Data Augmentation in Medical Image Datasets with DCGAN

Seha Ay, Adam Prys, Anand Patel, Thomas Jeong

Team 13

Seha Ay
- Wake Forest School of Medicine
- Biomedical Engineering, PhD Student
- Data privacy in ML and FL applications

Thomas Jeong
- Wake Forest School of Medicine
- Biomedical Engineering, PhD Student
- Finite element modelling, vehicle safety

Adam Prys
- Virginia Tech
- Master of Engineering in Computer Science
- Concentration in Software Development

Anand Patel
- Virginia Tech
- Master of Engineering in Computer Science
- Concentration in Software Development

Problem Description

Insufficient Medical Data for Healthcare Solutions

- AI-based healthcare solutions require significant data for effective training
- However, access to medical data containing PHI is restricted under HIPAA regulations
- Insufficient data leads to poor healthcare solutions
- DCGAN (Deep Convolutional Generative Adversarial Network) can generate synthetic medical data
- DCGAN can be used to augment real medical data and improve AI-based healthcare solutions

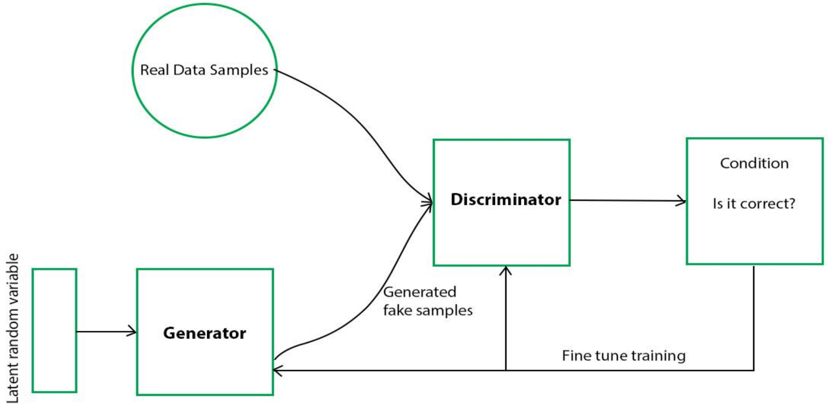
Approaches

Generative Adversarial Network (GAN)

- GAN is a DNN based model in adversarial settings to learn probabilistic distribution of the data to explain how data is generated.
- Purpose is to generate synthetic (fake) data based on the provided original samples.
- Consists of 2 networks:
  - **Generator (unsupervised):** Generate synthetic data
  - **Discriminator (supervised):** Predicts if the synthetic data is fake or real
- Synthetic data quality is improved in every other cycle between Generator and Discriminator.
- GAN model is concluded when Discriminator cannot distinguish synthetic data from the original samples.

GAN has various applications:

- ***Image (DCGAN)***
- Text
- Audio/Music
- Video

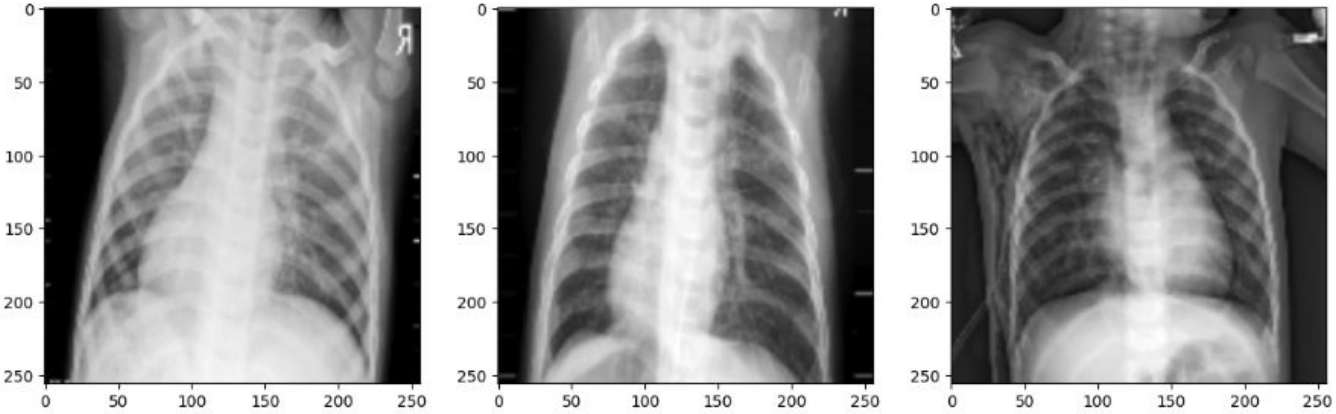


Deep Convolutional Generative Adversarial Network (DCGAN)

- Specific type of GAN that allows network to better understand the structure and spatial features of the input data.
  - Mainly adapted on image and video processing.

