Medical Image Synthesis using GANs for Chest X-rays

Dustin Pulver (20106164) , Matthew Boertjes (20101925), Ella Duffy (20109551)
Department of Electrical and Computer Engineering
Queen's University
December 16, 2022
ELEC825

Abstract

In this paper the implementation of Generative Adversarial Networks (GANs), and their use in the medical science field, is explored. Pediatric pneumonia detection would greatly benefit from artificial intelligence, as it would provide a ”second set of eyes” in finding irregularities. Chest X-ray radiograph datasets available have limitations; they are small and imbalanced, which poses an issue as typical classification methods implicitly assume equal distribution of classes. A simple Convolutional Neural Network (CNN), for classification purposes, on the original dataset achieved a test accuracy of 75.80% and a test loss of 4.0339. To better improve machine learning performance, 2,690 synthetic normal chest X-ray images are generated using DCGAN, LSGAN and WGAN-GP. Optimal models were established after running a variation of parameter tuning experiments on Google Colab Pro. To quantify the performance of each GAN model, the Fréchet inception distance (FID) of generator outputs and accuracy score from the simple CNN classifier using generated data were used. High-quality, synthetic chest X-ray images were produced from each model, with LSGAN outperforming the others. LSGAN achieved an FID score of 14.7034, and was able to increase the test accuracy of the classifier by 2.73%. Comparison between project outcomes and published results is explored as well as what could be done as next steps.

1 Introduction

1.1 Motivation

In the field of medical science, it is imperative that diagnosis be as accurate as possible since conditions left undiagnosed can progress beyond modern treatment. As such, medical diagnosis currently requires experts in the field with a vast amount of experience. The field is limited by availability of experts, availability of tools, and cost. These factors could be alleviated through deep learning techniques, which would greatly benefit the medical diagnosis field. Deep learning has the capability to detect diseases in medical images and enable quick and cost-efficient methods for diagnosing a patient. However, models such as convolution neural networks (CNNs) have not achieved suitable performance for clinical use largely due to training data limitations. These models tend to overfit because medical imaging datasets are not sufficiently large enough. There are many contributing reasons as to why medical imaging datasets are limited. Firstly, medical image data is costly to generate and time consuming to label. An expert is required to go through each sample and label the pathology present, or indicate if none are present. Secondly, datasets are often imbalanced towards the case where a disease is present. Lastly, the available datasets tend to be smaller because patient privacy must be preserved, thus access to large datasets are limited to research groups [1].

Pneumonia is a very common infection which can affect individuals healthy or not, but potentially becomes life-threatening for infants, those with other diseases, and the elderly. Young children exhibit first signs and symptoms at vastly different times depending on the type of pneumonia. In some cases pneumonia can appear suddenly and quickly cause severe effects to an individual if left unrecognized for long enough. Computer based detection approaches are important tools that can accompany experts to provide a faster and efficient diagnosis process. This will lead to a reduction in child mortality rates from pneumonia, especially in developing countries. [2]

Data augmentation is a common technique that is used in deep learning to artificially increase dataset size by applying transformations, rotations, intensity changes, and others on existing data. However, data augmentation has shown to be ineffective when applied to medical images and in particular to X-ray images [3]. This is because object location for normal computer vision problems usually does not correlate to a class label, thus applying a transformation can increase robustness. However, in X-ray images the location of an object is important to the structure of the body part being examined and what else may be present, thus possible data augmentation techniques are highly limited.

An alternative method to overcome data size limitations and fix class imbalances is the use of Generative Adversarial Networks (GANs) for augmenting data. This paper proposes a method to generate synthetic X-ray imaging data of children aged one to five, which can assist deep learning classification algorithms in detecting the presence of pneumonia.

1.2 Project Goal

The project goal is to implement GAN architectures to tackle the issues related to data volume, balance, and privacy by generating new medical samples from a smaller subset of readily available samples. Since large amounts of data are not always easy to obtain, especially in the medical field, the dataset is limited to a small, unbalanced dataset. This dataset includes 5,856 X-Ray images belonging to two categories, pneumonia and normal. To achieve equilibrium between the classes, the generator from GAN models are used to produce 2,690 additional normal images with the intent that deep learning for disease diagnosis is improved and made more feasible for clinical use.

