Self-Supervised Pretraining Enables High-Performance Chest X-Ray Interpretation Across Clinical Distributions

Authors: Niveditha S. Iyer[1*], Aditya Gulati[1*], Oishi Banerjee MS[2], Cécile Logé BS[1], Maha Farhat MD, MSc[2], Agustina D. Saenz MD, MPH[2,3*] & Pranav Rajpurkar PhD[2*]

Affiliations:

1. Department of Computer Science, Stanford University, Stanford, California, United States
2. Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts United States
3. Brigham and Women's Hospital, Boston, Massachusetts, United States

* These authors contributed equally:

Niveditha S. Iyer and Aditya Gulati share first authorship; Agustina D. Saenz and Pranav Rajpurkar share senior authorship.

Corresponding author:
Pranav Rajpurkar PhD
Assistant Professor of Biomedical Informatics
Harvard Medical School
pranav_rajpurkar@hms.harvard.edu

Abstract

Chest X-rays (CXRs) are a rich source of information for physicians – essential for disease diagnosis and treatment selection. Recent deep learning models aim to alleviate strain on medical resources and improve patient care by automating the detection of diseases from CXRs. However, shortages of labeled CXRs can pose a serious challenge when training models. Currently, models are generally pretrained on ImageNet, but they often need to then be finetuned on hundreds of thousands of labeled CXRs to achieve high performance. Therefore, the current approach to model development is not viable on tasks with only a small amount of labeled data. An emerging method for reducing reliance on large amounts of labeled data is self-supervised learning (SSL), which uses unlabeled CXR datasets to automatically learn features that can be leveraged for downstream interpretation tasks. In this work, we investigated whether self-supervised pretraining methods could outperform traditional ImageNet pretraining for chest X-ray interpretation. We found that SSL-pretrained models outperformed ImageNet-pretrained models on thirteen different datasets representing high diversity in geographies, clinical settings, and prediction tasks. We thus show that SSL on unlabeled CXR data is a promising pretraining approach for a wide variety of CXR interpretation tasks, enabling a shift away from costly labeled datasets.

Introduction

Chest X-rays (CXRs) are the most commonly used type of medical imaging globally, critical for the screening, diagnosis, and management of many life-threatening diseases. However, many medical systems currently lack the resources to accurately read CXRs, due to the time-consuming nature of CXR interpretation and an international shortage of radiologists[1,2]. Therefore, there have been significant efforts to create AI tools that can assist with CXR interpretation. Deep learning models have shown promise in this area, matching the performance of trained radiologists on a variety of tasks[3,4,5,6].

In current practice, AI tools for any given CXR interpretation task are usually trained using a large CXR dataset with labels specifically addressing that task. Unfortunately, obtaining CXR data can itself be a costly, time-consuming process. Furthermore, obtaining high-quality labels requires experienced radiologists to manually interpret and annotate each CXR, which is extremely expensive[7]. The resulting lack of reliably labeled CXR data has posed a significant challenge for CXR model development, especially for less common tasks unaddressed by labels in existing datasets. For reference, one of the most frequently used datasets is CheXpert, a single-center dataset of 224,316 CXRs. While this dataset offers lower-quality “weak” labels automatically extracted from the radiology reports, board-certified radiologists manually annotated only 700 images.[8] Our project aims to overcome the label shortage, addressing a key question: whether we train high-performing models without relying on large, densely annotated datasets.

We focus on transfer learning, a procedure that is widely used to improve model performance. In transfer learning, models are first pretrained on a large, general-purpose dataset and then finetuned on a more specific dataset addressing the target task. The pretraining step can compensate for weaknesses in downstream target datasets, such as when datasets are small or have weak labels. Over the past ten years, models for image interpretation have usually been pretrained specifically on ImageNet, a large dataset of natural images[9]; ImageNet pretraining has been successfully applied to a variety of medical image tasks, ranging from CXR interpretation to skin cancer classification to the diagnosis of retinal disease[10,11,12]. Unfortunately, the benefits of transfer learning lessen when the source data differs dramatically from the target data, as in the case of ImageNet images, which capture color images naturally occurring in the world, and CXRs, which are often taken in standard grayscale views with subtle local pixel differences accounting for differences in their classification.[13] ImageNet-pretrained models often need to be finetuned on large, labeled CXR datasets containing tens or hundreds of thousands of datapoints in order to achieve high performance on downstream tasks. We hypothesize that pretraining on source datasets that more closely resemble target CXR datasets can improve downstream performance, potentially allowing CXR models to use significantly less labeled data during finetuning.

Self-supervised approaches, which train models without labels, have shown promising results on medical imaging tasks.[14] For example, smaller scale studies have shown that self-supervised models pretrained on CheXpert had better performance and generalized better than models pretrained through supervised learning on ImageNet[15]. Furthermore, self-supervised models like MedAug have been shown to achieve high performance on specific chest X-ray pathology classification tasks using as little as 1% of CheXpert training dataset, making self-supervised learning (SSL) a promising approach to label-efficient training[16]. However, it is unknown whether self-supervised pretraining on unlabeled CXRs can outperform ImageNet-pretrained models across a range of CXR tasks.

To evaluate whether self-supervised models outperform ImageNet-pretrained models across clinical distributions, we implemented seven self-supervised learning approaches to pretrain

models on a large CXR dataset without using labels. We compared the performance of these models against traditional Imagenet-pretrained models on thirteen datasets covering a variety of tasks, such as lung nodule detection, line and tube placement, edema severity classification, pediatric pneumonia detection, pneumothorax detection and multi-class differential diagnosis across demographics and clinical settings. After performing supervised learning on our target datasets, we evaluated how models behave across different distributions covering datasets of different sizes with different labeling methodologies.

Our results demonstrate that self-supervised pretrained models may learn generalizable chest X-ray representations, which improve performance over ImageNet-pretrained models; using such models can enable a shift away from costly labeled datasets for a large variety of chest-X-ray interpretation tasks across different clinical distributions. By eliminating the need to train models on large labeled datasets that are difficult to obtain, this work will enable institutions to more rapidly train and evaluate models across distinct clinical tasks, getting a step closer to model deployment into clinical settings.

