Towards Generalizable Machine Learning for Chest X-ray Diagnosis with Multi-task learning

Salah Ghamizi [1] , **Beatriz Garcia Santa Cruz**[2] , **Paul Temple**[3] , **Maxime Cordy**[1] , **Gilles Perrouin**[3] , **Mike Papadakis**[1] and **Yves Le Traon**[1]

[1]University of Luxembourg; [2]Centre Hospitalier de Luxembourg; [3]University of Namur

{firstname.lastname}@uni.lu, garciasantacruzbeatriz@gmail.com, {firstname.lastname}@unamur.be

Abstract

Clinicians use chest radiography (CXR) to diagnose common pathologies. Automated classification of these diseases can expedite analysis workflow, scale to growing numbers of patients and reduce healthcare costs. While research has produced classification models that perform well on a given dataset, the same models lack generalization on different datasets. This reduces confidence that these models can be reliably deployed across various clinical settings. We propose an approach based on multitask learning to improve model generalization. We demonstrate that learning a (main) pathology together with an auxiliary pathology can significantly impact generalization performance (between -10% and +15% AUC-ROC). A careful choice of auxiliary pathology even yields competitive performance with state-of-the-art models that rely on fine-tuning or ensemble learning, using between 6% and 34% of the training data that these models required. We, further, provide a method to determine what is the best auxiliary task to choose without access to the target dataset. Ultimately, our work makes a big step towards the creation of CXR diagnosis models applicable in the real world, through the evidence that multitask learning can drastically improve generalization.

1 Introduction

The recent success of Machine Learning (ML) techniques for image analysis has extended to medical imaging, notably to assist radiologists when diagnosing pathologies. Yet, medical data solutions present two specific challenges that have to be solved before these solutions reliably run in clinical settings. First, medical images differ from classical ML datasets as they present lower contrast and signal-to-noise ratio, higher dimensionality and increased uncertainty in the labelling process [Bruno *et al.*, 2017]. Second, medical data tend to be not only scarce but also partially represented [Varoquaux and Cheplygina, 2021].

These challenges affect model *generalizability*. Generalizability is essential in the medical domain because practitioners need models that provide stable predictions and can efficiently adapt to new clinical settings (e.g. different hospitals, different population, etc.) at an affordable computation cost. Lack of generalization hampers the safe and accurate translation into clinical trials [Beede *et al.*, 2020].

Experimentally, a common way to assess model generalizability is to train a model on the training set of a *source* dataset (which represents the original population the model learns from) and evaluate its performance on the test set of a target dataset (which represents the population the model will be applied to) [Yao *et al.*, 2019]. In such settings that violate the i.i.d. data assumption [D'Amour and Heller, 2020], models hardly generalize – i.e. a model trained to predict pathologies on the source population has poor performance on the target population. This contrasts with most research works that benchmarks medical ML models on a single dataset, *e.g.* by winning the CheXpert competition [Irvin *et al.*, 2019] [1].

Our research aims to improve the study of model generalization in medical image analysis, in particular chest radiographs (CXRs), and brings three contributions. Our first contribution is the demonstration that learning multiple pathologies together – in a multi-task model – can significantly increase or decrease generalization performance (-10% to +10% AUC-ROC) on each individual pathology. To measure this, we train ML classifiers on all pairs of pathology among seven (using the source dataset) and assess their performance on the target dataset. Through our experiments, we reveal that some pathologies consistently improve generalization regardless of the pathology they are joint with. Based on these results, we propose *Auxiliary Pathology Learning*, a method to improve medical model generalization via a multi-task model that simultaneously learns a main pathology of interest with a well-chosen auxiliary pathology.

Our second contribution is to show that Auxiliary Pathology Learning, when it selects the most appropriate auxiliary tasks, achieves competitive generalization performance using a much lower number of data – viz. only 6% to 34% of the data that state-of-the-art approaches based on fine tuning and ensemble learning require.

Given the above, our third contribution is an approach to select an auxiliary pathology. We demonstrate that there is a strong correlation between the generalization performance

[1] https://stanfordmlgroup.github.io/competitions/chexpert/

of a multi-task model (on the given main pathology in the target dataset) and the test performance (in the source dataset) when fine-tuning a model from the auxiliary pathology to the main pathology. This means that one can determine the best auxiliary pathology (for a given main pathology) using only the source dataset.

Finally, we discuss factors (*e.g.*, availability of certain medical features) that can influence model generalizability. We evaluate the similarity between the layers of our models and identify patterns that could relate to better generalization.

2 Related Work

Domain shift and generalization. In domain shift literature, the challenge is to learn a machine learning model that can generalize to unseen data distributions and test environments. It has been thoroughly studied outside the medical context, in particular for computer vision tasks [Wang *et al.*, 2021].

The naive baseline is to use empirical risk minimization (ERM) to learn on a mixture of data across all training environments. Recent approaches involve augmenting the training data, for instance using adversarial data augmentations [Zhou *et al.*, 2020], or optimizing the learning strategies: Ensemble Learning [Wu and Gong, 2021], Meta-Learning [Kim *et al.*, 2021], and Self-supervised Learning [Carlucci *et al.*, 2019]. Meanwhile, Representation Learning has proven to be fruitful in increasing generalization performances. It encompasses techniques that aim for domain-invariant representation learning [Hu *et al.*, 2019; Jin *et al.*, 2020; Mitrovic *et al.*, 2020] and techniques that investigate feature disentanglement [Peng *et al.*, 2020].

Domain generalization in medical imaging. Domain shift and its impacts on generalization performance are even more acute when dealing with medical imaging context. A model trained on hospitals in one region may be deployed to another, but due to domain shift (*e.g.*, differences in the age of patients), prediction performances may drop leading to erroneous diagnoses. Tackling Chest X-ray classification, Zhang *et al.* [Zhang and al., 2021] investigate how subsampled datasets with varying label prevalence between genders can impact the generalization capabilities of the models and their impact on fairness and bias. Pooch *et al.* [Pooch *et al.*, 2020] evaluate how well models trained on a hospital-specific dataset generalize to unseen data from other hospitals. While it uncovers the drop in generalization performance of the models, it does not provide any insight into the causes or how to mitigate this drop (except by selecting one dataset over another). Cohen *et al.* [Cohen *et al.*, 2020] study the cross-domain performance and agreement between models. They identify discrepancies between performance and agreement of models. Then, they provide insights on the representation changes from one dataset to another. While their evaluation is the most exhaustive, it promotes the common assumption that generalization is attained by mixing different datasets, either through training, or ensemble. Our work is parallel to theirs, we do not study the latent representation change from *one dataset* to another, but the latent representation change from a *combination of pathologies* to another.

To the best of our knowledge, our work is the first to investigate the impact of multiple pathology learning and its implications on the test and generalization performances. Our work serves as a guidance for practitioners on how to target the pathologies to efficiently increase generalization performance with limited additional data.

3 Methods

3.1 Auxiliary Pathology Learning

We propose to envision multilabel CXR classification as a multitask binary problem where the main task is our target pathology and the auxiliary task is the pathology used for augmentation. Labelling with *0* means the pathology is not found, while *1* indicates the pathology is present. All pathologies share a common encoder, that extracts the features most relevant to the set of pathologies, and each pathology has a dedicated decoder that learns pathology-specific weights and outputs the final probability of this pathology. This approach offers two main advantages over multi-label learning: 1. each pathology can learn independent weights from the other pathologies with its dedicated decoder, and 2. the binary cross-entropy loss is computed for each head independently when the state of its pathology is certain (*1* or *0*) and back-propagated through all the network. When the label of the pathology is uncertain, which is common in CheXpert dataset [Irvin *et al.*, 2019], the loss is not back-propagated through its head (instead of being back-propagated as *0* in common multilabel classification, or as a third class *2* in others).

We denote by **main pathology** the pathology we aim to evaluate on the target population and **auxiliary pathology** the secondary pathology we include in as an auxiliary task.

3.2 Problem Definition

Formally, we consider CXR image x. We denote by $\bar{y}$ the corresponding ground-truth label, defined as $\bar{y} = (y_1, ..., y_i, y_M)$ where y_i is the corresponding ground truth for task i (*i.e.*, pathology i) for an input x. For a given population, x and $\bar{y}$ are drawn from some joint distribution $p(x, \bar{y})$.

Let $\mathscr{M}$ be a multitask model with tasks $\mathscr{T} = \{t_1, ..., t_M\}$. $\mathscr{M}$ is trained to estimate $p(\bar{y} \mid x)$ but may not generalize well when the joint distribution changes. For instance, when the population of patients changes, or the collection protocol (hospitals, machines, ...) varies. We hypothesize that $p(\bar{y} \mid x)$ is not consistent across datasets, and we propose to consider a more fine-grained problem: Training a multitask model $\mathscr{M}$ to learn to estimate $p(y_j \mid x, y_k)$ where $t_j, t_k \in \mathscr{T}^2$ two pathologies that are learned together by our model $\mathscr{M}$. Using an auxiliary pathology to learn a main one could mitigate the covariate shift that [Cohen *et al.*, 2020] previously uncovered.

3.3 Datasets

We run our evaluation on three large public chest X-ray datasets. NIH Chest X-ray14 [Wang *et al.*, 2017] (NIH) is a dataset of 112k images partially labelled automatically with the NegBio labeller. *CheXpert* [Irvin *et al.*, 2019] (Chex), a set of 224k chest radiographs labelled with a custom automated labeller over the NLP analysis of radiology reports. *PadChest* [Bustos *et al.*, 2020] (PC) is a 160k image dataset where, for 27% of them, the labels are extracted from radiographic reports manually annotated by trained physicians.

We restrict our evaluation to the seven pathologies common to the three datasets: Atelectasis (ATE), cardiomegaly (CAR), consolidation (CON), edema (EDE), effusion (EFF), pneumonia (PNE), and pneumothorax (PTX). The datasets display a diverse set of pathology distribution , gender balance and distribution of age of the evaluated subjects.

For each dataset, we restrict our study to the erect anteroposterior chest views (AP views) images. We use 80% of the images for the training and 20% for testing the performance as proposed in [Cohen *et al.*, 2020]. We present in Appendix A the properties of both datasets and their pathologies.

3.4 Network and training

Encoders We focus our evaluation on ResNet50 encoders, as they are commonly used in the Chest X-ray literature [Baltruschat *et al.*, 2019; Bressem *et al.*, 2020].