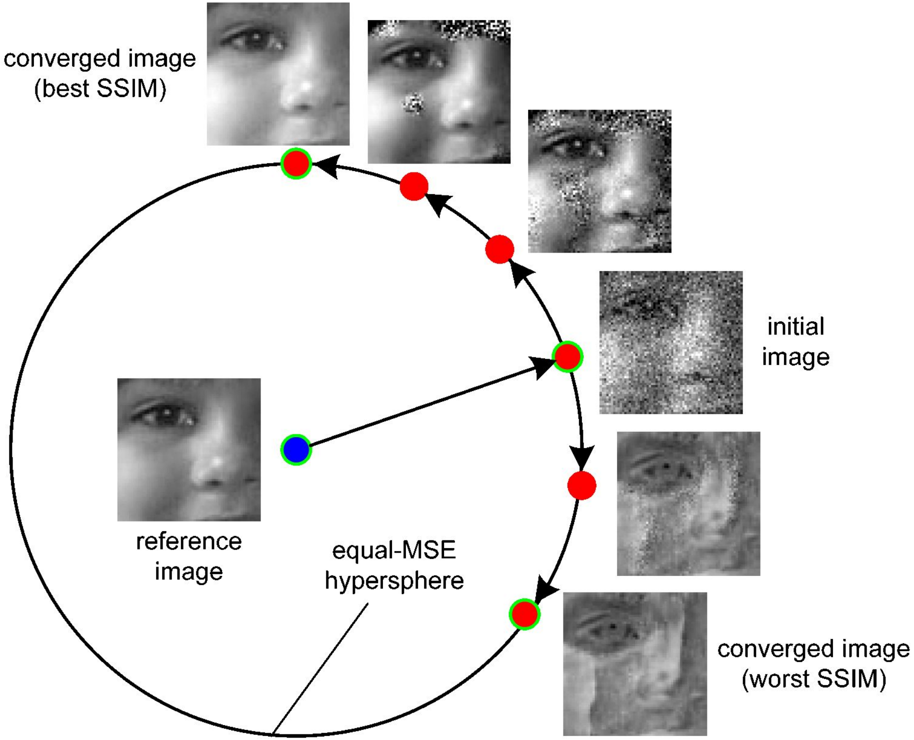
- We utilized this method for augmenting Chest X-ray data used in pneumonia detection.
  - This data is collected by Guangzhou Women and Children’s Medical Center and publicly accessible under a CC (Creative Commons) BY 4.0 license.

