Prediction of abnormalities on different chest X-ray datasets using Convolution Networks

Natalia Siwek

ABSTRACT

This project investigates how well deep learning models trained to detect diseases in chest X-rays generalize to unseen data from different sources. The aim is to assess whether models trained on a given dataset perform best on test data from the same source or outside ones, and whether manipulating the composition of training data through sampling and combining multiple datasets can improve generalization. These hypotheses in the binary classification of normal vs. abnormal chest X-rays. Based on conducted experiments, the papers showcases difficulties with building generalizable models but shows promise that through appropriate training the models can perform reasonably on combined datasets.

Introduction

Despite the rapid advancement of deep learning models in medical imaging, a persistent challenge is generalization: how well models trained on a specific dataset perform when applied to external data collected under different conditions, from different populations or using different labeling methods. This question is critical because healthcare data often vary across hospitals, geographic regions, and patient demographics. It's especially important because the hospitals who have resources to collect high-quality data, train and sustain the models are likely the one's with most resources, in the Western part of world, and treating patients not representing the general world population. In the same time, the hospitals that would benefit from working models the most (underfunded one, with not enough doctors and analysts), are not the one's whose data is used for training models. This means that a model that performs well only on data similar to its training source offers limited real-world utility and may exacerbate existing inequalities in healthcare delivery. On the other end, having models that would have high rates of detecting the status, could save plenty of resources and speed up the diagnosis process.

This project aims to evaluate two primary hypotheses:

1. Same-source advantage: Deep learning models trained to detect diseases in chest X-rays perform best on test data originating from the same source as the training data.
2. Improved generalization through dataset mixing: Training on appropriately sampled and mixed datasets improves model accuracy and generalizability, particularly when evaluated on external data sources.

We will assess these hypotheses by developing and testing binary classifiers tested on in sample and out of sample data and the

one's trained on datasets combined from multiple sources.

Related Work

The question of whether deep learning models can generalize well across different patient populations and clinical settings has become increasingly central to research in medical imaging.

One particularly relevant study tested whether models trained on a set of diseases could detect abnormalities from diseases they had never seen before. The results were promising, even though, performance got slightly worse for unseen diseases, it remained high overall, suggesting that deep learning models can learn general features of abnormality that transfer beyond specific cases[1]. This study highlights the potential of generalization at the disease level, but it doesn't address the more complex issue of generalizing across patient populations or clinical contexts.

The work by Almeida et al. concerns this challenge, as it focuses on detecting chronic obstructive pulmonary disease (COPD) in patients from different ethnic backgrounds[2]. The authors found that training on data from only one group (in that case African American patients) led to higher bias than training on a balanced dataset containing multiple ethnicities. This finding shows the importance of diversity in training data and the ethical and practical concerns related to models trained on narrow datasets in clinical contexts. It also reinforces the mentioned hypothesis that carefully mixing datasets can lead to more robust, fairer models.

A more direct parallel to this project comes from the work of Bajwa et al., who trained models using a combined dataset from Stanford (CheXpert dataset that this project uses too), NIH, and Shifa Hospital (in Israel)[3]. They found that the model trained on this merged data not only performed well on each of the individual test sets but also outperformed models trained on any one dataset alone. This suggests that generalization across different sources is not only possible but can be actively improved by blending data from diverse environments.

Yet the literature isn't optimistic in all cases . In another study, Oakden-Rayner et al. found that their model consistently performed best on test data from the same institution it was trained on[4]. Even when additional datasets were included in training, performance gains were mostly limited to those specific datasets, rather than leading to improved generalization overall. This shows that generazability is of models can be also dependent on datasets used, and impossible, or really hard to achieve in some cases.

These studies collectively offer an understanding of generalization in medical imaging. While some highlight effective strategies (for example, using balanced and diverse training data or some less popular models for training data) others showcase that regardless of specific model, performance often declines when applied to new data.

This project contributes to the ongoing discussion by systematically investigating how the chalenge of generazability plays out in the context of two datasets. This paper analyzes and compares performance on models on unseen data, and potential approaches to combining data to train one model. By focusing on two datasets not previously paired in related studies, CheXpert and PadChest, this paper aims to generate new insights into cross-dataset generalization in chest X-ray diagnostics.

