Project Proposal: Robust Domain-Generalizable AI for Chest X-Ray Disease Classification

Contributors: Yash Bansal, Kyoungeui Hong, Hanie Kang, Juann Kim

Abstract:

Medical AI models often show a significant performance drop when deployed in clinical environments different from the dataset they were trained on. This "generalization gap" is a major barrier to equitable healthcare, particularly in regards to disadvantaged institutions with older equipment or different patient demographics. Our project aims to address this challenge by developing and evaluating a robust deep learning model for multi-label thoracic disease classification. Initially, we will quantify the performance degradation of a standard AI model when transitioning from a high-quality, single-source dataset (MIMIC-CXR) to a more diverse, multi-source dataset (ChestX-ray14). Then, we will implement and validate a domain generalization strategy using advanced data augmentation to create a model that is resilient to variations in data quality and characteristics, which improves our model's robustness. This research aligns with UN Sustainable Development Goals 3 and 10 (Good Health and Well-being, and Reduced Inequalities, respectively) by contributing to the development of more equitable and reliable medical AI.

1. Introduction & Motivation

Chest X-rays are one of the most common diagnostic imaging tools worldwide. AI-powered classification models have shown great promise in detecting thoracic pathologies, but their real-world utility is often hampered by a lack of generalization. A model trained on pristine data from a state-of-the-art hospital may fail when applied to noisier images from a rural clinic's older equipment. This creates a significant decline in the quality of AI-assisted healthcare, which often results in performance failure.

Thus, we will address how to build a diagnostic AI that is fair, robust, and performs reliably across different hospital environments within a relatively small dataset. Inspired by the focus on "Generalization" and "Fairness" at top-tier conferences, including MICCAI and AAAI, our project pursues to create a model that bridges this gap, ensuring that the benefits of medical AI are accessible to a wide range of clinical settings.

2. Background & Related Work

The task of multi-label classification from chest X-rays is well-established, with standard CNN architectures like DenseNet-121 and ResNet-50 serving as common baselines. The primary challenge, however, lies in domain shift, the phenomenon where the statistical distribution of data differs between training (source) and testing (target) environments.

Recent research in domain generalization mainly focuses on learning domain-invariant features. While many advanced techniques exist (e.g., adversarial training, style transfer), a highly effective and feasible approach for a two-month project is the application of strong, realistic data augmentation and contrastive learning. By simulating the variations found in lower-quality imaging environments (e.g., changes in contrast, noise, resolution) during training, we can force the model to learn the fundamental pathological features of a disease rather than superficial characteristics of the training images. Furthermore, we will apply contrastive learning by creating positive–negative pairs to shape the embedding space, pulling together semantically similar views and pushing apart dissimilar samples, thereby producing more domain-invariant and robust representations.

3. Proposed Methodology

The project will be executed in two distinct phases:

Phase 1: Baseline Model & Problem Quantification

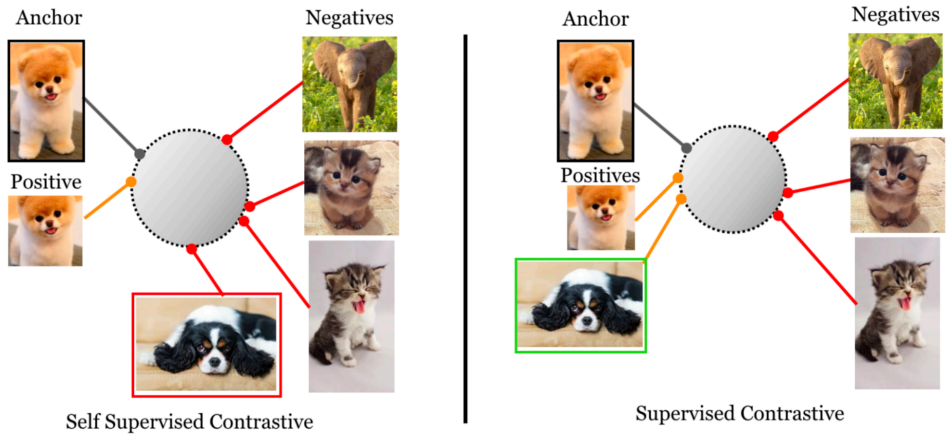
1. **Datasets:**
    - **Source Domain: MIMIC-CXR**, a large, high-quality dataset from a single U.S. medical center. Used for training and internal validation.
    - **Target Domain: ChestX-ray14 (NIH)**, a heterogeneous dataset from multiple sources, representing our "real-world" test case. Used *only* for final testing.
2. **Task:** Train a **DenseNet-121** model on the MIMIC-CXR dataset to perform 14-label thoracic disease classification (including Pneumonia, Cardiomegaly, etc.).
3. **Evaluation:** We will measure the performance (AUC score) of this baseline model on both the MIMIC-CXR test set and the ChestX-ray14 test set. We hypothesize a significant drop in AUC, thereby quantifying the "generalization gap."

