DEEP LEARNING FOR MEDICAL IMAGE INTERPRETATION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Pranav Rajpurkar
June 2021

Abstract

There have been rapid advances at the intersection of deep learning and medicine over the last few years, especially for the interpretation of medical images. In this thesis, I describe three key directions that present challenges and opportunities for the development of deep learning technologies for medical image interpretation. First, I discuss the development of algorithms for expert-level medical image interpretation, with a focus on transfer learning and self-supervised learning algorithms designed to work in low labeled medical data settings. Second, I discuss the design and curation of high-quality datasets and their roles in advancing algorithmic developments, with a focus on high-quality labeling with limited manual annotations. Third, I discuss the real-world evaluation of medical image algorithms with studies systematically analyzing performance under clinically relevant distribution shifts. Altogether this thesis summarizes key contributions and insights in each of these directions with key applications across medical specialties.

Acknowledgments

This PhD is the accomplishment of the people that nurtured, inspired, and supported me over my academic career. My parents, Samir Rajpurkar and Nilam Rajpurkar, instilled in me a love for learning at an early age, and have been my guiding lights for the past 26 years and counting. This thesis is dedicated to them.

My interest in Artificial Intelligence (AI) was sparked by a talk I attended during my undergraduate orientation by Professor Andrew Ng, who I call Professor Ng to this day. I would get my start in AI research later that year in Professor Ng's research lab, and continue there throughout my undergraduate years culminating in my honors thesis on autonomous driving. During this time, I had the fantastic mentorship of senior members of the lab, including Sameep Tandon, Tao Wang, Brody Huval, and Adam Coates. Among the many lessons I have learned from Professor Ng in the past nine years, his advice to me to think about the scale of positive impact in whatever I choose to do has had a profound influence on how I select and approach problems.

My interest in pursuing a PhD was cemented my sophomore summer when I did HCI research under the mentorship of Professor Michael Bernstein. During this time, Michael and my project mentor, Ethan Fast, swiftly shaped my know-how of the research process. Twice a week at our scheduled times, I would walk into Michael's office for discussions that were as exhilarating as they were inspiring. It was the same office I would walk into a year later when Michael shared the news of my acceptance into Stanford's PhD program, which I can comfortably characterize as one of the happiest moments of my life.

The first paper I wrote as a PhD student came towards the end of my first year under the mentorship of Professor Percy Liang. Among the many lessons I've learned from Percy over the years, the importance of drilling deeper and calibrating myself in the domain in which I was working would help me numerous times later. At the time, seeing the implementation of this lesson translate into the impact that the publication, SQuAD, with Percy had in the community was particularly important in giving me the confidence to dive into a new domain.

I want to additionally thank my co-advisors Professor Andrew Ng and Professor Percy Liang, in addition to Professor Michael Bernstein, for their advice on this thesis. I also want to thank Orals committee members Professor Mykel Kochenderfer and Professor Leanne Williams for their

mentorship and enthusiastic support.

My interest in working at the intersection of AI and medicine started towards the end of the second year of my PhD, and this interest blossomed over the next years as I collaborated with some incredible faculty including Professors Curt Langlotz, Sanjay Basu, Nigam Shah, Jeanne Shen, Bhavik Patel, Kristen Yeom, Leanne Williams, Utkan Demirci, Gozde Durmus, Sidhartha Sinha, Catherine Hogan, Sebastian Fernandez-Pol, Yaso Natkunam, Mitchell Rosen, Geoff Tison, Andrew Beam, David Kim, Nikhil Agarwal, Tobias Salz, and Eric Topol.

An early collaboration with Professor Matthew Lungren on chest X-ray interpretation evolved into a fruitful, multi-year effort to push towards the deployment of this technology. Over our numerous collaborations over the years, I hope that I have picked up some of Matt's boldness in the pursuit of projects and cheery nature that have made working with Matt such a pleasure.

As we were setting up early medical AI projects, Jeremy Irvin and I also got to set up new office desks and chairs in the Stanford ML Group HQ, Gates 108. In this office, I got to spend many memorable mornings, afternoons, evenings, and nights. I am grateful for having had the opportunity to work with labmates Awni Hannun, Anand Avati, Swati Dube, Sharon Zhou, Hao Sheng, Ziang Xie, and Robin Jia.

Outside of lab, I had the chance to spend a wonderful summer working with Bora Uyumazturk and Amir Kiani building out a Coursera course on AI for Medicine. I was fortunate to get to work with Emma Chen and Professor Eric Topol on building Doctor Penguin, a newsletter highlighting latest AI and medicine research. In the middle of the COVID pandemic, I was grateful to be able to start The AI Health Podcast with Adriel Saporta and Oishi Banerjee; I would only meet Adriel in person after a year of Zoom calls. I got to end my Stanford student career on a high note with an entrepreneurship class I took with the GloFlow team of Anirudh Joshi, Damir Vrabac, and Viswesh Krishna.

As someone who has called Stanford home for the past nine years, first as an undergrad and then as a PhD student, I want to thank all the people who made my time at Stanford the best years of my life. Brad Girardeau, who by some miracle got matched as my freshman year roommate, has been a best friend, inspiration, and support through all these years. My roommate, Alex Tamkin, has been my go-to source of seemingly infinite scientific wisdom, puns, and late-night chats. I also want to thank Anuj Pareek, Tina Diao, Siyu Shi, Saahil Jain, Akshay Smit, Henrik Marklund, Mars Huang, Aarti Bagul, Amita Kamath, Nathan Dass, Catherine Gu, Trisha Singh, Jordan Cotler, Daniel Ranard, Edward Mazenc, Omar Qureshi, Jared Quincy Davis, and Dillon Laird.

My most defining graduate experience has been leading the AI For Healthcare Bootcamp program at Stanford, where I have had the opportunity to mentor and work with over 118 students on research projects at the intersection of AI and Medicine. I want to thank the incredible people I got to work with and learn from, including Alex Gui, Alex Ke, Alex Wang, Allison Park, Amit Schechter,

Andrew Huang, Andrew Kondrich, Andy Kim, Ashwin Agrawal, Behzad Haghgoo, Ben Cohen-Wang, Brandon Yang, Bryan Gopal, Can Liu, Cécile Logé, Cheuk To Tsui, Chloe O'Connell, Chris Chute, Chris Lin, Chris Wang, Christian Rose, Dahlia Radif, Daisy Ding, Daniel Michael, David Dadey, Ekin Tiu, Ellie Talius, Emily Ross, Emily Wen, Erik Jones, Ethan Chi, Gautham Raghupathi, Gordon Chi, Grace He, Hari Sowrirajan, Hershel Mehta, Ishaan Malhi, Jason Li, Jessica Wetstone, Jiangshan Li, Jingbo Yang, Joe Lou, JC Peruzzi, Jon Braatz, Kaushik Ram Sadagopan, Kaylie Zhu, Kevin Tran, Mark Endo, Mark Sabini, Matthew Sun, Michael Bereket, Michael Ko, Michael Zhang, Minh Phu, Nhi Truong Vu, Nicholas Bien, Nick Phillips, Nidhi Manoj, Niranjan Balachandar, Nishit Asnani, Niveditha Iyer, Pratham Soni, Pujan Patel, Raj Palleti, Rayan Krishnan, Rebecca Gao, Rehaan Ahmad, Richard Wang, Robin Cheong, Ryan Chi, Ryan Han, Shubhang Desai, Silviana Ciurea-Ilcus, Soham Gadgil, Stephanie Zhang, Suvadip Paul, Tanay Kothari, Thao Nguyen, Tom Jin, Tony Duan, Vinjai Vale, William Ellsworth, Yancheng Li, Yifan Yu, Zach Harned, and Zihan Wang.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the years ahead, artificial intelligence (AI) is poised to reshape medicine. AI systems will be routinely used to detect illnesses earlier, improve prognosis and provide more successful, personalized treatment plans, even while saving time and cutting costs. In the near future, algorithms that can read chest X-rays or histopathology slides will manage worklists for medical doctors, enable decision support for clinicians without subspecialty training, and power AI-driven telehealth services. Beyond the hospital, AI technologies will be used to continuously monitor the health of millions of patients, and route patients to physician visits and follow ups with unprecedented scale.

In recent years, deep learning, a form of AI in which neural networks learn patterns directly from raw data, has achieved remarkable success in image classification [128]. Medical AI research has consequently blossomed in specialties that rely heavily on the interpretation of images, such as radiology, pathology and ophthalmology [137]. Much of this progress has been driven by advancements in algorithms and creation of data sets over the last few years. On the algorithmic front, the improvement of convolutional neural network architectures and training procedures has enabled progress for medical imaging applications. In addition, the success of these algorithms has been enabled by the curation of large label data sets for medical imaging. Some AI tools have moved past testing to deployment, clearing regulatory hurdles and winning administrative support [20]. The Center for Medicare and Medicaid Services, which approves public insurance reimbursement costs, has facilitated the adoption of AI in clinical settings by allowing some of the first reimbursements of AI tools for medical image diagnosis [69]. However, there still remains a large gap between the number of deep learning algorithms that have been shown to be successful at medical image interpretation on retrospective datasets and the number translated to clinical practice [116].

This thesis proposes that there have been three key technical challenges towards widespread deployment of deep learning algorithms for medical image interpretation. The first challenge facing the field is that the current development of algorithms focuses on tackling narrow tasks that require a lot of clean data, rather than on tackling a broader range of tasks with noisy or limited label

data commonly found in medicine. The second challenge facing the field is that datasets used to train and validate models are small, noisy, and homogeneous, rather than large, high quality, and heterogeneous. The third challenge facing the field is that current studies validate algorithms in the context of the dataset distributions on which the algorithms were trained, while clinical deployment requires evaluation of algorithm performance under clinically relevant distribution shifts.

Organization

This thesis covers advancements, challenges, and opportunities in the directions of algorithms, datasets, and studies.

Algorithms The past few years have produced the some of the first demonstrations of deep learning algorithms that can make clinically important diagnoses at the level of medical experts across medical specialties, including radiology, cardiology, dermatology, ophthalmology, and pathology [139]. In Chapter 2, I describe the development of an algorithm for detecting diseases in chest X-rays that we showed could perform at a level comparable to practicing radiologists. In Chapter 3, I describe the development of an algorithm for detecting abnormal heart rhythms from electrocardiograms at the level of practicing cardiologists. In both cases, I also describe the collection of large datasets that make it possible to train end-to-end deep learning algorithms. Together, these chapters describe the first demonstrations of expert-level performances on the chest x-ray interpretation and the arrhythmia detection tasks.

One of the main practical challenges towards the development of algorithms is their reliance on manual, time-consuming annotation of data. Especially for biomedical tasks, which require significant expertise for annotation, data labeling at scale required for development of supervised deep learning algorithms is especially challenging. For medical imaging, transfer learning using pre-trained ImageNet [55] models has been the standard approach for developing algorithms in limited labeled data settings [180]. In Chapter 4, I describe the first systematic investigation of the performance and efficiency of ImageNet architectures and weights for chest X-ray interpretation. In Chapter 5 and Chapter 6, I also describe how self-supervised contrastive learning may enable a paradigm shift for training models for medicine where a relatively small number of annotations can enable training of highly accurate models. Together, these chapters describe how transfer learning and self-supervised learning address algorithmic challenges in the limited labeled data medical settings.

Datasets Large, high-quality datasets have play a critical role in driving application and advancement of deep learning algorithms. In the context of medicine, dataset curation requires building partnerships with hospital administrators, frameworks for securely handling and de-identifying data, and strategies for data organization and annotation. In Chapter 7, I describe the curation of a dataset of photos of chest x-rays and synthetic transformations designed to evaluate algorithm performance on

photos of x-rays towards benchmarking robustness in real clinical settings. In Chapter 8, I describe the curation and annotation of a dataset containing tissue microarray slides accompanied by clinical and cytogenetic data from cancer cases towards the discovery of prognostic biomarkers.

For medical imaging datasets, annotation typically requires manual annotation, which can be costly and hard to obtain, and labels acquired through automated methods can be noisy. In the context of supervising computer vision models to interpret medical images, high-quality automatic extraction of medical conditions from free-text radiology reports is critical. In Chapter 9 and Chapter 10, I describe the process of building high-quality radiology report labelers that tackle both noise and the limited availability of expert annotations.

Studies While the majority of foundational work in medical image interpretation has evaluated algorithms on the same dataset distributions on which they were trained, the deployment of these algorithms requires understanding of their performance under clinically relevant distribution shifts. In Chapter 11, I describe a systematic evaluation of the performance of deep learning models in the presence of diseases not labeled for or present during training using chest x-ray interpretation as an example. In Chapter 12, I describe a systematic investigation of different chest X-ray models when applied to smartphone photos of chest X-rays and external datasets without any finetuning.

Overall, this thesis demonstrates progress in deep learning for medical image interpretation using the combination of advancements in (1) algorithms in the context of large and small labeled datasets, (2) datasets through clinically-informed curation and labeling, (3) and studies systematically evaluating the performance of algorithms under clinically relevant distribution shifts.

Chapter 2

Expert-level Deep Learning for Chest X-Rays

We present an algorithm that can interpret x-rays at the level of performance of radiologists. To enable high performance, we develop a large dataset that contains 224,316 chest radiographs of 65,240 patients. We design a labeler to automatically detect the presence of 14 observations in radiology reports, capturing uncertainties inherent in radiograph interpretation. We investigate different approaches to using the uncertainty labels for training convolutional neural networks that output the probability of these observations given the available frontal and lateral radiographs. On a validation set of 200 chest radiographic studies which were manually annotated by 3 board-certified radiologists, we find that different uncertainty approaches are useful for different pathologies. We then evaluate our best model on a test set composed of 500 chest radiographic studies annotated by a consensus of 5 board-certified radiologists, and compare the performance of our model to that of 3 additional radiologists in the detection of 5 selected pathologies. On Cardiomegaly, Edema, and Pleural Effusion, the model ROC and PR curves lie above all 3 radiologist operating points.

This chapter is based on [106].

2.1 Introduction

Chest radiography is the most common imaging examination globally, critical for screening, diagnosis, and management of many life threatening diseases. Automated chest radiograph interpretation at the level of practicing radiologists could provide substantial benefit in many medical settings, from improved workflow prioritization and clinical decision support to large-scale screening and global population health initiatives. For progress, there is a need for labeled datasets that (1) are large, (2) have strong reference standards, and (3) provide expert human performance metrics for

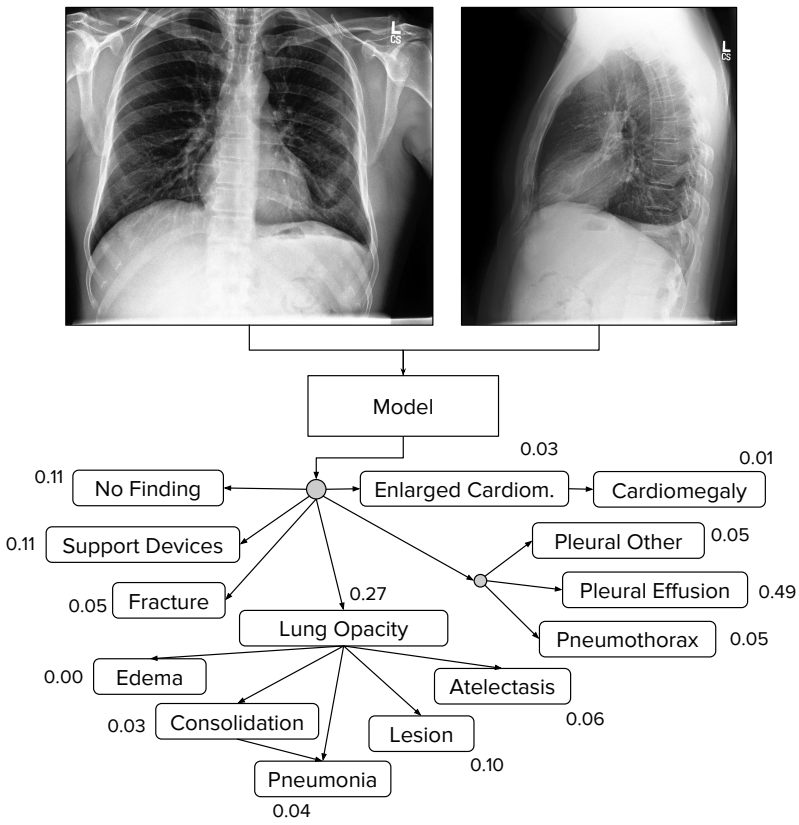


Figure 2.1: The CheXpert task is to predict the probability of different observations from multi-view chest radiographs.

comparison.

In this work, we present CheXpert (**Ch**est e**Xpert**), a large dataset for chest radiograph interpretation. The dataset consists of 224,316 chest radiographs of 65,240 patients labeled for the presence of 14 common chest radiographic observations. We design a labeler that can extract observations from free-text radiology reports and capture uncertainties present in the reports by using an uncertainty label.

The CheXpert task is to predict the probability of 14 different observations from multi-view chest radiographs (see Figure 2.1). We pay particular attention to uncertainty labels in the dataset, and investigate different approaches towards incorporating those labels into the training process. We assess the performance of these uncertainty approaches on a validation set of 200 labeled studies, where ground truth is set by a consensus of 3 radiologists who annotated the set using the radiographs. We evaluate the approaches on 5 observations selected based on their clinical significance and prevalence in the dataset, and find that different uncertainty approaches are useful for different observations.