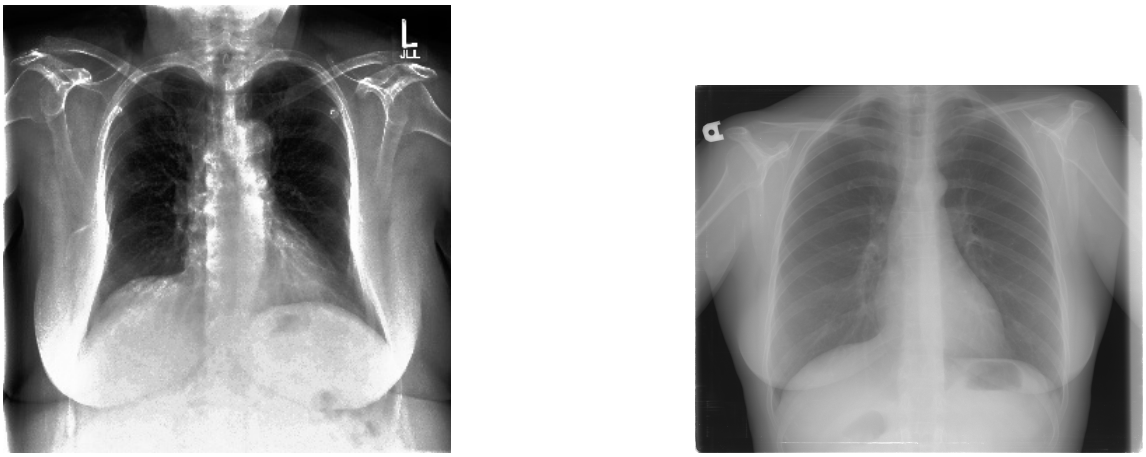
Set up and Methods

Datasets

As mentioned, this project uses 2 datasets: CheXpert and Padchest that are described below.

CheXpert

The CheXpert dataset is a widely used for developing and evaluating machine learning models in chest X-ray interpretation. It consists of 224,316 images from 65,240 patients and includes both frontal and lateral views of chests. The dataset features labels of diseases (from 14 types + "No Findings") with varying degrees of confidence and includes an evaluation set annotated by radiologists.[5]



(a) Example of image from CheXpert

(b) Example of image from Padchest

Figure 1. Images from used datasets

Padchest

PadChest is a X-ray dataset designed for advancing automated medical image interpretation. It contains over 160,000 images from 67,000 patients collected at Hospital San Juan in Spain between 2009 and 2017, encompassing six different view positions per instance. Each image is linked to a radiologist-generated report, which has been labeled with a rich set of annotations including 174 radiographic findings and 19 differential diagnoses. These labels are organized hierarchically and mapped to standard terminology. Dataset can be accessed after filling a form from BIMCV.[6]

Experiments

We use 3 experiments to evaluate the generazability of the models.

