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

Hypothesis

The aim of this project is to achieve more generalizable results in the detection of diseases through chest X-rays. This means creating a model that, based on a given dataset, produces accurate results not only on data gathered in a similar way but also on data from other contexts or collected in different environments. I aim to confirm the following hypotheses:

1. Models perform best on test data originating from the same source as the training data.
2. Manipulating the training dataset, through appropriate proportions and sampling, can improve accuracy.

I will evaluate these hypotheses in two scenarios:

- Detecting a specific disease (or absence thereof).
- Identifying whether any disease is present (i.e., normal vs. abnormal).

Literature Review & Importance

There is extensive research on detecting chest diseases and evaluating the generalizability of these models. Most research evaluates models trained on different datasets.

1. *Deep Learning for Distinguishing Normal versus Abnormal Chest Radiographs and Generalization to Unseen Diseases* examines whether models can generalize to detect normal/abnormal conditions for diseases not seen during training. This means the generalization occurs at the level of disease types rather than patient types. The study achieved results only slightly worse than trained radiologists, with similar discrepancies for both seen and unseen diseases.
2. *How Do Deep-Learning Models Generalize Across Populations? Cross-Ethnicity Generalization of COPD Detection* addresses the bias in AI models against African Americans and Latinos in healthcare. The study concludes that balanced datasets lead to better results and reduced bias, emphasizing the importance of dataset diversity. Notably, training on a mixed dataset reduced bias compared to training exclusively on data from African Americans.
3. *A Review of Recent Advances in Deep Learning Models for Chest Disease Detection Using Radiography* discusses various machine learning models and their application to chest disease detection. It highlights the difficulty humans face in differentiating diseases due to the visual similarity of X-rays and explores preprocessing techniques (e.g., contrast adjustment, Gabor filtering, CLAHE) and model ensembling for improved

performance. This paper emphasizes the advantages of ensembling multiple models and outlines techniques for improving input data quality.

4. *A Generalized Deep Learning Model for Multi-Disease Chest X-Ray Diagnostics* combines datasets from Stanford, NIH, and Shifa Hospital. The best performance was achieved using a combined dataset (from NIH and Shifa), reinforcing the idea that diverse datasets improve generalization. This paper is particularly relevant as it mirrors my approach by using multiple datasets to create a unified model. The findings suggest that training on a diverse dataset enhances generalization to other datasets compared to training on a single dataset.
5. *Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study* highlights the challenges in generalizing across datasets. Models perform better on data from sources similar to their training data, even when additional datasets are included. This paper reveals that even with increased dataset variety, performance tends to remain higher on datasets resembling the training source, emphasizing the difficulty of achieving true generalization.
6. *Multi-Population Generalizability of a Deep Learning-Based Chest Radiograph Severity Score for COVID-19* evaluates model performance across hospitals in the USA and Brazil. Despite differences in patient populations, the model generalized well, with correlation coefficients between model outputs and radiologist assessments ranging from 0.85 (in Brazil) to 0.9 (in a community hospital in the USA). This study supports the hypothesis that balanced and diverse datasets can improve model performance across different populations.

These sources highlight the importance of this research. Accurate detection of healthy chest X-rays can save radiologists' time and reduce human bias. Furthermore, training models on diverse datasets can promote equitable healthcare by improving diagnosis across different racial and geographic populations.

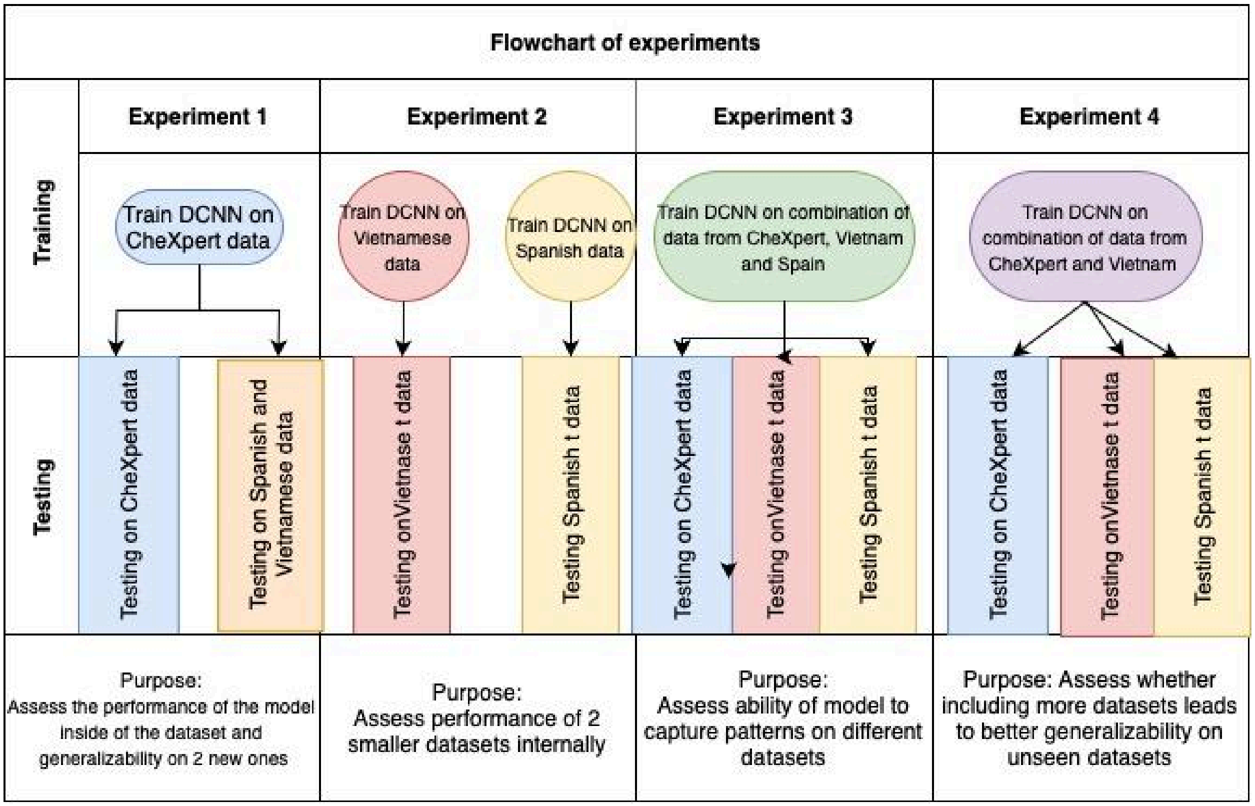
Out of all the above, the studies most aligned with my work are the fourth and fifth papers. I will replicate their approach by producing a classification model trained on different dataset combinations to evaluate generalization. The contrasting conclusions in these papers make this inquiry particularly valuable. The fourth paper supports the notion that combining datasets enhances generalization, while the fifth paper suggests that even with diverse training data, performance may still be limited to data resembling the training source.

