Closing the Gender Gap in Medical Imaging with Variational Autoencoders

Sarah Korb
Columbia University
sbk2176@columbia.edu

Zara Hall
Columbia University
zyh2000@columbia.edu

Abstract

Gender disparities in medical data are well-documented, with women historically underrepresented in datasets that inform modern diagnostic models. These imbalances contribute to biased clinical outcomes and under performance of AI systems on female patients. Given the high cost and privacy constraints of acquiring medical images, data augmentation via generative models is a promising strategy to address such imbalances.In this project, we propose using Variational Autoencoders (VAEs) to generate synthetic medical images that represent female populations. We apply this framework to the NIH ChestX-ray14 dataset and evaluate whether a CNN trained on a dataset balanced with our synthetic data performs as well as a CNN trained on balanced, real data, using accuracy metrics to assess impact. Our work hopes to build off of previous work (Larrazabal et al. [2020]) which showed that CNN's training on gender imbalanced X-ray datasets resulted in decreased performance. Our results show significant challenges in training CNN models on chest X-ray data, with baseline models exhibiting poor discriminative ability despite various optimization strategies.

1 Introduction

Medical imaging models trained on imbalanced datasets often exhibit performance disparities across demographic groups. Prior work has shown that convolutional neural networks (CNNs) trained on gender-skewed chest X-ray datasets perform significantly worse on underrepresented groups, particularly female patients Larrazabal et al. [2020]. However, collecting large amounts of gender-balanced medical image data is challenging due to privacy constraints, cost, and limited availability. With AI becoming integrated in the medical field, particularly in imaging and diagnosis, it is imperative that this disparity is addressed to avoid biased and potentially harmful outcomes to underrepresented communities.

Generative models, such as Variational Auto-encoders (VAE) and Generative Adversarial Networks (GAN) offer a promising solution to this issue. By training on and learning distributions from medical images of underrepresented groups, these models may help mitigate performance disparities in downstream applications by providing realistic, synthetic data. In this paper, we investigate whether a Variational Autoencoder (VAE), trained on female chest X-rays, can generate high-quality samples that improve diagnostic performance of a CNN when added to a gender-imbalanced dataset.

We evaluated this approach in the NIH ChestX-ray14 dataset, comparing classifier performance when trained on (a) balanced real data, and (b) balanced data augmented with synthetic female images. We assessed results using AUROC, sensitivity, specificity, and group fairness metrics.

2 Related Work

Several recent studies have found demographic bias in medical imaging classification models, and the performance pitfalls of these imbalances. Seyyed-Kalantari et al. [2021] analyzed state-of-the-art chest X-ray systems and showed that they systematically underdiagnose underserved populations. Similarly, Larrazabal et al. [2020] demonstrated that convolutional neural networks (CNNs) trained on gender-imbalanced datasets perform significantly worse on female patients compared to male patients.

Generative approaches such as VAEs and GANs are increasingly being explored for medical image synthesis and augmentation. EndoVAE, a VAE-based model for endoscopic image generation, has shown improved performance over GAN-only approaches [Diamantis et al., 2022]. More broadly, Rais et al. [2024] provide a comprehensive survey of VAE applications in biomedical imaging, highlighting their effectiveness to augment small data sets. However, there is limited work explicitly applying generative models to address demographic imbalances in medical image datasets. Our project aims to help close this gap by using a conditional VAE to generate synthetic female chest X-ray images for balancing gender disparities in training data.

3 Methodology

3.1 Dataset Preparation

We use a random sample from the NIH ChestX-ray14 dataset [Wang et al., 2017], which contains over 5000 frontal-view chest radiographs annotated with 14 disease labels and patient metadata, including gender. We chose to use a small subset due to computational and memory limitations. We first extract a subset consisting only of female patient images. This subset serves as the training set for our VAE.