Decoders For each dataset, we train single-pathology decoders (7 in total) and pairwise decoders (21). In addition, each of the single-pathology model is fine-tuned on each other pathology, with and without encoder freezing ($7{*}6{*}2 = 84$). Over the 3 datasets, we trained and evaluated a total of 336 models ($(7 + 21 + 84) \times 3$).

Training Models are trained for 250 epochs with a learning rate of 0.001 using Adam optimizer as proposed by [Cohen *et al.*, 2020]. When multiple decoders are used, they are weighted equally, following common practice [Zamir *et al.*, 2018]. The impact of task weighing on robustness is outside the scope of this work, but has been covered by [Ghamizi *et al.*, 2021]. We used a data augmentation using center cropping, a 15° random rotation, a 15% random scaling and translation [Cohen *et al.*, 2020]. For pathology decoders, the loss function is a binary cross-entropy.

For each dataset, we obtain the 84 fine-tuned models by starting from the best performing models on an auxiliary pathology, then training it for an additional 10 epochs on the main pathology following the same protocol as above.

3.5 Metrics

Model performance We report in our main evaluation the performance of the models using the area under the ROC curve metric (AUC) as it is the most standard metric for unbalanced binary classification [Cohen *et al.*, 2020; Irvin *et al.*, 2019]. We provide in Appendix B figures using the full ROC curves including the FPR and the FNR values.

Hidden representation analysis Neural network hidden representations are challenging to analyze because of the distribution of the neurons and their activation and interactions across the layers.Kornblith *et al.* [Kornblith *et al.*, 2019] proposed the **centered kernel alignment (CKA)** that provides a reliable quantitative measure of the similarity of neural network representation. We provide in appendix D a detailed explanation of this metric and how to compute it.

4 Results

4.1 Auxiliary Pathology Learning

Generalization performance is pathology-specific. Hereafter, we show that the choice of combinations of pathologies (*i.e.*, auxiliary learning) may impact the test target performances. For each of the 7 pathologies, we train, on CheXpert, a model to predict a pair of pathologies and evaluate their prediction performance (*i.e.*, AUC) on CheXpert (source) and on NIH (target). Appendix B gives statistics about the performance of the models for all the combinations. Overall, the standard deviation between pathology combinations is up to 8 times higher for the test performance on target than the test performance on the source (EDE). Edema and pneumothorax show the largest variance of performance on target when they are learned with another pathology. Due to space restrictions, we focus on these two pathologies and provide the remaining results in Appendix B. Figure 1(a,c) show the ROC curves of the source test performance while Figure 1(b,d) show the ROC curves of the target test performance.

Figure 1(a,c) show the selection of a combination of pathologies solely based on the source test performance is difficult due to very little performance differences.However, in Figure 1(b,d), larger differences are shown clearly suggesting that a combination is preferable to another. For instance, the source test performance of edema learned with Consolidation is 1% lower than the one of edema learned with effusion; but the difference is reversed when considering the target test performance: Edema with consolidation gets a 15% higher target test performance than edema with effusion. In the end, trying to learn with the combination edema-consolidation is likely to be more beneficial than edema-effusion when looking for a model that performs better on a target dataset but defining the right choice requires prior analysis.

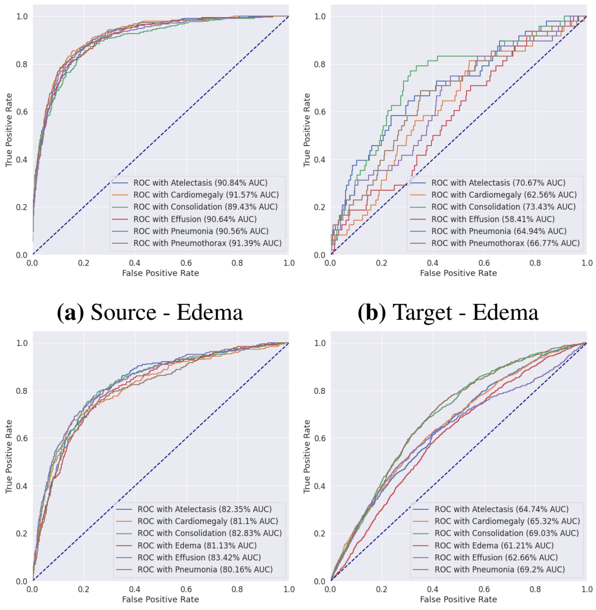
Some pathologies consistently improve generalization. Our previous evaluation shows that we cannot rely on source test performance to choose the best auxiliary pathology, we investigate however if given one training dataset (CheXpert), we can identify some auxiliary pathologies that consistently improve generalization on another dataset (NIH or PC).

We study the change in target test AUC for pairwise models trained on CheXpert dataset and evaluated on NIH dataset in Table 1 and models evaluated on PC in Table 2.

Our evaluation demonstrates that consolidation (CON) is the best auxiliary pathology to improve the generalization of models trained using the CheXpert dataset. It is the best in 5 out of 6 combinations to generalize to NIH and 3 out of 6 combinations to generalize to PC. Even when it is not first, it is a close second in the remaining combinations of Table 1.

For models trained on the NIH dataset and generalized to CheXpert or to PC datasets, we provide a complete evaluation in appendix B. Overall, effusion improves the best atelectasis, consolidation and cardiomegaly show the best improvement for each other, and effusion improve the best pneumonia.

Conclusion: Practitioners should not trust the source test performance to select the pathology combinations that maximize generalization performance.



(a) Source - Edema

(b) Target - Edema

(c) Source - Pneumothorax

(d) Target - Pneumothorax

Figure 1: ROC curves of source performance (CHEX→CHEX) and target performance (CHEX→NIH) for edema (top) and pneumothorax (bottom) when learned with the 6 other pathologies.

Main	ATE	CAR	CON	EDE	EFF	PNE	PTX	AVG
Aux								
ATE	0.00	0.65	1.26	1.99	-0.80	-1.05	8.04	1.44
CAR	3.73	0.00	3.31	-9.72	-2.76	-4.47	9.01	-0.13
CON	**4.97**	**0.81**	0.00	**5.97**	**-0.38**	**1.16**	15.20	**3.96**
EDE	3.86	-5.10	5.27	0.00	-0.96	-5.20	2.15	0.00
EFF	3.79	-4.40	**8.15**	-15.71	0.00	-3.22	4.57	-0.97
PNE	3.57	-4.42	3.58	-6.27	-3.77	0.00	**15.48**	1.17
PTX	1.84	-4.81	2.99	-3.64	-4.63	-4.08	0.00	-1.76
AVG	3.11	-2.47	3.51	-3.91	-1.90	-2.41	7.78	0.53

Table 1: Target AUC (Chex→NIH). In bold and underline the best (respectively worst) auxiliary pathology (AUX) per column.

4.2 Auxiliary Pathology Learning is competitive with SoTA models

We compare test and generalization performances of SoTA models and our proposed approach (column **AUX**) in Table 3. SoTA models are provided and pre-trained by Cohen *et al.* [Cohen *et al.*, 2020]. They introduce different strategies to improve performance: **DenseNet** combines all pathologies and uses a densenet-121 architecture. Bressem *et al.* [Bressem *et al.*, 2020] have shown that Densenet121 is the best performing architecture for the classification of chest radiographs. It remains more expensive to train than Resnet-50. **EnsembleNet** uses an ensemble of 30 densenet-121 models. Both models are trained on all the pathologies of the CheXpert dataset (*i.e.*, about 223k images). Model **MixtureNet** uses a ResNet-50 model, but is trained on a mixture of all the datasets (cf. Appendix A) and all the pathologies, totaling a training set of 900k images. The rightmost column of Table 3 refers to the number of images needed by Auxiliary Pathology Learning to achieve the performance in column **AUX**.

Main	ATE	CAR	CON	EDE	EFF	PNE	PTX	AVG
Aux								
ATE	0.00	**4.57**	2.03	-1.66	-0.70	7.52	-18.20	-0.92
CAR	0.44	0.00	7.11	-1.00	-0.92	12.96	-23.54	-0.71
CON	-1.78	4.06	0.00	1.53	**0.40**	**17.54**	**-11.54**	1.46
EDE	-1.76	0.36	5.84	0.00	-0.39	9.74	-20.48	-0.96
EFF	**1.23**	3.43	**11.72**	-0.90	0.00	-0.72	-15.13	-0.05
PNE	-0.11	0.15	2.34	1.19	-6.60	0.00	-19.29	-3.19
PTX	-4.36	1.05	5.53	**1.59**	-1.17	8.07	0.00	**1.53**
AVG	-0.91	1.95	4.94	0.11	-1.34	7.87	-15.45	-0.41

Table 2: Target AUC (Chex→PC).

All techniques suffer from generalization performance drop (up to 20% for atelectasis pathology), including ours, but it needs a fraction of the training budget. We outperform model DenseNet in the classification of cardiomegaly, edema and effusion using respectively only 24%, 28% and 32% of its training data. We outperform model A in the classification of atelectasis, consolidation, edema, effusion and pneumothorax using respectively only 6%, 6%, 7%, 8% and 14% of its training data. EnsembleNet still outperforms our approach, but it requires (1) training expensive ensemble models, and (2) 3 to 4 times more data than our approach.

Fine-tuning requires the same amount of data as our approach. Table 3 shows that fine-tuning provides competitive test performances, however, the generalization performance is always below our approach.

While our approach is designed to minimize the training data collection process and cost, it can be used in combination with the strategies proposed in DenseNet, EnsembleNet or MixtureNet when more data are available. Designing optimizations over the datasets or the pathologies to include, and the ensemble to build is a natural follow-up of this work.

Conclusion: Auxiliary Pathology Learning achieves similar generalization performance to SoTA models with a fraction of training data (between 6% and 34%).

4.3 Fine-tuning can guide Auxiliary Pathology selection

We demonstrated in Section 4.2 that Auxiliary Pathology Learning outperforms fine-tuning. We hypothesize, however, that both approaches are related: Source pathologies most beneficial to fine-tuning are the pathologies most beneficial to Auxiliary Pathology Learning. Hence, we can use fine-tuning performance to guide the selection of the best auxiliary pathology to be used within our Auxiliary Pathology Learning framework.

To confirm our hypothesis, we evaluate the generalization performances of single pathology models pre-trained on a source pathology, then fine-tuned on a target pathology. We compare two models: when the full model is fine-tuned (table 4); and when the encoder is frozen and only the decoder is fine-tuned (table 5). The tables report the relative change of AUC compared to full training on the target pathology.