Workflow for Experiments

1) **Data Acquisition**
    a) Download datasets: CheXpert, Vietnam dataset, PadChest (Spanish dataset)
    b) Connect datasets with corresponding labels (e.g., normal/abnormal classification, disease identifiers).

2) **Data Preprocessing**
    a) Handle missing values appropriately.
    b) Standardize disease identifiers across datasets.
    c) Perform image adjustments (e.g., contrast correction).
    d) Split datasets into training, validation and testing sets:
3) **Experiment 1: Baseline Model on CheXpert**
    a) Train a DCNN on CheXpert training data.
    b) Evaluate model performance on:
        i) CheXpert test data (in-domain evaluation).
        ii) Vietnam and PadChest test data (cross-domain evaluation).
        iii) Combined test sets (mixed-population evaluation).
    c) Compare normal/abnormal classification accuracy across datasets.
    d) Explore potential performance improvements through ensembling.
4) **Experiment 2: Dataset-Specific Models**
    a) Train separate models on Vietnamese dataset andPadChest dataset
    b) Compare performance against the CheXpert-trained model on Vietnamese and Spanish datasets.
5) **Experiment 3: Combined Dataset Training**
    a) Merge all datasets (CheXpert, Vietnam, PadChest) and train a unified model (consider balancing).
    b) Evaluate model generalization across all test datasets.
6) **Experiment 4: Combined Training on CheXpert and Vietnam**
    a) Combine CheXpert and Vietnam datasets and train a new DCNN model.
    b) Evaluate and compare generalization against previous experiments.
7) **Performance Comparison and Visualization**
    a) Visualize results using, ROC curves and bar plots for inter-experiment comparisons

All experiments will utilize the ResNet architecture.


Flowchart of experiments
Training
Testing
Experiment 1
Experiment 2
Experiment 3
Experiment 4
Train DCNN on CheXpert data
Train DCNN on Vietnamese data
Train DCNN on Spanish data
Train DCNN on combination of data from CheXpert, Vietnam and Spain
Train DCNN on combination of data from CheXpert and Vietnam
Testing on CheXpert data
Testing on Spanish and Vietnamese data
Testing onVietnase t data
Testing Spanish t data
Testing on CheXpert data
Testing onVietnase t data
Testing Spanish t data
Testing on CheXpert data
Testing onVietnase t data
Testing Spanish t data

Purpose:
Assess the performance of the model inside of the dataset and generalizability on 2 new ones

Purpose:
Assess performance of 2 smaller datasets internally

Purpose:
Assess ability of model to capture patterns on different datasets

Purpose: Assess whether including more datasets leads to better generalizability on unseen datasets

Results and Evaluation

Since the primary goal is to analyze generalizability, cross-experiment comparisons are essential. I will assess how much performance degrades when predicting on datasets not used for training and evaluate the impact of including new datasets on generalization.

Performance metrics will include:

- **Accuracy**: The proportion of correctly identified cases across all cases.
- **Precision**: The ratio of true positives to the sum of true positives and false positives (for abnormality detection).
- **Recall**: The ratio of true positives to the sum of true positives and false negatives (for abnormality detection).

Results will be visualized using ROC-AUC curves and comparative bar plots across experiments and testing sets.

Definition of Success & Expected Challenges

Success will be determined by the ability to identify trends in model generalization across datasets rather than achieving specific performance benchmarks. Demonstrating how different training strategies affect generalization will be a key outcome.

A significant challenge is the dataset size, as the combined datasets exceed 300,000 images, requiring substantial storage and processing time. If this becomes a bottleneck, I will implement balanced subsampling.

Preliminary Progress & Next Steps

So far, I have preprocessed two datasets (Vietnamese and Spanish), identified common disease labels, and merged images with patient descriptions. I am currently working on the third dataset, with the main challenge being its size (over 100GB), which has caused download issues.

Next steps include:

1. Completing preprocessing of the third dataset.
2. Training the baseline model on CheXpert and evaluating its performance.
3. Proceeding with subsequent experiments to gather sufficient information for meaningful conclusions.