Sample Images from Chest X-Ray Dataset



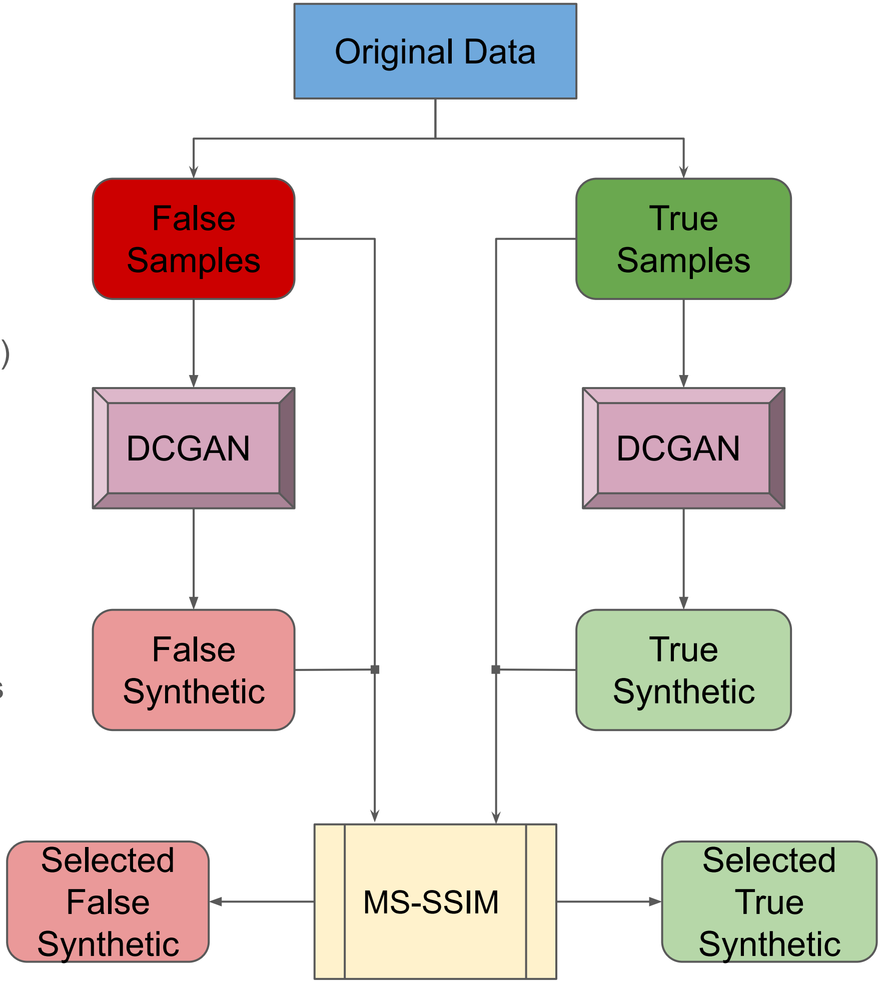
Multi-Scale Structural Similarity Index (MS-SSIM)

- Commonly adopted technique in image processing.
- Fine to coarse
- Scores between 0 to 1.
- Initial validation of the synthetic image quality for model training.



Procedure

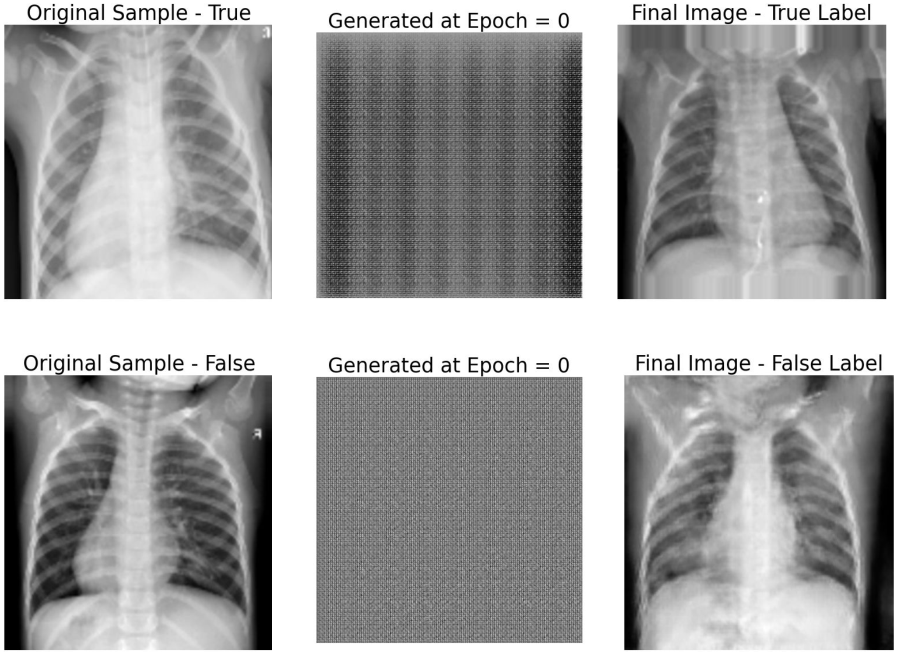
- Split dataset based on labels
  - 0 (FALSE) : Healthy Patients (~2000 Samples)
  - 1 (TRUE) : Patients with Pneumonia (~900 Samples)
- DCGAN training (individually)
  - To avoid overlapping between different classes
  - To ease labelling of synthetic images
- MS-SSIM calculation for initial validation of synthetic image quality
  - Minimum MS-SSIM score among original dataset was 0.0314
  - Our threshold MS-SSIM score 0.15