For both approaches, models fine-tuned for atelectasis or cardiomegaly have always lower generalization performances than fully trained models. Besides, fine-tuning from consolidation to edema or to pneumothorax leads to the highest increase of performance, mirroring the results in Table 1.

Main	Test AUC % (Source)					Test AUC % (Target)					AUX	
Pathology	DN	EN	MN	FT	**AUX**	DN	EN	MN	FT	**AUX**	Best combination	# images
Atelectasis (ATE)	90.53	93.07	78.34	86.88	89.59	69.95	71.10	68.26	67.10	69.25	ATE + CON	54.902
Cardiomegaly (CAR)	89.04	91.00	92.25	85.12	89.71	73.83	76.64	77.47	67.84	76.88	CAR +CON	54.758
Consolidation (CON)	87.64	90.99	75.77	84.97	86.74	71.20	74.64	69.11	68.56	71.13	CON + EFF	72.034
Edema (EDE)	90.48	92.62	78.05	88.55	91.57	71.27	82.44	65.82	72.85	75.91	EDE + CON	63.438
Effusion (EFF)	91.22	95.59	85.56	90.84	93.52	80.75	83.71	80.33	81.02	83.15	EFF + CON	72.034
Pneumonia (PNE)	84.94	N/A	65.78	77.29	80.06	63.27	N/A	62.12	60.81	62.50	PNE + CON	90.009
Pneumothorax (PTX)	82.10	N/A	82.82	80.25	83.42	73.81	N/A	59.29	67.34	68.17	PTX + PNE	124.230

Table 3: Comparison of AUC of SoTA models and our AUX approach. The first column denotes the target pathology, the following columns report Area Under Curve of each of the models for test performances on Source and Target datasets. *FT* stands for fine-tuning, *DN* for DenseNet, *EN* for EnsembleNet, and *MN* for MixtureNet. The right part presents the best performing pair for each pathology using our approach (*AUX*) and how much training data is required.

Main Auxiliary	ATE	CAR	CON	EDE	EFF	PNE	PTX
ATE	0.00	-13.77	**5.23**	-0.85	**-2.29**	-1.99	-5.67
CAR	-8.85	0.00	3.88	-15.87	-15.27	-0.47	-0.72
CON	-3.86	-9.25	0.00	**9.06**	-4.52	1.04	**7.47**
EDE	-3.46	**-7.44**	2.98	0.00	-12.61	1.79	-2.07
EFF	**-0.31**	-13.33	4.11	-0.25	0.00	1.40	3.59
PNE	-0.65	-8.05	-2.62	-0.84	-8.79	0.00	-7.50
PTX	-4.04	-15.42	-7.95	-5.96	-8.18	**5.46**	0.00

Table 4: (Chex→NIH) AUC change with a full model fine-tuning. The whole training is done on CheXpert and the evaluation on NIH. The model is pre-trained on the auxiliary pathology (row) then fine-tuned on the main pathology (column). The values are relative changes to the diagonal, where the models are pre-trained and fine-tuned on the main pathology.

Main Auxiliary	ATE	CAR	CON	EDE	EFF	PNE	PTX
ATE	0.00	-5.07	0.38	0.17	-1.59	1.18	-1.42
CAR	-0.21	0.00	-0.17	-0.55	-1.84	**1.25**	-0.49
CON	-0.36	-4.34	0.00	-0.13	**0.39**	0.66	0.26
EDE	-0.22	-6.25	0.18	0.00	-0.83	0.79	**1.30**
EFF	**-0.03**	-7.69	**0.45**	-1.64	0.00	1.02	-1.16
PNE	-0.07	**-1.61**	-0.36	**0.85**	-0.09	0.00	-0.23
PTX	-0.66	-6.25	-0.56	-0.59	-0.17	0.38	0.00

Table 5: (Chex→NIH) AUC change with a decoder fine-tuning. The whole training is done on CheXpert and the evaluation on NIH. The model is pre-trained on the auxiliary pathology (row) then **only** the decoder is fine-tuned on the main pathology (column).

	AUC 1	AUC 2	Corr	p-value
(a)	NIH→NIH CHEX→CHEX	NIH→CHEX CHEX→NIH	0.14 0.06	0.37 0.70
(b)	CHEXFinetune CHEXFreeze	CHEX→NIH CHEX→NIH	0.34 0.11	0.03 0.45
(c)	ΔTARGETFinetune ΔSOURCEFinetune	CHEX→NIH CHEX→NIH	0.57 0.47	7.1e-5 1.6e-3

Table 6: Spearman correlations between different AUC changes

We deepen this evaluation with statistical analysis in Table 6. In Table 6(a), we evaluate the Spearman correlation between the change of performance with an auxiliary pathology on source test performance (AUC1) and the change of performance on target test performance (AUC2). There are no clear correlations (r<0.2; p-value>5%)between pairs that improve source test performance and target test performance.

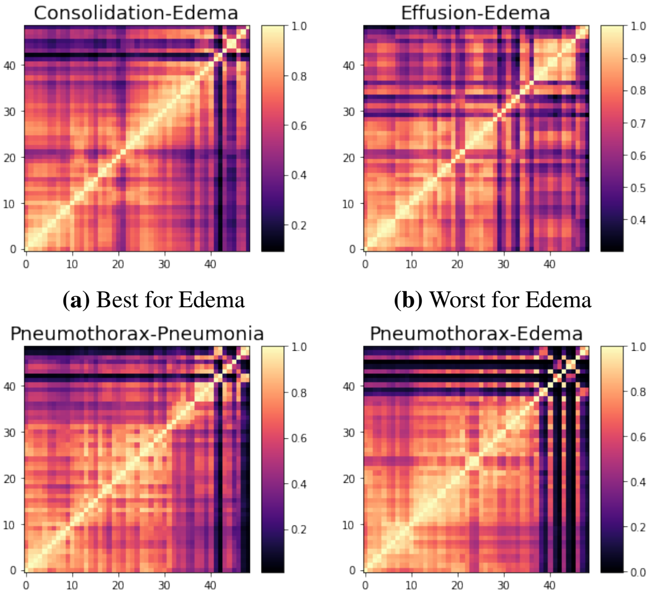
In Table 6(b), we evaluate the correlation between the change of performance with pre-training on the auxiliary pathology then fine-tuning on the target pathology (AUC1) and the change of generalization performance when training with both target and auxiliary pathology (AUC2). Fine-tuning a full model from the auxiliary pathology to the main pathology has a moderate correlation with the generalization performance of the combination of the main and auxiliary pathologies (r=0.34), while fine-tuning the decoder only shows no correlation.

Finally, we compute ΔTARGETFinetune, the difference between raw AUC values obtained in Table 4 and Table 5. It reflects the impact of fine-tuning the encoder part on the generalization performance of the fine-tuned models. We evaluate in Table 6(c) the correlation between ΔTARGETFinetune and the generalization performance with Auxiliary Pathology Learning framework. This evaluation yields a high correlation (r=0.57; p-value<5%). Similarly, we compute ΔSOURCEFinetune, the difference between AUC on the source dataset between fully fine-tuned models and decoder-only fined tuned models. The correlation between ΔSOURCEFinetune and the generalization performance of our framework remains high (r=0.47; p-value<0.05).

Conclusion: Practitioners can use fine-tuning performances either on source or target dataset to select the best combinations of pathologies to achieve the highest generalization performances with Auxiliary Pathology Learning.

5 Discussion

Our results demonstrate the key role of pathology interactions when training CXR models on a source population to generalize to a different target population. In this section, we provide insights into this phenomenon and discuss the impact of our method in medical practice. We analyze the interactions between pathologies through the prism of layer similarity: Using the CKA metric [Kornblith *et al.*, 2019], we evaluate the patterns of the activation of each of the 50 layers of our models. In the following, we focus on edema and pneumothorax, previously discussed in Figure 1.



(a) Best for Edema

(b) Worst for Edema

(c) Best for Pneumothorax

(d) Worst for Pneumothorax

Figure 2: Layer Similarity for best (left) and worst (right) generalization performances (Chex→NIH). The main pathology of the top row is edema and the main one for the second row is Pneumothorax.

Figure 2 shows in (a) and (c) the CKA for the combinations that lead to the highest generalization performance, while (b) and (d) report the CKA for the combination with the lowest generalization performance.

In these figures, large block structures (*e.g.*, from layer 22 to 42 in (a) or from layer 0 to 30 in (c)) suggest that these parts of the models rely on similar representations. On the contrary, smaller blocks (*e.g.*, centered around 10, 25, or 40 in (b) or centered around layer 8, layer 17, or layer 32 in (d)) suggest that smaller portions preserve a stable activation through the network. In addition, dark stripes are also present that means dissimilarity of the activations. These stripes are more present in (b) and (d) than in (a) and (c). These patterns are also present in other combinations (see Appendix D), in which, generally larger blocks appear in highly generalizable models while strides appear in less generalizable ones.

Large block structures primarily appear in over-parameterized models [Kornblith *et al.*, 2019] which may explain why the models generalize better. Therefore, we hypothesize that it is not the high level features but the lower level ones that contribute most to the generalization performance of pairs of pathologies.

While our research investigates the impact of pathologies on generalization, we concede that our findings, especially the patterns found in layer activation similarities, can be affected by other confounding factors (*e.g.*, ethnicity or age).

Building safe models for hospital settings entails several maintenance considerations, such as monitoring the deployed models for potential risk-sensitive events, *e.g.*, data drift derived from population changes, different distribution of features, new commodities and other public health events. Such changes may require re-calibration or re-training. However, getting to know the source and target populations not only requires feature knowledge but also metadata knowledge (*e.g.*, which devices were employed for data acquisition). Therefore, the life-cycle of a deployed model is not limited to sustained model work, but also entails constant data work. Data collection, especially in medical settings, entails heterogeneous data sources that may hamper an effective collection of representative datasets and increase the overall operational expenditure. Such associated high costs become hard to overcome in developing countries [Sambasivan and Kapania, 2021]. For these reasons, it is crucial to develop tools for effective evaluation and guidance of such processes.

Our focus on CXR is motivated by the fact that respiratory infections are the leading cause of death in developing countries [Ferkol and Schraufnagel, 2014]. It is especially dangerous for children, the geriatric population and immunocompromised patients, causing over 15% of deaths in children under 5 years old worldwide [Rui *et al.*, 2017]. We believe that using ML systems for CXR diagnosis has the most impact on under-serviced populations that have limited access to specialists.They are, however, also the most under-represented in the datasets used to build and benchmark these ML systems. The largest datasets are from NIH (NIH-14, USA), Stanford Hospital (CheXpert, USA), and San Juan Hospital (PadChest, Spain), which cover similar Caucasian populations and pathology distributions. Even if these datasets have close origins, our results already show a drop of about 20% AUC moving from one dataset to another.