Each image in the dataset may be associated with multiple disease labels. To represent this, we create a multi-hot encoded vector for each image, where each entry is a one hot encoding indicating the presence or absence of a particular condition represented at that index. There were 15 possible disease labels, including a 'No Finding' label.

3.2 Conditional VAE Training

As the foundation for our model implementation, we used an open-source repository that trained a Variational Autoencoder (VAE) to generate prostate images [Holsman, 2023]. Using this codebase as a scaffold, we modified the architecture to implement a conditional VAE, enabling the model to generate chest X-ray images conditioned on the disease(s) present. Without this conditioning mechanism, the VAE would model a single distribution across all training samples, regardless of pathology. This would likely result in generated images that blur disease-specific features, producing unrealistic samples that do not accurately represent any particular disease class.

We then trained our Conditional Variational Autoencoder (CVAE) to model the distribution of female chest X-ray images. The CVAE is conditioned on the multi-hot disease label vector, allowing the model to learn the joint distribution of image appearance and pathology. The encoder maps an image and its condition vector to a latent representation, and the decoder reconstructs the image from this latent code and the same condition. This setup allows us to sample from the latent space to generate completely new, synthetic female chest X-rays corresponding to specific diseases. We generated 144 synthetic chest X-ray samples using our trained conditional VAE, conditioning on the label vectors from the test set. This approach ensures that the generated images reflect real-world disease co-occurrence patterns, as opposed to randomly sampled one-hot or multi-hot vectors, which could result in clinically implausible combinations of conditions, and thus, highly improbable data examples.

3.3 Architecture overview

We implemented a Conditional Variational Autoencoder with 5 convolutional layers in the encoder and 4 transposed convolutional layers in the decoder, followed by a final upsampling block. We use mean squared error (MSE) for reconstruction loss and standard KL divergence for latent regularization.

4 Experiments

The generated female images of the CVAE are then used to augment male-dominated datasets, producing more balanced training distributions.

To evaluate the impact of this augmentation, we train a CNN classifier (CheXNet [Rajpurkar et al., 2017]) under three conditions:

- A balanced dataset with real female and male images.
- An imbalanced dataset augmented with synthetic female images.

We then compare the classification performance of the models trained on these different datasets, using metrics such as AUROC, sensitivity, and fairness scores.

5 Results

5.1 Synthetic Image Generation

Our VAE successfully learned to generate synthetic chest X-ray images that preserved basic anatomical structures. The generated images were blurry compared to real images, which is a common limitation of VAE models, however they preserved most of the essential features such as lung fields, heart silhouettes and rib cage structure.

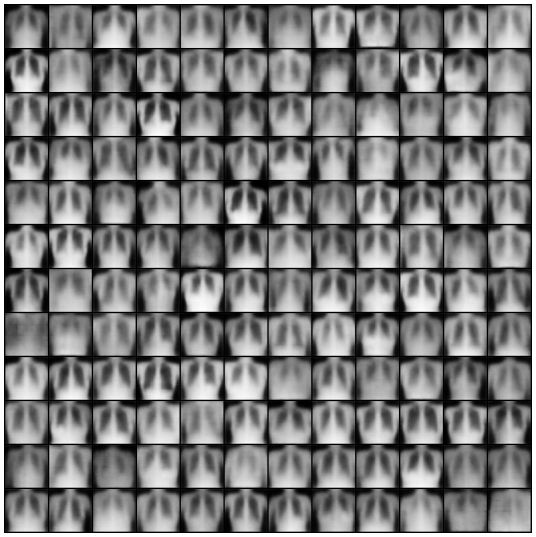

Figure 1: Grid of generated examples after final epoch

5.2 Challenges with Model Convergence

We implemented a training pipeline for both the VAE and the downstream CNN classifier but faced challenges with convergence. The classifier consistently predicted the majority class for all samples resulting in high accuracy but zero sensitivity for all pathologies Table 1 presents the evaluation metrics from our latest experimental run, highlighting these convergence challenges.