2 Related Work

2.1 GANs In The Medical Field

GANs have been used for medical image generation to increase the volume of images used for classification tasks. Y. Yang *et al.* [4] used GAN-based image generation through Health-IoT platforms to produce medical images. GAN-based systems continue to support real time diagnosis by helping facilitate the expanding industry of Health-IoT. Working with GANs in the medical imaging space supports the growth of datasets by increasing the volume of images and compensating for class imbalance present in many medical imaging applications. C. Han *et al.* [5] used GAN-based synthetic image generation to create brain magnetic resonance (MR) imaging. The synthetic MR images were created to improve the performance of predictive models using such data. Other applications in medical fields include P. Sedigh *et al.* [6] use of GAN-based image generation to improve CNN performance for skin cancer classification.

2.2 Chest X-Ray Classification

CNNs have been applied to pneumonia classification in chest X-rays with varying degrees of success. Transfer learning has been used to perform classification of pneumonia in chest X-rays, in which five pretrained models on ImageNet are used in an ensemble [7]. A few learnable weight layers are appended to the pretrained model which has its weights frozen. Then the weights of the appended layers are learned based on a smaller dataset of chest X-ray images, helping to overcome the need for a large dataset [7]. This method allowed the model to achieve excellent accuracy, outperforming many models at the time. Another approach for pneumonia classification proposes GNet which uses graph knowledge in feature reconstruction [8]. This method was able to increase the performance of simple neural network classifiers however, both methods discussed in this section are limited by small datasets and thus do not provide sufficient results for clinical use.

2.3 GANs for Chest X-Ray

GANs offer synthetic data augmentation, which is a promising area of research, especially in the medical domain. S. Sundaram *et al.* [9] demonstrated the efficacy of GAN-based data augmentation, when compared to standard data augmentation and no augmentation techniques in correcting class imbalances found in the Stanford CheXpert dataset. 14 pathologies are highlighted in each image (pneumonia, lung leisure, cardiomegaly, etc.) – a label of 1 corresponds to positive, 0 to negative, and -1 to uncertain. In this work, a label smoothing technique is used and allows for uncertain labels to be mapped to positive labels. To quantitatively assess the data augmentation techniques, a DenseNet-121 classification model is employed, and the ROC-AUC score is examined. This evaluation metric is the area under the curve of the true positive rate versus the false positive rate at different classification rates [10]. For GAN-based augmentation, the work uses a Conditional GAN with mirrored structures for both the generator and discriminator. Conditional GANs take advantage of labels during the training process, a novel idea that conditions the networks. The authors observed that the GAN augmentation caused a 0.03 AUC gain from the no augmentation for lung lesions and pleural other, and a 0.07 AUC gain for fracture. Using standard augmentation was an improvement from no augmentation, but did not show the same impressive results as GAN augmentation. The AUC performance gains prove that GAN-based data augmentation is an effective tool for correcting class-imbalanced medical datasets, specifically for chest X-ray images. Another work outlined in [11] also uses GAN architectures for augmentation purposes on chest X-ray data. An Inception-Augmentation GAN is proposed, which is different from the Conditional GAN discussed above. The generator for this model implements attention layers to capture long-range dependencies in the image. Also, using inception and residual components increases the GAN's ability to capture more details from training image space without losing spatial information. This intricate generator is evaluated with a simple discriminator made up of four CNN layers. Although a different approach was taken in this model, an AUC performance gain was also observed when GAN techniques were implemented.

3 Experiments

3.1 Dataset

This work made use of the dataset Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification [12]. This dataset is used in the training process of multiple GANs explored throughout this paper. The dataset includes 5,856 RGB X-ray images belonging to two categories, pneumonia and normal. There exists an underlying imbalance in the dataset, as more patients will seek medical assistance when they show symptoms of pneumonia. Therefore, in the dataset there are 4,273 X-rays from the pneumonia class and 1,583 from the normal class. These X-ray images were collected from pediatric patients between the ages of one to five. Expert physicians labeled the X-rays as either normal or pneumonia and removed low quality scans. The presented dataset was used by D. Kermany *et al.* [13] to test the generalizability of their model trained to predict the presence of treatable blinding retinal diseases.

3.2 Compute Resources

Experiments were conducted on the Google Colaboratory platform, which provides users access to cloud-based GPUs. This provided access to Tesla T4 or P100 GPU's for the project, accommodating effective training and testing of the GANs explained in Section 4. The singular limitation experienced with this platform was its inability to provide enough storage and resources to process higher resolution images.

3.3 Comparison Metrics

To quantify model performance of each GAN implementation, the FID of generator outputs and the accuracy score from a CNN classifier using the generated data is used.