Our work also shows that data collection and annotation can be optimized by focusing on a subset of pathologies to improve generalization. The choice of pathologies can be driven by empirical evaluations (as proposed in our approach) supported by domain-knowledge provided by practitioners.

Conclusion

This work focuses on the critical topic of generalizable ML systems for medical diagnosis. Ensuring that the systems that have been designed and tested by the research community are effective on very different populations is of utmost importance before we consider deploying them in practice.

One common practice to improve generalization performance is to launch large data collection campaigns to build datasets for training and fine-tuning across many pathologies. However, our study demonstrates that a better understanding of the pathologies and combination of only **two** pathologies can lead to models that using a fraction of the data manage to outperform models trained with large datasets, ensembling or fine-tuning. Our approach and our feature analyses to identify the right combination of pathologies to use is among the most effective and efficient ways to provide reliable and cost-effective ML systems for medical diagnosis. Especially for populations with limited medical facilities and resources.

Following our research, we advise ml practitioners to dedicate time and resources to understand the interactions between pathologies in target populations instead of building larger, complex and data-hungry models to tackle ML-based medical diagnosis. Our research in a nutshell champions the saying: *More data is good, smart data is better.*

Acknowledgment:

This work is mainly supported by the Luxembourg National Research Funds (FNR) through CORE project C18/IS/12669767/ STELLAR/LeTraon

References

[Baltruschat *et al.*, 2019] Ivo M. Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of deep learning approaches for multi-label chest x-ray classification, 2019.

[Beede *et al.*, 2020] Emma Beede, Elizabeth Baylor, and al. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[Bressem *et al.*, 2020] Keno K. Bressem, Lisa C. Adams, Christoph Erxleben, Bernd Hamm, Stefan M. Niehues, and Janis L. Vahldiek. Comparing different deep learning architectures for classification of chest radiographs. *Scientific Reports*, 10(1), Aug 2020.

[Bruno *et al.*, 2017] Michael A Bruno, Jonelle Petscavage-Thomas, and Hani H Abujudeh. Communicating uncertainty in the radiology report. *American Journal of Roentgenology*, 209(5):1006–1008, 2017.

[Bustos *et al.*, 2020] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, Dec 2020.

[Carlucci *et al.*, 2019] Fabio Maria Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles, 2019.

[Cohen *et al.*, 2020] Joseph Paul Cohen, Mohammad Hashir, Rupert Brooks, and Hadrien Bertrand. On the limits of cross-domain generalization in automated x-ray prediction, 2020.

[D'Amour and Heller, 2020] Alexander D'Amour and Katherine et al. Heller. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

[Ferkol and Schraufnagel, 2014] Thomas Ferkol and Dean Schraufnagel. The global burden of respiratory disease. *Annals of the American Thoracic Society*, 11(3):404–406, 2014.

[Ghamizi *et al.*, 2021] Salah Ghamizi, Maxime Cordy, Mike Papadakis, and Yves Le Traon. Adversarial robustness in multi-task learning: Promises and illusions, 2021.

[Hu *et al.*, 2019] Shoubo Hu, Kun Zhang, Zhitang Chen, and Laiwan Chan. Domain generalization via multidomain discriminant analysis, 2019.

[Irvin *et al.*, 2019] Jeremy Irvin, Pranav Rajpurkar, and al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019.

[Jin *et al.*, 2020] Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Feature alignment and restoration for domain generalization and adaptation, 2020.

[Kim *et al.*, 2021] Jin Kim, Jiyoung Lee, Jungin Park, Dongbo Min, and Kwanghoon Sohn. Self-balanced learning for domain generalization. *2021 IEEE International Conference on Image Processing (ICIP)*, Sep 2021.

[Kornblith *et al.*, 2019] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited, 2019.

[Mitrovic *et al.*, 2020] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms, 2020.

[Nguyen *et al.*, 2021] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth, 2021.

[Peng *et al.*, 2020] Xingchao Peng, Yichen Li, and Kate Saenko. Domain2vec: Domain embedding for unsupervised domain adaptation, 2020.

[Pooch *et al.*, 2020] Eduardo H. P. Pooch, Pedro L. Ballester, and Rodrigo C. Barros. Can we trust deep learning models diagnosis? the impact of domain shift in chest radiograph classification, 2020.

[Rui *et al.*, 2017] Pinyao Rui, K Kang, and Michael Albert. National hospital ambulatory medical care survey: 2015 emergency department summary tables. *National center for health statistics*, 2017.

[Sambasivan and Kapania, 2021] Nithya Sambasivan and Shivani et al. Kapania. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[Varoquaux and Cheplygina, 2021] Gaël Varoquaux and Veronika Cheplygina. How i failed machine learning in medical imaging–shortcomings and recommendations. *arXiv preprint arXiv:2103.10292*, 2021.

[Wang *et al.*, 2017] 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. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[Wang *et al.*, 2021] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Wenjun Zeng, and Tao Qin. Generalizing to unseen domains: A survey on domain generalization. *arXiv preprint arXiv:2103.03097*, 2021.

[Wu and Gong, 2021] Guile Wu and Shaogang Gong. Collaborative optimization and aggregation for decentralized domain generalization and adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6484–6493, October 2021.

[Yao *et al.*, 2019] Li Yao, Jordan Prosky, Ben Covington, and Kevin Lyman. A strong baseline for domain adaptation and generalization in medical imaging, 2019.

[Zamir *et al.*, 2018] Amir R. Zamir, Alexander Sax, William B. Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

[Zhang and al., 2021] Haoran Zhang and al. An empirical framework for domain generalization in clinical settings. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 279–290, 2021.

[Zhou *et al.*, 2020] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Deep domain-adversarial image generation for domain generalisation, 2020.

Appendix

Appendix A: Experimental protocol details

Datasets We show in table 7 the general properties of the datasets used in training our models. Table 8 ([Cohen *et al.*, 2020]) details the number of positive and negative examples with each label for each dataset. Our models are trained either on NIH, PC or CheXpert depending of the evaluation. Models *C1* and *C2* are trained on CheXpert dataset and model *A* is trained on all the 8 datasets.

	NIH	CheXpert
Number of patient radiographs	112,120	224,316
Number of patients	30,805	65,240
Age in years: mean (standard deviation)	46.9 (16.6)	60.7 (18.4)
Percentage of females (%)	43.5%	40.6%
Number of pathology labels	8	14

Table 7: Characteristics of NIH and CheXpert datasets using in our trained models.

Models We provide in table 9 the test and generalization performances of models trained on the *CheXpert* dataset obtained with Auxiliary (Aux) pathology learning compared to fine-tuning (Tune), single pathology learning (Single) and all pathology learning (All). The table presents absolute AUC values.

Table 10, table 11 and table 12 show the test AUC of models trained and evaluated on the same dataset. Respectively, on the dataset *CheXpert*, *NIH*, and *PadChest*.

We provide in table 13 and table 14 the generalization performances of models trained on the *CheXpert* dataset obtained with Auxiliary (Aux) pathology learning and tested on respectively NIH and PadChest datasets.

We provide in table 15 and table 16 the generalization performances of models trained on the *NIH* dataset obtained with Auxiliary (Aux) pathology learning and tested on respectively CheXpert and PadChest datasets.

Appendix B: Pathology selection has a significant impact on generalization

Evaluation of Resnet50 architectures (details of the main paper) We present in ROC curves for all pathologies for models:

- trained on CheXpert and evaluated on NIH in figure 3;
- trained on CheXpert and evaluated on PC in figure 4;
- trained on NIH and evaluated on CheXpert in figure 5;
- trained on NIH and evaluated on PC in figure 6;

For each figure, we provide for reference the test performance on the original dataset using in the training, then the test performance on the target dataset.

We compute the statistics of mean, standard deviation, maximum and minimum across all these values and present it in table 17.

Evaluation of other architectures We run the same experiments as Section 4.1, but using a Densenet121 architecture. Figure 7 shows that our claims, evaluated on Resnet50 in the main paper are confirmed on other architectures. While the AUC of Edema shows little variance across combinations of auxiliary pathologies on the source dataset (CheXpert, Figure 7i), the choice of the auxiliary pathology can have a significant impact on the AUC performance on the target dataset (NIH , Figure 7j).

Appendix C: Fine-tuning and encoder-freezing

In tables 18 and 19 we pre-train a model on an auxiliary pathology (rows) then fine-tuned all the model on the main pathology. Pre-training and fine-tuning are both done using the source dataset CheXpert. We then evaluate the performance of the main pathology on the CheXpert dataset in table 18 and on the NIH dataset in table 19.

In tables 20 and 21 we pre-train a model on an auxiliary pathology (rows) then fine-tuned on the decoder of each model on the main pathology, while the weights of the encoder are frozen. Pre-training and fine-tuning are both done using the source dataset CheXpert. We then evaluate the performance of the main pathology on the CheXpert dataset in table 20 and on the NIH dataset in table 21.

Appendix D: CKA patterns

[Kornblith *et al.*, 2019] proposed a novel metric, **Centered kernel alignment (CKA)** that provides a reliable quantitative measure of the similarity of neural network representation.

Following [Nguyen *et al.*, 2021], let $\mathbf{X} \in \mathbb{R}^{m\times n_1}$ and $\mathbf{Y} \in \mathbb{R}^{m\times n_2}$ the matrix representations of two layers, one with n_1 neurons and another n_2 neurons, to the same set of m examples. Each element of the $m \times m$ Gram matrices $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^\top$ and $\boldsymbol{L} = \boldsymbol{Y}\boldsymbol{Y}^\top$ reflects the similarities between a pair of examples according to the representations contained in $\boldsymbol{X}$ or $\boldsymbol{Y}$. Let $\boldsymbol{H} = \boldsymbol{I}_m - \frac{1}{m}\mathbf{1}\mathbf{1}^\top$ be the centering matrix. Then $\boldsymbol{K}' = \boldsymbol{H}\boldsymbol{K}\boldsymbol{H}$ and $\boldsymbol{L}' = \boldsymbol{H}\boldsymbol{L}\boldsymbol{H}$ reflect the similarity matrices with their column and row means subtracted.