Results

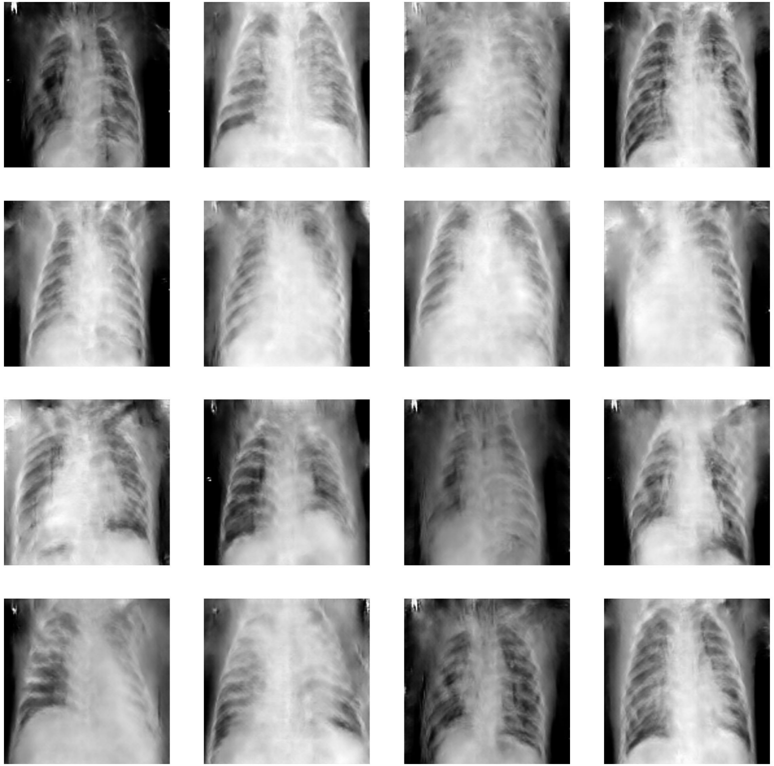
Results - DCGAN

- DCGAN model trained with TRUE and FALSE labeled images individually
    - 800 epochs for TRUE (Patients w/ Pneumonia)
    - 800 epochs for FALSE (Healthy Patients)
- Results illustrated on right
- Images generated at final epoch have a good visual quality for both cases