Results

Experimental Setup

To evaluate whether self-supervised CXR pretraining outperforms ImageNet pretraining, we selected seven SSL methods, three that only use CXR images (MedAug[16], S2MTS2[17], MoCo-CXR[18]) and four that use both chest X-ray images and corresponding radiology reports (CXR-RePaiR-CLIP[19], ConVIRT[20],REFERS[21], GLoRIA[22]). We give further details on these learning algorithms in the Methods section. Significantly, none of these methods used any labels, requiring only CXRs, reports, and metadata that were naturally produced as part of medical workflows.

To assess whether models can generalize across clinical distributions, we chose a wide variety of downstream CXR datasets that we used to finetune and validate our models. These datasets came from diverse sites, including Brazil[23], China[24,25], United States[,25,,26,27,28,29,30,31], Spain[32], Japan[33], Vietnam[34] and other countries in Eastern Europe and Central Asia.[35] They also addressed a large set of tasks, such as lung nodule detection, line and tube placement, edema severity classification, pediatric pneumonia detection, pneumothorax detection and multi-class differential diagnosis; some datasets address wide-ranging detection tasks for multiple pathologies. Furthermore, these datasets were collected in a variety of clinical settings, including a pediatric setting (Pediatric Pneumonia); a Tuberculosis screening program (Montgomery); an emergency department (MIMIC CXR); outpatient (Shenzhen) and in-hospital settings (see Table 1).

Dataset (No. of x-rays, No. of patients)	Country	Setting	Predictive Task	Task Type
RANZCR CLiP (30083, -)	United States	Chest X-ray images from MIMIC-CXR* and NIH ChestXray14.	Detect presence and position of catheters and lines on chest X-rays and categorize poorly placed tubes	Multi-label (11)
SIIM-ACR Pneumothorax (15252, -)	Spain	From NIH dataset with augmented annotations created by radiologists from Society for Imaging Informatics in Medicine (SIIM) and the Society of Thoracic Radiology.	Predict the presence of pneumothorax	Binary Classification
Pediatric Pneumonia (5856, -)	China	Pediatric patients (1-5yrs) at the Guangzhou Women and Children's Medical Center.	Predict the presence of pneumonia	Binary Classification
Shenzhen (662, 662)	China	Outpatient clinics at Shenzhen No. 3 People's Hospital in Shenzhen, China.	Predict the presence of tuberculosis (The Shenzhen and Montgomery datasets were combined for this task)	Binary Classification
Montgomery (138, 138)	United States	Montgomery County's Tuberculosis Screening Program.		
Pulmonary Edema Severity (17857, 1916)	United States	CXRs of patients identified to have congestive heart failure within the MIMIC-CXR dataset.	Predict the severity of pulmonary edema from a chest X-ray	Multi-class (4)
TB Portals (1532, 1532)	Eastern Europe/Central Asia (Azerbaijan, Belarus, Georgia, India, Kazakhstan, Moldova, Nigeria, Romania, Ukraine)	Medical sites around Eastern Europe/Central Asia with high percentages of drug resistant infections	Predict Smear Classification from a chest X-ray	Binary Classification
Indiana University (8121, -)	United States	Two large hospital systems within the Indiana Network for Patient Care database.	Identify normal structures and predict the presence of multiple pathologies from a Chest-X-ray	Multi-label (20)
JSRT (247, 247)	Japan, United States	Thirteen institutions in Japan and one in the United States.	Predict the subtlety of lung nodules present in a chest X-ray	Multi-class (6)
PadChest (160868, 69882)	Spain	Hospital Universitario de San Juan, Alicante (Spain) from 2009 to 2017	Detect multiple pathologies from chest X-rays	Multi-label (9)
REFLACX (3052, -)	United States	MIMIC-CXR filtered for including only CXRs that contained "ViewPosition" metadata with value "PA" or "AP"	Detect multiple pathologies from chest X-rays	Multi-label (9)
BRAX (40967, 19351)	Brazil	Records from the Picture Archiving and Communication System of Hospital Israelita Albert Einstein (HIAE).	Detect multiple pathologies from chest X-rays	Multi-label (14)
NIH ChestX-ray14 (112120, 30805)	United States	Records from the Picture Archiving and Communication System of the NIH Clinical Center.	Detect multiple pathologies from chest X-rays	Multi-label (14)
VinDr-CXR (18000, -)	Vietnam	Retrospectively collected from Hospital 108 (Vietnam) and the Hanoi Medical University Hospital	Detect multiple pathologies from chest X-rays	Multi-label (18)

* MIMIC-CXR consists of chest x-rays and reports from the archives of Beth Israel Deaconess Medical Center, Boston

Table 1: Overview of each target dataset: number of radiographs, number of unique patients and description of the patient population (country and setting), descriptions of prediction task and task type for each dataset. We selected thirteen distinct datasets that vary in size, patient population (age, gender, country) and the type of predictive task.

Once pretrained, models were adapted for each downstream task, using all available training data in the target dataset. In our first set of experiments, we fine tuned all model parameters and compared the performance of SSL CXR-pretrained models against the ImageNet-pretrained models, controlling for the model capacity and hyperparameter search (as detailed in Methods). In the second set of experiments, we performed linear probing, freezing the weights of our CXR pretrained models and training one final fully-connected classifier layer for each. These experiments directly evaluated the quality of the features learned through CXR self-supervision. After training, we tested the performance of our models using the testing data from each dataset.