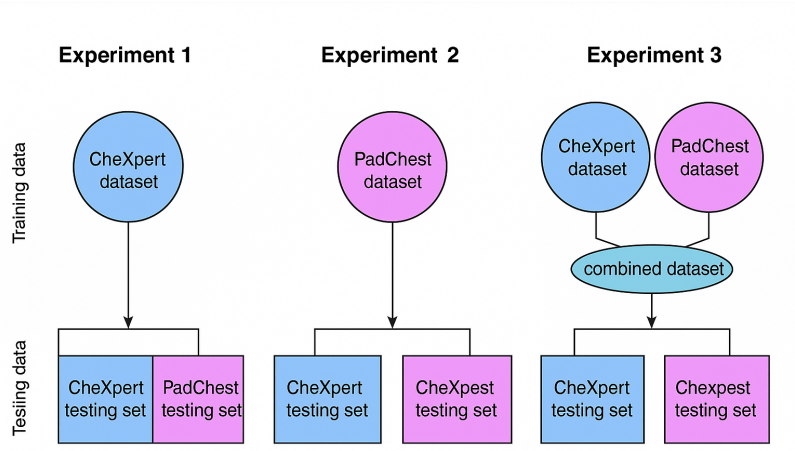


Figure 2. Experiments Setup

In the first experiment, we train a ResNet-18 model on the CheXpert dataset using a learning rate of 1×10^{-4} and a batch size of 32 (also tested for le $= 3 \times 10^{-5}$, le $= 3 \times 10^{-4}$ and batch sizes 16 and 64). The dataset is split into training, validation, and testing sets in a 70-15-15 ratio. Each of the sets is balanced, using undersampling from the original data. This means that all splits are balanced to contain an equal number of normal and abnormal chest X-rays (50-50 ratio). We evaluate this model on both in-domain (CheXpert) and out-of-domain (PadChest) test sets. This allows us to compare the model's performance on in-sample versus out-of-sample data. In the Github Repo, we can see this code in `experiment1.py`[7]

In the second experiment, we follow a similar approach and model design, but this time we train the model on the PadChest dataset and evaluate it on both PadChest and CheXpert. This enables us to assess how well a model trained on PadChest generalizes, compared to the model from the first experiment. In the Github Repo, we can see this code in `experiment2.py`[7]

In the third experiment, we combine the CheXpert and PadChest datasets into a single training dataset. We balance the combined dataset to have equal representation of normal and abnormal samples, as well as equal representation from both source datasets. We then train a model on this merged dataset and evaluate it on both CheXpert and PadChest. This setup helps us analyze the impact of combining datasets on generalizability and performance differences across the two domains. Code for this part is in `experiment3.py` and `experiment3_v2.py` (version 2 uses diffrent model design that will be explained later down the paper).[7]

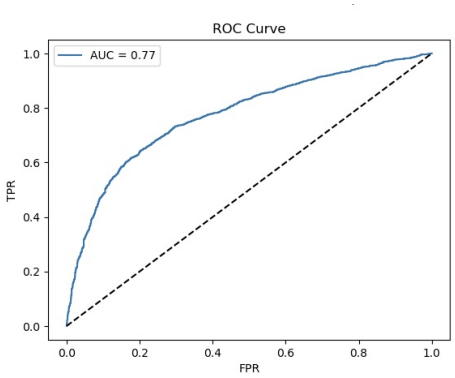
Results

Experiment 1

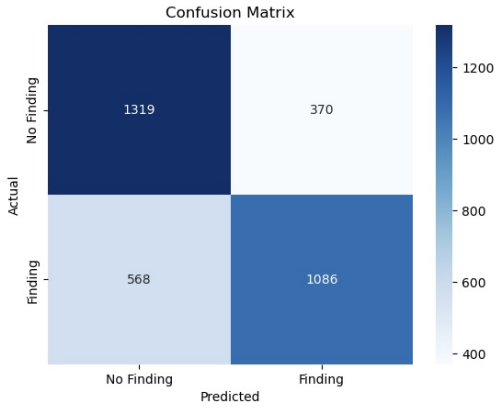
In the first experiment we get the following results.

Metric	CheXpert testing	Padchest testing
Accuracy	0.7194	0.5107
Precision	0.7459	0.5109
Recall	0.6566	0.9993
F1 Score	0.6984	0.6761
AUROC	0.7747	0.4740

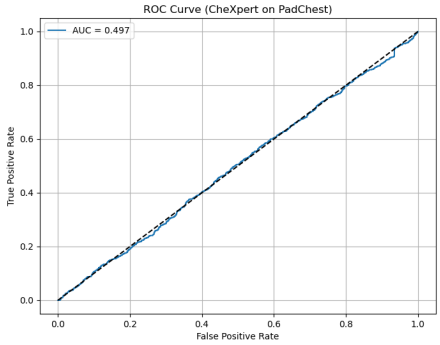
Table 1. Comparison of evaluation metrics both testing sets