Pathology	Positive (%)	Uncertain (%)	Negative (%)
No Finding	16627 (8.86)	0 (0.0)	171014 (91.14)
Enlarged Cardiom.	9020 (4.81)	10148 (5.41)	168473 (89.78)
Cardiomegaly	23002 (12.26)	6597 (3.52)	158042 (84.23)
Lung Lesion	6856 (3.65)	1071 (0.57)	179714 (95.78)
Lung Opacity	92669 (49.39)	4341 (2.31)	90631 (48.3)
Edema	48905 (26.06)	11571 (6.17)	127165 (67.77)
Consolidation	12730 (6.78)	23976 (12.78)	150935 (80.44)
Pneumonia	4576 (2.44)	15658 (8.34)	167407 (89.22)
Atelectasis	29333 (15.63)	29377 (15.66)	128931 (68.71)
Pneumothorax	17313 (9.23)	2663 (1.42)	167665 (89.35)
Pleural Effusion	75696 (40.34)	9419 (5.02)	102526 (54.64)
Pleural Other	2441 (1.3)	1771 (0.94)	183429 (97.76)
Fracture	7270 (3.87)	484 (0.26)	179887 (95.87)
Support Devices	105831 (56.4)	898 (0.48)	80912 (43.12)

Table 2.1: CheXpert label distribution.

We compare the performance of our final model to 3 additional board certified radiologists on a test set of 500 studies on which the consensus of 5 separate board-certified radiologists serves as ground truth. We find that on 4 out of 5 pathologies, the model ROC and PR curves lie above at least 2 of 3 radiologist operating points. We make our dataset publicly available to encourage further development of models.

2.2 Dataset

CheXpert is a large public dataset for chest radiograph interpretation, consisting of 224,316 chest radiographs of 65,240 patients labeled for the presence of 14 observations as positive, negative, or uncertain. We report the prevalences of the labels for the different obsevations in Table 2.1.

2.2.1 Data Collection and Label Selection

We retrospectively collected chest radiographic studies from Stanford Hospital, performed between October 2002 and July 2017 in both inpatient and outpatient centers, along with their associated radiology reports. From these, we sampled a set of 1000 reports for manual review by a board-certified radiologist to determine feasibility for extraction of observations. We decided on 14 observations based on the prevalence in the reports and clinical relevance, conforming to the Fleischner Society's recommended glossary [87] whenever applicable. "Pneumonia", despite being a clinical diagnosis, was included as a label in order to represent the images that suggested primary infection as the diagnosis. The "No Finding" observation was intended to capture the absence of all pathologies.

1. *unremarkable* cardiomediastinal silhouette

2. diffuse reticular pattern, which can be seen with an atypical infection **or** chronic fibrotic change. *no* focal consolidation.

3. *no* pleural effusion or pneumothorax

4. mild degenerative changes in the lumbar spine and old right rib fractures.

Observation	Labeler Output
No Finding	
Enlarged Cardiom.	0
Cardiomegaly	
Lung Opacity	1
Lung Lesion	
Edema	
Consolidation	0
Pneumonia	u
Atelectasis	
Pneumothorax	0
Pleural Effusion	0
Pleural Other	
Fracture	1
Support Devices	

Figure 2.2: CheXpert labeler output example.

2.2.2 Label Extraction from Radiology Reports

We developed an automated rule-based labeler to extract observations from the free text radiology reports to be used as structured labels for the images. Our labeler is set up in three distinct stages: mention extraction, mention classification, and mention aggregation.

Mention Extraction

The labeler extracts mentions from a list of observations from the *Impression* section of radiology reports, which summarizes the key findings in the radiographic study. A large list of phrases was manually curated by multiple board-certified radiologists to match various ways observations are mentioned in the reports.

Mention Classification

After extracting mentions of observations, we aim to classify them as negative ("no evidence of pulmonary edema, pleural effusions or pneumothorax"), uncertain ("diffuse reticular pattern may represent mild interstitial pulmonary edema"), or positive ("moderate bilateral effusions and bibasilar opacities"). The 'uncertain' label can capture both the uncertainty of a radiologist in the diagnosis as well as ambiguity inherent in the report ("heart size is stable"). The mention classification stage is a 3-phase pipeline consisting of pre-negation uncertainty, negation, and post-negation uncertainty. Each phase consists of rules which are matched against the mention; if a match is found, then the mention is classified accordingly (as uncertain in the first or third phase, and as negative in the second phase). If a mention is not matched in any of the phases, it is classified as positive.

	Mention F1		Negation F1		Uncertain F1	
Category	NIH	Ours	NIH	Ours	NIH	Ours
Atelectasis	0.976	**0.998**	0.526	**0.833**	0.661	**0.936**
Cardiomegaly	0.647	**0.973**	0.000	**0.909**	0.211	**0.727**
Consolidation	0.996	**0.999**	0.879	**0.981**	0.438	**0.924**
Edema	0.978	**0.993**	0.873	**0.962**	0.535	**0.796**
Pleural Effusion	0.985	**0.996**	0.951	**0.971**	0.553	**0.707**
Pneumonia	0.660	**0.992**	0.703	**0.750**	0.250	**0.817**
Pneumothorax	0.993	**1.000**	0.971	**0.977**	0.167	**0.762**
Enlarged Cardiom.	N/A	0.935	N/A	0.959	N/A	0.854
Lung Lesion	N/A	0.896	N/A	0.900	N/A	0.857
Lung Opacity	N/A	0.966	N/A	0.914	N/A	0.286
Pleural Other	N/A	0.850	N/A	1.000	N/A	0.769
Fracture	N/A	0.975	N/A	0.807	N/A	0.800
Support Devices	N/A	0.933	N/A	0.720	N/A	N/A
No Finding	N/A	0.769	N/A	N/A	N/A	N/A
Macro-average	N/A	0.948	N/A	0.899	N/A	0.770
Micro-average	N/A	0.969	N/A	0.952	N/A	0.848

Table 2.2: CheXpert labeler performance.

Rules for mention classification are designed on the universal dependency parse of the report. To obtain the universal dependency parse, we follow a procedure similar to [162]: first, the report is split and tokenized into sentences using NLTK [23]; then, each sentence is parsed using the Bllip parser trained using David McClosky's biomedical model [35, 142]; finally, the universal dependency graph of each sentence is computed using Stanford CoreNLP [52].

Mention Aggregation

We use the classification for each mention of observations to arrive at a final label for 14 observations that consist of 12 pathologies as well as the "Support Devices" and "No Finding" observations. Observations with at least one mention that is positively classified in the report is assigned a positive (1) label. An observation is assigned an uncertain (u) label if it has no positively classified mentions and at least one uncertain mention, and a negative label if there is at least one negatively classified mention. We assign (*blank*) if there is no mention of an observation. The "No Finding" observation is assigned a positive label (1) if there is no pathology classified as positive or uncertain. An example of the labeling system run on a report is shown in Figure 2.2.

2.3 Labeler Results

We evaluate the performance of the labeler and compare it to the performance of another automated radiology report labeler on a report evaluation set.

2.3.1 Report Evaluation Set

The report evaluation set consists of 1000 radiology reports from 1000 distinct randomly sampled patients that do not overlap with the patients whose studies were used to develop the labeler. Two board-certified radiologists without access to additional patient information annotated the reports to label whether each observation was mentioned as confidently present (1), confidently absent (0), uncertainly present (u), or not mentioned (blank), after curating a list of labeling conventions to adhere to. After both radiologists independently labeled each of the 1000 reports, disagreements were resolved by consensus discussion. The resulting annotations serve as ground truth on the report evaluation set.

2.3.2 Comparison to NIH labeler

On the radiology report evaluation set, we compare our labeler against the method employed in [162] which was used to annotate another large dataset of chest radiographs using radiology reports [239]. We evaluate labeler performance on three tasks: mention extraction, negation detection, and uncertainty detection. For the mention extraction task, we consider any assigned label (1, 0, or u) as positive and *blank* as negative. On the negation detection task, we consider 0 labels as positive and all other labels as negative. On the uncertainty detection task, we consider u labels as positive and all other labels as negative. We report the F1 scores of the labeling algorithms for each of these tasks.

Table 2.2 shows the performance of the labeling methods. Across all observations and on all tasks, our labeling algorithm achieves a higher F1 score. On negation detection, our labeling algorithm significantly outperforms the NIH labeler on Atelectasis and Cardiomegaly, and achieves notably better performance on Consolidation and Pneumonia. On uncertainty detection, our labeler shows large gains over the NIH labeler, particularly on Cardiomegaly, Pneumonia, and Pneumothorax.

We note three key differences between our method and the method of [239]. First, we do not the use automatic mention extractors like MetaMap or DNorm, which we found produced weak extractions when applied to our collection of reports. Second, we incorporate several additional rules in order to capture the large variation in the ways negation and uncertainty are conveyed. Third, we split uncertainty classification of mentions into pre-negation and post-negation, which allowed us to resolve cases of uncertainty rules double matching with negation rules in the reports. For example, the following phrase “cannot exclude pneumothorax.” conveys uncertainty in the presence of pneumothorax. Without the pre-negation stage, the ‘pneumothorax’ match is classified

	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion
U-Ignore	0.818 (0.759,0.877)	*0.828 (0.769,0.888)*	0.938 (0.905,0.970)	0.934 (0.893,0.975)	0.928 (0.894,0.962)
U-Zeros	*0.811 (0.751,0.872)*	0.840 (0.783,0.897)	0.932 (0.898,0.966)	0.929 (0.888,0.970)	0.931 (0.897,0.965)
U-Ones	**0.858 (0.806,0.910)**	0.832 (0.773,0.890)	0.899 (0.854,0.944)	0.941 (0.903,0.980)	0.934 (0.901,0.967)
U-SelfTrained	0.833 (0.776,0.890)	0.831 (0.770,0.891)	0.939 (0.908,0.971)	0.935 (0.896,0.974)	0.932 (0.899,0.966)
U-MultiClass	0.821 (0.763,0.879)	**0.854 (0.800,0.909)**	0.937 (0.905,0.969)	0.928 (0.887,0.968)	0.936 (0.904,0.967)

Table 2.3: CheXpert validation set scores using different approaches to using uncertainty labels.

as negative due to the 'exclude XXX' rule. However, by applying the 'cannot exclude' rule in the pre-negation stage, this observation can be correctly classified as uncertain.

2.4 Model

We train models that take as input a single-view chest radiograph and output the probability of each of the 14 observations. When more than one view is available, the models output the maximum probability of the observations across the views.

2.4.1 Uncertainty Approaches

The training labels in the dataset for each observation are either 0 (negative), 1 (positive), or u (uncertain). We explore different approaches to using the uncertainty labels during the model training.

Ignoring

A simple approach to handling uncertainty is to ignore the u labels during training, which serves as a baseline to compare approaches which explicitly incorporate the uncertainty labels. In this approach (called *U-Ignore*), we optimize the sum of the *masked* binary cross-entropy losses over the observations, masking the loss for the observations which are marked as uncertain for the study. Formally, the loss for an example X is given by

$$+ (1 - y_o) \log p(Y_o = 0|X)],$$

where X is the input image, y is the vector of labels of length 14 for the study, and the sum is taken over all 14 observations. Ignoring the uncertainty label is analogous to the listwise (complete case) deletion method for imputation [80], which is when all cases with a missing value are deleted. Such methods can produce biased models if the cases are not missing completely at random. In this dataset, uncertainty labels are quite prevalent for some observations: for Consolidation, the

uncertainty label is almost twice as as prevalent (12.78%) as the positive label (6.78%), and thus this approach ignores a large proportion of labels, reducing the effective size of the dataset.

Binary Mapping

We investigate whether the uncertain labels for any of the observations can be replaced by the 0 label or the 1 label. In this approach, we map all instances of u to 0 (*U-Zeroes* model), or all to 1 (*U-Ones* model).

These approaches are similar to zero imputation strategies in statistics, and mimic approaches in multi-label classification methods where missing examples are used as negative labels [121]. If the uncertainty label does convey semantically useful information to the classifier, then we expect that this approach can distort the decision making of classifiers and degrade their performance.

Self-Training

One framework for approaching uncertainty labels is to consider them as unlabeled examples, lending its way to semi-supervised learning [268]. Most closely tied to our setting is *multi-label learning with missing labels* (MLML) [245], which aims to handle multi-label classification given training instances that have a partial annotation of their labels.

We investigate a self-training approach (*U-SelfTrained*) for using the uncertainty label. In this approach, we first train a model using the *U-Ignore* approach (that ignores the u labels during training) to convergence, and then use the model to make predictions that re-label each of the uncertainty labels with the probability prediction outputted by the model. We do not replace any instances of 1 or 0s. On these relabeled examples, we set up loss as the mean of the binary cross-entropy losses over the observations.

Our work follows the approach of [254], who train a classifier on labeled examples and then predict on unlabeled examples labeling them when the prediction is above a certain threshold, and repeating until convergence. [179] build upon the self-training technique and remove the need for iteratively training models, predicting on transformed versions of the inputs instead of training multiple models, and output a target label for each unlabeled example; soft labels, which are continuous probability outputs rather than binary, have also been used [99, 134].

3-Class Classification

We finally investigate treating the u label as its own class, rather than mapping it to a binary label, for each of the 14 observations. We hypothesize that with this approach, we can better incorporate information from the image by supervising uncertainty, allowing the network to find its own representation of uncertainty on different pathologies. In this approach (*U-MultiClass* model), for each observation, we output the probability of each of the 3 possible classes $\{p_0, p_1, p_u\} \in [0, 1]$, $p_0 + p_1 + p_u = 1$. We set up the loss as the mean of the multi-class cross-entropy losses over the

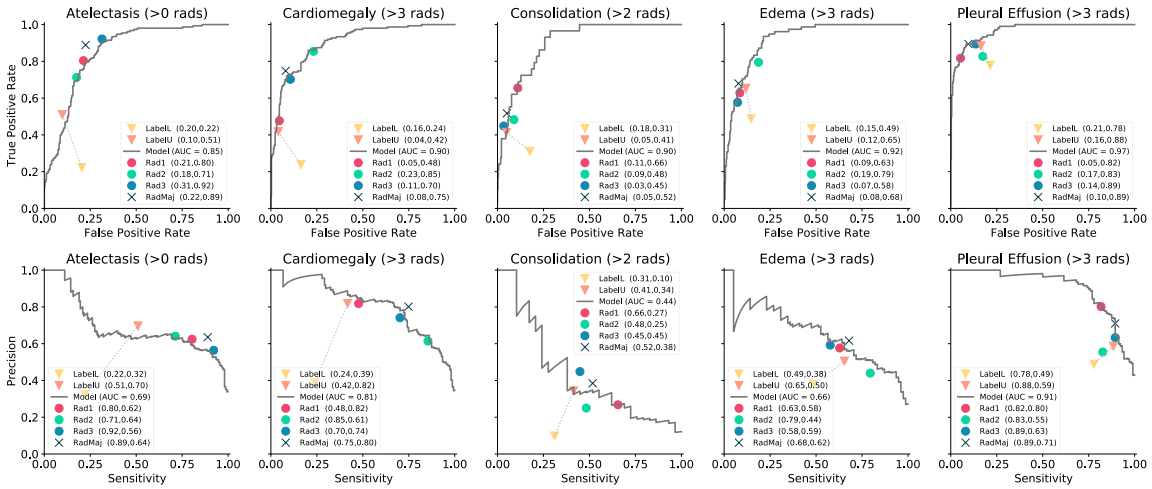


Figure 2.3: Comparison of performance to radiologists.

observations. At test time, for the probability of a particular observation, we output the probability of the positive label after applying a softmax restricted to the positive and negative classes.

2.4.2 Training Procedure

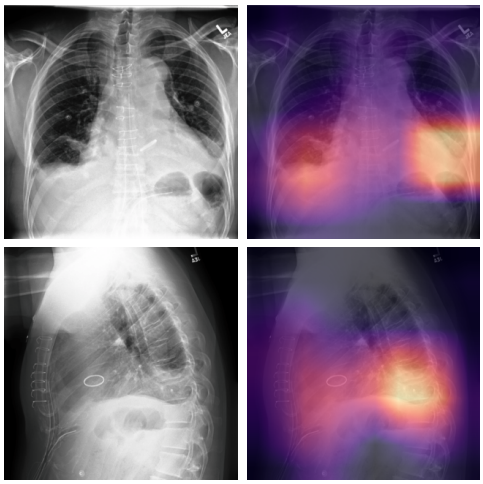
We follow the same architecture and training process for each of the uncertainty approaches. We experimented with several convolutional neural network architectures, specifically ResNet152, DenseNet121, Inception-v4, and SE-ResNeXt101, and found that the DenseNet121 architecture produced the best results. Thus we used DenseNet121 for all our experiments. The weights of the network are initialized with weights from a model pretrained on ImageNet. Images are fed into the network with size 320×320 pixels and normalized using the mean and standard deviation of images in the ImageNet training set. We use the Adam optimizer with default β-parameters of $\beta_1 = 0.9$, $\beta_2 = 0.999$ and learning rate 1×10^{-4} which is fixed for the duration of the training. Batches are sampled using a fixed batch size of 16 images. We train for 3 epochs, saving checkpoints every 4800 examples.

2.5 Validation Results

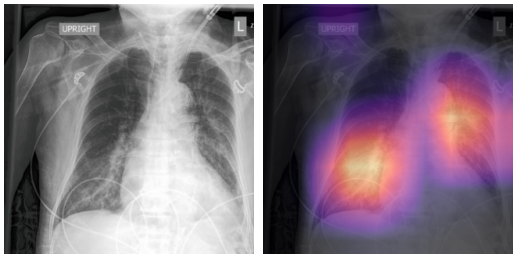
We compare the performance of the different uncertainty approaches on a validation set on which the consensus of radiologist annotations serves as ground truth.

2.5.1 Validation Set

The validation set contains 200 studies from 200 patients randomly sampled from the full dataset with no patient overlap with the report evaluation set. Three board-certified radiologists individually



(a) Frontal and lateral radiographs of the chest in a patient with bilateral pleural effusions; the model localizes the effusions on both the frontal (top) and lateral (bottom) views, with predicted probabilities $p = 0.936$ and $p = 0.939$ on the frontal and lateral views respectively.



(b) Single frontal radiograph of the chest demonstrates bilateral mid and lower lung interstitial predominant opacities and cardiomegaly most consistent with cardiogenic pulmonary edema. The model accurately classifies the edema by assigning a probability of $p = 0.824$ and correctly localizes the pulmonary edema. Two independent radiologist readers misclassified this examination as negative or uncertain unlikely for edema.