Phase 2: Improved Model with Domain Generalization

1. **Technique:** We will re-train the same DenseNet-121 architecture on the same MIMIC-CXR data, but with the addition of heavy data augmentation. This includes:
    - Aggressive brightness and contrast adjustments.
    - Addition of Gaussian noise.
    - Simulated low-resolution effects.

Application of supervised contrastive learning: Based on the augmented data, we will treat each chest X-ray and its augmented variants as positive pairs, while images from different patients and different disease labels will serve as negative pairs. By pulling augmented views of the same image closer in the embedding space, and pushing away embeddings from different classes, the model will learn domain-invariant and disease-relevant features. This approach encourages the network to focus on true pathological signals rather than superficial imaging characteristics, thereby improving generalization across different hospitals and imaging conditions.

2. **Hypothesis:** This technique will force the model to become invariant to the superficial image "style" and focus on true pathological indicators.
3. **Final Evaluation:** The improved model's performance will be compared against the baseline on the ChestX-ray14 dataset. Our primary success metric will be a statistically significant reduction in the performance gap between the source and target domains. We will also use Grad-CAM visualizations to qualitatively assess if the improved model focuses more accurately on relevant lung regions in noisy images.



[Figure 1] Example of Supervised Contrastive Learning.

4. Project Plan & Timeline (8 Weeks)

- Week 6-7: Data Collection & Pre-Processing
- Week 8-10: Model Training & Initial Experiments
- Week 11-12: Preliminary Paper Draft Due
- Week 13-14: Finalize Experiments & Analyze Results
- Week 15: Project Presentation & Final Report Submission

5. Expected Outcomes & Impact

The expected outcome of this project is to build a robust deep learning model for chest X-ray analysis that generalizes better than a standard baseline. Beyond accuracy gains, this work will serve as a reproducible case study on using targeted data augmentation and supervised contrastive learning to address domain shift. By showing how these methods reduce bias and improve fairness, we aim to provide insights for designing equitable medical AI. This aligns not only with SDG 3 (Good Health and Well-being) and SDG 10 (Reduced Inequalities), but also with SDG 9 (Industry, Innovation and Infrastructure) by promoting resilient healthcare technologies and SDG 17 (Partnerships for the Goals) by encouraging global collaboration to ensure equitable access to medical AI. Ultimately, the results should guide future research and support diagnostic models that can be reliably deployed across diverse clinical environments.

References

[1] Zhang, Li, et al. “Generalizing Deep Learning for Medical Image Analysis.” *PMC*, 2020. PMC

[2] Yao, Li, Jordan Prosky, Ben Covington, and Kevin Lyman. “A Strong Baseline for Domain Adaptation and Generalization in Medical Imaging.” *arXiv*, 2019. arXiv

[3] Huang, S. C., et al. “Self-supervised learning for medical image classification.” *Nature Digital Medicine*, 2023. Nature

[4] Zhou, Wenshuo, et al. “Contrastive Centroid Supervision Alleviates Domain Shift in Medical Image Classification.” *arXiv*, 2022. arXiv