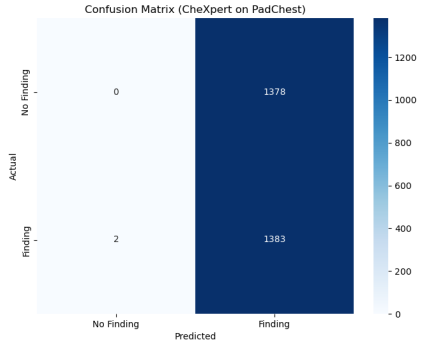
(a) ROC for CheXpert testing



(b) Confusion Matrix for CheXpert testing



(c) ROC for Padchest testing



(d) Confusion Matrix for Padchest testing

Figure 3. Results for experiment 1

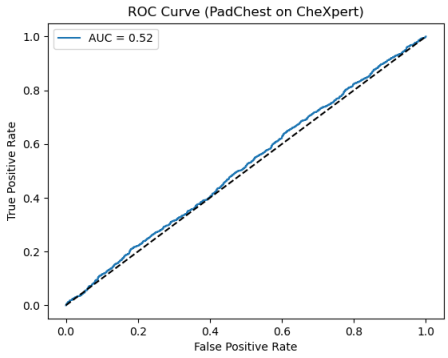
We can notice that in the case of in-domain evaluation, the model achieves 0.72 accuracy, which is not fully satisfactory, but promising. However, in the case of Padchest testing, the model is performing almost as badly as random choice, as it guesses "Findings" for almost all of the samples in the dataset. This results in 0.51 accuracy and almost perfect recall, but shows that the model is not able to perform accurately, and it didn't "learn" traits that would help it recognize the status of images in Padchest dataset.

Experiment 2

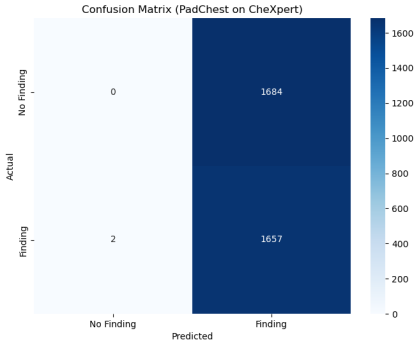
Now, we want to compare performance on the model trained on different dataset - Padchest.

Metric	Padchest	CheXpert
Accuracy	0.6124	0.5005
Precision	0.5938	0.5009
Recall	0.7013	0.9986
F1 Score	0.6431	0.6671
AUROC	0.6589	0.4927

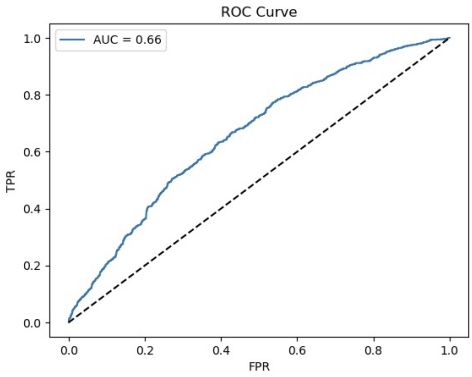
Table 2. Performance comparison between testing on CheXpert and Padchest data



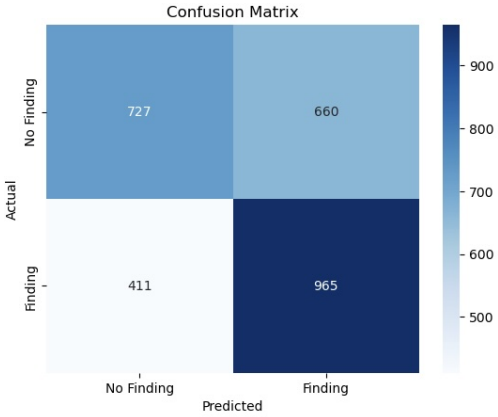
(a) ROC for CheXpert testing



(b) Confusion Matrix for CheXpert testing



(c) ROC for Padchest testing



(d) Confusion Matrix for Padchest testing

Figure 4. Results for experiment 2

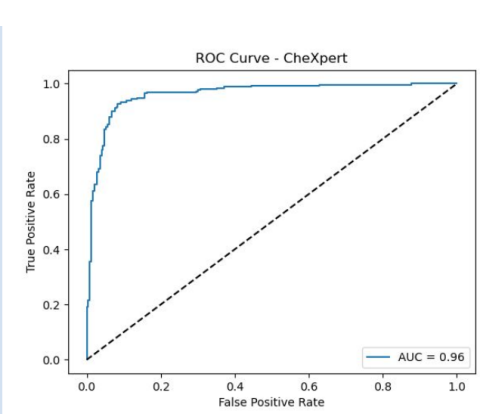
In the case of model trained on the Padchest data, we note 0.62 for accuracy in the in domain testing and 0.5 in out of domain testing. We should note that the accuracy for in-domain testing is lower than as in experiment 1. In addition, as in the experiment 1, here we also note that out of sample result is very similar to a random chance. This shows the similar problem with out-of-domain testing as in the first experiment.