5.3 Model Performance Evaluation

Table 1 presents the evaluation metrics from our baseline model (without VAE augmentation), highlighting the convergence challenges we encountered.

Table 2 shows the results for the model trained with VAE-augmented data, illustrating similar convergence issues despite the data augmentation.

Table 1: Baseline Model Performance Metrics Across Pathologies

Condition	AUC	Accuracy	Sensitivity	Specificity	Class Ratio
Atelectasis	0.499	0.937	0.000	1.000	0.063
Cardiomegaly	0.499	0.969	0.000	1.000	0.031
Effusion	0.501	0.946	0.000	1.000	0.054
Infiltration	0.511	0.896	0.000	1.000	0.104
Mass	0.501	0.963	0.000	1.000	0.037
Nodule	0.501	0.951	0.000	1.000	0.049
Pneumonia	0.501	0.993	0.000	1.000	0.007
Pneumothorax	0.501	0.987	0.000	1.000	0.013
Consolidation	0.501	0.984	0.000	1.000	0.016
Edema	0.500	0.993	0.000	1.000	0.007
Emphysema	0.499	0.993	0.000	1.000	0.007
Fibrosis	0.500	0.979	0.000	1.000	0.021
Pleural Thickening	0.499	0.977	0.000	1.000	0.023
Hernia	0.501	0.996	0.000	1.000	0.004

Table 2: VAE-Augmented Model Performance Metrics Across Pathologies

Condition	AUC	Accuracy	Sensitivity	Specificity	Class Ratio
Atelectasis	0.501	0.937	0.000	1.000	0.063
Cardiomegaly	0.501	0.969	0.000	1.000	0.031
Effusion	0.501	0.946	0.000	1.000	0.054
Infiltration	0.489	0.896	0.000	1.000	0.104
Mass	0.499	0.963	0.000	1.000	0.037
Nodule	0.499	0.951	0.000	1.000	0.049
Pneumonia	0.500	0.993	0.000	1.000	0.007
Pneumothorax	0.500	0.987	0.000	1.000	0.013
Consolidation	0.501	0.984	0.000	1.000	0.016
Edema	0.500	0.993	0.000	1.000	0.007
Emphysema	0.501	0.993	0.000	1.000	0.007
Fibrosis	0.501	0.979	0.000	1.000	0.021
Pleural Thickening	0.501	0.977	0.000	1.000	0.023
Hernia	0.501	0.996	0.000	1.000	0.004

5.4 Interventions and Attempted Solutions

To address these convergence issues, we implemented several interventions:

1. **Removal of Early Stopping**: We initially hypothesized that early stopping might be preventing the model from learning effectively. We modified the training pipeline to remove early stopping and allow the model to train for the full number of epochs.

2. **Extended Training Duration**: We increased the number of training epochs from 50 to 150 and fine-tuning epochs from 20 to 50, giving the model significantly more time to learn patterns in the data.

3. **Learning Rate Adjustments**: Our training logs showed learning rate reductions over time, with the rate decreasing from 1e-4 to 1e-6 through the ReduceLROnPlateau callback. Despite these adjustments, the model showed minimal improvement in validation loss (from 0.16289 to 0.15392).

4. **Class Weighting**: We implemented class weighting with a multiplier of 2.0 for positive examples to address the significant class imbalance present in medical imaging datasets.