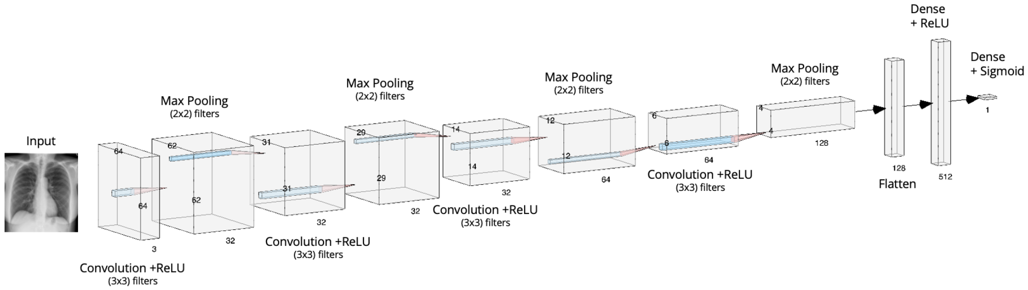


Figure 1: Classification Model Architecture

3.3.1 FID Score

Equation 1 was used to compute the FID score between real X-ray images and generated fake images from each model. The mean, μ, and variance, Σ, are the multivariate normal distribution estimated by Inception v3 where the subscript r refers to real and subscript g refers to generated images [14]. FID summarizes how similar two sets of images are, with lower scores indicating higher similarity. By using this metric, the quality of generated fake images can be assessed as it reveals how close to real images they are. FID is found to be consistent with human judgment in assessing quality of images and is more robust then metrics which predate it, such as inception score [15].

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

3.3.2 Classification Model

Secondly, the predictive accuracy of a simple CNN, that uses generated fake images to balance the training dataset, will be compared for all models. The purpose of this metric is to determine the change in accuracy that including fake generated images has on a simple classifier. Accuracy will be compared from the baseline classifier, which uses the original dataset as input, to the classifiers being run on the GAN-balanced datasets. The architecture of the simple CNN used in this project can be found in Figure 1. Taking in a three dimensional image, the input is passed through a series of convolutional layers and max pooling layers. Finishing off with two fully-connected layers, the model is compiled using a $sigmoid$ activation, the Adam optimizer and binary cross entropy (BCE) loss. The original dataset, with no GAN-based augmentation techniques, achieved a test accuracy of 75.80% and a test loss of 4.0339.

4 Methodology

This section presents multiple GAN architectures explored in this paper and determines the best candidate to solve the problem at hand. The model tuning process explained in Section 4.2 outlines the optimized models explored here. This work was implemented $Python$ 3.9.12 in a $Google$ $Colaboratory$ notebook using the $PyTorch$ package.

4.1 GAN Models

GANs were designed in 2014 by Ian Goodfellow and his colleagues - they are considered to be some of the most interesting machine learning discoveries in recent years. The proposed framework for estimating generative models via an adversarial process, trains two models simultaneously – a generator and a discriminator. The role of the generative model, G, is to capture the data distribution, and a discriminative model, D, estimates the probability that a sample came from the training data rather than G. GANs are often referred to as a minimax two-player game in which G is to maximize the probability of D making a mistake[16]. One agent's gain is another agent's loss. In this paper, three architectures are explored: DCGAN, LSGAN and WGAN-GP. Each being variations on the vanilla GAN architecture, with DCGAN being the foundation. This baseline architecture can be found in Figure 2 – the latter models build from this as well.

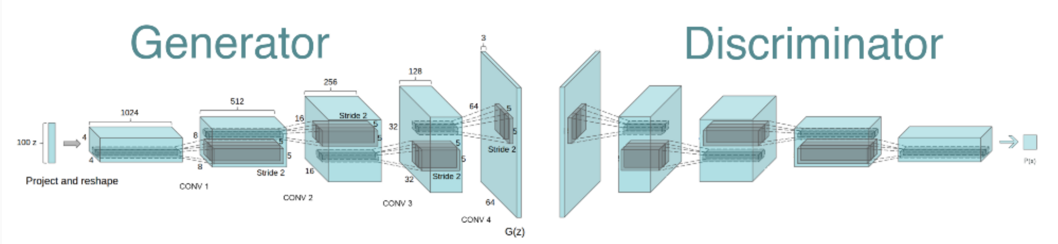


Figure 2: GAN Baseline Architecture - DCGAN [18]

