Detecting Disease in Chest X-rays and Investigating Generalization to a New Clinical Population

Hypothesis

We can achieve satisfactory results in recognizing abnormal chest X-rays when operating within a similar context (for example, after training on data from hospital A, it will work well for future patients in hospital A). However, performance is likely to degrade when applying the model to different populations or contexts, such as urban vs. rural areas or different countries. We aim to explore how varying the dataset impacts performance across different contexts. Additionally, we will investigate whether generalizability is more effective in detecting abnormalities in general or in identifying specific diseases.

Expected Outcomes

- We anticipate improved chest abnormality detection when increasing the diversity of datasets.
- We expect generalizability to be more effective in detecting abnormalities than in identifying specific diseases.

Datasets

To enhance the generalizability of our model, we will include datasets from various geographic regions and institutions:

1. **ChestX (Stanford)**
   - 224,316 chest radiographs with disease labels from 65,240 patients
2. **NIH Chest X-rays (U.S.)**
   - 112,120 X-ray images with disease labels (14 categories) from 30,805 unique patients
3. **VinBigData Chest X-ray (Vietnam)**
   - 18,000 images
4. **MIMIC-CXR (Harvard Medical School)**
   - Over 377,000 images
5. **PadChest (Spain)**
   - 160,000 images from 67,000 patients
   - *Unlike the other datasets, this one is not publicly available and requires access approval. If needed, I can provide access to the downloaded data.*

Preprocessing Considerations

- The number of images is large, which may pose computational challenges.
- To ensure consistency, we will standardize disease categories across datasets. The PadChest dataset includes 19 diseases, whereas the others include 14; we will limit our scope to the common 14.
- If necessary, we will randomly subsample from the datasets to manage computational load.

Model Details

Input

- The model will be trained on images with captions and tested using images only.

Output

- The model will output either disease labels or a binary classification (normal/abnormal).

Model Architecture

- I will use either ResNet or AlexNet