HSIC is defined as the similarity of these centered similarity matrices by reshaping them to vectors and taking the dot product between these vectors, $\mathrm{HSIC}(\boldsymbol{K},\boldsymbol{L}) = \frac{\mathrm{vec}(\boldsymbol{K}')\mathrm{vec}(\boldsymbol{L}')}{(m-1)^2}$. HSIC is invariant to orthogonal transformations of the representations and to permutation of neurons, but it is not invariant to scaling of the original representations. CKA further normalizes HSIC to produce a similarity index between 0 and 1 that is invariant to isotropic scaling,

$$\mathrm{CKA}(\boldsymbol{K},\boldsymbol{L}) = \frac{\mathrm{HSIC}(\boldsymbol{K},\boldsymbol{L})}{\sqrt{\mathrm{HSIC}(\boldsymbol{K},\boldsymbol{K})\mathrm{HSIC}(\boldsymbol{L},\boldsymbol{L})}}.$$

[Kornblith *et al.*, 2019] showed that linear CKA between layers of architecturally identical networks, trained from different initialization, reliably identifies architecturally corresponding layers. We implement linear CKA following the mini-batch split proposed by [Nguyen *et al.*, 2021].

We show in figure 8, for each of our seven pathologies, the CKA of the combinations that lead to the best and the worst generalization performance when the models are trained on the CheXpert dataset.

Dataset	NIH	PadChest	CheXpert	Google	MIMIC_CH	MIMIC_NB	OpenI	Kaggle
Atelectasis	1702/29103	2441/59674	12691/14317	-	4077/30954	4048/32058	271/2996	-
Cardiomegaly	767/30038	5390/56725	9099/17765	-	3743/32312	3275/33431	185/3082	-
Consolidation	427/30378	494/61621	5390/22504	-	816/32297	762/33564	-	-
Edema	82/30723	108/62007	14929/20615	-	1157/33610	1121/34731	50/3217	-
Effusion	1280/29525	1637/60478	20640/23500	-	3713/33401	3595/34489	120/3147	-
Emphysema	265/30540	546/61569	-	-	-	-	84/3183	-
Enlarged Cardio	-	-	5181/20506	-	692/31505	660/32641	-	-
Fibrosis	571/30234	341/61774	-	-	-	-	17/3250	-
Fracture	-	1665/60450	4250/14948	60/1635	972/30961	696/32320	78/3189	-
Hernia	83/30722	988/61127	-	-	-	-	41/3226	-
Infiltration	3604/27201	4438/57677	-	-	-	-	66/3201	-
Lung Lesion	-	-	4217/14422	-	1321/31033	1271/32187	3/3264	-
Lung Opacity	-	-	30873/15675	601/1094	5426/31175	5301/32371	327/2940	9555/20672
Mass	1280/29525	507/61608	-	-	-	-	6/3261	-
Nodule	1661/29144	2194/59921	-	-	-	-	68/3199	-
Pleural Thickening	763/30042	2076/60039	-	-	-	-	30/3237	-
Pneumonia	168/30637	2051/60064	2822/14793	-	2176/33347	2042/34479	68/3199	9555/20672
Pneumothorax	269/30536	98/62017	4311/32685	72/1623	560/33651	500/34760	14/3253	-

Table 8: Samples distributions across each pathology and dataset. Each cell shows the number of positive/negative samples of the label. In bold the pathologies we use in training our models. Those are the 7 common pathologies in NIH, PC and Chexpert datasets.

	Test AUC % on CheXpert				Test AUC % on NIH			
Path	Fine-tuning	Auxiliary Task Learning	Single	All	Fine-tuning	Auxiliary Task Learning	Single	All
Atelectasis	86.88	89.59	87.97	89.17	67.10	69.25	70.77	67.82
Cardiomegaly	85.12	89.71	89.16	89.09	67.84	76.88	75.40	70.67
Consolidation	84.97	86.74	85.35	88.61	68.56	71.13	66.52	66.49
Edema	88.55	91.57	90.64	90.64	72.85	75.91	69.88	68.54
Effusion	90.84	93.52	93.73	91.96	81.02	83.15	83.66	79.81
Pneumonia	77.29	80.06	77.73	83.42	60.81	62.50	59.85	67.78
Pneumothorax	80.25	83.42	81.47	83.71	67.34	68.17	58.69	60.08

Table 9: AUC Performance of models trained on each pathology on the CheXpert dataset, tested on CheXpert (middle) and tested on NIH (right). Tune represents the fine-tuned model; Aux when learned with an Auxiliary ; Single represents a model training on a single pathology; and All when all the pathologies are learnt at once. Aux and Tune report only the best performing models.

Main pathology	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Pneumonia	Pneumothorax	Average
Auxiliary								
Atelectasis	0.88	0.90	0.87	0.91	0.93	0.80	0.82	0.87
Cardiomegaly	0.89	0.89	0.87	0.92	0.93	0.78	0.81	0.87
Consolidation	0.89	0.90	0.85	0.89	0.93	0.79	0.83	0.87
Edema	0.89	0.89	0.84	0.91	0.93	0.79	0.81	0.87
Effusion	0.90	0.89	0.86	0.91	0.94	0.80	0.83	0.87
Pneumonia	0.88	0.89	0.86	0.91	0.92	0.78	0.80	0.86
Pneumothorax	0.89	0.89	0.86	0.91	0.94	0.80	0.81	0.87
Average	0.89	0.89	0.86	0.91	0.93	0.79	0.82	0.87

Table 10: Test AUC of models trained on CheXpert and evaluated on CheXpert

Main pathology	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Pneumonia	Pneumothorax	Average
Auxiliary								
Atelectasis	0.79	0.92	0.75	0.85	0.88	0.65	0.85	0.81
Cardiomegaly	0.77	0.91	0.74	0.76	0.87	0.65	0.85	0.79
Consolidation	0.77	0.92	0.72	0.81	0.87	0.66	0.84	0.80
Edema	0.77	0.90	0.72	0.69	0.87	0.58	0.84	0.77
Effusion	0.80	0.91	0.75	0.79	0.87	0.66	0.85	0.81
Pneumonia	0.78	0.91	0.71	0.72	0.87	0.59	0.83	0.77
Pneumothorax	0.77	0.91	0.74	0.78	0.87	0.64	0.84	0.79
Average	0.78	0.91	0.73	0.77	0.87	0.63	0.84	0.79

Table 11: Test AUC of models trained on NIH and evaluated on NIH

Main pathology	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Pneumonia	Pneumothorax	Average
Auxiliary								
Atelectasis	0.80	0.94	0.83	0.96	0.95	0.79	0.82	0.88
Cardiomegaly	0.82	0.94	0.86	0.97	0.94	0.81	0.83	0.88
Consolidation	0.76	0.94	0.81	0.96	0.93	0.78	0.78	0.85
Edema	0.71	0.94	0.82	0.96	0.94	0.76	0.78	0.85
Effusion	0.82	0.94	0.85	0.97	0.95	0.79	0.82	0.87
Pneumonia	0.80	0.94	0.84	0.96	0.95	0.79	0.84	0.87
Pneumothorax	0.72	0.93	0.82	0.96	0.94	0.75	0.78	0.83
Average	0.78	0.94	0.83	0.96	0.94	0.78	0.81	0.86

Table 12: Test AUC of models trained on PC and evaluated on PC

Main pathology	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Pneumonia	Pneumothorax
Auxiliary							
Atelectasis	0.66	0.76	0.67	0.71	0.81	0.60	0.65
Cardiomegaly	0.69	0.75	0.68	0.63	0.80	0.58	0.65
Consolidation	0.70	0.76	0.66	0.73	0.82	0.62	0.69
Edema	0.69	0.72	0.70	0.69	0.81	0.58	0.61
Effusion	0.69	0.72	0.72	0.58	0.82	0.59	0.63
Pneumonia	0.69	0.72	0.69	0.65	0.79	0.61	0.69
Pneumothorax	0.68	0.72	0.68	0.67	0.78	0.58	0.60

Table 13: Test AUC of models trained on CheXpert and evaluated on NIH

Main pathology	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Pneumonia	Pneumothorax
Auxiliary							
Atelectasis	0.69	0.87	0.74	0.91	0.90	0.60	0.63
Cardiomegaly	0.69	0.84	0.78	0.92	0.90	0.63	0.59
Consolidation	0.67	0.87	0.72	0.94	0.91	0.65	0.69
Edema	0.68	0.84	0.77	0.93	0.91	0.61	0.62
Effusion	0.70	0.86	0.81	0.92	0.91	0.55	0.66
Pneumonia	0.69	0.84	0.74	0.94	0.85	0.55	0.63
Pneumothorax	0.66	0.84	0.76	0.94	0.90	0.60	0.78

Table 14: Test AUC of models trained on CheXpert and evaluated on PC

Main pathology Auxiliary	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Pneumonia	Pneumothorax
Atelectasis	0.72	0.72	0.77	0.56	0.84	0.55	0.48
Cardiomegaly	0.65	0.64	0.78	0.66	0.60	0.58	0.52
Consolidation	0.77	0.82	0.66	0.67	0.79	0.60	0.60
Edema	0.59	0.76	0.65	0.68	0.69	0.58	0.51
Effusion	0.80	0.59	0.74	0.63	0.81	0.61	0.59
Pneumonia	0.77	0.75	0.70	0.76	0.74	0.58	0.67
Pneumothorax	0.50	0.64	0.67	0.49	0.76	0.52	0.60

Table 15: Test AUC of models trained on NIH and evaluated on CheXpert

Main pathology Auxiliary	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Pneumonia	Pneumothorax
Atelectasis	0.75	0.89	0.76	0.74	0.91	0.66	0.74
Cardiomegaly	0.73	0.87	0.81	0.89	0.87	0.62	0.80
Consolidation	0.74	0.90	0.71	0.79	0.92	0.63	0.73
Edema	0.73	0.89	0.68	0.79	0.89	0.55	0.77
Effusion	0.77	0.87	0.77	0.84	0.89	0.64	0.75
Pneumonia	0.73	0.90	0.70	0.85	0.87	0.54	0.75
Pneumothorax	0.73	0.88	0.71	0.57	0.90	0.55	0.70