Dataset	ImageNet-Pretrained Baseline	Maximum Improvement over ImageNet AUC with Linear Probing and Finetuning	
		Linear Probing	*Finetuning*
RANZCR CLiP	0.813 [0.791, 0.833]	0.024 [0.002, 0.049]*	0.071 [0.045, 0.097]*
SIIM-ACR	0.917 [0.905, 0.929]	0.001 [-0.011, 0.016]	0.017 [0.006, 0.029]*
Pediatric Pneumonia	0.930 [0.898, 0.953]	0.064 [0.040, 0.094]*	0.053 [0.029, 0.081]*
Shenzhen & Montgomery	0.923 [0.876, 0.967]	0.033 [-0.014, 0.076]	0.046 [-0.001, 0.099]
MIMIC CXR	0.713 [0.685, 0.740]	0.078 [0.058, 0.098]*	0.089 [0.060, 0.119]*
TB Portals	0.707 [0.665, 0.766]	0.008 [-0.069, 0.074]	0.008 [-0.060, 0.070]
Indiana	0.727 [0.650, 0.798]	0.095 [0.009, 0.176]*	0.072 [-0.001, 0.150]
JSRT	0.543 [0.347, 0.709]	0.044 [-0.175, 0.257]	0.015 [-0.182, 0.204]
PadChest	0.635 [0.573, 0.694]	-0.052 [-0.137, 0.035]	0.024 [-0.041, 0.096]
REFLACX	0.744 [0.674, 0.804]	0.109 [0.037, 0.190]*	0.103 [0.035, 0.181]*
BRAX	0.779 [0.734, 0.821]	0.042 [0.000, 0.087]	0.041 [0.003, 0.077]*
NIH ChestX-ray14	0.706 [0.683, 0.731]	0.091 [0.061, 0.118]*	0.108 [0.082, 0.133]*
VinDR-CXR	0.814 [0.780, 0.847]	0.064 [0.027, 0.102]*	0.065 [0.028, 0.103]*

Table 2: Maximum AUC improvements over ImageNet pretraining achieved across the seven self-supervised models for each dataset. Starred entries indicate statistically significant improvements ($p < 0.05$) over the ImageNet Pretrained baseline. SSL models outperformed ImageNet pretrained models across datasets with distinct clinical tasks and patient populations.

Self-supervised CXR pretrained models outperform ImageNet pretrained models across clinical distributions

We found that finetuned SSL models consistently provided improvements over ImageNet pretraining across datasets (see Table 2). The complete tables showing these results can be found in Supplementary Note 3. In particular, the highest improvements in the average AUC achieved across tasks were seen on the NIH dataset with ConVIRT (0.108 improvement; [95 % CI 0.082, 0.1333]) and on the REFLACX dataset, also with ConVIRT (0.103 improvement, [95 % CI [0.035,0.181]). On both datasets, the highest AUC for each task (each pathology/condition in a dataset) was obtained by a self-supervised model; on the NIH dataset, the highest improvement was seen when identifying atelectasis (0.250 improvement [95% CI 0.224,0.278] with ConVIRT), while on the REFLACX dataset, the highest improvement was seen for pneumothorax detection (AUC improvement of 0.256 [95% CI 0.118, 0.437] with GLoRIA).
Figure 1 shows a per-task comparison across all multi-label datasets between each dataset's best SSL model and the ImageNet-pretrained model.

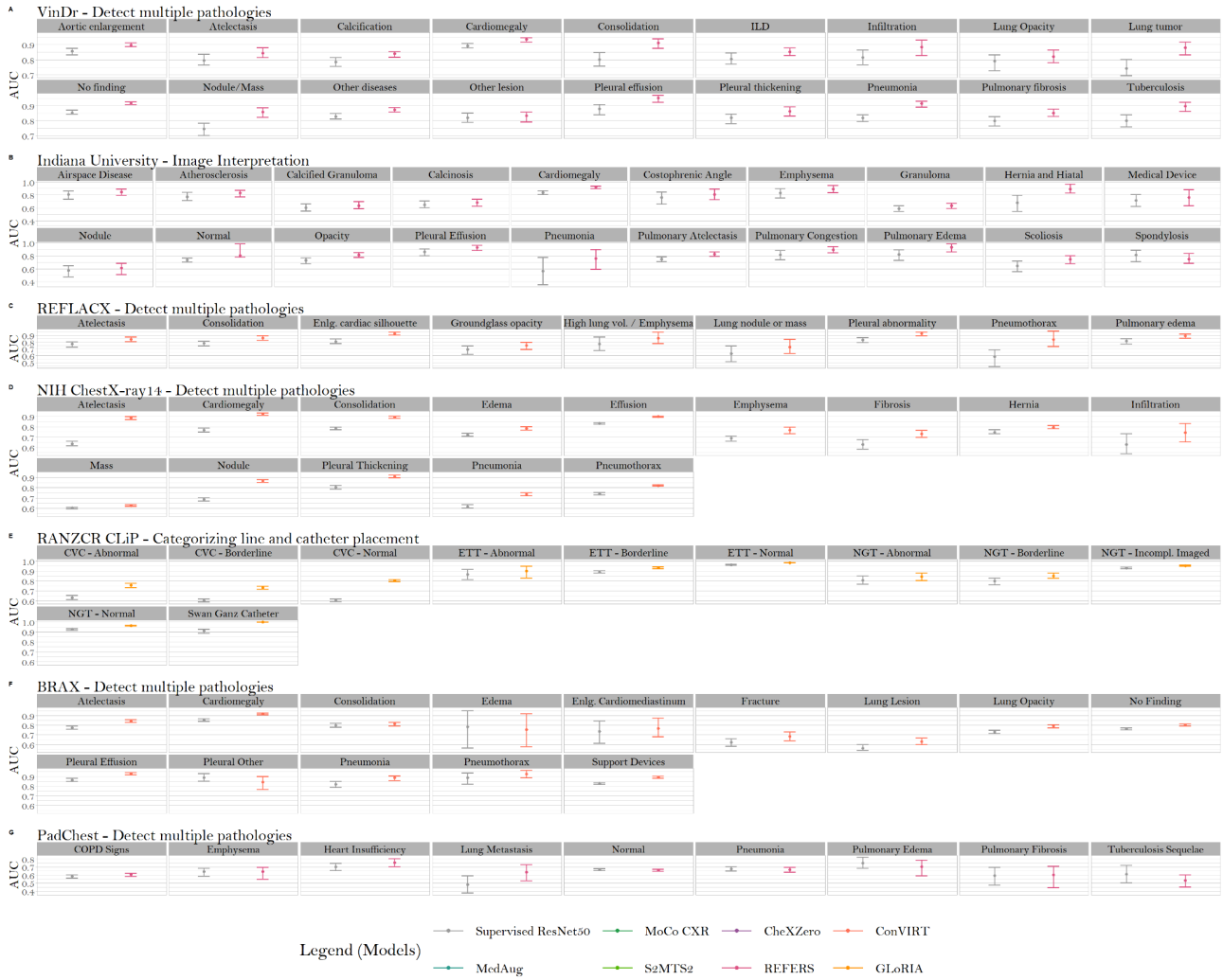

Figure 1. AUC score breakdown by task for the seven multi-label datasets with finetuning. On each task, we compare the AUC of the dataset's overall best-performing model against that of the ImageNet-pretrained model. Across datasets, SSL models had a great performance detecting cardiomegaly, pleural effusion, pneumothorax and ETT placement.