Figure 2.4: Gradient-weighted Class Activation Mappings with radiologist interpretation.

annotated each of the studies in the validation set, classifying each observation into one of present, uncertain likely, uncertain unlikely, and absent. Their annotations were binarized such that all present and uncertain likely cases are treated as positive and all absent and uncertain unlikely cases are treated as negative. The majority vote of these binarized annotations is used to define a strong ground truth [84].

2.5.2 Comparison of Uncertainty Approaches

Procedure

We evaluate the approaches using the area under the receiver operating characteristic curve (AUC) metric. We focus on the evaluation of 5 observations which we call the competition tasks, selected based of clinical importance and prevalence in the validation set: (a) Atelectasis, (b) Cardiomegaly, (c) Consolidation, (d) Edema, and (e) Pleural Effusion. We report the 95% two-sided confidence intervals of the AUC using the non-parametric method by DeLong [53, 221]. For each pathology, we also test whether the AUC of the best-performing approach is significantly greater than the AUC of the worst-performing approach using the one-sided DeLong's test for two correlated ROC curves [53]. We control for multiple hypothesis testing using the Benjamini-Hochberg procedure [21]; an adjusted p-value < 0.05 indicates statistical significance.

Model Selection

For each of the uncertainty approaches, we choose the best 10 checkpoints per run using the average AUC across the competition tasks. We run each model three times, and take the ensemble of the 30 generated checkpoints on the validation set by computing the mean of the output probabilities over the 30 models.

Results

The validation AUCs achieved by the different approaches to using the uncertainty labels are shown in Table 2.3. There are a few significant differences between the performance of the uncertainty approaches. On Atelectasis, the *U-Ones model* (AUC=0.858) significantly outperforms ($p = 0.03$) the *U-Zeros model* (AUC=0.811). On Cardiomegaly, we observe that the *U-MultiClass* model (AUC=0.854) performs significantly better ($p < 0.01$) than the *U-Ignore* model (AUC=0.828). On Consolidation, Edema and Pleural Effusion, we do not find the best models to be significantly better than the worst.

Analysis

We find that ignoring the uncertainty label is not an effective approach to handling uncertainty in the dataset, and is particularly ineffective on Cardiomegaly. Most of the uncertain Cardiomegaly cases are borderline cases such as "minimal cardiac enlargement", which if ignored, would likely cause the model to perform poorly on cases which are difficult to distinguish. However, explicitly supervising the model to distinguish between borderline and non-borderline cases (as in the *U-MultiClass* approach) could enable the model to better disambiguate the borderline cases. Moreover, assignment of the Cardiomegaly label when the heart is mentioned in the impression are difficult to categorize in many cases, particularly for common mentions such as "unchanged appearance of the heart" or "stable cardiac contours" either of which could be used in both enlarged and non-enlarged cases. These cases were classified as uncertain by the labeler, and therefore the binary assignment of 0s and 1s in this setting fails to achieve optimal performance as there is insufficient information conveyed by these modifications.

In the detection of Atelectasis, the *U-Ones* approach performs the best, hinting that the uncertainty label for this observation is effectively utilized when treated as positive. We expect that phrases such as "possible atelectasis" or "may be atelectasis," were meant to describe the most likely findings in the image, rather than convey uncertainty, which supports the good performance of *U-Ones* on this pathology. We suspect a similar explanation for the high performance of *U-Ones* on Edema, where uncertain phrases like "possible mild pulmonary edema" in fact convey likely findings. In contrast, the *U-Ones* approach performs worst on the Consolidation label, whereas the *U-Zeros* approach performs the best. We also note that Atelectasis and Consolidation are often mentioned

together in radiology reports. For example, the phrase "findings may represent atelectasis versus consolidation" is very common. In these cases, our labeler assigns uncertain for both observations, but we find that in the ground truth panel review that many of these sorts of uncertainty cases are often instead resolved as Atelectasis-positive and Consolidation-negative.

2.6 Test Results

We compare the performance of our final model to radiologists on a test set. We selected the final model based on the best performing ensemble on each competition task on the validation set: *U-Ones* for Atelectasis and Edema, *U-MultiClass* for Cardiomegaly and Pleural Effusion, and *U-SelfTrained* for Consolidation.

2.6.1 Test Set

The test set consists of 500 studies from 500 patients randomly sampled from the 1000 studies in the report test set. Eight board-certified radiologists individually annotated each of the studies in the test set following the same procedure and post-processing as described for the validation set. The majority vote of 5 radiologist annotations serves as a strong ground truth: 3 of these radiologists were the same as those who annotated the validation set and the other 2 were randomly sampled. The remaining 3 radiologist annotations were used to benchmark radiologist performance.

2.6.2 Comparison to Radiologists

Procedure

For each of the 3 individual radiologists and for their majority vote, we compute sensitivity (recall), specificity, and precision against the test set ground truth. To compare the model to radiologists, we plot the radiologist operating points with the model on both the ROC and Precision-Recall (PR) space. We examine whether the radiologist operating points lie below the curves to determine if the model is superior to the radiologists. We also compute the performance of the labels extracted automatically from the radiology report using our labeling system against the test set ground truth. We convert the uncertainty labels to binary labels by computing the upper bound of the labels performance (by assigning the uncertain labels to the ground truth values) and the lower bound of the labels (by assigning the uncertain labels to the opposite of the ground truth values), and plot the two operating points on the curves, denoted *LabelU* and *LabelL* respectively. We also measure calibration of the model before and after applying post-processing calibration techniques, namely isotonic regression [259] and Platt scaling [170], using the scaled Brier score [220].

Results

Figure 2.3 illustrates these plots on all competition tasks. The model achieves the best AUC on Pleural Effusion (0.97), and the worst on Atelectasis (0.85). The AUC of all other observations are at least 0.9. The model achieves the best AUPRC on Pleural Effusion (0.91) and the worst on Consolidation (0.44). On Cardiomegaly, Edema, and Pleural Effusion, the model achieves higher performance than all 3 radiologists but not their majority vote. On Consolidation, model performance exceeds 2 of the 3 radiologists, and on Atelectasis, all 3 radiologists perform better than the model. On all competition tasks, the lower bound of the report labels lies below the model curves. On all tasks besides Atelectasis, the upper bound of the report label lies on or below the model operating curves. On most of the tasks, the upper bound of the labeler performs comparably to the radiologists. The average scaled Brier score of the model before post-processing calibration is 0.110, after isotonic regression is 0.107, and after platt scaling is 0.101.

Limitations

We acknowledge two limitations to performing this comparison. First, neither the radiologists nor the model had access to patient history or previous examinations, which has been shown to decrease diagnostic performance in chest radiograph interpretation [173, 22]. Second, no statistical test was performed to assess whether the difference between the performance of the model and the radiologists is statistically significant.

2.6.3 Visualization

We visualize the areas of the radiograph which the model predicts to be most indicative of each observation using Gradient-weighted Class Activation Mappings (Grad-CAMs) [207]. Grad-CAMs use the gradient of an output class into the final convolutional layer to produce a low resolution map which highlights portions of the image which are important in the detection of the output class. Specifically, we construct the map by using the gradient of the final linear layer as the weights and performing a weighted sum of the final feature maps using those weights. We upscale the resulting map to the dimensions of the original image and overlay the map on the image. Some examples of the Grad-CAMs are illustrated in Figure 2.4.

2.7 Existing Chest Radiograph Datasets

One of the main obstacles in the development of chest radiograph interpretation models has been the lack of datasets with strong radiologist-annotated groundtruth and expert scores against which researchers can compare their models. There are few chest radiographic imaging datasets that are publicly available, but none of them have test sets with strong ground truth or radiologist

performances. The Indiana Network for Patient Care hosts the OpenI dataset [54] consisting of 7,470 frontal-view radiographs and radiology reports which have been labeled with key findings by human annotators . The National Cancer Institute hosts the PLCO Lung dataset [76] of chest radiographs obtained during a study on lung cancer screening . The dataset contains 185,421 full resolution images, but due to the nature of the collection process, it is has a low prevalence of clinically important pathologies such as Pneumothorax, Consolidation, Effusion, and Cardiomegaly. The MIMIC-CXR dataset [196] has been recently announced but is not yet publicly available.

The most commonly used benchmark for developing chest radiograph interpretation models has been the ChestX-ray14 dataset [239]. Due to the introduction of this large dataset, substantial progress has been made towards developing automated chest radiograph interpretation models [252, 183, 135, 123, 240, 82, 253]. However, using the NIH dataset as a benchmark on which to compare models is problematic as the labels in the test set are extracted from reports using an automatic labeler. A few recent studies have trained models on very large datasets of chest radiographs and evaluated them against radiologists [127, 176], but the datasets have not been made publicly available. The CheXpert dataset that we introduce features radiologist-labeled validation and test sets which serve as strong reference standards, as well as expert scores to allow for robust evaluation of different algorithms.

2.8 Conclusion

We present a large dataset of chest radiographs called CheXpert, which features uncertainty labels and radiologist-labeled reference standard evaluation sets. We investigate a few different approaches to handling uncertainty and validate them on the evaluation sets. On a test set with a strong ground truth, we find that our best model outperforms at least 2 of the 3 radiologists in the detection of 4 clinically relevant pathologies. We hope that the dataset will help development and validation of chest radiograph interpretation models towards improving healthcare access and delivery worldwide.

2.9 Appendix

2.9.1 Data Collection

The imaging studies were pulled in DICOM format from the Stanford Hospital Picture Archiving and Communication System (PACS) and the reports were obtained from the Stanford Hospital Epic System. The studies were selected to contain one or more keywords of pathologies from a prespecified list, which likely led to a higher disease prevalence population and high proportion of AP radiographs. Each of the selected studies were filtered for (1) cases which contain at least one frontal-view chest radiograph, (2) non-pediatric cases, and (3) cases without protected health

Pathology	Validation (%)	Test (%)
No Finding	26 (13.0)	62 (12.4)
Enlarged Cardiom.	105 (52.5)	253 (50.6)
Cardiomegaly	66 (33.0)	151 (30.2)
Lung Lesion	116 (58.0)	264 (52.8)
Lung Opacity	1 (0.5)	8 (1.6)
Edema	42 (21.0)	78 (15.6)
Consolidation	32 (16.0)	29 (5.8)
Pneumonia	8 (4.0)	11 (2.2)
Atelectasis	75 (37.5)	153 (30.6)
Pneumothorax	6 (3.0)	9 (1.8)
Pleural Effusion	64 (32.0)	104 (20.8)
Pleural Other	1 (0.5)	4 (0.8)
Fracture	0 (0.0)	5 (1.0)
Support Devices	99 (49.5)	261 (52.2)

Table 2.4: We report the number of studies which contain each of the 14 observations in the validation set (200 studies) and test set (500 studies) according to radiologist consensus. The studies in both sets are labeled as positive or negative for each observation.

information in the images. Finally we tried to remove any cases which referenced previous studies in the report.

The resulting dataset, which we call CheXpert, contains 224,316 chest radiographs from 188,341 studies of 65,240 patients. Each image is annotated with the patient sex and age as well as if the image is frontal-view or lateral-view. Additionally, if the image is frontal-view, it is labeled as anteroposterior (AP) view or posteroanterior (PA) view. Patients who were older than 89 were capped at age 90 to ensure patient deidentification. The patients are 55.4% male and on average 60.3 years of age after capping age. The dataset consists of 85.5% frontal-view radiographs and 14.5% lateral-view radiographs. Of the frontal-view radiographs, 84.6% are AP and 15.4% are PA. The average number of images per study is 1.19 and the average number of studies per patient is 3.45. The images were corrected for contrast inversion as well as contrast level using histogram equalization. We split the dataset into training (64,540 patients, 187,641 studies, 223,414 images), valid (200 patients, 200 studies, 234 images), and test (500 patients, 500 studies, 668 images). Table 2.4 describes the prevalence of each category in the validation and test sets.

2.9.2 Label Description

For the "Lung Lesion" observation, we condensed the accepted terms "Nodule" and "Mass" together so as to avoid misclassification based purely on the size (3 cm) criteria that defines the boundary between these two terms. In the "Support Devices" label we considered devices including central lines, pacemakers, and ET tubes. "Enlarged Cardiomediastinum" was intended to include any

enlargement of the mediastinum, heart, and/or hila, while "Cardiomegaly" referred to enlargement of the heart only.

Each category is labeled as 1 (mentioned and classified as positive), 0 (mentioned and classified as negative), -1 (mentioned and classified as uncertain), or blank (not mentioned).

2.9.3 Radiologist Setup

The radiologists viewed the images at full resolution within the original DICOM files using used a freely available image viewer with capabilities for PACS features such as zoom, window leveling, and contrast adjustment. The radiologists did not have access to disease prevalence or patient history.

Chapter 3

Expert-level Deep Learning for Ambulatory ECGs

We develop an algorithm which exceeds the performance of board certified cardiologists in detecting a wide range of heart arrhythmias from electrocardiograms recorded with a single-lead wearable monitor. To enable end-to-end learning, we build a dataset with more than 500 times the number of unique patients than previously studied corpora. On this dataset, we train a 34-layer convolutional neural network which maps a sequence of ECG samples to a sequence of rhythm classes. Committees of board-certified cardiologists annotate a gold standard test set on which we compare the performance of our model to that of 6 other individual cardiologists. We exceed the average cardiologist performance in both recall (sensitivity) and precision (positive predictive value).

This chapter is based on [181, 86].

3.1 Introduction

We develop a model which can diagnose irregular heart rhythms, also known as arrhythmias, from single-lead ECG signals better than a cardiologist. Key to exceeding expert performance is a deep convolutional network which can map a sequence of ECG samples to a sequence of arrhythmia annotations along with a novel dataset two orders of magnitude larger than previous datasets of its kind.

Many heart diseases, including Myocardial Infarction, AV Block, Ventricular Tachycardia and Atrial Fibrillation can all be diagnosed from ECG signals with an estimated 300 million ECGs recorded annually [96]. We investigate the task of arrhythmia detection from the ECG record. This is known to be a challenging task for computers but can usually be determined by an expert from a single, well-placed lead.

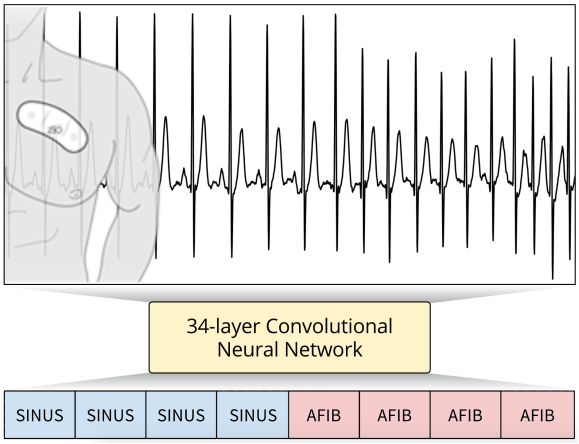


Figure 3.1: Modeling overview.

Arrhythmia detection from ECG recordings is usually performed by expert technicians and cardiologists given the high error rates of computerized interpretation. One study found that of all the computer predictions for non-sinus rhythms, only about 50% were correct [209]; in another study, only 1 out of every 7 presentations of second degree AV block were correctly recognized by the algorithm [83]. To automatically detect heart arrhythmias in an ECG, an algorithm must implicitly recognize the distinct wave types and discern the complex relationships between them over time. This is difficult due to the variability in wave morphology between patients as well as the presence of noise.

We train a 34-layer convolutional neural network (CNN) to detect arrhythmias in arbitrary length ECG time-series. Figure 3.1 shows an example of an input to the model. In addition to classifying noise and the sinus rhythm, the network learns to classify and segment twelve arrhythmia types present in the time-series. The model is trained end-to-end on a single-lead ECG signal sampled at 200Hz and a sequence of annotations for every second of the ECG as supervision. To make the optimization of such a deep model tractable, we use residual connections and batch-normalization [94, 105]. The depth increases both the non-linearity of the computation as well as the size of the context window for each classification decision.

We construct a dataset 500 times larger than other datasets of its kind [149, 77]. One of the most popular previous datasets, the MIT-BIH corpus contains ECG recordings from 47 unique patients. In contrast, we collect and annotate a dataset of about 30,000 unique patients from a pool of nearly 300,000 patients who have used the Zio Patch monitor[1] [229]. We intentionally select patients exhibiting abnormal rhythms in order to make the class balance of the dataset more even and thus the likelihood of observing unusual heart-activity high.

We test our model against board-certified cardiologists. A committee of three cardiologists serve

[1] iRhythm Technologies, San Francisco, California

as gold-standard annotators for the 336 examples in the test set. Our model exceeds the individual expert performance on both recall (sensitivity), and precision (positive predictive value) on this test set.

3.2 Model

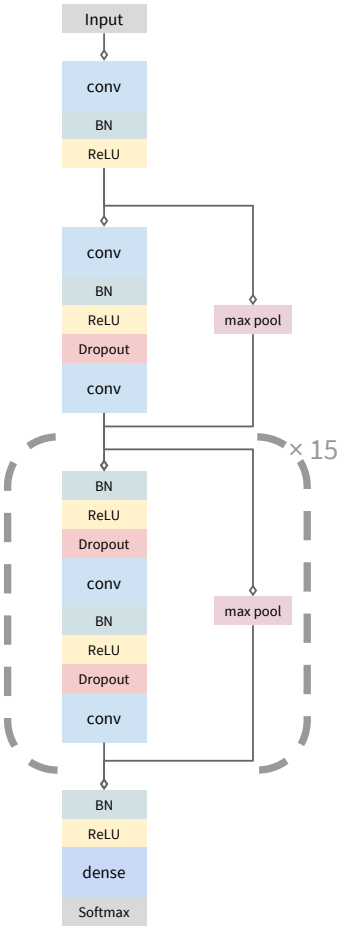


Figure 3.2: Network Architecture.