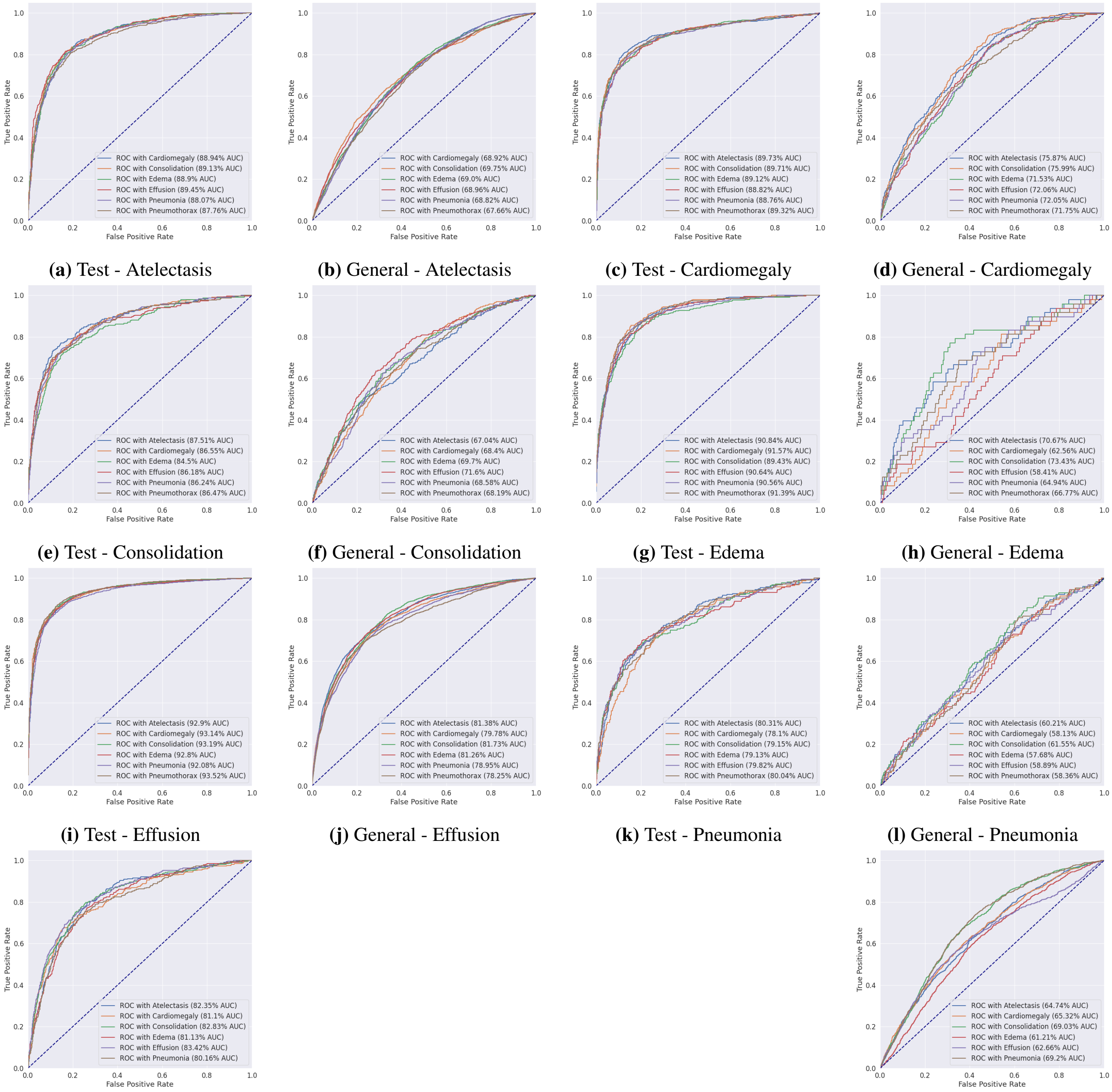
Table 16: Test AUC of models trained on NIH and evaluated on PC

	CHEX → CHEX						CHEX → NIH					
	mean	std	min	max	single	all	mean	std	min	max	single	all
ATE	88.83	0.57	87.80	89.59	87.97	89.17	68.39	0.52	67.73	69.25	70.77	67.82
CAR	89.21	0.37	88.76	89.71	89.16	89.09	74.08	1.87	71.90	76.88	75.40	70.67
CON	86.12	0.78	84.50	86.74	85.35	88.61	67.76	2.00	65.04	71.13	66.52	66.49
EDE	90.78	0.73	89.43	91.57	90.64	90.60	64.17	6.06	57.05	75.91	69.88	68.54
EFF	92.93	0.47	92.08	93.52	93.73	91.96	81.66	1.26	80.15	83.15	83.66	79.81
PNE	79.39	0.72	78.10	80.06	77.73	83.42	58.90	1.93	56.99	62.50	59.85	67.78
PNX	81.85	1.19	80.16	83.42	81.47	83.71	63.91	3.29	59.71	68.17	58.69	60.08

Table 17: Statistic of AUC performance computed for different combinations of models trained on the CheXpert dataset and evaluated on CheXpert (left) and NIH (right)

	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Pneumonia	Pneumothorax
Atelectasis	0.00	-6.57	-2.51	-2.93	-2.69	-2.10	-3.77
Cardiomegaly	-3.59	0.00	-11.85	-4.89	-12.23	-8.34	-13.29
Consolidation	-2.02	-9.60	0.00	-3.17	-3.06	-0.19	-1.71
Edema	-3.28	-5.03	-5.07	0.00	-6.07	-2.48	-8.22
Effusion	-1.31	-7.75	-2.18	-5.79	0.00	-4.35	-1.89
Pneumonia	-2.18	-5.62	-3.48	-3.45	-5.38	0.00	-4.57
Pneumothorax	-6.61	-9.63	-2.75	-6.24	-3.27	-2.17	0.00

Table 18: Change of AUC performance on CheXpert dataset when fine-tuning the whole model using CheXpert dataset.



(a) Test - Atelectasis

(b) General - Atelectasis

(c) Test - Cardiomegaly

(d) General - Cardiomegaly

(e) Test - Consolidation

(f) General - Consolidation

(g) Test - Edema

(h) General - Edema

(i) Test - Effusion

(j) General - Effusion

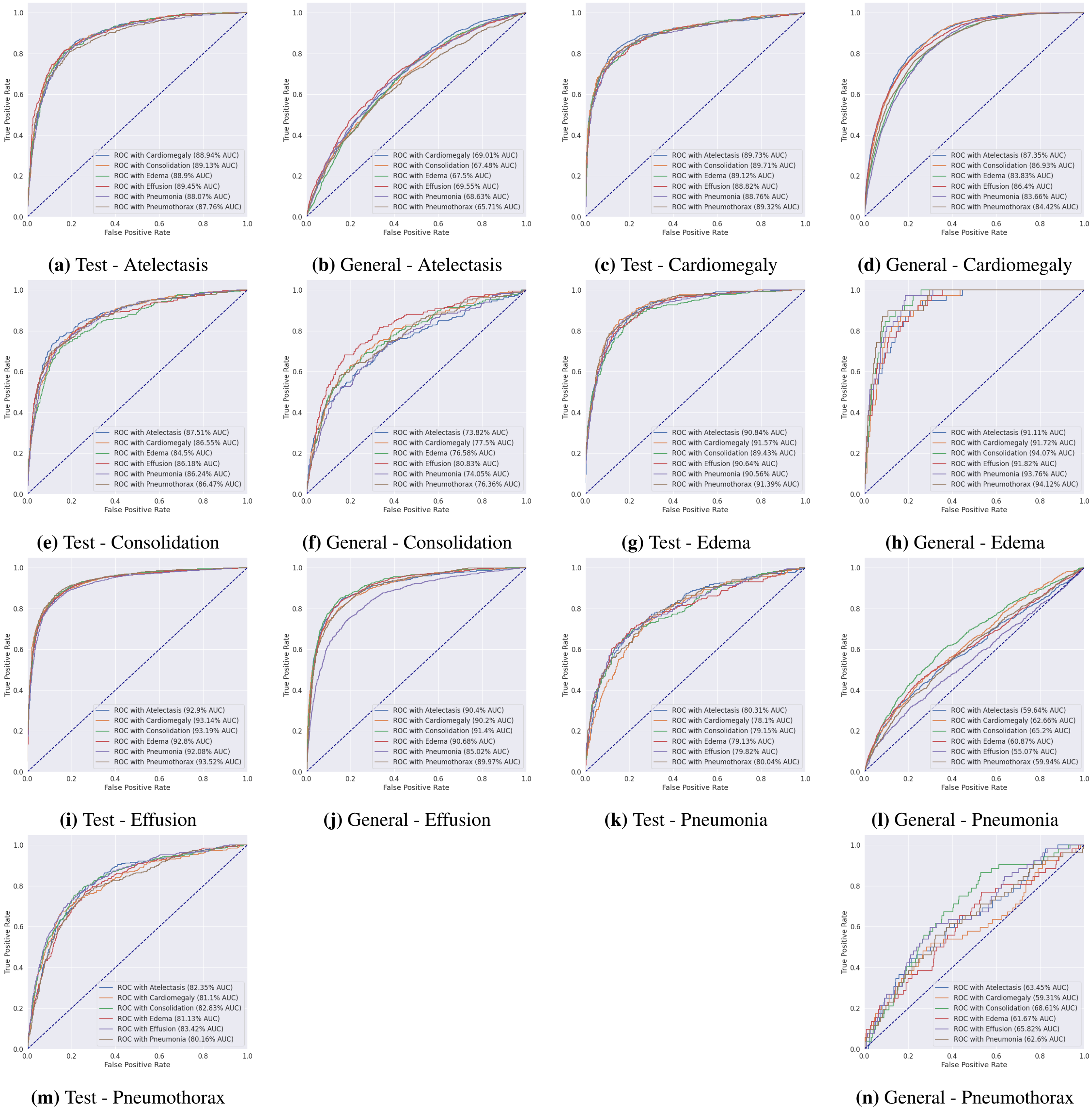
(k) Test - Pneumonia

(l) General - Pneumonia

(m) Test - Pneumothorax

(n) General - Pneumothorax

Figure 3: ROC curves of test performance (source) (CHEX→CHEX) and test performance (target) (CHEX→NIH) for each of the 6 pathologies when learned with the 6 other pathologies.