We found that with linear probing, SSL models outperformed the finetuned ImageNet model on twelve of thirteen datasets (see Table 2). The largest average improvements in performance across tasks were seen on the REFLACX dataset with GLoRIA (0.109 difference over the supervised model, [95% CI 0.037, 0.190]) and on the NIH ChestX-ray14 dataset with REFERS (0.091 difference, [95% CI 0.061,0.118]). On both datasets, the highest AUC per task was achieved by a self supervised CXR pretrained model. On the pathologies within REFLACX, the largest performance improvement was achieved by GLoRIA when recognizing pneumothorax (0.215 improvement, [95% CI 0.068,0.412] over the ImageNet-pretrained model after finetuning). Within the NIH dataset, the largest improvement was seen with CXR-RePaiR-CLIP when identifying atelectasis (0.244 improvement, [95% CI 0.221,0.266]). PadChest is the only dataset where the finetuned ImageNet pretrained model outperformed all self-supervised models after linear probing, though its improvement was not statistically significant: the AUC of the best-performing self-supervised model (CXR-RePaiR-CLIP) was lower by only -0.005 [95 % CI -0.088,0.073].

Comparing Different SSL Methods

To evaluate whether pretraining on images and text together yielded an advantage over image-only methods, we compared the average performance across all datasets of image-text with image-only SSL models with finetuning and linear probing. To further assess whether finetuning significantly boosted SSL models' performance, for both image-text and image-only methods, we compared each model's average performance across all datasets between finetuning and linear probing.

We found that finetuned SSL models that used both images and text outperformed the image-only self-supervised methods (See Figure 2 and Table 4 of Supplementary Note 3). On four of the thirteen datasets, the highest average AUC was achieved by GLoRIA and ConVIRT. The best-performing (highest average finetuning improvement across all datasets) image-text methods were ConVIRT (AUC of 0.810) and REFERS (AUC of 0.810), closely followed by GLoRIA (AUC of 0.799). The best-performing image-only method was S2MTS2 (AUC of 0.781).

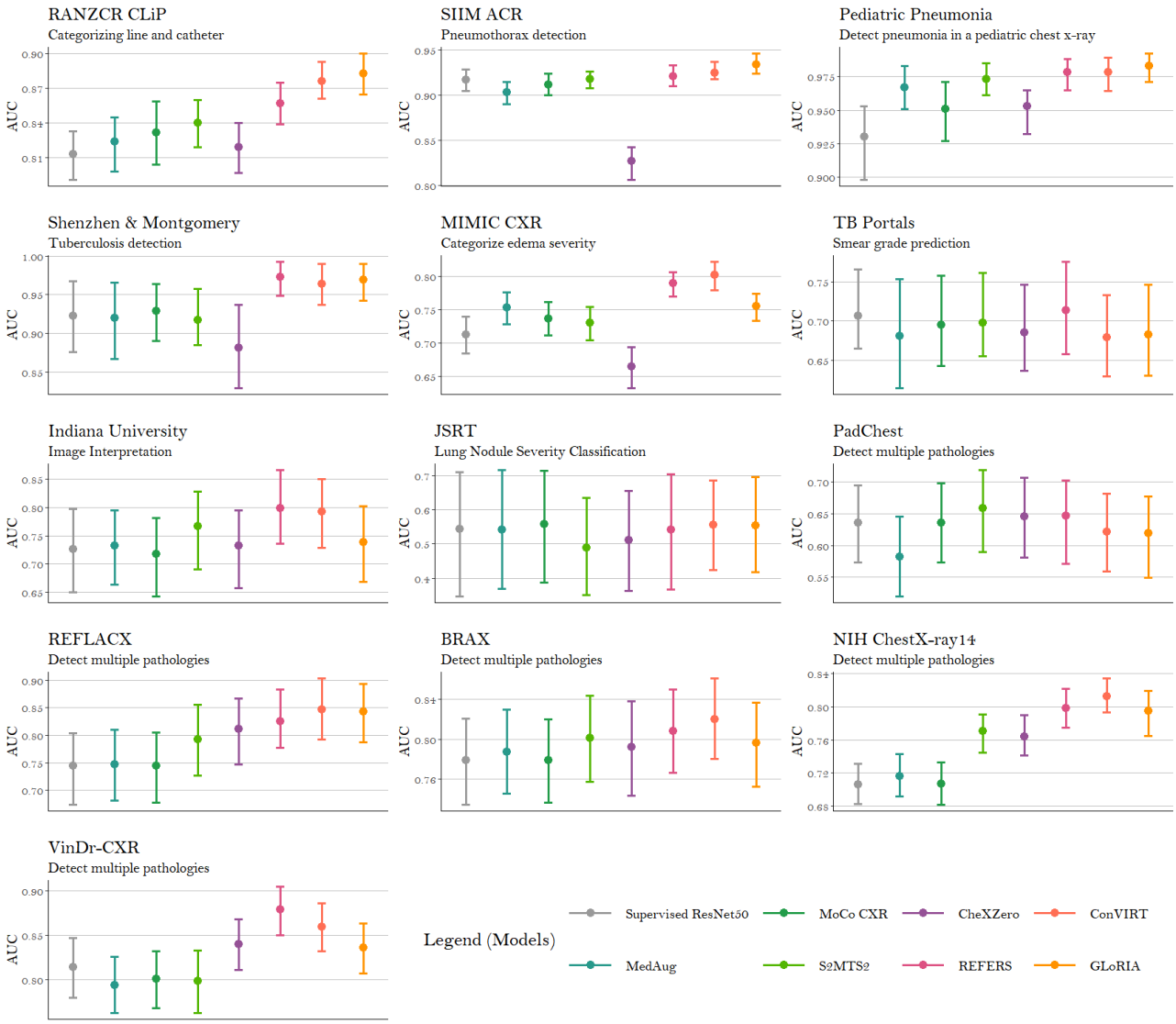

Figure 2. Finetuning AUC (averaged across tasks for datasets containing multiple tasks) achieved by finetuning of each ImageNet-pretrained model and seven self-supervised models on thirteen datasets. The figure shows the AUC of each SSL model and the ImageNet-pretrained model. SSL models that used both image and text outperformed image only SSL models.