Problem Formulation

The ECG arrhythmia detection task is a sequence-to-sequence task which takes as input an ECG signal $X = [x_1, ..x_k]$, and outputs a sequence of labels $r = [r_1, ...r_n]$, such that each r_i can take on one of m different rhythm classes. Each output label corresponds to a segment of the input. Together the output labels cover the full sequence.

For a single example in the training set, we optimize the cross-entropy objective function

$$\mathcal{L}(X, r) = \frac{1}{n} \sum_{i=1}^{n} \log p(R = r_i \mid X)$$

where $p(\cdot)$ is the probability the network assigns to the i-th output taking on the value r_i.

Model Architecture and Training

We use a convolutional neural network for the sequence-to-sequence learning task. The high-level architecture of the network is shown in Figure 3.2. The network takes as input a time-series of raw ECG signal, and outputs a sequence of label predictions. The 30 second long ECG signal is sampled at 200Hz, and the model outputs a new prediction once every second. We arrive at an architecture which is 33 layers of convolution followed by a fully connected layer and a softmax.

In order to make the optimization of such a network tractable, we employ shortcut connections in a similar manner to those found in the Residual Network architecture [95]. The shortcut connections between neural-network layers optimize training by allowing information to propagate well in very deep neural networks. Before the input is fed into the network, it is normalized using a robust normalization strategy. The network consists of 16 residual blocks with 2 convolutional layers per block. The convolutional layers all have a filter length of 16 and have $64k$ filters, where k starts out as 1 and is incremented every 4-th residual block. Every alternate residual block subsamples its inputs by a factor of 2, thus the original input is ultimately subsampled by a factor of 2^8. When a residual block subsamples the input, the corresponding shortcut connections also subsample their input using a Max Pooling operation with the same subsample factor.

Before each convolutional layer we apply Batch Normalization [105] and a rectified linear activation, adopting the pre-activation block design [94]. The first and last layers of the network are special-cased due to this pre-activation block structure. We also apply Dropout [219] between the convolutional layers and after the non-linearity. The final fully connected layer and softmax activation produce a distribution over the 14 output classes for each time-step.

We train the networks from scratch, initializing the weights of the convolutional layers as in [93]. We use the Adam [118] optimizer with the default parameters and reduce the learning rate by a factor of 10 when the validation loss stops improving. We save the best model as evaluated on the validation set during the optimization process.

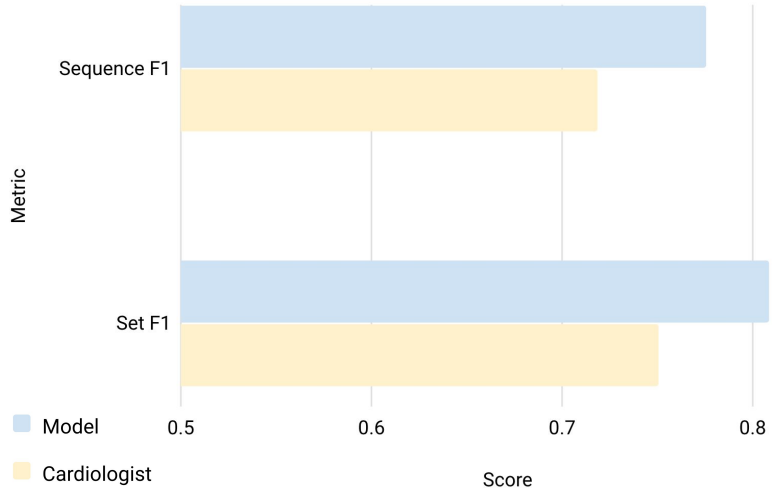


Figure 3.3: Comparison to experts on the test set.

3.3 Data

Training

We collect and annotate a dataset of 64,121 ECG records from 29,163 patients. The ECG data is sampled at a frequency of 200 Hz and is collected from a single-lead, noninvasive and continuous monitoring device called the Zio Patch which has a wear period up to 14 days [229]. Each ECG record in the training set is 30 seconds long and can contain more than one rhythm type. Each record is annotated by a clinical ECG expert: the expert highlights segments of the signal and marks it as corresponding to one of the 14 rhythm classes.

The 30 second records were annotated using a web-based ECG annotation tool designed for this work. Label annotations were done by a group of Certified Cardiographic Technicians who have completed extensive training in arrhythmia detection and a cardiographic certification examination by Cardiovascular Credentialing International. The technicians were guided through the interface before they could annotate records. All rhythms present in a strip were labeled from their corresponding onset to offset, resulting in full segmentation of the input ECG data. To improve labeling consistency among different annotators, specific rules were devised regarding each rhythm transition.

We split the dataset into a training and validation set. The training set contains 90% of the data. We split the dataset so that there is no patient overlap between the training and validation sets (as well as the test set described below).

Testing

We collect a test set of 336 records from 328 unique patients. For the test set, ground truth annotations for each record were obtained by a committee of three board-certified cardiologists; there are three committees responsible for different splits of the test set. The cardiologists discussed each individual record as a group and came to a consensus labeling. For each record in the test set we also collect 6 individual annotations from cardiologists not participating in the group. This is used to assess performance of the model compared to an individual cardiologist.

Rhythm Classes

We identify 12 heart arrhythmias, sinus rhythm and noise for a total of 14 output classes. The arrhythmias are characterized by a variety of features. Table 3.2 in the Appendix shows an example of each rhythm type we classify. The noise label is assigned when the device is disconnected from the skin or when the baseline noise in the ECG makes identification of the underlying rhythm impossible.

The morphology of the ECG during a single heart-beat as well as the pattern of the activity of the heart over time determine the underlying rhythm. In some cases the distinction between the rhythms can be subtle yet critical for treatment. For example two forms of second degree AV Block, Mobitz I (Wenckebach) and Mobitz II (here referred to as AVB_TYPE2) can be difficult to distinguish. Wenckebach is considered benign and Mobitz II is considered pathological, requiring immediate attention [62].

Table 3.2 in the Appendix also shows the number of unique patients in the training (including validation) set and test set for each rhythm type.

3.4 Results

Evaluation Metrics

We use two metrics to measure model accuracy, using the cardiologist committee annotations as the ground truth.

Sequence Level Accuracy (F1): We measure the average overlap between the prediction and the ground truth sequence labels. For every record, a model is required to make a prediction approximately once per second (every 256 samples). These predictions are compared against the ground truth annotation.

Set Level Accuracy (F1): Instead of treating the labels for a record as a sequence, we consider the set of unique arrhythmias present in each 30 second record as the ground truth annotation. Set Level Accuracy, unlike Sequence Level Accuracy, does not penalize for time-misalignment within a record. We report the F1 score between the unique class labels from the ground truth and those from the model prediction.

	Seq		Set	
	Model	Cardiol.	Model	Cardiol.
Class-level F1 Score				
AFIB	**0.604**	0.515	**0.667**	0.544
AFL	**0.687**	0.635	**0.679**	0.646
AVB_TYPE2	**0.689**	0.535	**0.656**	0.529
BIGEMINY	**0.897**	0.837	**0.870**	0.849
CHB	**0.843**	0.701	**0.852**	0.685
EAR	**0.519**	0.476	**0.571**	0.529
IVR	**0.761**	0.632	**0.774**	0.720
JUNCTIONAL	0.670	**0.684**	**0.783**	0.674
NOISE	**0.823**	0.768	**0.704**	0.689
SINUS	**0.879**	0.847	**0.939**	0.907
SVT	**0.477**	0.449	**0.658**	0.556
TRIGEMINY	**0.908**	0.843	**0.870**	0.816
VT	0.506	**0.566**	0.694	**0.769**
WENCKEBACH	**0.709**	0.593	**0.806**	0.736
Aggregate Results				
Precision (PPV)	**0.800**	0.723	**0.809**	0.763
Recall (Sensitivity)	**0.784**	0.724	**0.827**	0.744
F1	**0.776**	0.719	**0.809**	0.751

Table 3.1: Sequence and the Set F1 metrics for model and experts.

In both the Sequence and the Set case, we compute the F1 score for each class separately. We then compute the overall F1 (and precision and recall) as the class-frequency weighted mean.

Model vs. Cardiologist Performance

We assess the cardiologist performance on the test set. Recall that each of the records in the test set has a ground truth label from a committee of three cardiologists as well as individual labels from a disjoint set of 6 other cardiologists. To assess cardiologist performance for each class, we take the average of all the individual cardiologist F1 scores using the group label as the ground truth annotation.

Table 3.1 shows the breakdown of both cardiologist and model scores across the different rhythm classes. The model outperforms the average cardiologist performance on most rhythms, noticeably outperforming the cardiologists in the AV Block set of arrhythmias which includes Mobitz I (Wenckebach), Mobitz II (AVB_Type2) and complete heart block (CHB). This is especially useful given the severity of Mobitz II and complete heart block and the importance of distinguishing these two from Wenckebach which is usually considered benign.

Table 3.1 also compares the aggregate precision, recall and F1 for both model and cardiologist compared to the ground truth annotations. The aggregate scores for the cardiologist are computed by taking the mean of the individual cardiologist scores. The model outperforms the cardiologist average in both precision and recall.

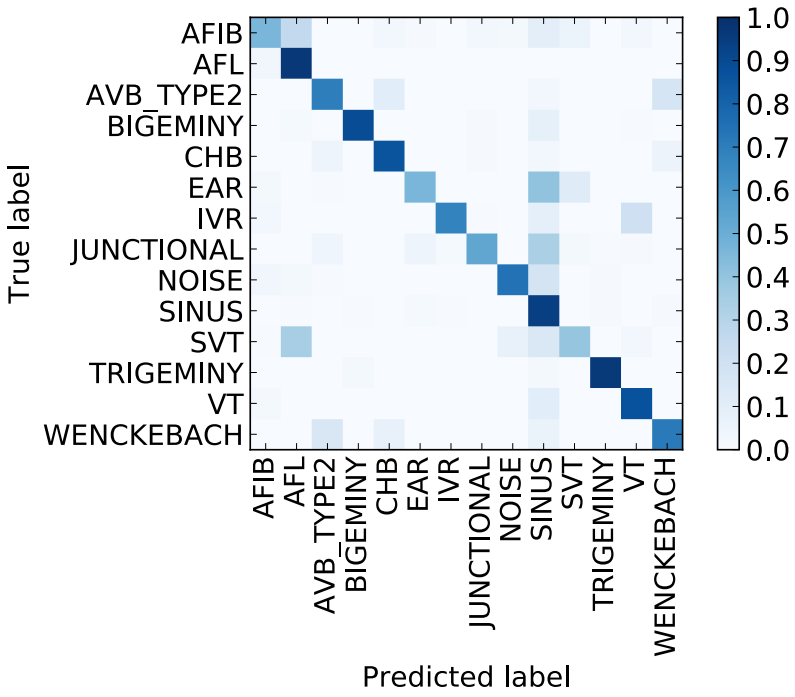


Figure 3.4: Confusion matrix for model predictions.

3.5 Analysis

The model outperforms the average cardiologist score on both the sequence and the set F1 metrics. Figure 3.4 shows a confusion matrix of the model predictions on the test set. Many arrhythmias are confused with the sinus rhythm. We expect that part of this is due to the sometimes ambiguous location of the exact onset and offset of the arrhythmia in the ECG record.

Often the mistakes made by the model are understandable. For example, confusing Wenckebach and AVB_Type2 makes sense given that the two rhythms in general have very similar ECG morphologies. Similarly, Supraventricular Tachycardia (SVT) and Atrial Fibrillation (AFIB) are often confused with Atrial Flutter (AFL) which is understandable given that they are all atrial arrhythmias. We also note that Idioventricular Rhythm (IVR) is sometimes mistaken as Ventricular Tachycardia (VT), which again makes sense given that the two only differ in heart-rate and are difficult to distinguish close to the 100 beats per minute delineation.

One of the most common confusions is between Ectopic Atrial Rhythm (EAR) and the sinus rhythm. The main distinguishing criteria for this rhythm is an irregular P wave. This can be subtle to detect especially when the P wave has a small amplitude or when noise is present in the signal.

3.6 Related Work

Automatic high-accuracy methods for R-peak extraction have existed at least since the mid 1980's [160]. Current algorithms for R-peak extraction tend to use wavelet transformations to compute features from the raw ECG followed by finely-tuned threshold based classifiers [132, 141]. Because accurate estimates of heart rate and heart rate variability can be extracted from R-peak features, feature-engineered algorithms are often used for coarse-grained heart rhythm classification, including detecting tachycardias (fast heart rate), bradycardias (slow heart rate), and irregular rhythms. However, such features alone are not sufficient to distinguish between most heart arrhythmias since features based on the atrial activity of the heart as well as other features pertaining to the QRS morphology are needed.

Much work has been done to automate the extraction of other features from the ECG. For example, beat classification is a common sub-problem of heart-arrhythmia classification. Drawing inspiration from automatic speech recognition, Hidden Markov models with Gaussian observation probability distributions have been applied to the task of beat detection [49]. Artificial neural networks have also been used for the task of beat detection [144]. While these models have achieved high-accuracy for some beat types, they are not yet sufficient for high-accuracy heart arrhythmia classification and segmentation. For example, [10] train a neural network to distinguish between Atrial Fibrillation and Sinus Rhythm on the MIT-BIH dataset. While the network can distinguish between these two classes with high-accuracy, it does not generalize to noisier single-lead recordings or classify among the full range of 15 rhythms available in MIT-BIH. This is in part due to insufficient training data, and because the model also discards critical information in the feature extraction stage.

The most common dataset used to design and evaluate ECG algorithms is the MIT-BIH arrhythmia database [149] which consists of 48 half-hour strips of ECG data. Other commonly used datasets include the MIT-BIH Atrial Fibrillation dataset [148] and the QT dataset [126]. While useful benchmarks for R-peak extraction and beat-level annotations, these datasets are too small for fine-grained arrhythmia classification. The number of unique patients is in the single digit hundreds or fewer for these benchmarks. A recently released dataset captured from the AliveCor ECG monitor contains about 7000 records [48]. These records only have annotations for Atrial Fibrillation; all other arrhythmias are grouped into a single bucket. The dataset we develop contains 29,163 unique patients and 14 classes with hundreds of unique examples for the rarest arrhythmias.

Machine learning models based on deep neural networks have consistently been able to approach and often exceed human agreement rates when large annotated datasets are available [6, 247, 93]. These approaches have also proven to be effective in healthcare applications, particularly in medical imaging where pretrained ImageNet models can be applied [70, 84]. We draw on work in automatic speech recognition for processing time-series with deep convolutional neural networks and recurrent neural networks [198], and techniques in deep learning to make the optimization of these models tractable [94, 95, 105].

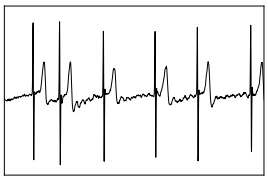
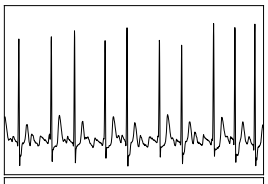
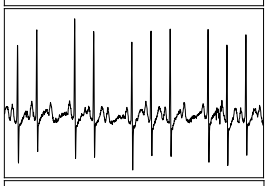
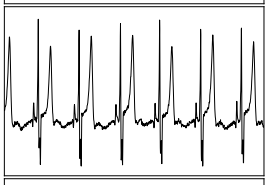
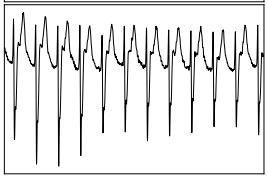
3.7 Conclusion

We develop a model which exceeds the cardiologist performance in detecting a wide range of heart arrhythmias from single-lead ECG records. Key to the performance of the model is a large annotated dataset and a very deep convolutional network which can map a sequence of ECG samples to a sequence of arrhythmia annotations.

On the clinical side, future work should investigate extending the set of arrhythmias and other forms of heart disease which can be automatically detected with high-accuracy from single or multiple lead ECG records. For example we do not detect Ventricular Flutter or Fibrillation. We also do not detect Left or Right Ventricular Hypertrophy, Myocardial Infarction or a number of other heart diseases which do not necessarily exhibit as arrhythmias. Some of these may be difficult or even impossible to detect on a single-lead ECG but can often be seen on a multiple-lead ECG.

Given that more than 300 million ECGs are recorded annually, high-accuracy diagnosis from ECG can save expert clinicians and cardiologists considerable time and decrease the number of misdiagnoses. Furthermore, we hope that this technology coupled with low-cost ECG devices enables more widespread use of the ECG as a diagnostic tool in places where access to a cardiologist is difficult.

3.8 Appendix

Class	Description	Example	Train + Val Patients	Test Patients
AFIB	Atrial Fibrillation		4638	44
AFL	Atrial Flutter		3805	20
AVB_TYPE2	Second degree AV Block Type 2 (Mobitz II)		1905	28
BIGEMINY	Ventricular Bigeminy		2855	22
CHB	Complete Heart Block		843	26
EAR	Ectopic Atrial Rhythm		2623	22
IVR	Idioventricular Rhythm		1962	34

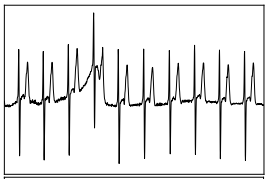
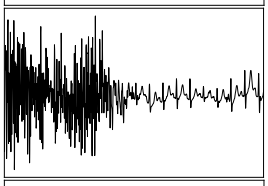
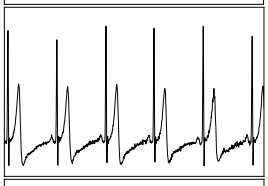
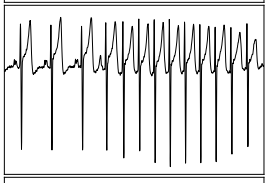
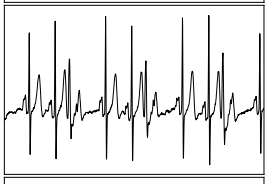
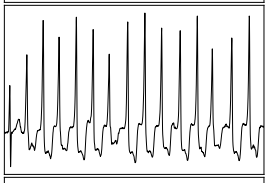
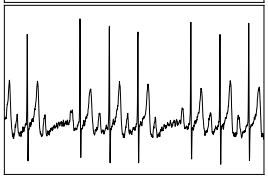
Class	Description	Example	Train + Val Patients	Test Patients
JUNCTIONAL	Junctional Rhythm		2030	36
NOISE	Noise		9940	41
SINUS	Sinus Rhythm		22156	215
SVT	Supraventricular Tachycardia		6301	34
TRIGEMINY	Ventricular Trigeminy		2864	21
VT	Ventricular Tachycardia		4827	17
WENCKEBACH	Wenckebach (Mobitz I)		2051	29

Table 3.2: A list of all of the rhythm types which the model classifies. For each rhythm we give the label name, a more descriptive name and an example chosen from the training set. We also give the total number of patients with each rhythm for both the training and test sets.