(a) Test - Atelectasis

(b) General - Atelectasis

(c) Test - Cardiomegaly

(d) General - Cardiomegaly

(e) Test - Consolidation

(f) General - Consolidation

(g) Test - Edema

(h) General - Edema

(i) Test - Effusion

(j) General - Effusion

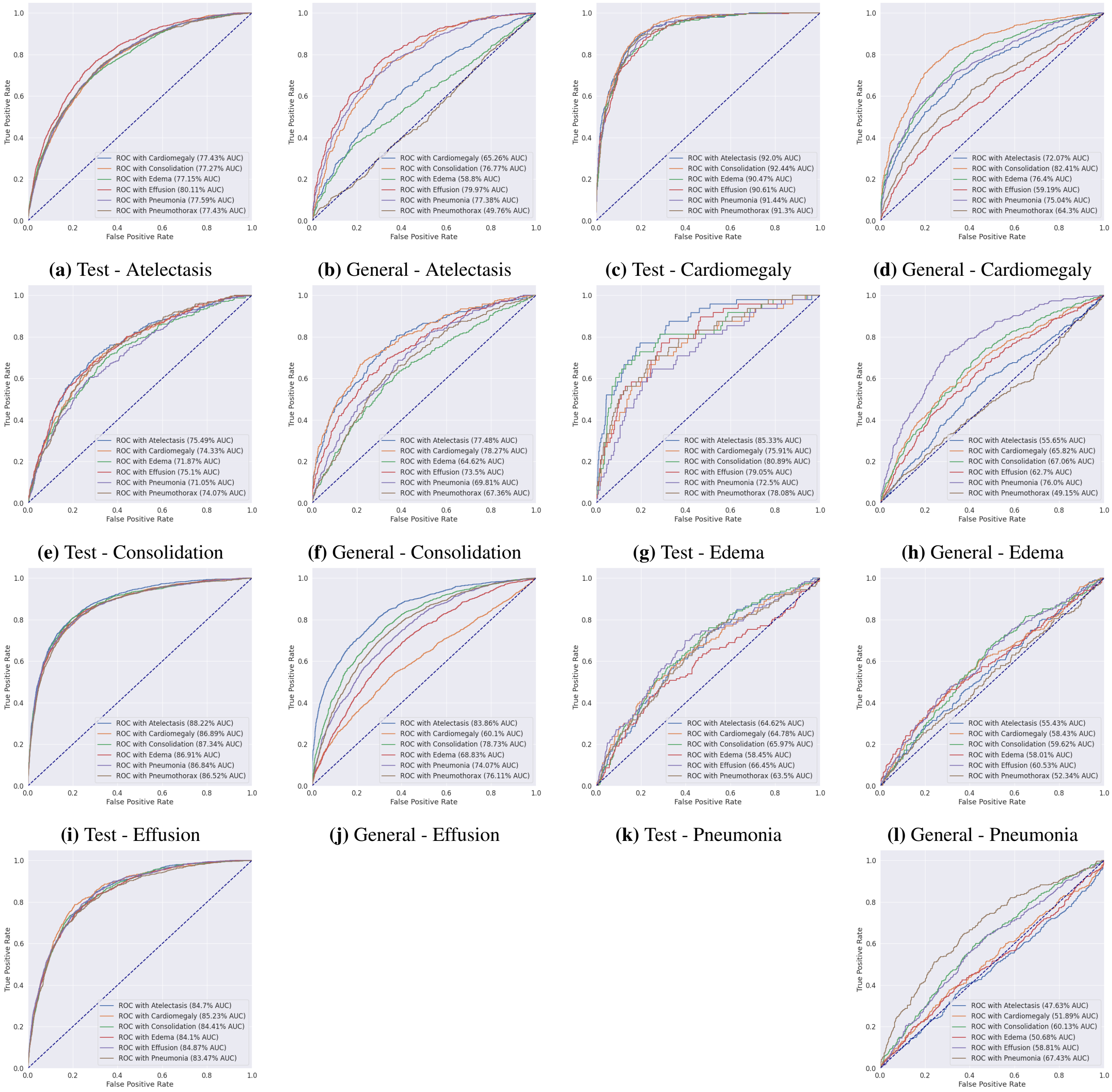
(k) Test - Pneumonia

(l) General - Pneumonia

(m) Test - Pneumothorax

(n) General - Pneumothorax

Figure 4: ROC curves of test performance (source) (CHEX→CHEX) and test performance (target) (CHEX→PC) for each of the 6 pathologies when learned with the 6 other pathologies.



(a) Test - Atelectasis

(b) General - Atelectasis

(c) Test - Cardiomegaly

(d) General - Cardiomegaly

(e) Test - Consolidation

(f) General - Consolidation

(g) Test - Edema

(h) General - Edema

(i) Test - Effusion

(j) General - Effusion

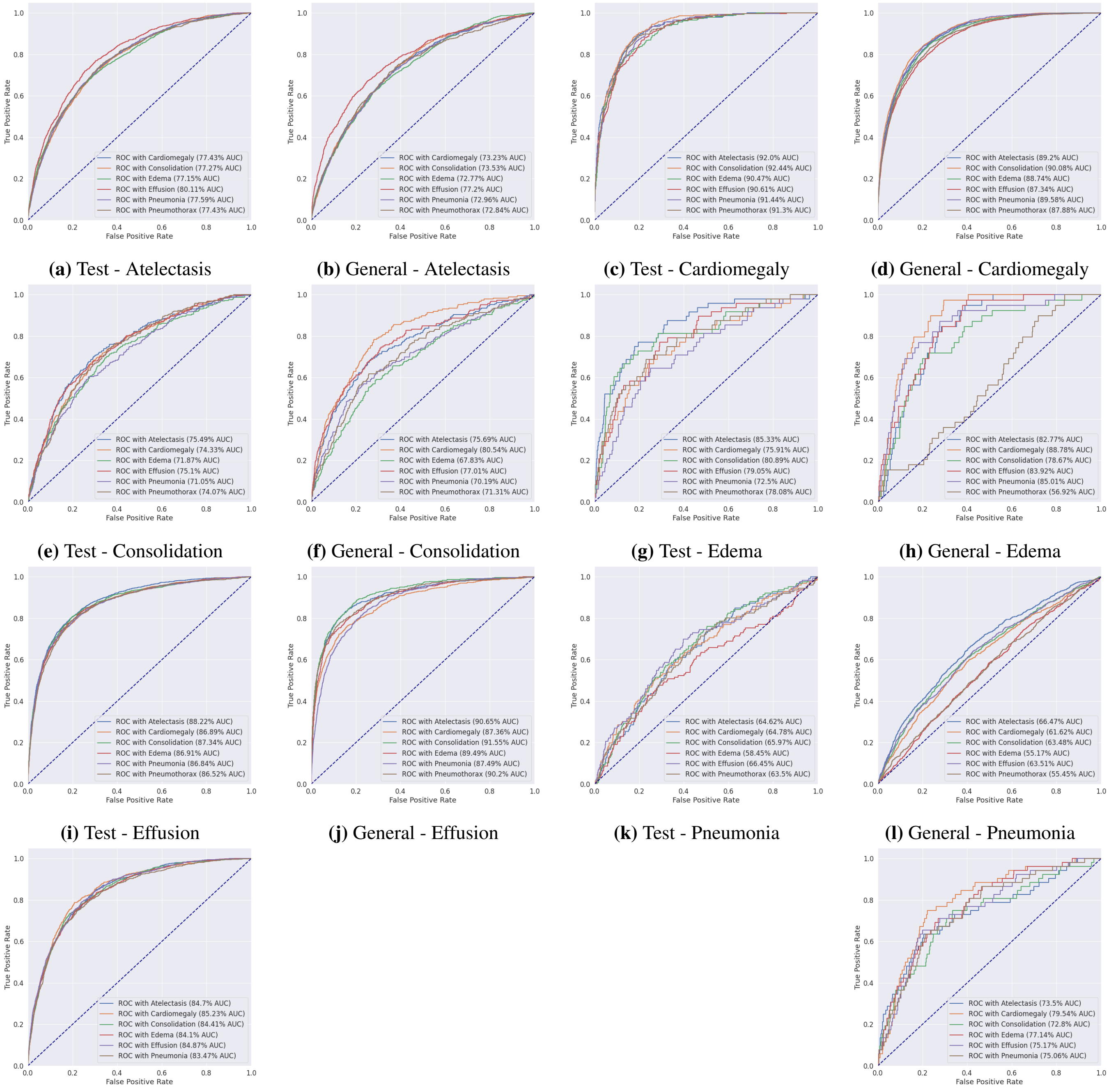
(k) Test - Pneumonia

(l) General - Pneumonia

(m) Test - Pneumothorax

(n) General - Pneumothorax

Figure 5: ROC curves of test performance (source) (NIH→NIH) and test performance (target) (NIH→CHEX) for each of the 6 pathologies when learned with the 6 other pathologies.