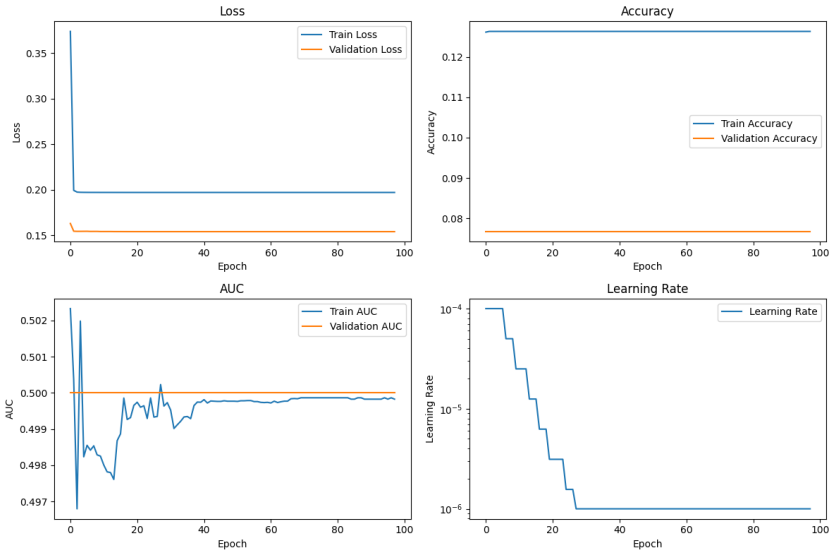

Figure 2: Training and validation loss across epochs for VAE augmented model, showing minimal improvement despite extended training and learning rate adjustments.

5.5 Training Analysis

In Figure 2 we show the training and validation loss of our model. Analyzing these logs, we noticed several key issues:

- The model quickly optimized to predict the majority class which led to low validation loss but poor discriminative ability
- Despite training for 100 epochs the AUC remained near 0.5 consistently
- The accuracy metric was flat the entire time
- These logs suggest there was a fundamental issue with the model that we used or the training data

The failure could be due to several factors, such as the extreme class imbalance inherent in medical datasets, the subtle nature of pathological features in chest X-rays, the limited size of our training dataset due to computational constraints, and potential limitations in the vanilla CheXNet architecture for imbalanced multi-label classification.

6 Discussion

Our results highlight both the promise and limitations of using synthetic data to mitigate demographic imbalances in medical imaging datasets. While our CVAE model successfully generated anatomically plausible female chest X-rays, the classifier trained on synthetic-augmented datasets struggled to achieve meaningful performance gains across all metrics. In particular, the AUROC values remained close to chance, and sensitivity scores were consistently zero, suggesting the model was unable to identify positive cases. Importantly, we observed that the classifier had the same issues even when we trained on just real data. With real data, the AUROC values hovered around 0.5 and there was zero sensitivity across all conditions. We attempted many different interventions, such as extended training, learning rate scheduling, and class weighting, and the model still failed to learn discriminative features effectively. Thus, we hypothesize this indicates a need for better synthetic image realism from improved VAE architectures, and an improved CNN architecture and training pipeline for image classification. Furthermore, alternative models for generation that typically have better graphic quality (such as GANs) may be a more appropriate (though more computationally expensive) choice. Indeed, indicators of disease, especially in grayscale images, can be very nuanced and low contrast.

7 Conclusion

In this study, we explored the use of a Conditional Variational Autoencoder (CVAE) to generate synthetic female chest X-rays for the purpose of mitigating gender imbalance in medical imaging datasets. While the model was able to generate anatomically reasonable samples, downstream classification performance did not improve significantly, highlighting challenges in convergence and generalization when using synthetic data for fairness interventions. This work contributes to the growing research area on addressing disparities in medical AI systems. We faced challenges however we think that this approach could be adapted for future applications. Being able to generate high quality synthetic medical images could eventually be a valuable tool for researchers and clinicians working with imbalanced datasets. At the same time, we believe that these challenges highlight the difficulty of developing methods for medical image classification.

7.1 Limitations

Our work is subject to several limitations. First, due to limited computational resources, we were constrained in model size and training time. As a result, many of the generated images exhibit blurriness or lack of anatomical detail, potentially leading to issues in disease classification in he downstream task. Second, due to memory limitations, we had access only to a public sample of the NIH ChestX-ray14 dataset, which restricted both the diversity and volume of training data. This constraint may affect the generalizability of our model, specifically across broader female patient populations and less common disease combinations.