Chapter 4

Pretraining Using Transfer Learning

As we have seen in the previous chapters, medical images typically require manual annotation, which can be costly and hard to obtain. Deep learning methods for chest X-ray interpretation typically rely on pretrained models developed for ImageNet. This paradigm assumes that better ImageNet architectures perform better on chest X-ray tasks and that ImageNet-pretrained weights provide a performance boost over random initialization. In this work, we compare the transfer performance and parameter efficiency of 16 popular convolutional architectures on a large chest X-ray dataset (CheXpert) to investigate these assumptions. First, we find no relationship between ImageNet performance and CheXpert performance for both models without pretraining and models with pretraining. Second, we find that, for models without pretraining, the choice of model family influences performance more than size within a family for medical imaging tasks. Third, we observe that ImageNet pretraining yields a statistically significant boost in performance across architectures, with a higher boost for smaller architectures. Fourth, we examine whether ImageNet architectures are unnecessarily large for CheXpert by truncating final blocks from pretrained models, and find that we can make models 3.25x more parameter-efficient on average without a statistically significant drop in performance. Our work contributes new experimental evidence about the relation of ImageNet to chest x-ray interpretation performance.

This chapter is based on [115].

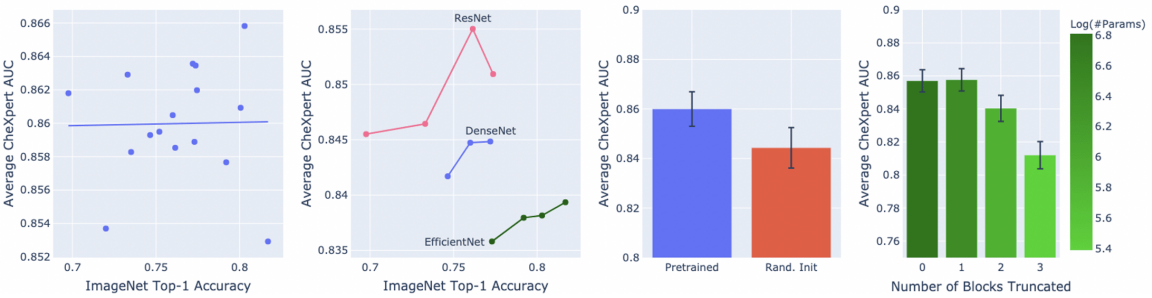


Figure 4.1: Visual summary of our contributions. From left to right: scatterplot and best-fit line for 16 pretrained models showing no relationship between ImageNet and CheXpert performance, CheXpert performance relationship varies across architecture families much more than within, average CheXpert performance improves with pretraining, models can maintain performance and improve parameter efficiency through truncation of final blocks. Error bars show one standard deviation.

4.1 Introduction

Deep learning models for chest X-ray interpretation have high potential for social impact by aiding clinicians in their workflow and increasing access to radiology expertise worldwide [187, 186]. Transfer learning using pretrained ImageNet [55] models has been the standard approach for developing models not only on chest X-rays [239, 183, 9] but also for many other medical imaging modalities [147, 262, 133, 51, 71]. This transfer assumes that better ImageNet architectures perform better and pretrained weights boost performance on their target medical tasks. However, there has not been a systematic investigation of how ImageNet architectures and weights both relate to performance on downstream medical tasks.

In this work, we systematically investigate how ImageNet architectures and weights both relate to performance on chest X-ray tasks. Our primary contributions are:

1. For models without pretraining and models with pretraining, we find *no relationship between ImageNet performance and CheXpert performance* (Spearman $\rho = 0.08$, $\rho = 0.06$ respectively). This finding suggests that architecture improvements on ImageNet may not lead to improvements on medical imaging tasks.

2. For models without pretraining, we find that within an architecture family, the largest and smallest models have small differences (ResNet 0.005, DenseNet 0.003, EfficientNet 0.004) in CheXpert AUC, but different model families have larger differences in AUC (> 0.006). This finding suggests that *the choice of model family influences performance more than size within a family* for medical imaging tasks.

3. We observe that *ImageNet pretraining yields a statistically significant boost in performance* (average boost of 0.016 AUC) across architectures, with a higher boost for smaller architectures

(Spearman $\rho = -0.72$ with number of parameters). This finding supports the ImageNet pretraining paradigm for medical imaging tasks, especially for smaller models.

4. We find that by truncating final blocks of pretrained models, we can make models *3.25x more parameter-efficient on average without a statistically significant drop in performance*. This finding suggests model truncation may be a simple method to yield lighter pretrained models by preserving architecture design features while reducing model size.

Our study, to the best of our knowledge, contributes the first systematic investigation of the performance and efficiency of ImageNet architectures and weights for chest X-ray interpretation. Our investigation and findings may be further validated on other datasets and medical imaging tasks.

4.2 Related Work

4.2.1 ImageNet Transfer

[122] examined the performance of 16 convolutional neural networks (CNNs) on 12 image classification datasets. They found that using these ImageNet pretrained architectures either as feature extractors for logistic regression or fine tuning them on the target dataset yielded a Spearman $\rho = 0.99$ and $\rho = 0.97$ between ImageNet accuracy and transfer accuracy respectively. However, they showed ImageNet performance was less correlated with transfer accuracy for some fine-grained tasks, corroborating [92]. They found that without ImageNet pretraining, ImageNet accuracy and transfer accuracy had a weaker Spearman $\rho = 0.59$. We extend [122] to the medical setting by studying the relationship between ImageNet and CheXpert performance.

[180] explored properties of transfer learning onto retinal fundus images and chest X-rays. They studied ResNet50 and InceptionV3 and showed pretraining offers little performance improvement. Architectures composed of just four to five sequential convolution and pooling layers achieved comparable performance on these tasks as ResNet50 with less than 40% the parameters. In our work, we find pretraining does not boost performance for ResNet50, InceptionV3, InceptionV4, and MNASNet but does boost performance for the remaining 12 architectures. Thus, we were able to replicate [180]'s results, but upon studying a broader set of newer and more popular models, we reached the opposite conclusion that ImageNet pretraining yields a statistically significant boost in performance.

4.2.2 Medical Task Architectures

[106] compared the performance of ResNet152, DenseNet121, InceptionV4, and SEResNeXt101 on CheXpert, finding that DenseNet121 performed best. In a recent analysis, all but one of the top ten CheXpert competition models used DenseNets as part of their ensemble, even though they have

been surpassed on ImageNet [184]. Few groups design their own networks from scratch, preferring to use established ResNet and DenseNet architectures for CheXpert [27]. This trend extends to retinal fundus and skin cancer tasks as well, where Inception architectures remain popular [147, 262, 133, 51]. The popularity of these older ImageNet architectures hints that there may be a disconnect between ImageNet performance and medical task performance for newer architectures generated through architecture search. We verify that these newer architectures generated through search (EfficientNet, MobileNet, MNASNet) underperform older architectures (DenseNet, ResNet) on CheXpert, suggesting that search has overfit to ImageNet and explaining the popularity of these older architectures in the medical imaging literature.

[27] postulated that deep CNNs that can represent more complex relationships for ImageNet may not be necessary for CheXpert, which has greyscale inputs and fewer image classes. They studied ResNet, DenseNet, VGG, SqueezeNet, and AlexNet performance on CheXpert and found that ResNet152, DenseNet161, and ResNet50 performed best on CheXpert AUC. In terms of AUPRC, they showed that smaller architectures like AlexNet and VGG can perform similarly to deeper architectures on CheXpert. Models such as AlexNet, VGG, and SqueezeNet are no longer popular in the medical setting, so in our work, we systematically investigate the performance and efficiency of 16 more contemporary ImageNet architectures with and without pretraining. Additionally, we extend [27] by studying the effects of pretraining, characterizing the relationship between ImageNet and CheXpert performance, and drawing conclusions about architecture design.

4.2.3 Truncated Architectures

The more complex a convolutional architecture becomes, the more computational and memory resources are needed for its training and deployment. Model complexity thus may impede the deployment of CNNs to clinical settings with less resources. Therefore, efficiency, often reported in terms of the number of parameters in a model, the number of FLOPS in the forward pass, or the latency of the forward pass, has become increasingly important in model design. Low-rank factorization [107, 47], transferred/compact convolutional filters [44], knowledge distillation [99], and parameter pruning [216] have all been proposed to make CNNs more efficient.

Layer-wise pruning is a type of parameter pruning that locates and removes layers that are not as useful to the target task [194]. Through feature diagnosis, a linear classifier is trained using the feature maps at intermediate layers to quantify how much a particular layer contributes to performance on the target task [40]. In this work, we propose model truncation as a simple method for layer-wise pruning where the final pretrained layers after a given point are pruned off, a classification layer is appended, and this whole model is finetuned on the target task.

4.3 Methods

4.3.1 Training and Evaluation Procedure

We train chest X-ray classification models with different architectures with and without pretraining. The task of interest is to predict the probability of different pathologies from one or more chest X-rays. We use the CheXpert dataset consisting of 224,316 chest X-rays of 65,240 patients [106] labeled for the presence or absence of 14 radiological observations. We evaluate models using the average of their AUROC metrics (AUC) on the five CheXpert-defined competition tasks (Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion) as well as the No Finding task to balance clinical importance and prevalence in the validation set.

We select 16 models pretrained on ImageNet from public checkpoints implemented in PyTorch 1.4.0: DenseNet (121, 169, 201) and ResNet (18, 34, 50, 101) from [161], Inception (V3, V4) and MNASNet from [30], and EfficientNet (B0, B1, B2, B3) and MobileNet (V2, V3) from [242]. We finetune and evaluate these architectures with and without ImageNet pretraining.

For each model, we finetune all parameters on the CheXpert training set. If a model is pretrained, inputs are normalized using mean and standard deviation learned from ImageNet. If a model is not pretrained, inputs are normalized with mean and standard deviation learned from CheXpert. We use the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with learning rate of 1×10^{-4}, a batch size of 16, and a cross-entropy loss function. We train on up to four Nvidia GTX 1080 with CUDA 10.1 and Intel Xeon CPU ES-2609 running Ubuntu 16.04. For one run of an architecture, we train for three epochs and evaluate each model every 8192 gradient steps. We train each model and create a final ensemble model from the ten checkpoints, which achieved the best average CheXpert AUC across the six tasks on the validation set. We report all our results on the CheXpert test set.

We use the nonparametric bootstrap to estimate 95% confidence intervals for each statistic. 1,000 replicates are drawn from the test set, and the statistic is calculated on each replicate. This procedure produces a distribution for each statistic, and we report the 2.5 and 97.5 percentiles as a confidence interval. Significance is assessed at the $p = 0.05$ level.

4.3.2 Truncated Architectures

We study truncated versions of DenseNet121, MNASNet, ResNet18, and EfficientNetB0. DenseNet121 and MNASNet were chosen because we found they have the greatest efficiency (by AUC per parameters) on CheXpert of the models we profile, ResNet18 was chosen because of its popularity as a compact model for medical tasks, and EfficientNetB0 was chosen because it is the smallest current-generation model of the 16 we study. DenseNet121 contains four dense blocks separated by transition blocks before the classification layer. By pruning the final dense block and associated transition block, the model now only contains three dense blocks, yielding DenseNet121Minus1. Similarly, pruning two dense blocks and associated transition blocks yields DenseNet121Minus2, and pruning

three dense blocks and associated transition blocks yields DenseNet121Minus3. For MNASNet, we remove up to the four of the final MBConv blocks to produce MNASNetMinus1 through MNASNetMinus4. For ResNet18, we remove up to the three of the final residual blocks with a similar method to produce ResNet18Minus1 through ResNet18Minus3. For EfficientNet, we remove up to two of the final MBConv6 blocks to produce EfficientNetB0Minus1 and EfficientNetB0Minus2.

After truncating a model, we append a classification block containing a global average pooling layer followed by a fully connected layer to yield outputs of the correct shape. We initialize the model with ImageNet pretrained weights, except the randomly initialized classification block, and finetune using the same training procedure as the 16 ImageNet models.

4.3.3 Class Activation Maps

We compare the class activation maps (CAMs) among a truncated DenseNet121 family to visualize their higher resolution CAMs. We generate CAMs using the Grad-CAM method [208], using a weighted combination of the model's final convolutional feature maps, with weights based on the positive partial derivatives with respect to class score. This averaged map is scaled by the outputted probability so more confident predictions appear brighter. Finally, the map is upsampled to the input image resolution and overlain onto the input image, highlighting image regions that had the greatest influence on a model's decision.

4.4 Experiments

4.4.1 ImageNet Transfer Performance

We investigate whether higher performance on natural image classification translates to higher performance on chest X-ray classification. We display the relationship between the CheXpert AUC, with and without ImageNet pretraining, and ImageNet top-1 accuracy in Figure 4.2.

When models are trained without pretraining, we find no monotonic relationship between ImageNet top-1 accuracy and average CheXpert AUC, with Spearman $\rho = 0.082$ at $p = 0.762$. Model performance without pretraining would describe how a given architecture would perform on the target task, independent of any pretrained weights. When models are trained with pretraining, we again find no monotonic relationship between ImageNet top-1 accuracy and average CheXpert AUC with Spearman $\rho = 0.059$ at $p = 0.829$.

Overall, we find no relationship between ImageNet and CheXpert performance, so models that succeed on ImageNet do not necessarily succeed on CheXpert. These relationships between ImageNet performance and CheXpert performance are much weaker than the relationships between ImageNet performance and performance on various natural image tasks reported by [122].

Figure 4.2: Average CheXpert AUC vs. ImageNet Top-1 Accuracy. The left plot shows results obtained without pretraining, while the right plot shows results with pretraining. There is no monotonic relationship between ImageNet and CheXpert performance without pretraining (Spearman $\rho = 0.08$) or with pretraining (Spearman $\rho = 0.06$).

We compare the CheXpert performance within and across architecture families. Without pretraining, we find that ResNet101 performs only 0.005 AUC greater than ResNet18, which is well within the confidence interval of this metric (Figure 4.2). Similarly, DenseNet201 performs 0.004 AUC greater than DenseNet121 and EfficientNetB3 performs 0.003 AUC greater than EfficientNetB0. With pretraining, we continue to find minor performance differences between the largest model and smallest model that we test in each family. We find AUC increases of 0.002 for ResNet, 0.004 for DenseNet and -0.006 for EfficientNet. Thus, increasing complexity within a model family does not yield increases in CheXpert performance as meaningful as the corresponding increases in ImageNet performance.

Without pretraining, we find that the best model studied performs significantly better than the worst model studied. Among models trained without pretraining, we find that InceptionV3 performs best with 0.866 (0.851, 0.880) AUC, while MobileNetV2 performs worst with 0.814 (0.796, 0.832) AUC. Their difference in performance is 0.052 (0.043, 0.063) AUC. InceptionV3 is also the third largest architecture studied and MobileNetV2 the smallest. We find a significant difference in the CheXpert performance of these models. This difference again hints at the importance of architecture

Figure 4.3: Average CheXpert AUC vs. Model Size. The left plot shows results obtained without pretraining, while the right plot shows results with pretraining. The logarithm of the model size has a near linear relationship with CheXpert performance when we omit pretraining (Spearman $\rho = 0.79$). However once we incorporate pretraining, the monotonic relationship is weaker (Spearman $\rho = 0.56$).

design.

4.4.2 CheXpert Performance and Efficiency

We examine whether larger architectures perform better than smaller architectures on chest X-ray interpretation, where architecture size is measured by number of parameters. We display these relationships in Figure 4.3 and Table 4.1.

Without ImageNet pretraining, we find a positive monotonic relationship between the number of parameters of an architecture and CheXpert performance, with Spearman $\rho = 0.791$ significant at $p = 2.62 \times 10^{-4}$. With ImageNet pretraining, there is a weaker positive monotonic relationship between the number of parameters and average CheXpert AUC, with Spearman $\rho = 0.565$ at $p = 0.023$.

Although there exists a positive monotonic relationship between the number of parameters of an architecture and average CheXpert AUC, the Spearman ρ does not highlight the increase in parameters necessary to realize marginal increases in CheXpert AUC. For example, ResNet101 is

Model	CheXpert AUC	#Params (M)
DenseNet121	0.859 (0.846, 0.871)	6.968
DenseNet169	0.860 (0.848, 0.873)	12.508
DenseNet201	0.864 (0.850, 0.876)	18.120
EfficientNetB0	0.859 (0.846, 0.871)	4.025
EfficientNetB1	0.858 (0.844, 0.872)	6.531
EfficientNetB2	0.866 (0.853, 0.880)	7.721
EfficientNetB3	0.853 (0.837, 0.867)	10.718
InceptionV3	0.862 (0.848, 0.876)	27.161
InceptionV4	0.861 (0.846, 0.873)	42.680
MNASNet	0.853 (0.839, 0.866)	4.38
MobileNetV2	0.854 (0.839, 0.869)	2.242
MobileNetV3	0.859 (0.847, 0.872)	4.220
ResNet101	0.863 (0.848, 0.876)	44.549
ResNet18	0.862 (0.847, 0.875)	11.690
ResNet34	0.863 (0.849, 0.875)	21.798
ResNet50	0.859 (0.843, 0.871)	25.557

Table 4.1: CheXpert AUC (with 95% Confidence Intervals) and Number of Parameters for 16 ImageNet-Pretrained Models.