Experiment 3

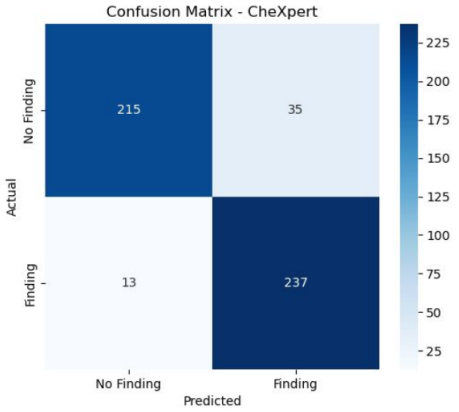
As none of the models trained on a singular dataset are able to generalize and perform well on multiple datasets for now the project aims to combine datasets and through it train a model that would perform better on data coming from multiple sources.

Metric	Result on CheXpert	Result on PadChest
Accuracy	0.904	0.500
Precision	0.870	0.000
Recall	0.948	0.000
F1 Score	0.910	0.000
AUROC	0.960	0.505

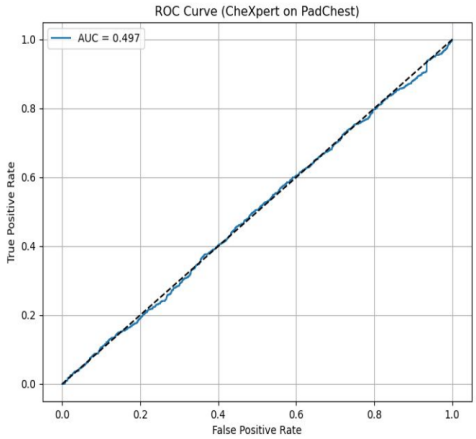
Table 3. Model performance comparison between CheXpert and PadChest datasets



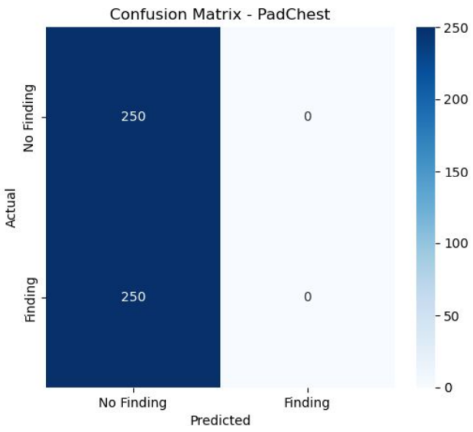
(a) ROC for CheXpert testing



(b) Confusion Matrix for CheXpert testing



(c) ROC for Padchest testing



(d) Confusion Matrix for Padchest testing

Figure 5. Results for experiment 1

In this case we see 0.9 accuracy in this case for the CheXpert dataset, but only 0.5 accuracy in the Padchest dataset. This result doesn't align with literature (as the papers claim that combining datasets leads to improved accuracy on in-domain data, so in this case on both of the datasets). This suggests we need to improve the architecture and set up. We use similar approach as described by Bajwa et al. to build the next experiment.[3].

Experiment 3 version 2

In the new model we implement data augmentation (to train model on less "similar" images as we include some rotations) and use the Dense-Net 121 Model. This model was introduced by Huang et al.[8] and it employs dense connectivity, where each layer receives input from all preceding layers which encourages feature reuse. We follow similar approach as in other experiments with balancing datasets for training, validation and testing data and testing multiple hyperparameters.