We found that, with linear probing, all image-text SSL models outperformed image-only SSL models averaged across datasets. (See Figure 3 and Table 5 of Supplementary Note 3). REFERS had the best average performance (AUC 0.802), followed closely by ConVIRT (average AUC 0.792), CXR-RePaiR-CLIP (average AUC 0.789) and GloRIA (average AUC 0.783). All three image-only SSL models underperformed the finetuned ImageNet-pretrained model. The best image-only SSL model was MedAug, which achieved an average AUC of 0.724 across datasets.

We found that the best linear probing models outperformed the best finetuned models on six out of thirteen datasets. However, model performance on average across datasets was better after finetuning. Only CXR-RePaiR-CLIP had a higher performance with linear probing (0.025 average improvement over finetuning). The three models that benefited the most from finetuning were pretrained only on images: S2MTS2 (0.069 difference over linear probing), MoCo-CXR (0.059), and MedAug (0.041).

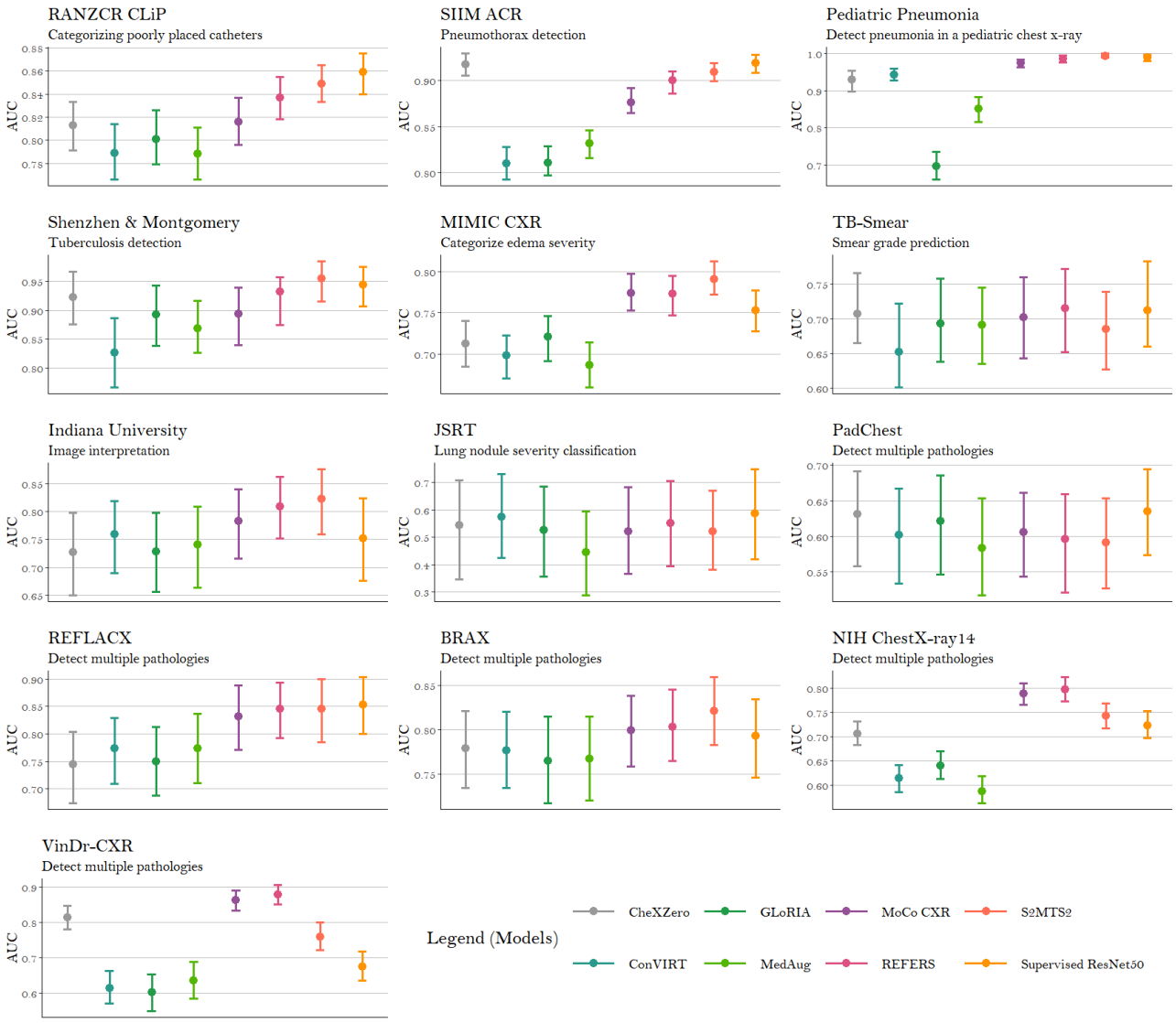

Figure 3. Linear Probing AUC (averaged across tasks for datasets containing multiple tasks) achieved through linear probing of each ImageNet-pretrained model and seven self-supervised models on thirteen datasets. (A) to (M) figure shows the AUC of each SSL model and the ImageNet-pretrained model. SSL models that used both image and text outperformed image only SSL models.