11.1x larger than EfficientNetB0, but with only increase of 0.005 in CheXpert AUC with pretraining.

Within a model family, increasing the number of parameters does not lead to meaningful gains in CheXpert AUC. We see this relationship in all families studied without pretraining (EfficientNet, DenseNet, and ResNet) in Figure 4.3. For example, DenseNet201 has an AUC 0.003 greater than DenseNet121, but is 2.6x larger. EfficientNetB3 has an AUC 0.004 greater than EfficientNetB0, but is 1.9x larger. Despite the positive relationship between model size and CheXpert performance across all models, bigger does not necessarily mean better within a model family.

Since within a model family there is a weaker relationship between model size and CheXpert performance than across all models, we find that CheXpert performance is influenced more by the macro architecture design than by its size. Models within a family have similar architecture design choices but different sizes, so they perform similarly on CheXpert. We observe large discrepancies in performance between architecture families. For example DenseNet, ResNet, and Inception typically outperform EfficientNet and MobileNet architectures, regardless of their size. EfficientNet, MobileNet, and MNASNet were all generated through neural architecture search to some degree, a process that optimized for performance on ImageNet. Our findings suggest that this search could have overfit to the natural image objective to the detriment of chest X-ray tasks.

4.4.3 ImageNet Pretraining Boost

We study the effects of ImageNet pretraining on CheXpert performance by defining the pretraining boost as the CheXpert AUC of a model initialized with ImageNet pretraining minus the CheXpert

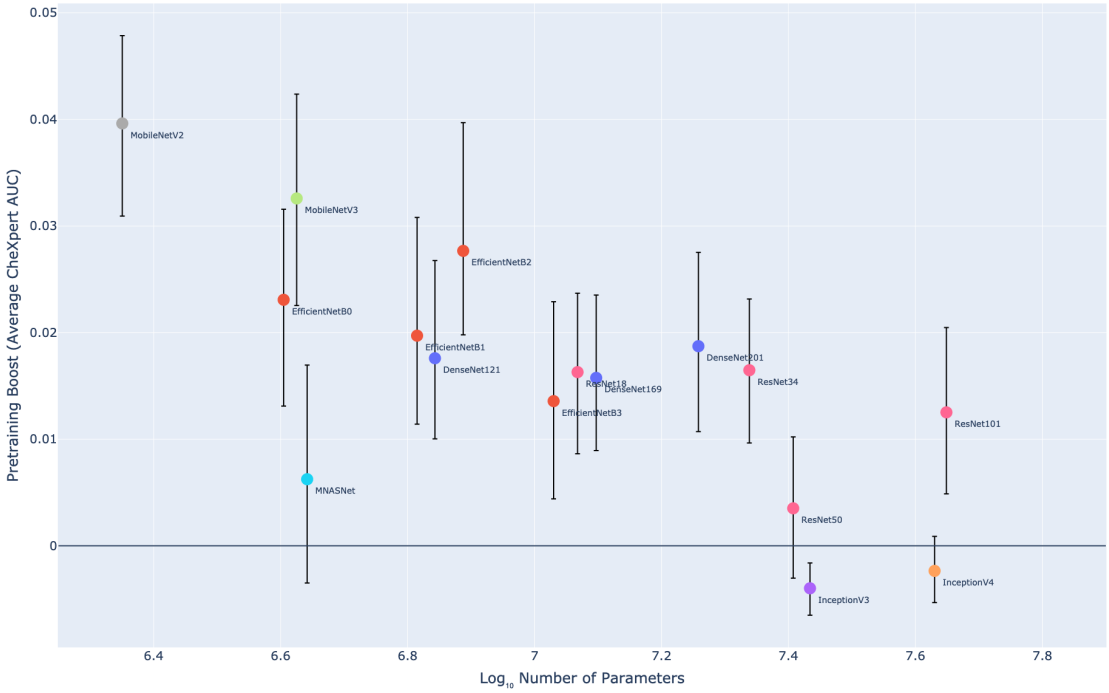


Figure 4.4: Pretraining Boost vs. Model Size. We define pretraining boost as the increase in the average CheXpert AUCs achieved with pretraining vs. without pretraining. Most models benefit significantly from ImageNet pretraining. Smaller models tend to benefit more than larger models (Spearman $\rho = -0.72$).

AUC of its counterpart without pretraining. The pretraining boosts of our architectures are reported in Figure 4.4.

We find that ImageNet pretraining provides a meaningful boost for most architectures (on average 0.015 AUC). We find a Spearman $\rho = -0.718$ at $p = 0.002$ between the number of parameters of a given model and the pretraining boost. Therefore, this boost tends to be larger for smaller architectures such as EfficientNetB0 (0.023), MobileNetV2 (0.040) and MobileNetV3 (0.033) and smaller for larger architectures such as InceptionV4 (-0.002) and ResNet101 (0.013). Further work is required to explain this relationship.

Within a model family, the pretraining boost also does not meaningfully increase as as model size increases. For example, DenseNet201 has a pretraining boost only 0.002 AUC greater than DenseNet121 does. This finding supports our earlier conclusion that model families perform similarly on CheXpert regardless of their size.

Model	AUC Change	Times-Smaller
EfficientNetB0	0.00%	1x
EfficientNetB0Minus1	0.15%	1.4x
EfficientNetB0Minus2	-0.45%	4.7x
MNASNet	0.00%	1x
MNASNetMinus1	0.55%	2.1x
MNASNetMinus2*	-1.38%	9.2x
MNASNetMinus3*	-1.55%	16.5x
MNASNetMinus4*	-4.90%	93.3x
DenseNet121	0.00%	1x
DenseNet121Minus1	-0.04%	1.6x
DenseNet121Minus2*	-1.33%	5.3x
DenseNet121Minus3*	-4.73%	20.0x
ResNet18	0.00%	1x
ResNet18Minus1	0.24%	4.2x
ResNet18Minus2*	-3.70%	17.1x
ResNet18Minus3*	-8.33%	73.8x

Table 4.2: Efficiency Trade-Off of Truncated Models. Pretrained models can be truncated without significant decrease in CheXpert AUC. Truncated models with significantly different AUC from the base model are denoted with an asterisk.

4.4.4 Truncated Architectures

We truncate the final blocks of DenseNet121, MNASNet, ResNet18, and EfficientNetB0 with pretrained weights and study their CheXpert performance to understand whether ImageNet models are unnecessarily large for the chest X-ray task. We express efficiency gains in terms of Times-Smaller, or the number of parameters of the original architecture divided by the number of parameters of the truncated architecture: intuitively, how many times larger the original architecture is compared to the truncated architecture. The efficiency gains and AUC changes of model truncation on DenseNet121, MNASNet, ResNet18, and EfficientNetB0 are displayed in Table 4.2.

For all four model families, we find that truncating the final block leads to no significant decrease in CheXpert AUC but can save 1.4x to 4.2x the parameters. Notably, truncating the final block of ResNet18 yields a model that is not significantly different (difference -0.002 (-0.008, 0.004)) in CheXpert AUC, but is 4.2x smaller. Truncating the final two blocks of an EfficientNetB0 yields a model that is not significantly different (difference 0.004 (-0.003, 0.009)) in CheXpert AUC, but is 4.7x smaller. However, truncating the second block and beyond in each of MNASNet, DenseNet121, and ResNet18 yields models that have statistically significant drops in CheXpert performance.

Model truncation effectively compresses models performant on CheXpert, making them more parameter efficient while still using pretrained weights to capture the pretraining boost. Parameter efficient models are able to lighten the computational and memory burdens for deployment to low-resource environments such as portable devices. In the clinical setting, the simplicity of our model

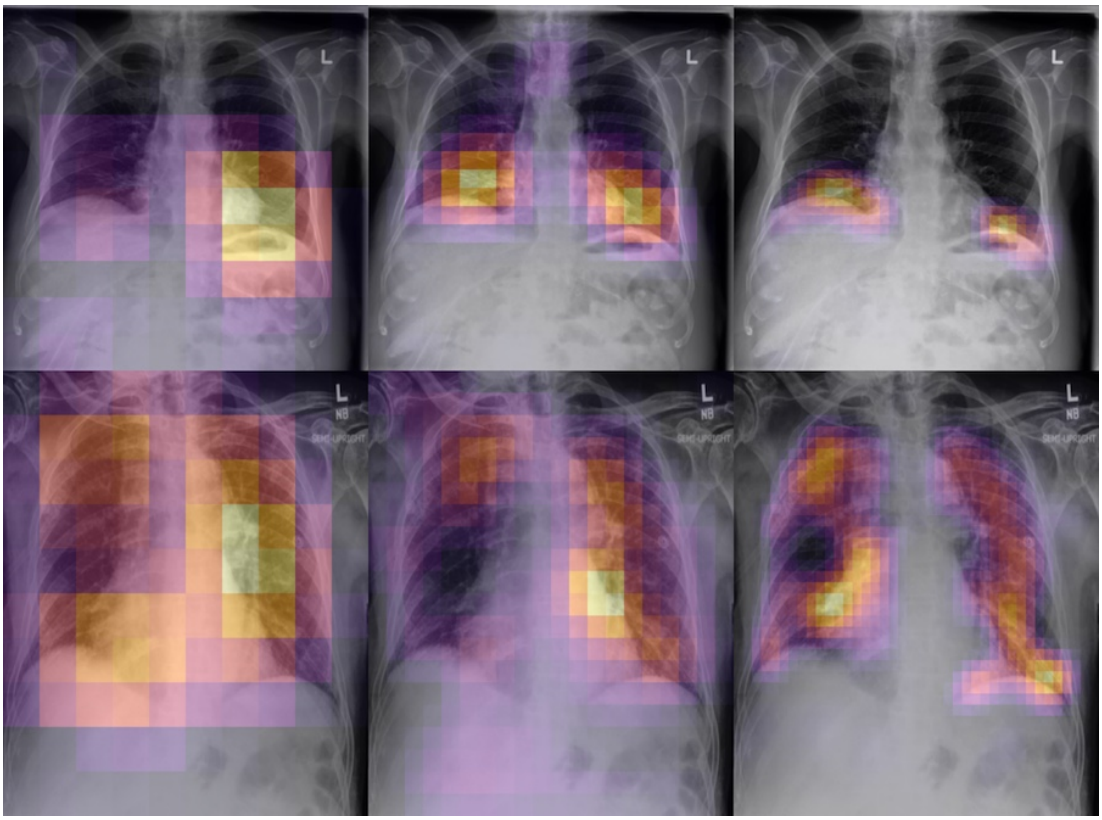


Figure 4.5: Comparison of Class Activation Maps Among Truncated Model Family. CAMs yielded by models, from left to right, DenseNet121, DenseNet121Minus1, and DenseNet121Minus2. Displays frontal chest X-ray demonstrating Atelectasis (top) and Edema (bottom). Further truncated models more effectively localize the Atelectasis, as well as tracing the hila and vessel branching for Edema.

truncation method encourages its adoption for model compression.

This finding corroborates [180] and [27], which show simpler models can achieve performance comparable to more complex models on CheXpert. Our truncated models can use readily-available pretrained weights, which may allow these models to capture the pretraining boost and speed up training. However, we do not study the performance of these truncated models without their pretrained weights.

As an additional benefit, architectures that truncate pooling layers will also produce higher-resolution class activation maps, as shown in Figure 4.5. The higher-resolution class activation maps (CAMs) may more effectively localize pathologies with little to no decrease in classification performance. In clinical settings, improved explainability through better CAMs may be useful for validating predictions and diagnosing mispredictions. As a result, clinicians may have more trust in models that provide these higher-resolution CAMs.

4.5 Discussion

In this work, we study the performance and efficiency of ImageNet architectures for chest x-ray interpretation.

Is ImageNet performance correlated with CheXpert? No. We show no statistically significant relationship between ImageNet and CheXpert performance. This finding extends [122]—which found a significant correlation between ImageNet performance and transfer performance on typical image classification datasets—to the medical setting of chest x-ray interpretation. This difference could be attributed to unique aspects the chest X-ray interpretation task and data attributes. The chest X-ray interpretation task differs from natural image classification in that (1) disease classification may depend on abnormalities in a small number of pixels, (2) chest X-ray interpretation is a multi-task classification setup, and (3) there are far fewer classes than in many natural image classification datasets. Second, the data attributes for chest X-rays differ from natural image classification in that X-rays are greyscale and have similar spatial structures across images (always either anterior-posterior, posterior-anterior, or lateral).

Does model architecture matter? Yes. For models without pretraining, we find that the choice of architecture family may influence performance more than model size. Our findings extend [180] beyond the effect of ImageNet weights, since we show that architectures that succeed on ImageNet do not necessarily succeed on medical imaging tasks. A notable finding of our work is that newer architectures generated through search on ImageNet (EfficientNet, MobileNet, MNASNet) underperform older architectures (DenseNet, ResNet) on CheXpert. This finding suggests that search may have overfit to ImageNet to the detriment of medical task performance, and ImageNet may not be an appropriate benchmark for selecting architectures for medical imaging tasks. Instead, medical imaging architectures could be benchmarked on CheXpert or other large medical datasets. Architectures derived from selection and search on CheXpert and other large medical datasets may be applicable to similar medical imaging modalities including other x-ray studies, or CT scans. Thus architecture search directly on CheXpert or other large medical datasets may allow us to unlock next generation performance for medical imaging tasks.

Does ImageNet pretraining help? Yes. We find that ImageNet pretraining yields a statistically significant boost in performance for chest x-ray classification. Our findings are consistent with [180], who find no pretraining boost on ResNet50 and InceptionV3, but we find pretraining does boost performance for 12 out of 16 architectures. Our findings extend [92]—who find models without pretraining had comparable performance to models pretrained on ImageNet for object detection and image segmentation of natural images—to the medical imaging setting. Future work may

investigate the relationship between network architectures and the impact of self-supervised pretraining for chest x-ray interpretation as has recently been developed by [214, 11, 218].

Can models be smaller? Yes. We find that by truncating final blocks of ImageNet-pretrained architectures, we can make models 3.25x more parameter-efficient on average without a statistically significant drop in performance. This method preserves the critical components of architecture design while cutting its size. This observation suggests model truncation may be a simple method to yield lighter models, using ImageNet pretrained weights to boost CheXpert performance. In the clinical setting, truncated models may provide value through improved parameter-efficiency and higher resolution CAMs. This change may enable deployment to low-resource clinical environments and further develop model trust through improved explainability.

In closing, our work contributes to the understanding of the transfer performance and parameter efficiency of ImageNet models for chest X-ray interpretation. We hope that our new experimental evidence about the relation of ImageNet to medical task performance will shed light on potential future directions for progress.

Chapter 5

Pretraining Using Simple Contrastive Learning

In the previous chapter, we have seen how ImageNet pre-training may be leveraged in limited labeled data settings. In this chapter, we will look at contrastive learning, which is a form of self-supervision that can leverage unlabeled data to produce pretrained models. While contrastive learning has demonstrated promising results on natural image classification tasks, its application to medical imaging tasks like chest X-ray interpretation has been limited. In this work, we propose MoCo-CXR, which is an adaptation of the contrastive learning method Momentum Contrast (MoCo), to produce models with better representations and initializations for the detection of pathologies in chest X-rays. In detecting pleural effusion, we find that linear models trained on MoCo-CXR-pretrained representations outperform those without MoCo-CXR-pretrained representations, indicating that MoCo-CXR-pretrained representations are of higher-quality. End-to-end fine-tuning experiments reveal that a model initialized via MoCo-CXR-pretraining outperforms its non-MoCo-CXR-pretrained counterpart. We find that MoCo-CXR-pretraining provides the most benefit with limited labeled training data. Finally, we demonstrate similar results on a target Tuberculosis dataset unseen during pretraining, indicating that MoCo-CXR-pretraining endows models with representations and transferability that can be applied across chest X-ray datasets and tasks.

This chapter is based on [214].

5.1 Introduction

Self-supervised approaches such as Momentum Contrast (MoCo) [91, 43] can leverage unlabeled data to produce pretrained models for subsequent fine-tuning on labeled data. Contrastive learning of visual representations has emerged as the front-runner for self-supervision and has demonstrated

superior performance on downstream tasks. In addition to MoCo, these include frameworks such as SimCLR [41, 42] and PIRL [146]. All contrastive learning frameworks involve maximizing agreement between positive image pairs relative to negative/different images via a contrastive loss function; this pretraining paradigm forces the model to learn good representations. These approaches typically differ in how they generate positive and negative image pairs from unlabeled data and how the data are sampled during pretraining. While MoCo and other contrastive learning methods have demonstrated promising results on natural image classification tasks, their application to medical imaging settings has been limited [180, 45].

Chest X-ray is the most common imaging tool used in practice, and is critical for screening, diagnosis, and management of diseases. The recent introduction of large datasets (size 100k-500k) of chest X-rays [106, 111, 29] has driven the development of deep learning models that can detect diseases at a level comparable to radiologists [187, 186]. Because there is an abundance of unlabeled chest X-ray data [189], contrastive learning approaches represent a promising avenue for improving chest X-ray interpretation models.

Chest X-ray interpretation is fundamentally different from natural image classification in that (1) disease classification may depend on abnormalities in a small number of pixels, (2) data attributes for chest X-rays differ from natural image classification because X-rays are larger, grayscale and have similar spatial structures across images (always either anterior-posterior, posterior-anterior, or lateral), (3) there are far fewer unlabeled chest X-ray images than natural images. These differences may limit the applicability of contrastive learning approaches, which were developed for natural image classification settings, to chest X-ray interpretation. For instance, MoCo uses a variety of data augmentations to generate positive image pairs from unlabeled data; however, random crops and blurring may eliminate disease-covering parts from an augmented image, while color jittering and random gray scale would not produce meaningful transformations for already grayscale images. Furthermore, given the availability of orders of magnitude fewer chest X-ray images than natural images, and larger image sizes, it remains to be understood whether retraining may improve on the traditional paradigm for automated chest X-ray interpretation, in which models are fine-tuned on labeled chest X-ray images from ImageNet-pretrained weights.