Metric	Result on CheXpert	Result on Padchest
Accuracy	0.7478	0.5986
Precision	0.7685	0.6210
Recall	0.7210	0.5159
F1 Score	0.7440	0.5636
AUROC	0.8090	0.6421

Table 4. Detailed metrics comparison between CheXpert and PadChest datasets for experiment 3 version 2

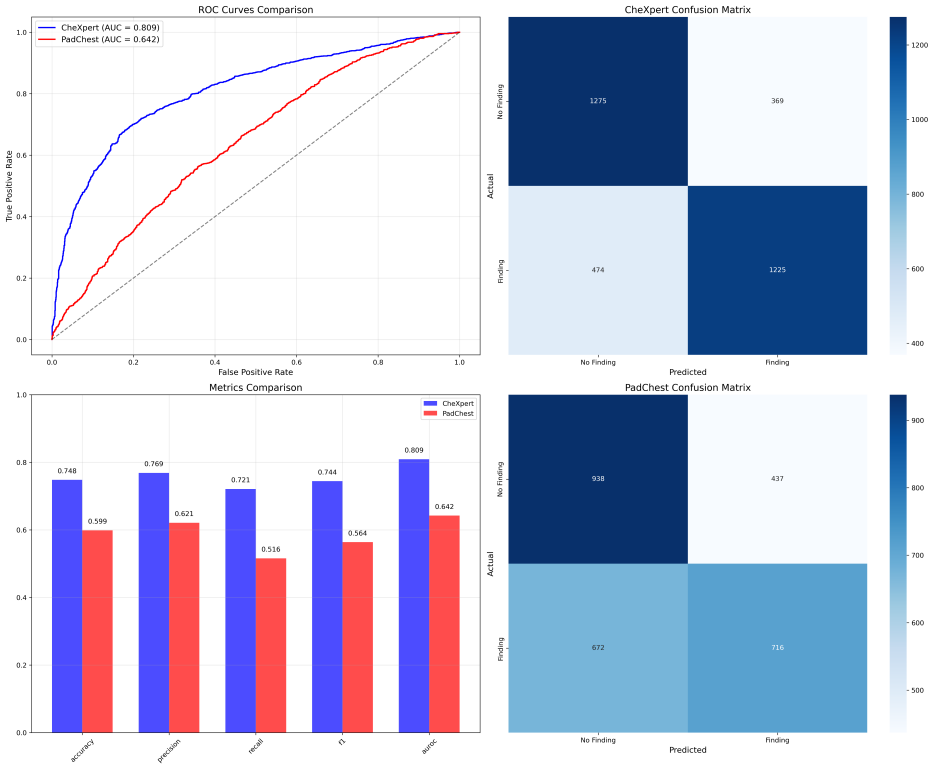


Figure 6. Results for Dense-Net 121 model

In this case, we can note that the performance of the model is more similar across two of the datasets. On CheXpert the model reaches the accuracy 0.75 and on Padchest around 0.6. This means worse performance for CheXpert than in the first version of this experiment but better performance for Padchest. Notably, this model reaches better than random performance for both datasets. Additional discussion about this experiment is presented in the demo video.

Discussion

Our results confirm a hypothesis about the same-source advantage: models trained on CheXpert and PadChest each achieved substantially higher performance on their respective in-domain test sets than on the external dataset. In Experiment 1, the ResNet-18 trained on CheXpert achieved an accuracy of 0.72 and AUROC of 0.77 on CheXpert test images, but fell to near-random performance (accuracy around 0.51, AUROC around 0.47) on PadChest. Conversely, in Experiment 2, the ResNet-18 trained on PadChest achieved 0.61 accuracy and 0.66 AUROC in-domain, but again collapsed to chance (accuracy around 0.50) on CheXpert. This clear decrease in out-of-sample performance shows that the dataset-specific biases are clearly present in standard CNN models. This aligns with prior findings that institution-specific images style and patient demographics can dominate learned features and by this hinder transferability of the models for other cases.

The attempt to improve cross-domain generalization by simply merging the two datasets (Experiment 3 version 1) actually increased this imbalance: although in-domain CheXpert accuracy rose to 0.90, performance on PadChest remained at chance. This suggests that naive dataset concatenation may bias the model toward the model that is "easier" to train rather than encouraging a balanced feature representation. Only after introducing extensive data augmentation and switching to a DenseNet-121 architecture (Experiment 3 version 2) did we observe more similar performance: CheXpert accuracy stabilized at 0.75 with AUROC 0.81, while PadChest accuracy reached 0.60 with AUROC 0.64. This improvement indicates that richer connectivity patterns (as in DenseNet) and training on augmented images better equip the model to learn invariant features across protocols and populations. We should however note, that the performance of this experiment is worse than in the case of cited paper. In that case, across 3 used dataset we have accuracy between 0.8 and 0.84. This means that the generazability is not only impacted by the model selection, but also the specific datasets used.