The generator takes in an uniform noise distribution Z, which is reshaped into a 4-dimensional tensor. This input is also referred to as the latent vector, which was set to 100. Next the stack of convolutions start, which all share similar components. BN ensures that all inputs have zero mean and unit variance, which stabilizes the network. This is a critical element of the architecture as it was found that BN prevents the generator from collapsing all samples to a single point. But, sample oscillation was observed by DCGAN creators which was solved by removing BN from the generator output and discriminator input. In all layers of the generator, the $ReLU$ activation is integrated with the exception of the output which uses $Tanh$. The justification is as follows... ”We observed that using a bounded activation allowed the model to learn more quickly to saturate and cover the color space of the training distribution” [17]. The $LeakyReLU$ activation is used in the discriminator instead, because it was found to work well for high resolution modeling. The last important aspect of DCGAN is that inside the convolution nets, it uses strided convolutions instead of pooling layers. That means no fully connected or pooling layers are included in the architecture found in Figure 2. The result of our generator is a 3x64x64 RGB image, which the discriminator will use as input to decide if it is ”Real” or ”Fake.” The loss function of the discriminator is what makes DCGAN, LSGAN and WGAN-GP uniquely different and is discussed in the sections below. To arrive at optimal models for each, parameters were varied and discussed in 4.2.

4.1.1 DCGAN

Deep Convolutional GAN (DCGAN), as seen in Figure 2, is a variation on the vanilla GAN structure with the primary difference being the use of convolutional and transposed convolutional layers in its architecture [17]. The discriminator is equivalent to a standard convolutions neural network with the aim of outputting a scalar probability that the input comes from real data [18]. The generator however learns to convert latent inputs into images with the help of transposed convolutions. A critical distinction of DCGAN is the inclusion of BN layers after transposed convolutional layers in the generator. This enabled a deeper model and helped with gradient propagation, preventing generators from collapsing samples to a single point which is common failure of GANs [17]. DCGAN makes use of the BCE loss function, thus it uses a $sigmoid$ activation layer at the end of the discriminator to normalize outputs between 0 and 1.

4.1.2 LSGAN

As mentioned, the DCGAN discriminator utilizes the $sigmoid$ cross entropy loss function, which leads to the vanishing gradient problem during the learning process. To remedy this problem, the Least Squares GAN (LSGAN) was proposed. This architecture adopts the least squares loss function for the discriminator, which allows for improved generated images and stable performance during the learning process. Since generated samples match the statistics of the real data, the $sigmoid$ activation in the discriminator is also not required for this architecture. The objective functions for LSGANs are highlighted below in equations 2 and 3 [19].

$$min_D V_{LSGAN}(D) = \frac{1}{2}E_{x\sim p_{data}(x)}[(D(x)-b)^2] + \frac{1}{2}E_{z\sim p_z(z)}[(D(G(z))-a)^2] \tag{2}$$

$$min_G V_{LSGAN}(G) = \frac{1}{2}E_{z\sim p_z(z)}[(D(G(z))-c)^2] \tag{3}$$

4.1.3 WGAN-GP

GANs tend to suffer from lack of convergence during training and mode collapse, these issues have been addressed through the creation of the Wasserstein Generative Adversarial Network with gradient penalty (WGAN-GP) [20]. WGAN-GP proposes a new cost function, an alternative to those seen in DCGAN and LSGAN. This cost function uses Wasserstein distance which was found to have a smoother gradient no matter the performance of the generator. WGAN-GP had no sign of mode collapse through the experimentation presented in [20]. In order to use the Wasserstein distance in the loss function, the model's discriminator needs to be 1-Lipschitz continuous. This means the norm of the gradient must maintain a value of at most one. The model introduces a gradient penalty (GP) which enforces this continuity with its addition to the loss function. Thus enforcing a constraint to the loss function to allow for proper implementation of the Wasserstein distance.

4.2 Experiment List

To arrive at the optimal models for DCGAN, LSGAN and WGAN-GP, experiments were conducted and are outlined in Table 1. Each model was trained only with the normal class images to ensure the generated images were from the normal class. For each experiment, the Fréchet inception distance (FID), outlined in Section 3.3, was used to evaluate how the parameter tuning affected the model. As our initial experiment, experiments from the origin papers were followed and changes were made as necessary.

