Towards Robust Chest X-Ray Diagnostics via Generative AI

Minhyek Jeon, Jonathan Zhu
Carnegie Mellon University, Computational Biology Department
`minhyekj@andrew.cmu.edu; jazhu@andrew.cmu.edu`

Abstract

Despite considerable advancements in deep learning and computer vision algorithms, the data on which they train is often lacking and class-imbalanced, especially in the case of chest x-ray recognition to detect thoracic diseases, which inherently leads to overfitting and/or poor model performance. As such, we utilized a combination of models, including the foundation model CheXFound, GLoRI, Demons registration, and flow matching to inpaint healthy chest x-ray images with specific diseases. The generation of new images allowed for improvement in image classification tasks as well as addressing the problem of overfitting.

1 Introduction

For a variety of diseases in the thoracic region of the human body, diagnosis often requires more than simply external visible symptoms; one of the most powerful tools in diagnosing a variety of conditions in this region is the chest radiograph, known more commonly as a chest x-ray. While chest x-rays primarily examine the lungs, the thoracic region of the body also contains the human heart and key blood vessels. This allows for physicians to examine not only respiratory conditions such as pneumothorax (collapsed lung) [13] and pleural effusion (accumulation of fluid in the lungs) [5], but also heart failure [3] and blood vessel abnormalities, including aortic enlargement [8].

An estimated 70 million chest x-rays are performed in the United States every year [4], implying a large supply of imaging data. Training physicians to accurately analyze a chest x-ray can be a difficult task [11], so it makes sense to apply computer vision and deep learning approaches to chest x-ray recognition. While the body of previous work is large, the problem of image recognition with chest x-rays inherently suffers from severe class imbalance; most chest x-rays show healthy or clinically irrelevant results, and many thoracic diseases have low prevalence [4]. Collecting more data is both impractical and unethical (for example, it is unethical to intentionally give a patient a collapsed lung simply for collecting more pneumothorax x-rays).

To solve this problem, this project aims to generate synthetic data to enhance model performance in diverse clinical settings with limited datasets. Our aim is to leverage latent diffusion and foundation models to create high-fidelity, disease-relevant images that augment underrepresented classes and mitigate data scarcity. By improving data diversity, this approach strengthens models' abilities to generalize across different imaging protocols, equipment variations, and population demographics, ultimately enhancing their clinical applicability and reliability.

2 Methods

2.1 Dataset

The training and testing dataset is sourced from the National Institute of Health (NIH) Clinical Center, comprising over 60,000 anonymized chest X-ray images from more than 30,000 patients. The dataset includes 15 label classifications, 14 of which correspond to lung diseases, while the remaining label represents "No findings". The full list of disease classes and their image count is given in Table A1.

2.2 Classification Models

All classification tasks were performed with EfficientNet-B4 [9] implemented in Pytorch. Models were fine-tuned from a set of pretrained weights for 10 epochs utilizing BCEWithLogits loss and the Adam optimizer. We used a base learning rate of 10^{-4} and a ReduceLROnPlateau learning rate scheduler.

2.3 Image Generation

To synthetically generate images, we utilized a combination of various foundation models to generate attention maps for each of the diseases within the dataset. These attention maps are then used in an image registration process to determine regions of healthy images to inpaint with disease characteristics, and the inpainting itself is performed with flow matching. More specifics about these processes are described as follows.

2.3.1 CheXFound and GLoRI

Yang et al. [12] describe their foundation model CheXFound as a self-supervised model that learns representations of chest x-rays. Specifically, they use DINOv2 [7] and a teacher-student framework; the student takes in various local crops of the input images while the teacher takes in the global crops. This allows for the patches of the image to be tokenized.

Following this, the tokens are then put through a secondary model called Global and Local Representations Integration (GLoRI) that projects the CheXFound tokens into a GLoRI space in combination with specific disease queries. In short, this allows for the retrieval of image representations through prompting.

2.3.2 Demons Registration

Thirion [10] first proposed the idea of Demons registration for image matching; this aims to alter an input image to be as close to a reference image as possible. However, our version of Demons registration is more customized to the task at hand. Specifically, we take a query image F and a reference image R; for both of these images, we determine an intensity gradient field $\nabla F, \nabla R$ and use both images to compute a difference mapping $\alpha(R - F)$. All three of these images are then combined to determine a deformation field, which informs a warping of the query image to closely match the reference image, resulting in a registered image. This process is repeated for multiple iterations by taking the registered image as the new query image.