Discussion

Performance across distribution shifts

Several studies have demonstrated a significant drop in performance across a variety of CXR applications when deep learning systems were tested on datasets outside their training domain[36,37,38] . Machine learning models trained through supervised learning often achieve high performance on only a single dataset due to multiple factors, such as technical differences in how CXR images are captured across datasets and varying disease distributions across different populations[37]. We demonstrated that the best self-supervised models maintained their performance across datasets from different geographies, populations and tasks. Although there was not a single

model that maintained high performance across all thirteen datasets, our results are promising, suggesting that self-supervised pretraining can broadly improve generalizability. We believe that labeled data is still essential for high performance and generalization, since both supervised and self-supervised models had consistently poor performance across PadChest and JSRT both for finetuning and linear probing.

Self-supervised models learn better representations using both images and reports

Image-text models learned higher-quality, more generalizable representations than image-only models, across both our finetuning and linear probing experiments. Amongst the image-text models, ConVIRT and REFERS were generally the best-performing models after both finetuning and linear probing. However, they did not achieve the highest performance on all datasets, indicating that while certain models can learn more useful representations on average, they may still falter on specific tasks. Our results did not find any one SSL model that consistently outperforms other models on every new dataset / new task, and work remains to be done on this front.

Clinical Implications

Although the performance of CXR self-supervised models was overall better than ImageNet-pretrained supervised models, results were not homogeneous across different tasks and datasets.

For tasks that assess the placement of lines and tubes, as seen on the the RANCZR CLiP dataset, CXR-pretrained models generally did better than the ImageNet-pretrained model. Both kinds of models could successfully detect NGT or ETT placement, but neither model did well enough to evaluate central line placements. The latter is a more complicated task, because this dataset includes different types of central catheters (central line, peripherally inserted central catheter, Ports, Hickman's) and because these catheters can be inserted at different levels and from either the right or left side of the body. Performance likely dropped due to the increased complexity of the task.

For complex tasks with multiple pathologies, we observed that our models performed poorly on datasets with multiple labels that did not clearly distinguish between a specific finding and a diagnostic impression. Specifically, they underperformed on PadChest when distinguishing between heart insufficiency and pulmonary edema or emphysema and chronic obstructive pulmonary disease (COPD), even though PadChest is a relatively large dataset with more than 160,000 images. These labels can be particularly difficult to apply, and even the most experienced radiologists disagree on the correct label due to ambiguity in instructions. These findings contribute to the hypothesis that size is not the only important feature of a dataset; it is also necessary to have clear, specific labels, especially for complex tasks.

When the task involves a more subjective appraisal, as when determining the severity of pulmonary edema or of subtlety of a pulmonary nodule, we found that neither set of models did well. These results mirror clinical practice, where board-certified radiologists and emergency medicine physicians demonstrate a sensitivity of 77% and 59% respectively, when detecting the presence of edema, and their inter-agreement is also low[39,40]. Similarly, when detecting lower-subtlety pulmonary nodules, performance decreases even among experienced radiologists[41].

Although more research is necessary, self-supervised pretraining is a promising approach to image interpretation tasks, especially for tasks where datasets are small but contain clear, specific labels. A recent study found that pretraining the data on natural images, followed by pretraining on large weakly labeled CXR datasets, and finally, task-specific training on small labeled datasets could significantly reduce label requirements on the target dataset[42]; thus the composition of large-scale out-of-domain pretraining and self-supervised pre training may be a fruitful direction for future work.

Overall, CXR self-supervised pretrained models did better than ImageNet-pretrained models, especially for image-text models. There was no task where the ImageNet-pretrained model consistently did better than models pretrained through CXR self-supervision. Tasks where CXR self-supervised models did particularly well included the detection of pneumothorax, cardiomegaly, pleural effusion, and tube placement. Within these tasks, models maintained high performance across different populations and clinical settings, indicating that these models could be ready for deployment, assisting clinicians in identifying those critical conditions instantly. It is also important to note that the quality of the labels may influence performance more than the number of images and labels in a specific dataset; this is an important consideration when creating future datasets.

Conclusion

In this study, we tested whether the emerging paradigm of self-supervised pretraining would outperform the traditional paradigm of ImageNet pretraining on different chest x-ray interpretation tasks. Traditionally, ImageNet pretraining has been the dominant approach to tackling medical imaging tasks with limited data. However, our results indicate that self-supervised pretraining on medical data could take over as the traditional paradigm for low-data medical tasks. In addition to significantly improving performance across downstream tasks, this study represents a step forward in learning generalizable representations of chest X-rays. Such self-supervised pretrained models could be easily adapted to and deployed in diverse clinical settings.

Methods

Model Descriptions

We benchmarked the performance of seven SSL methods. The methods were primarily divided into two categories, depending on whether the pretraining tasks used images alone or used both images and reports. The image-only models (MedAug, S2MTS2, MoCo-CXR) were pretrained on CXR images, whereas image+reports models (CXR-RePaiR-CLIP, ConVIRT, REFERS, GLoRIA) were pretrained on the CXR images along with corresponding reports. These models used different backbone architectures (ResNet-50[43], CLIP-ViT[44], Dense121[45]) and underlying pretraining frameworks (contrastive learning[46], CLiP, transformers[47], self-attention[48], student-teacher[49], cross-supervision[50]). A detailed description of these methods is given in Table 3.