In this work, we demonstrate that our proposed MoCo-CXR method leads to better representations and initializations than those acquired without MoCo-pretraining (solely from ImageNet) for chest X-ray interpretation. The MoCo-CXR pipeline involved first a modified MoCo-pretraining on CheXpert [106], where we adapted initialization, data augmentations, and learning rate scheduling of this pretraining step for successful application on chest X-rays. This was then followed by supervised fine-tuning experiments using different fractions of labeled data. We showed that MoCo-CXR-pretrained representations are of higher quality than ImageNet-pretrained representations by evaluating the performance of a linear classifier trained on pretrained representations on a chest

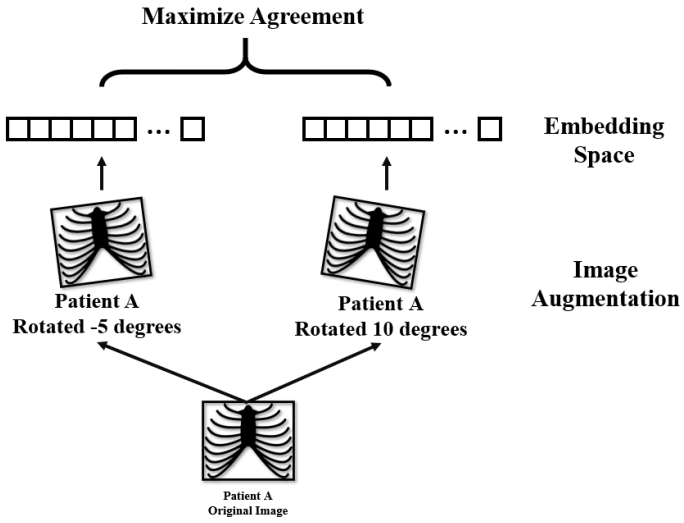


Figure 5.1: Contrastive learning maximizes agreement of embeddings generated by different augmentations of the same chest X-ray image.

X-ray interpretation task. We also demonstrated that a model trained end-to-end with MoCo-CXR-pretrained initialization had superior performance on the X-ray interpretation tasks, and the advantage was especially apparent at low labeled data regimes. Finally, we also showed that MoCo-CXR-pretrained representations from the source (CheXpert) dataset transferred to another small chest X-ray dataset (Shenzhen) with a different classification task [108]. Our study demonstrates that MoCo-CXR provides high-quality representations and transferable initializations for chest X-ray interpretation.

5.2 Related Work

Self-supervised learning Self-supervision is a form of unsupervised pretraining that uses a formulated pretext task on unlabeled data as the training goal. Handcrafted pretext tasks include solving jigsaw puzzles [156], relative patch prediction [59] and colorization [263]. However, many of these tasks rely on ad-hoc heuristics that could limit the generalization and transferability of learned representations for downstream tasks. Consequently, contrastive learning of visual representations has emerged as the front-runner for self-supervision and has demonstrated superior performance on downstream tasks [43, 42].

Contrastive learning for chest X-rays Prior work using contrastive learning on chest X-rays is limited in its applicability to unlabeled data and evidence of transferability. A controlled approach is to explicitly contrast X-rays with pathologies against healthy ones using attention network [138];

however, this approach is supervised. There has also been a proposed domain-specific strategy of extracting contrastive pairs from MRI and CT datasets using a combination of localized contrastive loss function and global loss function during pretraining [33]. However, the method is highly dependent on the volumetric nature of MRI and CT scans, as the extraction of similar image pairs is based on taking 2D image slices of a single volumetric image. Thus, the technique would have limited applicability to chest X-rays. Work applying broader self-supervised techniques to medical imaging is more extensive. For example, encoding shared information between different imaging modalities for ophthalmology was shown to be an effective pretext task for pretraining diabetic retinopathy classification models [101]. Other proposed pretext tasks in medical imaging include solving a Rubik's cube [269, 267], predicting the position of anatomical patches [14], anatomical reconstruction [266], and image patch distance estimation [215].

ImageNet transfer for chest X-ray interpretation The dominant computer vision approach of starting with an ImageNet-pretrained model has been proven to be highly effective at improving model performance in diverse settings such as object detection and image segmentation [101]. Although high performance deep learning models for chest X-ray interpretation use ImageNet-pretrained weights, [222] found that common regularization techniques limit ImageNet transfer learning benefits and that ImageNet features are less general than previously believed. Moreover, [269] showed that randomly-initialized models are competitive with their ImageNet-initialized counterparts on a vast array of tasks with sufficient labeled data, and that pretraining merely speeds up convergence. [180] further investigated the efficacy of ImageNet pretraining, observing that simple convolutional architectures are able to achieve comparable performance as larger ImageNet model architectures.

5.3 Methods

5.3.1 Chest X-ray datasets and diagnostic tasks

We used a large source chest X-ray dataset for pretraining and a smaller external chest X-ray dataset for the evaluation of model transferability. The source chest X-ray dataset we select is CheXpert, a large collection of chest X-ray images labeled for the presence or absence of several diseases [106]. CheXpert consists of 224k chest X-rays collected from 65k patients. Chest X-ray images included in the CheXpert dataset are of size 320×320. We focused on identifying the presence of pleural effusion, a clinically important condition that has high prevalence in the dataset (with 45.63% of all images labeled as positive or uncertain). We performed follow-up experiments with other CheXpert competition tasks (cardiomegaly, consolidation, edema and atelectasis) from [106] to verify that our method worked on different pathologies. In addition, we use the Shenzhen Hospital X-ray set [108] for evaluation of model transferability to an external target dataset. Chest X-ray images included in

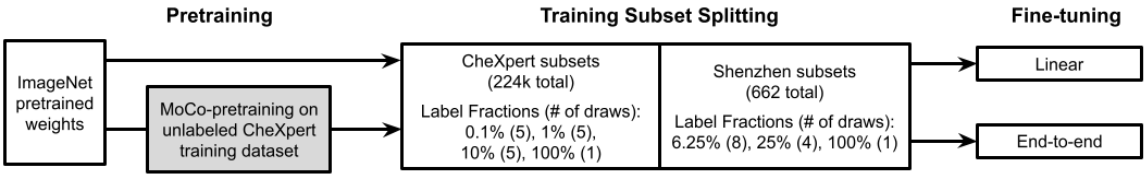


Figure 5.2: MoCo-CXR training pipeline. MoCo acts as self-supervised training agent. The model is subsequently tuned using chest X-ray images.

the Shenzhen dataset are of size 4020 × 4892 and 4892 × 4020. The Shenzhen dataset contains 662 X-ray images, of which 336 (50.8%) are abnormal X-rays that have manifestations of tuberculosis. All images in both datasets are resized to 224 × 224 for MoCo-CXR.

5.3.2 MoCo-CXR Pretraining for Chest X-ray Interpretation

We adapt the MoCo-pretraining procedure to chest X-rays. MoCo is a form of self-supervision that utilizes contrastive learning, where the pretext task is to maximize agreement between different views of the same image (positive pairs) and to minimize agreement between different images (negative pairs). Figure 5.1 illustrates how data augmentations are used to generate views of a particular image and are subsequently contrasted to learn embeddings in an unsupervised fashion.

Our choice to use MoCo is driven by two constraints in medical imaging AI: (1) large image sizes, and (2) the cost of large computational resources. Compared to other self-supervised frameworks such as SimCLR [41], MoCo requires far smaller batch sizes during pretraining [43]. The MoCo implementation used a batch size of 256 and achieved comparable performance on ImageNet as the SimCLR implementation, which used a batch size of 4096; in contrast, SimCLR experienced lower performance at a batch size of 256 [43]. MoCo's reduced dependency on mini-batch size is achieved by using a momentum updated queue of previously seen samples to generate contrastive pair encodings. An illustration of MoCo's momentum encoding framework has been added as Figure 5.6. Using MoCo, we were able to conduct experiments on a single NVIDIA GTX 1070 with a batch size of 16.

We performed MoCo-pretraining on the entire CheXpert training dataset. We chose to apply MoCo-pretraining on ImageNet-initialized models to leverage possible convergence benefits [180]. Due to the widespread availability of ImageNet-pretrained weights, there is no extra cost to initialize models with ImageNet weights before MoCo-pretraining.

We modified the data augmentation strategy used to generate views suitable for the chest X-ray interpretation task. Data augmentations used in self-supervised approaches for natural images may not be appropriate for chest X-rays. For example, random crop and Gaussian blur could change the disease label for an X-ray image or make it impossible to distinguish between diseases. Furthermore,

color jittering and random grayscale do not represent meaningful augmentations for grayscale X-rays. Instead, we use random rotation (10 degrees) and horizontal flipping (Figure 5.7), a set of augmentations commonly used in training chest X-ray models [106, 183] driven by experimental findings in the supervised setting and clinical domain knowledge. Future work should investigate the impact of various additional augmentations and their combinations.

The overall training pipeline with MoCo-CXR-pretraining and the subsequent fine-tuning with CheXpert and Shenzhen datasets is illustrated in Figure 5.2. We maintained hyperparameters related to momentum, weight decay, and feature dimension from MoCo [43]. Checkpoints from top performing epochs were saved for subsequent checkpoint selection and model evaluation. In the subsequent fine-tuning step, we selected hyperparameters based on performance of linear evaluations. We used two backbones, ResNet18 and DenseNet121, to evaluate the consistency of our findings across model architectures. We experimented with initial learning rates of 10^{-2}, 10^{-3}, 10^{-4} and 10^{-5}, and investigated their effect on performance. We also experimented with milestone and cosine learning rate schedulers.

5.3.3 MoCo-CXR Model Fine-tuning

We fine-tuned models on different fractions of labeled training data. We also conducted baseline fine-tuning experiments with ImageNet-pretrained models that were not subjected to MoCo-CXR-pretraining. We use label fraction to represent the ratio of data with its labels retained during training. For a model trained with 1% label fraction, the model will only have access to 1% of the all labels, while the remaining 99% of labels are hidden from the model. The use of label fraction is a proxy for the real world, where large amounts of data remain unlabelled and only a small portion of well-labelled data can be used toward supervised training. As presented in Figure 5.2, the label fractions of training sets are 0.1%, 1%, 10% and 100% for the CheXpert dataset and 6.25%, 25%, 100% for the external Shenzhen dataset. Fine-tuning experiments on small label fractions are repeated multiple times with different random samples and averaged to guard against anomalous, unrepresentative training splits.

To evaluate the transfer of representations, we froze the backbone model and trained a linear classifier on top using the labeled data (MoCo-CXR/ImageNet-pretrained Linear Models). In addition, we unfreeze all layers and fine-tune the entire model end-to-end using the labeled data to assess transferability on the overall performance (MoCo-CXR/ImageNet-pretrained end-to-end Models). Our models were fine-tuned using the same configurations as fully-supervised models designed for CheXpert [106], which has determined an optimal batch size, learning rate and other hyper-parameters. To be specific, we use a learning rate of 3×10^{-5}, batch size of 16 and number of epochs that scale with the size of labeled data. For the CheXpert dataset, these are 220, 95, 41, 18 epochs for the 4 label fractions respectively.

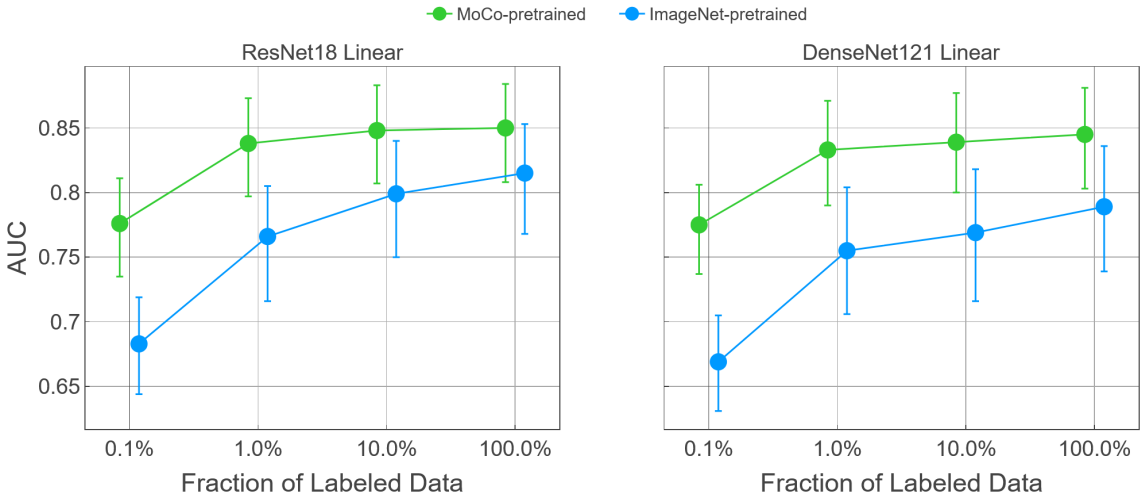


Figure 5.3: AUC on pleural effusion task for linear models with MoCo-CXR-pretraining is consistently higher than AUC of linear models with ImageNet-pretraining, showing that MoCo-CXR-pretraining produces higher quality representations than ImageNet-pretraining does.

5.3.4 Statistical analysis

We compared the performance of the models trained with and without MoCo-CXR-pretraining using the area under the receiver operating characteristic curve (AUC). To assess whether MoCo-CXR-pretraining significantly changed the performance, we computed the difference in AUC on the test set with and without MoCo-CXR-pretraining. The non-parametric bootstrap was used to estimate the variability around model performance. A total of 500 bootstrap replicates from the test set were drawn, and the AUC and its corresponding differences were calculated for the MoCo-CXR-pretrained model and non-Moco-CXR pretrained model on these same 500 bootstrap replicates. This produced a distribution for each estimate, and the 95% bootstrap percentile intervals were computed to assess significance at the $p = 0.05$ level.

5.4 Experiments

5.4.1 MoCo-CXR-pretrained representations on CheXpert

We investigated whether representations acquired from MoCo-CXR-pretraining are of higher quality than those transferred from ImageNet. To evaluate the representations, we used the linear evaluation protocol [159, 263, 122, 13], where a linear classifier is trained on a frozen base model, and test performance is used as a proxy for representation quality. We visualize the performance of MoCo-CXR-pretrained and ImageNet-pretrained linear models at various label fractions in Figure 5.3 and tabulate the corresponding AUC improvements in Table 5.1.

Trained on small label fractions, the ResNet18 MoCo-CXR-pretrained linear model demonstrated statistically significant performance gains over its ImageNet counterpart. With 0.1% label fraction, the improvement in performance is 0.096 (95% CI 0.061, 0.130) AUC; the MoCo-CXR-pretrained and ImageNet-pretrained linear models achieved performances of 0.776 and 0.683 AUC respectively. These findings support the hypothesis that MoCo-representations are of superior quality, and are most apparent when labeled data is scarce.

With larger label fractions, the MoCo-CXR-pretrained linear models demonstrated clear yet diminishing improvements over the ImageNet-pretrained linear models. Training with 100% of the labeled data, the ResNet18 MoCo-CXR-pretrained linear model yielded a performance gain of 0.034 (95% CI -0.009, 0.080). Both backbones were observed to have monotonically decreasing performance gains with MoCo as we increase the amount of labeled training data. These results provide evidence that MoCo-CXR-pretrained representations retain their quality at all label fractions, but less significantly at larger label fractions. We generally observe similar performance gains with MoCo-CXR on the CheXpert competition pathologies, as seen in Table 5.6.

5.4.2 End-to-end MoCo-CXR-pretrained models on CheXpert

We investigated whether MoCo-CXR-pretraining translates to higher performance for models fine-tuned end-to-end. We visualize the performance of the MoCo and ImageNet-pretrained end-to-end models at different label fractions in Figure 5.4. AUC improvements of using a MoCo-CXR-pretrained linear model over an ImageNet-pretrained linear is tabulated in Table 5.1.

We found that MoCo-CXR-pretrained end-to-end models outperform their ImageNet-pretrained counterparts more at small label fractions than at larger label fractions. With the 0.1% label fraction, the ResNet18 MoCo-CXR-pretrained end-to-end model achieved an AUC of 0.813 while the ImageNet-pretrained end-to-end model achieves an AUC of 0.775, yielding a statistically significant AUC improvement of 0.037 (95% CI 0.015, 0.062). The AUC improvement with the 1.0% label fraction was also statistically significant at 0.027 (95% CI 0.006, 0.047). Both pretraining approaches converge to an AUC of 0.942 with the 100% label fraction.

These results demonstrate that MoCo-CXR-pretraining yields performance boosts for end-to-end training, and further substantiate the quality of the pretrained initialization, especially for smaller label fractions. This finding is consistent with behavior of SimCLR [41], which also saw larger performance gains for self-supervised models trained end-to-end on smaller label fractions of ImageNet. We generally observe similar performance gains with MoCo-CXR end-to-end on the CheXpert competition pathologies, as seen in Table 5.7.

5.4.3 Transfer benefit of MoCo-CXR-pretraining on an external dataset

We conducted experiments to test whether MoCo-CXR-pretrained chest X-ray representations acquired from a source dataset (CheXpert) transfer to a small target dataset (Shenzhen Dataset for

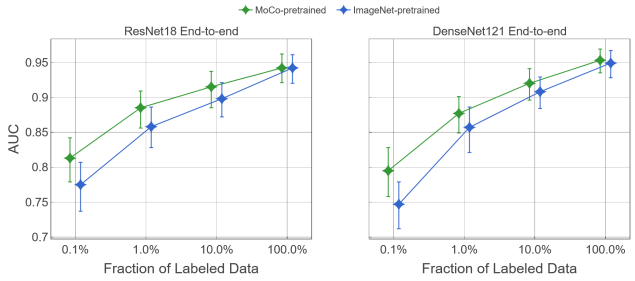


Figure 5.4: AUC on pleural effusion task for models fine-tuned end-to-end with MoCo-CXR-pretraining is consistently higher than those without MoCo-CXR-pretraining, showing that MoCo-CXR-pretraining representations are more transferable than those produced by ImageNet-pretraining only.