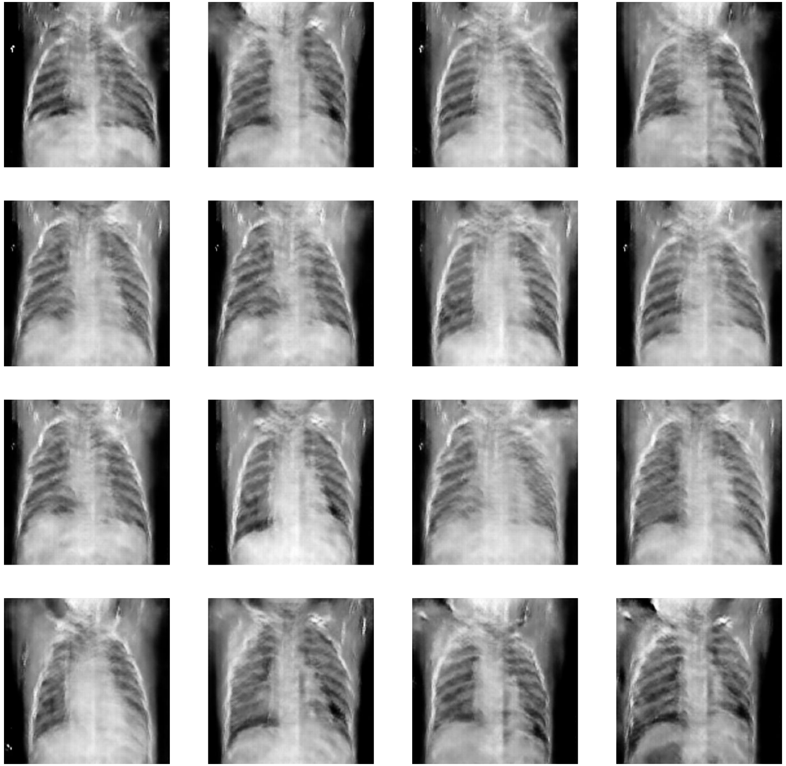


Results - Multiple Outputs from DCGAN

True Labels - Pneumonia



False Labels - Healthy

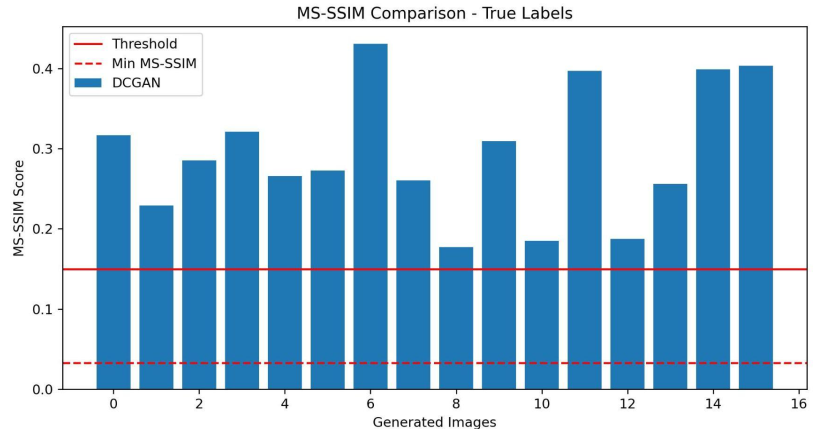


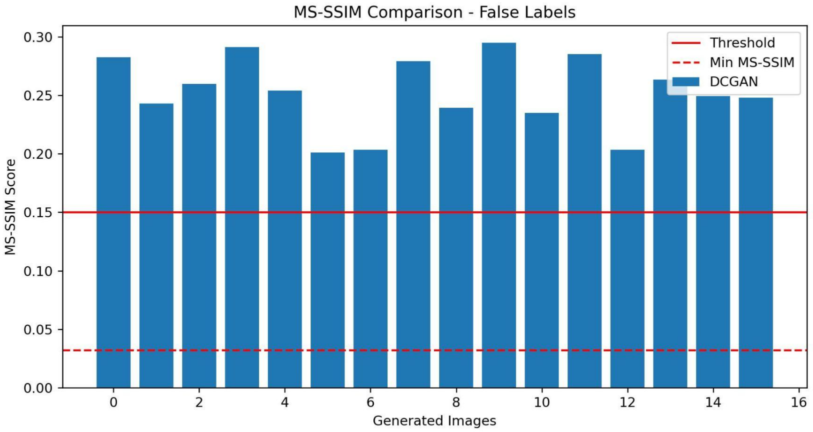
Results - MS-SSIM

- MS-SSIM scores as illustrated
  - 15 out of 15 > threshold (0.15)
  - 15 out of 15 > min MS-SSIM (0.0314)

- TRUE (Pneumonia) Synthetic:
  - Max MS-SSIM ~ 0.42
  - Min MS-SSIM ~ 0.20

- FALSE (Healthy) Synthetic:
  - MAX MS-SSIM ~ 0.30
  - Min MS-SSIM ~ 0.20





Lesson Learned

- In this project we identified a solution for augmenting data for tasks with insufficient training data.
- Specifically we used a medical image dataset for pneumonia detection, one of the most prevalent problems in the healthcare space.
- Our results show promising approach for such problems
- However,
    - High computational-cost
    - Requires high memory allocations
    - May be time inefficient
        - Depending on number of synthetic data requirements
    - May require manual sampling of images depending on the data variation

Future Work

- Initial observations shows promising results, but room for improvement
- In this study :
    - Only used a medical image dataset
    - Task was data augmentation
    - Good for increasing dataset size
- Next:
    - Test augmented dataset in CNN based Pneumonia Detection
    - Apply these procedures in non-image datasets (text, audio, video)
    - Try tackling different problems under same concept:
        - Balancing unbalanced class samples among datasets
        - Mitigation bias among dataset
            - Sampling Bias
                - Gender, Underrepresented populations, minority ethnic groups

Thank You