Table 1: Experiments conducted to determine the optimal models

Parameter	Experiment	Finding
Batch Size	16, 32, 64	For the batch size, the rule of thumb for a good initial choice is 32. The paper [21] concluded that there is a significant degradation of models when a larger batch size is used. Due to this, the size of our dataset, and other parameters, it was observed that a batch size of 16 led to optimal results.
Epochs	16, 40, 70, 100	Each model was run on the computer of the individual who developed it. Based off the model and the accessible computation resources (Google Colab Pro or Google Colab), different epochs provided different results. As a rule of thumb, a larger epoch created higher-quality X-ray images. DCGAN used 16, LSGAN used 70, and WGAN-GP used 100.
Image Size	64x64, 128x128	The images from the original dataset are quite large, therefore the initial attempt was for the GANs to process a larger input. Unfortunately, 64x64 images continued to produce better results which is due to the computation resources available.
Learning Rate	0.0002, 0.0001, 0.001	The original papers for our models suggested that a learning rate of 0.001 was too high, thus lower values were explored. It is understood that a high learning rate can skip the optimal solution and a low learning rate can take too long to converge [22]. Using 0.0002 was a happy medium for the models.
β_1	0.5, 1.0	Using 0.5 instead of 1.0 for the momentum term Beta1 resulted in less model oscillation.

5 Results & Analysis

This section describes the results of the explored GAN architectures for synthetic image generation. The proposed system generates minority class synthetic X-ray images with the goal of balancing the class distribution in the presented dataset [11]. Both quantitative and qualitative evaluation were conducted to best demonstrate the main contributions of this paper. The combination of the evaluation techniques provide conflation of visual and numeric data to validate the findings.

Table 2: Results from Hyper-Parameter Tuning, using evaluation metrics to determine the best model

Optimal Model Parameters				Evaluation Metrics	
Name	Batch Size	# of Epochs	Learning Rate	FID	Classifier [test_loss,test_acc]
DCGAN	16	16	0.002	18.0763	[2.1663,0.7788]
LSGAN	**16**	**70**	**0.002**	**14.7034**	**[1.0482,0.7853]**
WGAN-GP	16	100	0.005	112.202	[3.6990,0.7596]