(a) Test - Atelectasis

(b) General - Atelectasis

(c) Test - Cardiomegaly

(d) General - Cardiomegaly

(e) Test - Consolidation

(f) General - Consolidation

(g) Test - Edema

(h) General - Edema

(i) Test - Effusion

(j) General - Effusion

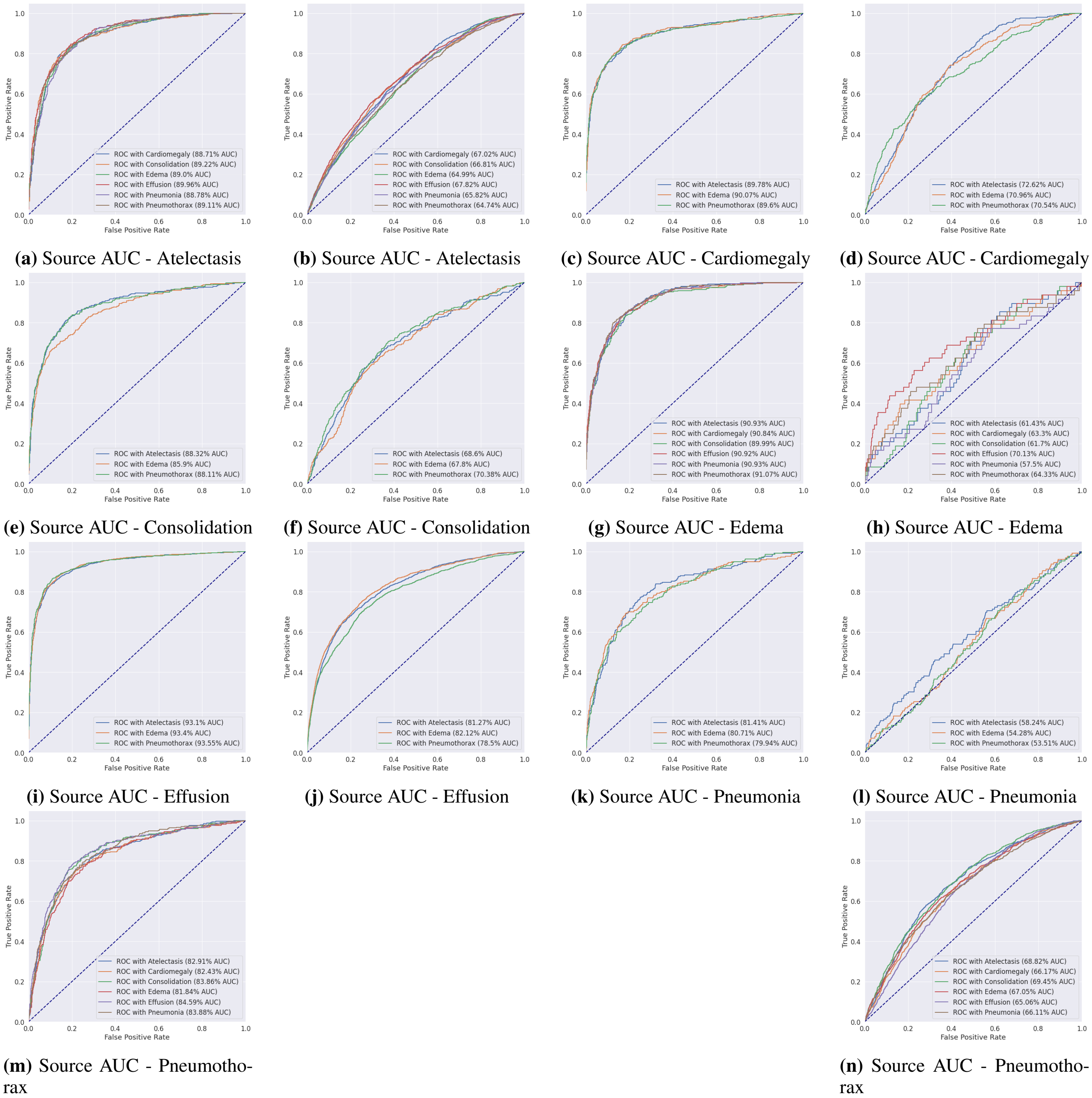
(k) Test - Pneumonia

(l) General - Pneumonia

(m) Test - Pneumothorax

(n) General - Pneumothorax

Figure 6: ROC curves of test performance (source) (NIH→NIH) and test performance (target) (NIH→PC) for each of the 6 pathologies when learned with the 6 other pathologies.



(a) Source AUC - Atelectasis

(b) Source AUC - Atelectasis

(c) Source AUC - Cardiomegaly

(d) Source AUC - Cardiomegaly

(e) Source AUC - Consolidation

(f) Source AUC - Consolidation

(g) Source AUC - Edema

(h) Source AUC - Edema

(i) Source AUC - Effusion

(j) Source AUC - Effusion

(k) Source AUC - Pneumonia

(l) Source AUC - Pneumonia

(m) Source AUC - Pneumothorax

(n) Source AUC - Pneumothorax

Figure 7: ROC curves of test performance (source) (CHEX→CHEX) and test performance (target) (CHEX→NIH) for each of the 6 pathologies when learned with the 6 other pathologies. The models use a DenseNet architecture

	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Pneumonia	Pneumothorax
Atelectasis	0.00	-13.77	5.23	-0.85	-2.29	-1.99	-5.67
Cardiomegaly	-8.85	0.00	3.88	-15.87	-15.27	-0.47	-0.72
Consolidation	-3.86	-9.25	0.00	9.06	-4.52	1.04	7.47
Edema	-3.46	-7.44	2.98	0.00	-12.61	1.79	-2.07
Effusion	-0.31	-13.33	4.11	-0.25	0.00	1.40	3.59
Pneumonia	-0.65	-8.05	-2.62	-0.84	-8.79	0.00	-7.50
Pneumothorax	-4.04	-15.42	-7.95	-5.96	-8.18	5.46	0.00

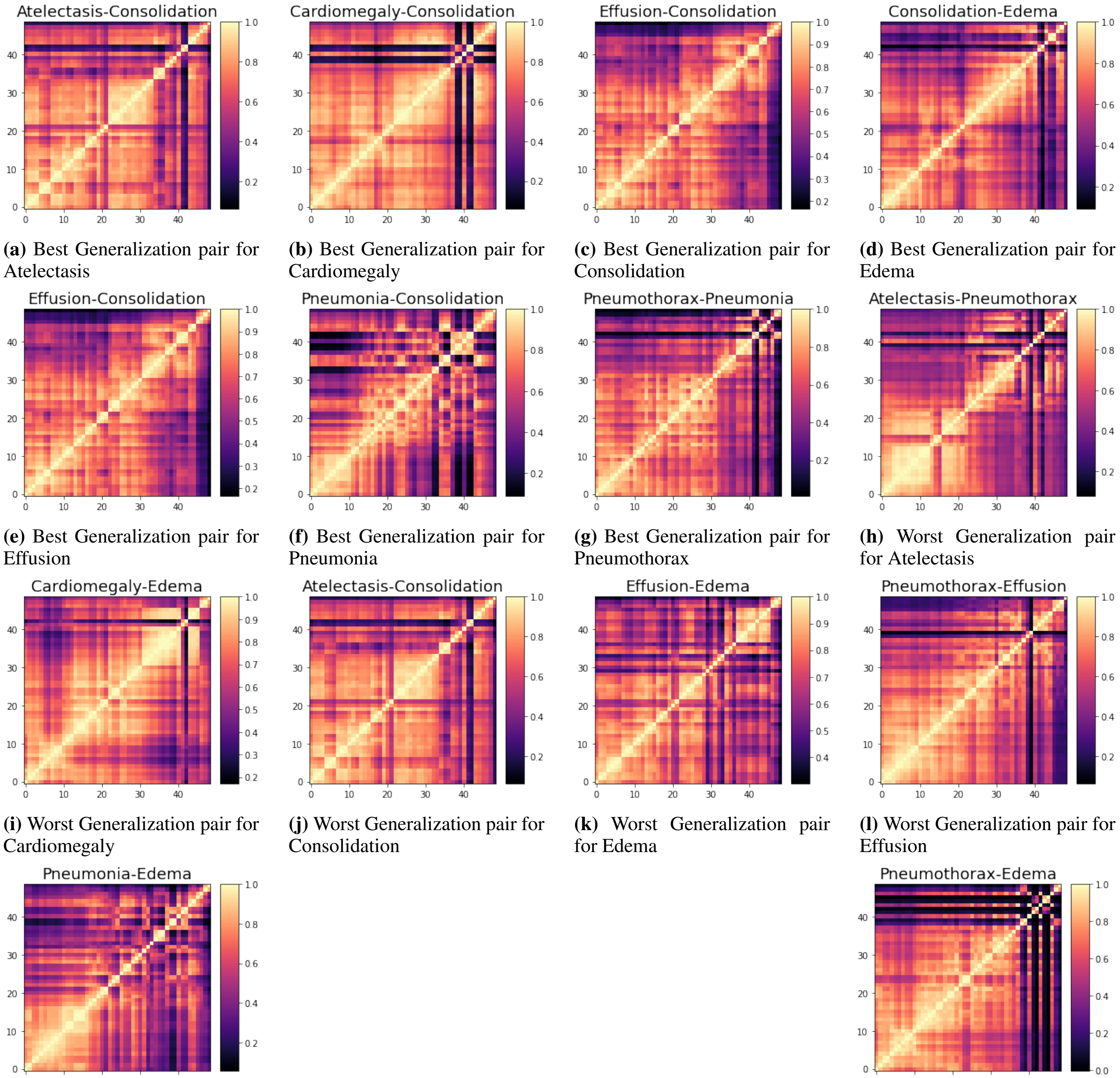
Table 19: Change of AUC performance on NIH dataset when fine-tuning the whole model using CheXpert dataset.

	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Pneumonia	Pneumothorax
Atelectasis	0.00	-0.03	-0.28	-0.11	-0.13	-0.21	0.02
Cardiomegaly	0.06	0.00	-0.09	0.08	-0.62	0.15	-0.29
Consolidation	-0.35	-0.75	0.00	-0.04	-0.24	0.36	-0.29
Edema	-0.43	-0.13	0.38	0.00	-0.76	0.26	-0.89
Effusion	0.17	-0.45	-0.11	-0.10	0.00	0.12	-0.05
Pneumonia	-0.19	-0.32	0.13	-0.34	-0.35	0.00	-0.11
Pneumothorax	-0.96	-1.27	0.00	-0.82	-0.79	0.37	0.00

Table 20: Change of AUC performance on CheXpert dataset when fine-tuning only the decoder using CheXpert dataset.

	Atelectasis	Cardiomegaly	Consolidation	Edema	Effusion	Pneumonia	Pneumothorax
Atelectasis	0.00	-5.07	0.38	0.17	-1.59	1.18	-1.42
Cardiomegaly	-0.21	0.00	-0.17	-0.55	-1.84	1.25	-0.49
Consolidation	-0.36	-4.34	0.00	-0.13	0.39	0.66	0.26
Edema	-0.22	-6.25	0.18	0.00	-0.83	0.79	1.30
Effusion	-0.03	-7.69	0.45	-1.64	0.00	1.02	-1.16
Pneumonia	-0.07	-1.61	-0.36	0.85	-0.09	0.00	-0.23
Pneumothorax	-0.66	-6.25	-0.56	-0.59	-0.17	0.38	0.00

Table 21: Change of AUC performance on NIH dataset when fine-tuning only the decoder using CheXpert dataset.



(a) Best Generalization pair for Atelectasis

(b) Best Generalization pair for Cardiomegaly

(c) Best Generalization pair for Consolidation

(d) Best Generalization pair for Edema

(e) Best Generalization pair for Effusion

(f) Best Generalization pair for Pneumonia

(g) Best Generalization pair for Pneumothorax

(h) Worst Generalization pair for Atelectasis

(i) Worst Generalization pair for Cardiomegaly

(j) Worst Generalization pair for Consolidation

(k) Worst Generalization pair for Edema

(l) Worst Generalization pair for Effusion

(m) Worst Generalization pair for Pneumonia

(n) Worst Generalization pair for Pneumothorax

Figure 8: Layer Similarity for combination of models trained on CheXpert dataset