2.3.3 Flow Matching

Lipman et al. [6] describe the general process of flow matching as a means of image generation. In short, flow matching defines data points $x = (x^1, \ldots, x^d) \in \mathbb{R}^d$ and defines a time-dependent probability density path function $[0, 1] \times \mathbb{R}^d \to \mathbb{R}_{>0}$, with p_t as the corresponding PDF. It similarly defines a time-dependent vector field $v : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$. A vector field can then be used to similarly make a time-dependent diffeomorphic map, also known as a *flow*, defined as $\phi : [0, 1] \times \mathbb{R}^d \to \mathbb{R}^d$. In our case, v_t defines the vector field at time t, while u_t defines the training vector field. In this case, flow can be defined in the form of the ordinary differential equation (ODE) $\frac{d}{dt}\phi_t(x) = v_t(\phi_t(x))$, with $\phi_0 = x$. This indicates that flow change is simply the vector field difference at each time step, and that the initial flow is simply the sample. Thus, we establish a framework that allows us to go from a sample, through an ODE solver, to a target distribution.

The flow matching framework allows for high-quality *de novo* image generation, but this is unsuitable for the task of inpainting a chest x-ray with a specific disease. As such, we follow the framework of Albergo et al. [2] that, in loose terms, links base and target distributions to allow for higher performance on generative tasks when parts of an image are already given. This allows for specific inpainting of images where portions of the image are removed, which is what we do to inpaint disease symptoms onto healthy images.

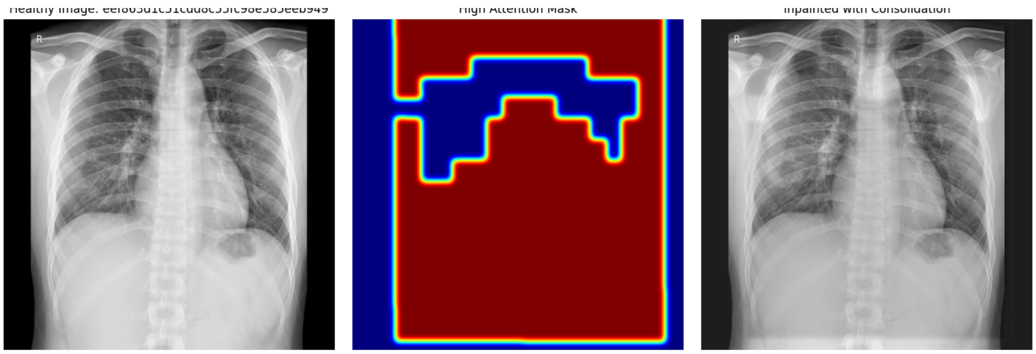

Figure 1: Inpainting process and intermediary results, showing an example of a healthy chest x-ray (left), an attention map for consolidation (middle), and the resulting inpainted image (right).

3 Results

For brevity, we only comment on partial image generation results here; to visualize the full results, please refer to the section on code and data availability.

3.1 Image Generation

Across 100 epochs of the image generation pipeline, we extracted 638 generated images. Each image yielded an increase in classification probability for the generated disease. An example is given in Figure 1, showing the reference healthy x-ray, the derived attention map, and the resulting inpainted image. For all images, the resulting generation is subtle due to many disease symptoms being subtle.

3.2 Classifier Results

Our classifier trained on the base data yielded a top validation accuracy of 0.9651 with a minimum validation loss of 0.0908. When generated images were added to the data pool and the classifier was retrained, the validation accuracy peaked at 0.9644 with a minimum loss of 0.0918. Despite seemingly worse metrics, the augmented data helped to address overfitting; after 10 epochs, the difference in training and validation loss was 0.0083 on the base data but 0.0076 on the augmented data. Loss curves are given in Figure A1, showcasing the difference in overfitting.

4 Discussion

Our image generation pipeline showcases a viable starting point to improve methodologies to improving image quality for training computer vision tools on medical tasks, specifically in recognizing diseases in the human thoracic region. This generation pipeline can be helpful for some rarer diseases, such as pneumothorax, with approximately 10 cases per every 100,000 people per year [1]. While our generated images can show subtle differences to the naked eye, they highlight the areas of focus for computer vision models and enable more robust training.

A key limitation of our work is that we were unable to leverage large amounts of processing power. Time and space constraints prevented us from generating a large number of images, and the increase in dataset quality for this project is still quite small. As such, future work would involve leveraging larger processing power with larger training datasets and experimenting with larger foundation models for high-fidelity image generation. A further limitation of our work is the inclusion of only 14 thoracic conditions; image recognition models cannot recognize conditions not present in their training data, and our pipeline cannot generate images with conditions not present in its training data. As such, the acquisition of x-rays featuring conditions such as heart failure, bone fractures, and hiatal hernias and generating x-ray images with those conditions would constitute a promising area of future research. We also envision that models such as ours can also be applied to generate images for other fields, medicine or otherwise, that rely on imaging data yet have difficulties in its collection and labeling, including CT and ultrasound scans.

Code and Data Availability

All code is submitted on Gradescope, and all code, data, and results can be accessed at https://drive.google.com/drive/folders/1Gbc5X2w96jo94kvvWG7_C379ZW8VaYuJ?usp=drive_link.