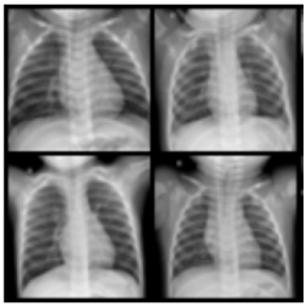


Figure 3: Normal X-Ray Samples

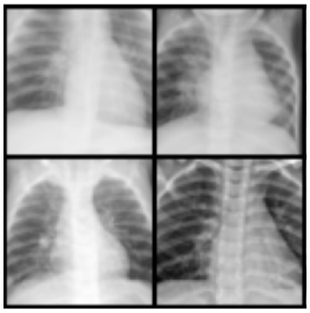


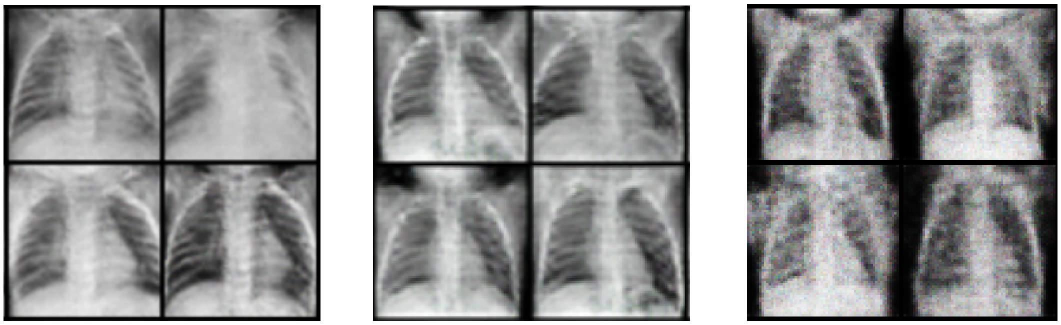
Figure 4: Pneumonia X-Ray Samples

5.1 Quantitative Evaluation

Seen in Table 2 each model is summarized based on their optimal model parameters and elevation metrics described in section 3.3. The model with the best performance, LSGAN, is emboldened in Table 2. LSGAN achieved a FID score of 14.7, proving the proposed architecture can generate normal class X-Ray images which are difficult to distinguish between real and synthetic. When comparing the classifier test accuracy of LSGAN to the baseline test accuracy of 75.80%, an increase of 2.73% in test accuracy has been observed. Thus quantitatively proving the contribution of the proposed solution to improving the performance of medical image classification. The optimal model parameters followed similar trends for all three models. The optimal batch size for all models was found to be 16, in addition the learning rate for both DCGAN and LSGAN was the same. The increased performance observed between DCGAN and LSGAN was an expected trend, as explained in section 4. The performance of WGAN-GP was unexpected, as explained in section 4 , the model makes improvements on others such as DCGAN and LSGAN with respect to training performance. The results show the opposite occurred, with WGAN-GP producing an almost 10 times worse FID score than LSGAN, seen in Table 2. This discrepancy can be further seen in section 5.2 and will be discussed in further detail.

5.2 Qualitative Evaluation

To qualitatively evaluate the quality of images generated by each model architecture, four random samples will be taken from each. Qualitative inspection will be made based upon human judgment for how realistic the samples are. Generated samples should appear similar to normal X-ray samples from the dataset, seen in Figure 3, and dissimilar to X-ray samples with pneumonia, seen in Figure 4. To mimic normal chest X-rays, samples should show clear distinctions between bone structures and empty elements like the chest cavity. Pneumonia samples often show white cloudiness in the chest cavity and faint white discolourations where the disease is present. As such, to distinguish from pneumonia samples features should be of high contrast. There should be minimal grey transitions between pixels which are white and black which could be seen as cloudiness from pneumonia. Together these metrics should distinguish features normal samples from unique features found in pneumonia samples. Thus, causing less confusion between class samples presented to a classification model.



(a) DCGAN Samples (b) LSGAN Samples (c) WGAN-GP Samples

Figure 5: Generated X-Ray Samples From All GAN Architectures

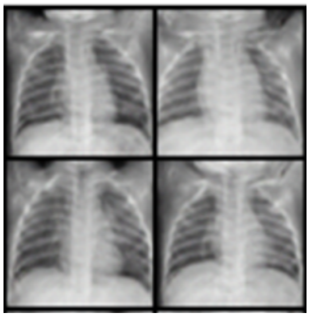


Figure 6: Extended Training Time DCGAN Samples

Figure 5a shows the random outputs from the DCGAN architecture. It can be seen that the output images are fairly realistic. In general, the features of an X-ray image are present and some cases produce very bold differences between the bone structure and the background. However, samples appear to have some cloudiness to them, especially seen in the top right of Figure 5a, with a white hue over top of the entire X-ray leading to a harder distinction of features. Because the presence of pneumonia also causes some cloudiness on top of the chest bones in an X-ray, this could easily cause confusion in the model leading to false positives.

From the quantitative analysis, LSGAN outperformed the two other models so it is expected that it will produce the best results from human judgment. As can be seen in 5b, LSGAN produces very realistic synthetic X-ray images. Outputs show clear contrasts between bone structures and the background with little noise present. Some cases do still exhibit a white hue over top of the entire X-ray however, features are still distinguishable and it is not as severe as in the case of DCGAN.

For the final model architecture, WGAN-GP, generated images can be seen in Figure 5c. WGAN was the worst performing model in the quantitative analysis and it follows that the model produces the worst outputs based on human judgement. As seen in 5c, outputs show WGAN performs poorly in this task. Generated images are very noisy, and show little contrast between background and bone structures. Noise is also manifesting in samples by showing non black and white pixel values that can be faintly seen. Some samples are poor enough that the rib cage is barely visible or the heart is missing. It can be concluded that WGAN synthetic images provide insufficient information to aid a classification algorithm, and therefore won't provide a performance gain.

Of notable interest, while optimizing model parameters, the DCGAN architecture was initialized with a number of training epochs equal to 100. The produced images from this setup are found in Figure 6. From inspection, model outputs appear very similar to real X-Ray images however the FID score of these generated images was higher than the DCGAN architecture at 16 training epochs. This model architecture also produced a classification accuracy which was worse than the performance of the classifier without generated samples. This is unexpected since both metrics aim to asses how realistic synthetic images are. A possible explanation is that extended training

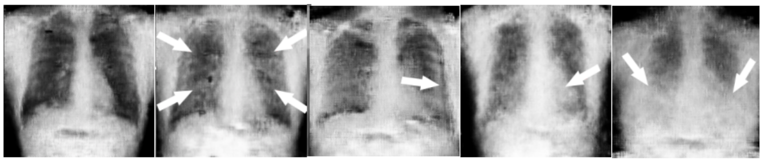


Figure 7: Published Generated X-Ray Images using GANs [19]

time caused the generator to learn features that fool the discriminator but are not shared with real samples, thus overfitting.