7.2 Future Work

We plan to continue iterating on our current VAE framework by exploring alternative model architectures and training techniques to improve the quality and realism of generated images. In particular, we aim to investigate the use of more expressive decoders and improved regularization strategies in our VAE. Additionally, we are interested in extending our work to Generative Adversarial Networks (GANs), which, while more computationally demanding, have been shown to produce sharper and more detailed medical images in prior studies. Furthermore, we hope to improve the classification component of our pipeline by experimenting with different CNN architectures and configurations.

References

Irem Çetin, Maialen Stephens, Oscar Camara, and Miguel A. Gonzalez Ballester. Attri-VAE: Attribute-based interpretable representations of medical images with variational autoencoders. *Comput. Med. Imaging Graph.*, 104:102158, 2023. doi: 10.1016/j.compmedimag.2022.102158.

Abadh K. Chaurasia, Stuart MacGregor, Jamie E. Craig, David A. Mackey, and Alex W. Hewitt. Assessing the efficacy of synthetic optic disc images for detecting glaucomatous optic neuropathy using deep learning. *Transl. Vis. Sci. Technol.*, 13(6):1, 2024. doi: 10.1167/tvst.13.6.1.

Dimitrios E. Diamantis, Panagiota Gatoula, and Dimitris K. Iakovidis. Endovae: Generating endoscopic images with a variational autoencoder. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5, 2022. doi: 10.1109/IVMSP54334.2022.9816329.

Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018. doi: 10.1016/j.neucom.2018.09.013.

Ben Glocker, Caius Jones, Matthias Bernhardt, and Sebastian Winzeck. Algorithmic encoding of protected characteristics in chest x-ray disease detection models. *EBioMedicine*, 89:104467, 2023. doi: 10.1016/j.ebiom.2023.104467.

Max Holsman. Vae for medical image generation. https://github.com/maxholsman/VAE-for-Medical-Image-Generation, 2023. GitHub repository, Accessed: 2024-05-10.

Aghiles Kebaili, Jérôme Lapuyade-Lahorgue, and Su Ruan. Deep learning approaches for data augmentation in medical imaging: A review. *Journal of Imaging*, 9(4):81, 2023. doi: 10.3390/jimaging9040081.

Bardia Khosravi, Frank Li, Theo Dapamede, Pouria Rouzrokh, Cooper U. Gamble, Hari M. Trivedi, Cody C. Wyles, Andrew B. Sellergren, Saptarshi Purkayastha, Bradley J. Erickson, and Judy W. Gichoya. Synthetically enhanced: unveiling synthetic data's potential in medical imaging research. *EBioMedicine*, 104:105174, 2024. doi: 10.1016/j.ebiom.2024.105174.

Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc. Natl. Acad. Sci. U.S.A.*, 117(23):12592–12594, 2020. doi: 10.1073/pnas.1919012117.

National Institutes of Health. Nih chest x-ray dataset (sample). https://www.kaggle.com/datasets/nih-chest-xrays/sample. Accessed: 2024-05-10.

Khadija Rais, Mohamed Amroune, Abdelmadjid Benmachiche, and Mohamed Yassine Haouam. Exploring variational autoencoders for medical image generation: A comprehensive study. *arXiv preprint arXiv:2411.07348*, 2024. Available at https://arxiv.org/abs/2411.07348.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017. URL https://arxiv.org/abs/1711.05225.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, 2021. doi: 10.1038/s41591-021-01595-0.

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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017. doi: 10.1109/CVPR.2017.369.

Zikang Xu, Jun Li, Qingsong Yao, Han Li, Mingyue Zhao, S. Kevin Zhou, et al. Addressing fairness issues in deep learning-based medical image analysis: a systematic review. *npj Digital Medicine*, 7:286, 2024. doi: 10.1038/s41746-024-01276-5.

7.3 Appendix

The CVAE implementation and image generation pipeline are available at: https://github.com/sarahkorb/VAE-for-Medical-Image-Generation. The main CVAE model code can be found in the file: models/cvae.py.