It should be noted that despite the gains in experiment 3, the gap between datasets remains undeniable. PadChest performance, although improved, still is behind CheXpert. This may reflect differences in image resolution, label noise, and disease popularity and severity between Spain and the US. Moreover, our binary normal–abnormal label can lead to struggles with understanding different forms of "findings" that can impact understanding of what constitutes normal and abnormal.

If we go back to the hypothesis, we can confirm the first one - there is significant advantage in being a part of the training test. In terms of the second hypothesis, we can notice a sort of the trade-off, that based on the model selection, we can either reach great performance on one of the datasets or have moderate performance on both. This shows the difficulty with creating models that will perform well on diverse data, they need to differ between the traits that indicate the image label, from differences related to belonging to different datasets.

Conclusions

The work shows the difficulties with building models that will generalize on new datasets. In our case the difference in quality in images, and type of abnormalities that were presented on the images. This means that without adding the second dataset, none of the models was able to perform visibly better than the random choice.

After adding the second dataset, based on which of the models was decided we either achieve great performance for one of the datasets (in version 1 of experiment 3) or more similar, medium-level performance in both of them. This again connects the choices made in terms of model design to the key ethical and practical questions related to populations we want to help, and the places where accurate models are needed the most.

Further Work

The findings of this study suggest several important directions for future research. First, while our experiments confirmed the limited generalizability of models trained on single datasets, they also highlighted the potential of architectural changes and data augmentation to improve cross-domain performance. Building on this, future work could explore more advanced domain adaptation techniques that could explicitly reduce the distribution gap between datasets like CheXpert and Padchest.

Secondly, our study used a binary classification framework (normal vs. abnormal), which, while simplifying the problem, may obscure more nuanced generalization failures across specific disease categories. Future research should consider multi-label or hierarchical classification approaches that reflect the real-world complexity of chest pathology. In this case, it would be necessary to reliably map labels from one dataset to another.

Third, it was observed that combining datasets did not consistently yield the expected improvement in generalization. To better understand and control for this, future work could include detailed dataset analysis—quantifying label distributions, imaging protocols, and demographic composition—to develop sampling strategies that promote balance not only in label frequency, but also in clinical and technical characteristics. Future work should therefore explore multi-label classification or hierarchical disease taxonomies to disentangle these factors and more precisely target transfer failures. Additionally, while DenseNet-121 outperformed ResNet-18 in the mixed-data setting, other state-of-the-art models (such as Vision Transformers or self-supervised pretraining on large-scale unlabeled medical data) may further enhance generalizability. These models may be better suited to learning domain-invariant features, particularly when trained with contrastive or masked image modeling objectives.

Finally, while we focused on CheXpert and PadChest, a broader evaluation on a wider range of datasets could strengthen the external validity of the conclusions. Inclusion of additional metadata (like scanner type, patient age, or hospital location) could also help analyzing impacts of those on generazability.

Ultimately, developing models that generalize across institutions, populations, and imaging conditions is essential to fulfill the potential of AI-assisted diagnostics. This work is an initial step, and future efforts must go beyond accuracy to prioritize fairness, interpretability, and robustness in real-world clinical use.

References

1. Lakhani, P. & Sundaram, B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* **284**, 574–582 (2017).

2. Almeida, S. How do deep learning models generalize across populations? a cross-ethnicity investigation of generalization in copd detection. *Insights Imaging 15, 198* **15** (2024).

3. Bajwa, N. *et al.* A generalized deep learning model for multi-disease chest x-ray diagnostics (2020).

4. Oakden-Rayner, L., Dunnmon, J., Carneiro, G. & Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *arXiv* (2019).

5. Chexpert-v1.0-small dataset. *https://www.kaggle.com/datasets/ashery/chexpert/discussion?sort=undefined* (2020).

6. Padchest dataset. *https://bimcv.cipf.es/bimcv-projects/padchest/* (2017).

7. Github repository with project code. *https://github.com/nsiwek1/neuro140project.git* .

8. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. 4700–4708 (2017).