5.3 Comparison to Published Results

Quality of generated images has also been assessed by comparing to published results. Figure 7 shows generated X-ray images from a published GAN model with the same objective as this work and also using a DCGAN architecture [23]. This configuration however produces normal X-ray images as well as X-rays with the presence of pneumonia. The left most image in Figure 7 shows the published generated normal X-ray and the rest have pneumonia present. From inspection, the results produced by LSGAN and the extended training time results from DCGAN appear to be of much better resolution than the published results. The published results show a lot of noise and poor feature definition which is similar to what is seen in with the WGAN model. This indicates that the DCGAN and LSGAN architecture can serve as a better method than published results for producing realistic synthetic images to assist pneumonia diagnosis with deep learning. It should be noted, that the published result manages to increase a classifiers accuracy from around 70% to 90% despite having lower resolution images. The method proposed by the published result produces generated images for all classes and doubles the size of the largest class, which is the target to balance to. This indicates that performance can be greatly increased if a combination between both approaches is taken. Generating samples from all classes can increase classification accuracy while using an architecture from this work can produce more visually realistic samples.

6 Conclusion

The paper presents a solution to the issues in the field of medical machine learning such as issues of data volume and class imbalance. In 2019, Pneumonia accounted for 14% of all deaths in children under 5 years old - 740 180 young lives [24]. Using GANs, we can generate synthetic X-ray image data for pediatric patients with the objective of improving classification results. An improved machine learning technique allows for the disease to be detected faster, without direct doctor assessment, resulting in treatment beginning promptly. Three GAN models were explored throughout this project, these being DCGAN, LSGAN and WGAN-GP. LSGAN outperformed the others by improving the classification results from the baseline with 75.80% test accuracy to 78.53%. This model also presented an exceptional FID score of 14.7034, which highlights the quality of the generated images. The other two models also outperformed the classifier on the original dataset when their generated images were added. DCGAN had a test accuracy of 77.88% and WGAN-GP had a test accuracy of 75.96%. From the quantitative and qualitative results, it can be concluded that even simple GAN models have the power to generate high-quality images. Therefore, the future advancements in this area are promising.

6.1 Future Work

The main constraints for the project were knowledge, time and computation resources. Since this was an introductory exploration of GANs, simpler model implementations were feasible. Future efforts should be put towards deploying more complex and intricate models with the goal of obtaining better results. Also, exploring conditional GANs was an are of interest for the team, had there been more time. Since we expected WGAN-GP to outperform the others and it didn't, this is an area to explore. In Section 5.3, it was proposed that images are generated for both classes in the dataset. This concept seemed to perform well, and therefore should be investigated for the application at hand. The main focus of the project was not on X-ray classification, but since it is a comparison metric used, more attention should be devoted to this model. Implementing ResNet or VGG might lead to better assessment of the produced outputs from the GAN models. In Table 1 the issue of trying to implement a larger image size input for the models is articulated. Future work includes exploring this structure modification as it would facilitate more real-world applications. Furthermore, investigating the use of higher resolution images may allow GANs to capture and create more complex features, leading to more realistic synthetic images and therefore be a greater contribution to the field of medical machine learning. Investing in better computation resources would also excel future project success.

6.2 Group Member Contribution

All group members contributed equally to the term project.

Dustin Pulver: Responsible for the implementation of WGAN-GP and the simple CNN classifier. Responsible for writing the following sections: GANs in the Medical Field, Dataset, Compute Resources, WGAN-GP, and Quantitative Evaluation. Contributed to editing the document.

Matthew Boertjes: Responsible for the implementation of DCGAN. Researched various techniques for evaluating GAN performance and implemented the function for calculating the FID score. Responsible for writing the following sections: motivation, chest X-ray related work, DCGAN, comparison metrics, qualitative analysis and comparison to published results. Contributed to editing the document.

Ella Duffy: Responsible for the implementation of LSGAN and exploring the 128x128 image input GAN. Responsible for writing the following sections: abstract, project goal, GANs for Chest X-Ray, experiment list, classification model, GAN models, LSGAN, conclusion, and future work. Contributed to editing the document.

