, and Robert H. Deng. Privacy and ownership preserving of outsourced medical data. In *Proceedings of the 21st International Conference on Data Engineering*, ICDE ’05, pages 521–532, Washington, DC, USA, 2005. IEEE Computer Society.
3. Der-Chiang Li, Chiao-Wen Liu, and Susan C. Hu. A learning method for the class imbalance problem with medical data sets. *Comput. Biol. Med.*, 40(5):509–518, May 2010.
4. Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.
5. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
6. Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017.
7. Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman. Learning to diagnose from scratch by exploiting dependencies among labels. *CoRR*, abs/1710.10501, 2017.
8. Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P. Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *CoRR*, abs/1711.05225, 2017.
9. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
10. 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. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*

(ICASSP), pages 990–994, 2018.

11. A. Madani, M. Moradi, A. Karargyris, and T. Syeda-Mahmood. Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1038–1042, April 2018.
12. Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of residual networks using large learning rates. *CoRR*, abs/1708.07120, 2017.