Method	Backbone Architecture	Pretrained Dataset	Method Pre-training type	Description	Number of tunable Parameters (Linear Probing per output node, Finetuning)
MedAug	ResNet-50	MIMIC-CXR	CXR Images only	MoCo based contrastive learning based SSL method utilizing patient metadata for pairing positive/negative examples.	(2048, 23M)
CXR-RePaiR-CLIP	CLIP - ViT-B32	MIMIC-CXR	CXR Images + Reports	CLiP based zero-shot learning based SSL method pretrained on CXR images and reports. CheXzero generalizes to unseen classes without any explicit manual or annotated labels.	(512, 151M)
REFERS	ViT-B16	MIMIC-CXR	CXR Images + Reports	Cross-supervised methodology reviewing free-text reports for supervision, which takes supervision cues from original radiology report supplementing the CXRs. REFERS uses a vision transformer to learn joint representations from multiple views within every patient study.	(768, 86M)
ConVIRT	ResNet-50	MIMIC-CXR	CXR Images + Reports	SSL method to learn CXR visual representations from naturally occurring pairing of images and reports. ConVIRT pretrains ResNet50 image encoder with paired text data via a domain -agnostic bidirectional contrastive objective between the two modalities.	(2048, 23M)
GLoRIA	ResNet-50	CheXpert	CXR Images + Reports	Self-attention based method for learning global and local representations by contrastive learning between CXR image sub-regions and words in the paired report.	(2048, 23M)
S2MTS2	Dense121	CheXpert	CXR Images only	S2MTS2 is a self-supervised mean-teacher pre-training method based on joint contrastive learning with semi-supervised fine-tuning. It utilises infinite number of pairs of positive query and key features to improve the mean-teacher representation.	(1000, 25M)
MoCo-CXR	ResNet-50	CheXpert	CXR Images only	Momentum Contrast (MoCo) based SSL method pretrained on CXR images using contrastive learning with no fixed metadata based pairing like MedAug.	(2048, 23M)

Table 3. Overview of SSL methods with backbone architecture, pretraining dataset, pretraining type, description and number of tunable parameters during linear probing and finetuning.

Adapting Strategies

We adapted self-supervised models in two ways: linear probing and finetuning. For linear probing, we added a logistic regression layer on top of the last layer of the backbones for 100 epochs, using the L2 norm and Limited-memory Broyden–Fletcher–Goldfarb–Shanno solver. For finetuning, we tuned all the model parameters with SGD optimizer for 100 epochs with a learning rate of 0.001, a batch size of 32, and a learning rate scheduling step of 200 with cosine annealing. These sets of hyperparameters are given in the official implementation of respective methods. Model parameter selection was based on the model with the highest AUC on the validation set. We ran the finetuning of SSL methods on a single NVIDIA RTX A4000 GPU card.

Dataset Selection

We evaluated CXR self-supervised models across a range of tasks, using fourteen distinct datasets that vary in size, patient population, geography, setting, and type of predictive task. These tasks are a mix of binary classification (5 datasets), multi-label (7 datasets) and multi-class (2 datasets) tasks. The tasks that we have selected are described in Table 1.

Dataset Processing

We split every dataset except the Pediatric Pneumonia and VinDr-CXR datasets into training, validation and test sets with 60%, 20%, and 20% of datapoints respectively. We ensured that there was no patient overlap between the three sets. The Pediatric Pneumonia and VinDr-CXR datasets were split into (80%, 10%, 10%) and (66%, 16%, 16%) respectively, as we retained the test split provided by the creators of the dataset.

Statistical Analysis

We evaluated the performance of models using AUC scores, representing the area under the receiver operating characteristic curve. We reported the variability in our estimate of this measure using the percentile bootstrap: we constructed bootstrap confidence intervals by taking 1000 replicates of sampling test data points with replacement.

The nonparametric bootstrap was used to estimate the variability around each of the performance measures; 1000 bootstrap replicates from the test set were drawn, and each performance measure was calculated for the ImageNet-pretrained baseline model and the seven self-supervised models on these same 1000 bootstrap replicates. This process produced a distribution for each estimate, and the 95% bootstrap percentile intervals (2.5th and 97.5th percentiles) are reported. To find the performance improvement provided by the best SSL model over the ImageNet supervised learning model, we computed the differences of 1000 replicated AUCs and reported the 95% bootstrap percentile intervals[51]. The analyses were performed using statsmodels, scikit-learn and SciPy packages in Python.

Package description - *cxrlearn*

The SSL methods implemented in this work were developed by multiple research groups across different platforms with multiple dependencies. We have integrated all SSL codebases into one reusable library called *cxrlearn*. *cxrlearn* provides documented functions for dataset conversion, self-supervised learning, model object generation, finetuning, linear probing, and evaluation. Our open-source library codebase is available here.

Data availability

Most datasets used in this study are public and can be accessed through their respective websites. Requests concerning the TB portals dataset should be addressed to Maha Farhat, (Maha_Farhat@hms.harvard.edu).

References

1. Rimmer, Abi. "Radiologist shortage leaves patient care at risk, warns royal college." *BMJ: British Medical Journal (Online)* 359 (2017).
2. Mahajan, Vidur, and Vasanth Venugopal. "Audit of Artificial Intelligence Algorithms and Its Impact in Relieving Shortage of Specialist Doctors." *Artificial Intelligence: Applications in Healthcare Delivery* (2020): 207.
3. Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. (2018) Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 15(11): e1002686. https://doi.org/10.1371/journal.pmed.1002686
4. P Rajpurkar, J Irvin, K Zhu, *et al.* Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning arXiv (2017),published online Nov 14.https://arxiv.org/abs/1711.05225
5. JT Wu, KCL Wong, Y Gur, *et al.* Comparison of chest radiograph interpretations by artificial intelligence algorithm vs radiology residents JAMA Netw Open, 3 (2020), Article e2022779
6. S Jang, H Song, YJ Shin, *et al.* Deep learning-based automatic detection algorithm for reducing overlooked lung cancers on chest radiographs Radiology, 296 (2020), pp. 652-661
7. Willemink, Martin J., et al. "Preparing medical imaging data for machine learning." *Radiology* 295.1 (2020): 4-15.
8. Irvin, Jeremy, et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. No. 01. 2019.
9. Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.
10. Esteva, Andre, et al. "Dermatologist-level classification of skin cancer with deep neural networks." *nature* 542.7639 (2017): 115-118.
11. De Fauw, Jeffrey, et al. "Clinically applicable deep learning for diagnosis and referral in retinal disease." *Nature medicine* 24.9 (2018): 1342-1350.
12. Mitani, Akinori, et al. "Detection of anaemia from retinal fundus images via deep learning." *Nature Biomedical Engineering* 4.1 (2020): 18-27.
13. Raghu, Maithra, et al. "Transfusion: Understanding transfer learning for medical imaging." *Advances in neural information processing systems* 32 (2019).
14. Krishnan, Rayan, Pranav Rajpurkar, and Eric J. Topol. "Self-supervised learning in medicine and healthcare." *Nature Biomedical Engineering* (2022): 1-7.
15. Azizi, Shekoofeh, et al. "Big self-supervised models advance medical image classification." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
16. Nhi Truong Vu, Yen, et al. "MedAug: Contrastive learning leveraging patient metadata improves representations for chest X-ray interpretation." *arXiv e-prints* (2021): arXiv-2102.
17. Liu, Fengbei, et al. "Self-supervised mean teacher for semi-supervised chest x-ray classification." *International Workshop on Machine Learning in Medical Imaging*. Springer, Cham, 2021.
18. Sowrirajan, Hari, et al. "Moco pretraining improves representation and transferability of chest X-ray models." *Medical Imaging with Deep Learning*. PMLR, 2021.
19. Endo, Mark, et al. "Retrieval-Based Chest X-Ray Report Generation Using a Pre-trained Contrastive Language-Image Model." *Machine Learning for Health*. PMLR, 2021.
20. Zhang, Yuhao, et al. "Contrastive learning of medical visual representations from paired images and text." *arXiv preprint arXiv:2010.00747* (2020).
21. Zhou, Hong-Yu, et al. "Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports." *Nature Machine Intelligence* (2022): 1-9.