References

[1] E. Bertino, B. C. Ooi, Y. Yang, and R. H. Deng, “Privacy and ownership preserving of Outsourced Medical Data,” 21st International Conference on Data Engineering (ICDE’05), Apr. 2005.
2

[2] J. A. Scott, W. A. Brooks, J. S. M. Peiris, D. Holtzman, and E. K. Mulhollan, “Pneumonia research to reduce childhood mortality in the developing world,” Journal of Clinical Investigation, vol. 118, no. 4, pp. 1291–1300, 2008.

[3] M. Elgendi, M. U. Nasir, Q. Tang, D. Smith, J.-P. Grenier, C. Batte, B. Spieler, W. D. Leslie, C. Menon, R. R. Fletcher, N. Howard, R. Ward, W. Parker, and S. Nicolaou, “The effectiveness of image augmentation in deep learning networks for detecting COVID-19: A geometric transformation perspective,” Frontiers in Medicine, vol. 8, 2021.

[4] Y. Yang et al., ”GAN-Based Semi-Supervised Learning Approach for Clinical Decision Support in Health-IoT Platform,” in IEEE Access, vol. 7, pp. 8048-8057, 2019.

[5] C. Han et al., ”GAN-based synthetic brain MR image generation,” 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), 2018.

[6] P. Sedigh, R. Sadeghian and M. T. Masouleh, ”Generating Synthetic Medical Images by Using GAN to Improve CNN Performance in Skin Cancer Classification,” 2019 7th International Conference on Robotics and Mechatronics (ICRoM), 2019.

[7] V. Chouhan, S. K. Singh, A. Khamparia, D. Gupta, P. Tiwari, C. Moreira, R. Damaševičius, and V. H. de Albuquerque, “A novel transfer learning based approach for pneumonia detection in chest X-ray images,” Applied Sciences, vol. 10, no. 2, p. 559, 2020.

[8] X. Yu, S.-H. Wang, and Y.-D. Zhang, “CGNet: A graph-knowledge embedded convolutional neural network for detection of pneumonia,” Information Processing & Management, vol. 58, no. 1, p. 102411, 2021.

[9] S. Sundaram and N. Hulkund, ”GAN-based Data Augmentation for Chest X-ray Classification,” 7 July 2021. [Online]. Available: https://arxiv.org/pdf/2107.02970.pdf.

[10] Google Developers, ”Classification: ROC Curve and AUC,” Machie Learning Crash Course, 18 July 2022. [Online]. Available: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

[11] 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,” 22 November 2021. [Online].

[12] Kermany, Daniel; Zhang, Kang; Goldbaum, Michael (2018), “Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification”, Mendeley Data, V2.

[13] Kermany, Daniel, K. Zhang, and M. Goldbaum, “Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification,” data.mendeley.com, vol. 2, Jun. 2018.

[14] J. Brownlee, “How to implement the Frechet Inception Distance (FID) for evaluating Gans,” MachineLearningMastery.com, 10-Oct-2019. [Online]. Available: https://machinelearningmastery.com/how-to-implement-the-frechet-inception-distance-fid-from-scratch/.

[15] A. Borji, “Pros and cons of gan evaluation measures,” Computer Vision and Image Understanding, vol. 179, pp. 41–65, 2019.

[16] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, ”Generative Adversarial Networks,” arXiv, 2014.

[17] A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” arXiv, Nov. 2015.

[18] N. Inkawhich, “DCGAN Tutorial,” PyTorch, 2018. [Online]. Available: https://pytorch.org/tutorials/beginner/dcgan_faces_tutorial.html.

[19] X. Mao, Q. Li, H. Xie, R. Lau, Z. Wang and S. P. Smolley, ”Least Squares Generative Adversarial Networks,” arXiv, 2016.

[20] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, “Improved Training of Wasserstein GANs,” arXiv:1704.00028 [cs, stat], Dec. 2017.

[21] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy and P. T. P. Tang, ”On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima,” arXiv, 2016.

[22] A. Rakhecha, ”Understanding Learning Rate,” Towards Data Science, 28 June 2019. [Online]. Available: https://towardsdatascience.com/https-medium-com-dashingaditya-rakhecha-understanding-learning-rate-dd5da26bb6de#: :text=If%20the%20learning%20rate%20is,good%20learning%20rate%20is%20crucial..

[23] 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,” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Feb. 2018.

[24] World Health Organization, ”Pneumonia in children,” World Health Organization, 11 November 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/pneumonia.