References

[1] Adeel Ahmad Khan, Muhammad Zahid, Mousa Ahmad Alhiyari, Abdulrahman Ahmad Al-Andulmalek, Unwam Ekpuk Jumbo, Mohammad Naser Kloub, Muhammad Muslim, Muhammad Naeem, Zohaib Yousaf, Rashid Mazhar, et al. Demographics, clinical characteristics, and recurrence rate of patients with primary spontaneous pneumothorax at a tertiary care center in qatar. *Qatar Medical Journal*, 2022(4):56, 2022.

[2] Michael S Albergo, Mark Goldstein, Nicholas M Boffi, Rajesh Ranganath, and Eric Vanden-Eijnden. Stochastic interpolants with data-dependent couplings. *arXiv preprint arXiv:2310.03725*, 2023.

[3] Cândida Fonseca, Teresa Mota, Humberto Morais, Fernando Matias, Catarina Costa, António G Oliveira, Fátima Ceia, and EPICA Investigators. The value of the electrocardiogram and chest x-ray for confirming or refuting a suspected diagnosis of heart failure in the community. *European journal of heart failure*, 6(6):807–812, 2004.

[4] Lisa Iyeke, Rachel Moss, Rochelle Hall, Jeffrey Wang, Laiba Sandhu, Brendan Appold, Enessa Kalontar, Demetra Menoudakos, Mityanand Ramnarine, Sean P LaVine, et al. Reducing unnecessary 'admission'chest x-rays: An initiative to minimize low-value care. *Cureus*, 14(10), 2022.

[5] Berthold Jany and Tobias Welte. Pleural effusion in adults—etiology, diagnosis, and treatment. *Deutsches Ärzteblatt International*, 116(21):377, 2019.

[6] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[7] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[8] Tom Rosenwasser, Ronit Lain, and Miri Weiss Cohen. Aortic enlargement detection using chest x-rays to identify potential marfan syndrome. *Procedia Computer Science*, 207:2125–2133, 2022.

[9] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[10] J-P Thirion. Image matching as a diffusion process: an analogy with maxwell's demons. *Medical image analysis*, 2(3):243–260, 1998.

[11] Bram Van Ginneken, BM Ter Haar Romeny, and Max A Viergever. Computer-aided diagnosis in chest radiography: a survey. *IEEE Transactions on medical imaging*, 20(12):1228–1241, 2001.

[12] Zefan Yang, Xuanang Xu, Jiajin Zhang, Ge Wang, Mannudeep K Kalra, and Pingkun Yan. Chest x-ray foundation model with global and local representations integration. *arXiv preprint arXiv:2502.05142*, 2025.

[13] Paul Zarogoulidis, Ioannis Kioumis, Georgia Pitsiou, Konstantinos Porpodis, Sofia Lampaki, Antonis Papaiwannou, Nikolaos Katsikogiannis, Bojan Zaric, Perin Branislav, Nevena Secen, et al. Pneumothorax: from definition to diagnosis and treatment. *Journal of thoracic disease*, 6(Suppl 4):S372, 2014.

A Appendix

Table A1: Disease classes and the number of chest x-rays labeled as that class present in the image dataset of anonymized chest x-rays from the NIH.

Disease Class	**Number of X-Rays**
No Findings	31,818
Aortic Enlargement	7,162
Cardiomegaly	5,427
Pleural Thickening	4,842
Pulmonary Fibrosis	4,655
Nodule/Mass	2,580
Lung Opacity	2,483
Pleural Effusion	2,476
Other Lesion	2,203
Infiltration	1,247
ILD	1,000
Calcification	960
Consolidation	556
Atelectasis	279
Pneumothorax	226

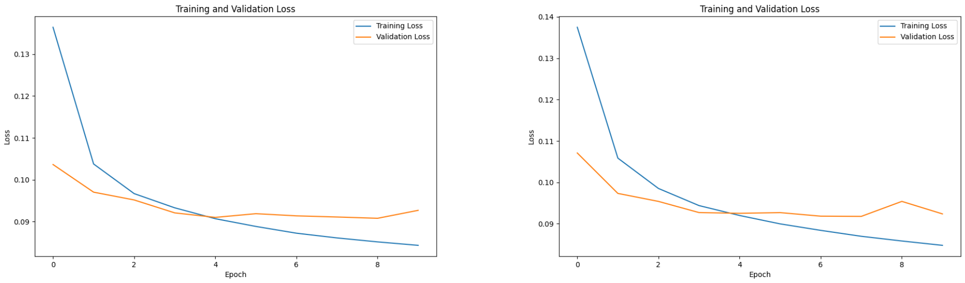

Figure A1: Training and validation loss curves over 10 epochs for EfficientNet-B4 trained on unaltered data (left) and data augmented with generated diseased chest x-rays (right). Critically, the augmented data showcases a decrease in overfitting despite increasing the dataset size by 1%.