22. Huang, Shih-Cheng, et al. "GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-Efficient Medical Image Recognition." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
23. Reis, Eduardo Pontes. "BRAX, a Brazilian labeled chest X-ray dataset."
24. Kermany, Daniel S., et al. "Identifying medical diagnoses and treatable diseases by image-based deep learning." *Cell* 172.5 (2018): 1122-1131.
25. Jaeger, Stefan, et al. "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases." *Quantitative imaging in medicine and surgery* 4.6 (2014): 475.
26. Tang, Jennifer SN, et al. "CLiP, catheter and line position dataset." *Scientific Data* 8.1 (2021): 1-7.
27. https://www.kaggle.com/jesperdramsch/siimacrpneumothorax-segmentation-data.
28. Lanfredi, Ricardo Bigolin, et al. "REFLACX, a dataset of reports and eye-tracking data for localization of abnormalities in chest X-rays." *arXiv preprint arXiv:2109.14187* (2021).
29. Wang, Xiaosong, et al. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
30. Demner-Fushman, Dina, et al. "Preparing a collection of radiology examinations for distribution and retrieval." *Journal of the American Medical Informatics Association* 23.2 (2016): 304-310.
31. Johnson, Alistair EW, et al. "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports." *Scientific data* 6.1 (2019): 1-8.
32. Bustos, Aurelia, et al. "Padchest: A large chest X-ray image dataset with multi-label annotated reports." *Medical image analysis* 66 (2020): 101797.
33. Junji, Shiraishi, et al. "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules." *American Journal of Roentgenology* 174.1 (2000): 71-74.
34. Nguyen, Ha Q., et al. "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations." *arXiv preprint arXiv:2012.15029* (2020).
35. Gabrielian, Andrei, et al. "TB DEPOT (Data Exploration Portal): A multi-domain tuberculosis data analysis resource." *Plos one* 14.5 (2019): e0217410.
36. Zech, John R., et al. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study." *PLoS medicine* 15.11 (2018): e1002683.
37. Cohen, Joseph Paul, et al. "On the limits of cross-domain generalization in automated X-ray prediction." *Medical Imaging with Deep Learning*. PMLR, 2020.
38. Lam, Carson, et al. "Automated detection of diabetic retinopathy using deep learning." *AMIA summits on 18translational science proceedings* 2018 (2018): 147.
39. Hammon, Matthias, et al. "Improving diagnostic accuracy in assessing pulmonary edema on bedside chest radiographs using a standardized scoring approach." *BMC anesthesiology* 14.1 (2014): 1-9.
40. Kennedy, Sarah, et al. "Ability of physicians to diagnose congestive heart failure based on chest X-ray." *The Journal of emergency medicine* 40.1 (2011): 47-52.
41. Robinson, John W., et al. "The impact of radiology expertise upon the localization of subtle pulmonary lesions." *Medical Imaging 2016: Image Perception, Observer Performance, and Technology Assessment*. Vol. 9787. International Society for Optics and Photonics, 2016.
42. Sellergren, Andrew B., et al. "Simplified Transfer Learning for Chest Radiography Models Using Less Data." *Radiology* (2022): 212482.
43. He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
44. Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International Conference on Machine Learning*. PMLR, 2021.

45. Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
46. Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." *International conference on machine learning*. PMLR, 2020.
47. Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).
48. Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
49. Abbasi, Sajjad, et al. "Modeling teacher-student techniques in deep neural networks for knowledge distillation." *2020 International Conference on Machine Vision and Image Processing (MVIP)*. IEEE, 2020.
50. Gupta, Saurabh, Judy Hoffman, and Jitendra Malik. "Cross modal distillation for supervision transfer." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
51. Efron, Bradley, and Robert J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Acknowledgements

PR was supported by a Harvard Data Science Initiative Competitive Research Fund Grant and by the NHLBI BioData Catalyst Fellowship.

Author information

Department of Computer Science, Stanford University, Stanford, United States
Niveditha Iyer, Aditya Gulati, Cécile Logé
Brigham and Women's Hospital, Boston, Massachusetts, United States
Agustina Saenz, MD, MPH
Department of Biomedical Informatics, Harvard Medical School, Boston, United States
Agustina Saenz MD, MPH, Oishi Banerjee MS, Maha Farhat, MD, MSc, Pranav Rajpurkar PhD

Contributions

Concept and design: PR, NI, AG. Acquisition, analysis, or interpretation of data: AG, NI, PR, AS, CL, MF. Drafting of the manuscript: NI, AS, AG, OB. Critical revision of the manuscript for important intellectual content: All authors. Statistical analysis: AG, NI. Obtained funding: PR. Administrative, technical, or material support: PR. Supervision: PR. All authors approved the final version.

Corresponding author

Correspondence to Pranav Rajpurkar, PhD (pranav_rajpurkar@hms.harvard.edu).

Ethics declarations

The authors declare no competing interests.