Architecture	MoCo-CXR-pretrained	ImageNet-pretrained	0.1%	1.0%	10.0%	100%
ResNet18	End-to-End	End-to-End	0.037(0.015, 0.062)	0.027(0.006, 0.047)	0.017(0.003, 0.031)	0.000(-0.009, 0.009)
ResNet18	Linear Model	Linear Model	0.096(0.061, 0.130)	0.070(0.029, 0.112)	0.049(0.005, 0.094)	0.034(-0.009, 0.080)
ResNet18	Linear Model	End-to-End	0.001(-0.024, 0.025)	-0.022(-0.051, 0.009)	-0.050(-0.083, -0.018)	-0.094(-0.127, -0.062)
DenseNet121	End-to-End	End-to-End	0.048(0.023, 0.074)	0.019(0.001, 0.037)	0.012(0.000, 0.023)	0.003(-0.006, 0.013)
DenseNet121	Linear Model	Linear Model	0.107(0.075, 0.142)	0.078(0.035, 0.121)	0.067(0.023, 0.111)	0.055(0.008, 0.102)
DenseNet121	Linear Model	End-to-End	0.029(0.002, 0.055)	-0.024(-0.050, -0.003)	-0.070(-0.109, -0.036)	-0.107(-0.141, -0.073)

Table 5.1: AUC improvements on pleural effusion task achieved by MoCo-CXR-pretrained models against models without MoCo-CXR-pretraining on the CheXpert dataset.

Tuberculosis, with 662 X-rays). Results of these experiments are presented in Figure 5.5.

We first examined whether MoCo-CXR-pretrained linear models improve AUC on the external Shenzhen dataset. With 6.25% label fraction, which is approximately 25 images, the ResNet18 MoCo-CXR-pretrained model outperformed the ImageNet-pretrained one by 0.054 (95%, CI 0.024, 0.086). AUC improvement with the 100% label fraction was 0.018 (95% CI -0.011, 0.053). This is similar to the trend observed on the CheXpert dataset discussed previously. These observations suggest that representations learned from MoCo-CXR-pretraining are better suited for an external target chest X-ray dataset with a different task than representations learned from ImageNet-pretraining.

Next, we tested whether MoCo-CXR-pretrained models with end-to-end training also perform well on the external Shenzhen dataset. With the 100% label fraction, the ResNet18 MoCo-CXR-pretrained model was able to achieve an AUC of 0.974. However, the corresponding AUC improvement of only 0.003 (95% CI -0.014, 0.020) is much less than the improvement observed for linear models. Since the Shenzhen dataset is limited in size, it is possible that training end-to-end quickly saturates learning potential at low label fractions. Regardless, the non-zero improvement still suggests that MoCo-CXR-pretrained initializations can transfer to an external dataset. This echoes previous investigations of self-supervised and unsupervised learnings, which found that unsupervised pretraining pushes the model towards solutions with better generalization to tasks that are in the same domain [222, 68].

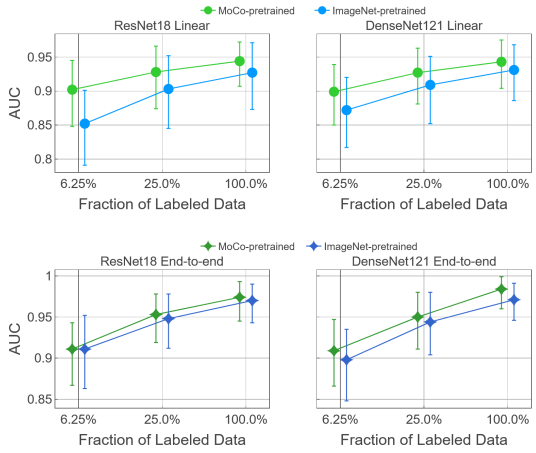


Figure 5.5: AUC on the Shenzhen tuberculosis task for models with and without MoCo-CXR-pretraining shows that MoCo pretraining still introduces significant improvement despite being fine-tuned on an external dataset.

Learning Rate	AUC
10^{-2}	0.786 (0.699, 0.861)
10^{-3}	0.908 (0.853, 0.955)
10^{-4}	0.944 (0.907, 0.972)
10^{-5}	0.939 (0.891, 0.975)

Table 5.2: AUC of MoCo pretrained ResNet18 on Shenzhen dataset at different pretraining learning rates with 100% label fraction.

5.5 Conclusion

We found that our MoCo-CXR method provides high-quality representations and transferable initializations for chest X-ray interpretation. Despite many differences in the data and task properties between natural image classification and chest X-ray interpretation, MoCo-CXR was a successful adaptation of MoCo pretraining to chest X-rays. These suggest the possibility for broad application of self-supervised approaches beyond natural image classification settings.

To the best of our knowledge, this work is the first to show the benefit of MoCo-pretraining across label fractions for chest X-ray interpretation, and also investigate representation transfer to a target dataset. All of our experiments are run on a single NVIDIA GTX 1070, demonstrating accessibility of this method. Our success in demonstrating improvements in model performance over the traditional supervised learning approach, especially on low label fractions, may be broadly extensible to other medical imaging tasks and modalities, where high-quality labeled data is expensive, but unlabeled data is increasingly easier to access.

5.6 Supplementary Details for the MoCo-CXR Method

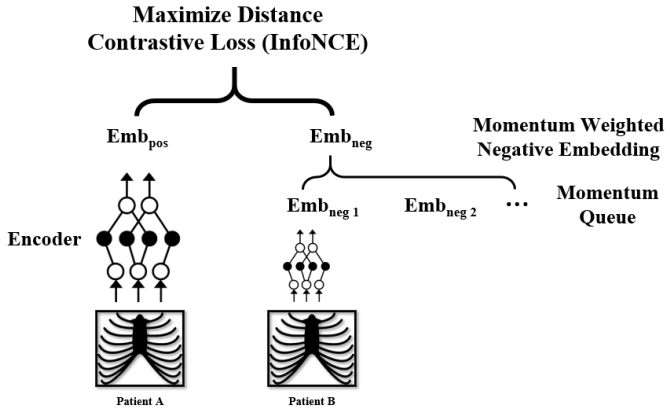


Figure 5.6: The MoCo framework generates negative embeddings in a momentum-weighted manner using a queue of negative embeddings. This setup reduces dependency on batch size, therefore has more relaxed hardware constraint compared to other self-supervised learning frameworks.

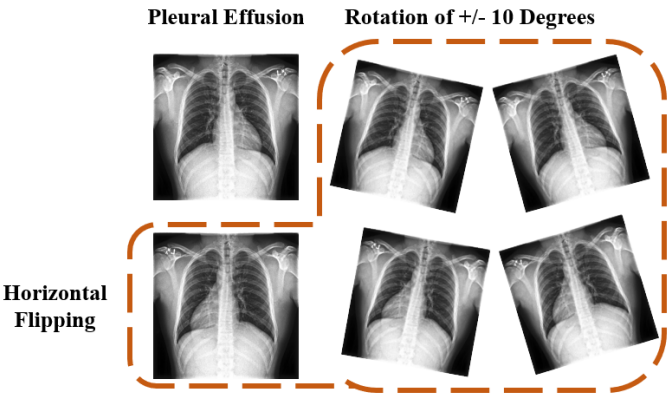


Figure 5.7: Illustration of data augmentation methods used for MoCo-CXR, which are horizontal flip and random rotations for data augmentation.

5.7 Supplementary Details for MoCo-CXR Performance

Pretraining	Architecture	Fine Tuning	0.1%	1.0%	10.0%	100.0%
MoCo	ResNet18	Linear Model	0.776(0.735, 0.811)	0.838(0.797, 0.873)	0.848(0.807, 0.883)	0.850(0.808, 0.884)
ImageNet	ResNet18	Linear Model	0.683(0.644, 0.719)	0.766(0.716, 0.805)	0.799(0.750, 0.840)	0.815(0.768, 0.853)
MoCo	DenseNet121	Linear Model	0.775(0.737, 0.806)	0.833(0.790, 0.871)	0.839(0.800, 0.877)	0.845(0.803, 0.881)
ImageNet	DenseNet121	Linear Model	0.669(0.631, 0.705)	0.755(0.706, 0.804)	0.769(0.716, 0.818)	0.789(0.739, 0.836)
MoCo	ResNet18	End-to-end	0.813(0.779, 0.842)	0.885(0.856, 0.909)	0.915(0.885, 0.937)	0.942(0.921, 0.962)
ImageNet	ResNet18	End-to-end	0.775(0.737, 0.807)	0.858(0.828, 0.886)	0.898(0.872, 0.921)	0.942(0.920, 0.961)
MoCo	DenseNet121	End-to-end	0.795(0.758, 0.828)	0.877(0.849, 0.901)	0.920(0.896, 0.941)	0.953(0.935, 0.969)
ImageNet	DenseNet121	End-to-end	0.747(0.712, 0.779)	0.857(0.821, 0.886)	0.908(0.884, 0.929)	0.949(0.928, 0.967)

Table 5.3: Table corresponding to Main Figure 5.3 and Figure 5.4. AUC of models trained to detect pleural effusion on the CheXpert dataset.

Pretraining	Architecture	Fine Tuning	6.25%	25.0%	100.0%
MoCo	ResNet18	Linear Model	0.902(0.848, 0.945)	0.928(0.874, 0.966)	0.944(0.907, 0.972)
ImageNet	ResNet18	Linear Model	0.852(0.791, 0.901)	0.903(0.845, 0.952)	0.927(0.873, 0.971)
MoCo	DenseNet121	Linear Model	0.899(0.850, 0.939)	0.927(0.881, 0.963)	0.943(0.904, 0.975)
ImageNet	DenseNet121	Linear Model	0.872(0.817, 0.920)	0.909(0.852, 0.951)	0.931(0.886, 0.968)
MoCo	ResNet18	End-to-end	0.911(0.867, 0.943)	0.953(0.919, 0.978)	0.974(0.945, 0.993)
ImageNet	ResNet18	End-to-end	0.911(0.863, 0.952)	0.948(0.912, 0.978)	0.970(0.943, 0.990)
MoCo	DenseNet121	End-to-end	0.909(0.866, 0.947)	0.950(0.911, 0.980)	0.984(0.960, 0.999)
ImageNet	DenseNet121	End-to-end	0.898(0.848, 0.935)	0.944(0.904, 0.980)	0.971(0.946, 0.991)

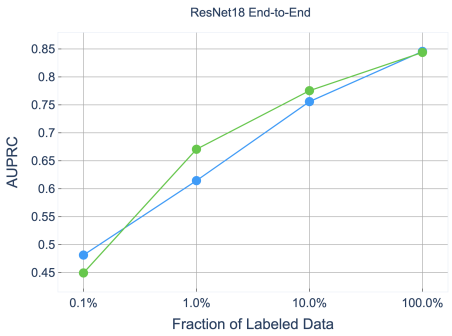
Table 5.4: Table corresponding to Main Figure 5.5. AUC of models trained to detect tuberculosis on the Shenzhen dataset.

Architecture	MoCo-CXR-pretrained	ImageNet-pretrained	6.25%	25.0%	100%
ResNet18	End-to-End	End-to-End	0.001(-0.022, 0.027)	0.005(-0.012, 0.027)	0.003(-0.014, 0.020)
ResNet18	Linear Model	Linear Model	0.054(0.024, 0.086)	0.026(-0.001, 0.056)	0.018(-0.011, 0.053)
ResNet18	Linear Model	End-to-End	-0.007(-0.029, 0.015)	-0.020(-0.040, -0.003)	-0.026(-0.052, -0.005)
DenseNet121	End-to-End	End-to-End	0.011(-0.006, 0.028)	0.006(-0.010, 0.023)	0.013(-0.003, 0.033)
DenseNet121	Linear Model	Linear Model	0.024(-0.001, 0.050)	0.016(-0.011, 0.043)	0.013(-0.014, 0.041)
DenseNet121	Linear Model	End-to-End	-0.001(-0.023, 0.019)	-0.016(-0.035, 0.001)	-0.027(-0.053, -0.003)

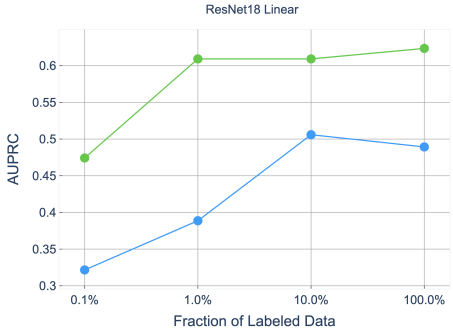
Table 5.5: AUC improvements achieved by MoCo-CXR-pretrained models against ImageNet-pretrained models on the Shenzhen tuberculosis task.

AUPRC Performance of MoCo-CXR For all following figures, the green line represents MoCo-CXR pretrained models, whereas the blue line represents baseline Imagenet-pretrained models. Figures on the left compare models trained end-to-end and figures on the right compare performance of the corresponding linear models.

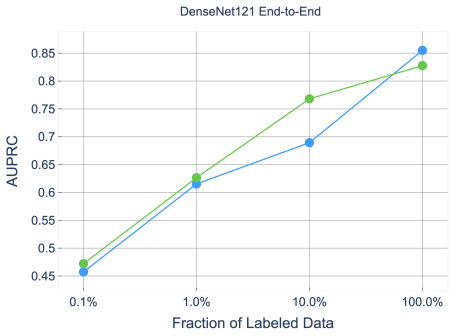
The first four images are AUPRC performance for the Pleural Effusion task from the CheXpert dataset. Here, we again observed that linear models based on MoCo-CXR pretrained representations consistently outperform those based on ImageNet-pretrained representations. In contrast, performance gains for end-to-end trained models are less significant.



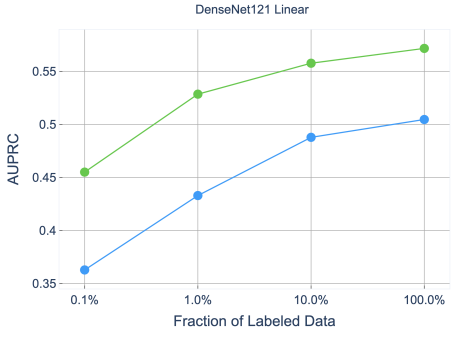
ResNet18 trained end-to-end



ResNet18 with linear fine-tuning



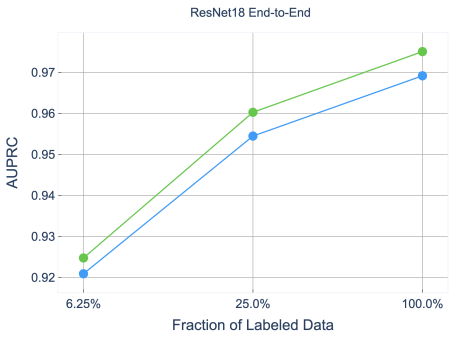
DenseNet121 trained end-to-end



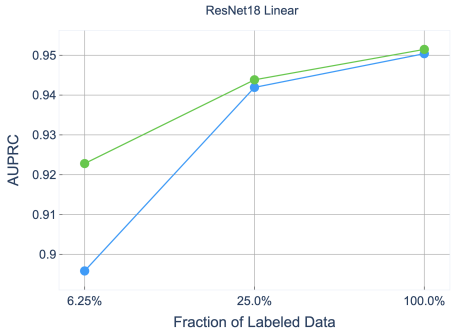
DenseNet121 with linear fine-tuning

Figure 5.8: Comparison of AUPRC performances for ResNet18-based and DenseNet121-based models on the Pleural Effusion task from the CheXpert dataset.

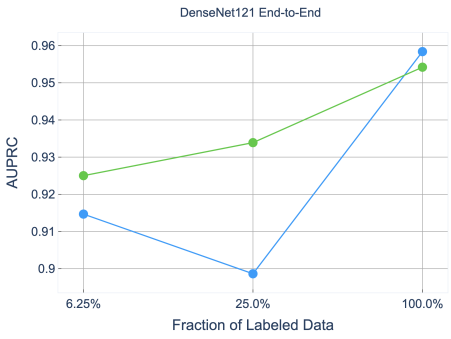
AURPC performance gains as visualized below for the Tuberculosis task on the Shenzhen dataset is roughly consistent with those observed on CheXpert and in line with AUROC performance discussed in Section 5.4.2.



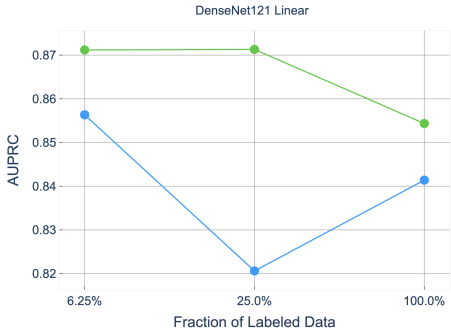
ResNet18 trained end-to-end



ResNet18 with linear fine-tuning



DenseNet trained end-to-end



DenseNet with linear fine-tuning

Figure 5.9: Comparison of AUPRC performances for ResNet18-based and DenseNet121-based models on the tuberculosis task from the Shenzhen dataset.

5.8 MoCo-CXR Performance on Other CheXpert Tasks

To reinforce our results for pleural effusion as presented in the main paper, we also conducted follow-up experiments evaluating MoCo-CXR performance on the CheXpert competition task pathologies. These experiments were done exclusively with the ResNet18 backbone.

MoCo-CXR pretrained models outperforms ImageNet-pretrained models in linear evaluation on all five CheXpert competition tasks (Table 5.6). For these tasks, the most significant performance gain is observed at 0.1% label fraction and gains diminish with increasing label fraction, the same behavior as observed for the Pleural Effusion task. We also observe statistically significant performance gains with MoCo-CXR in end-to-end fine-tuning on all the CheXpert competition tasks (Table 5.7). This was observed at most label fractions for each pathology.

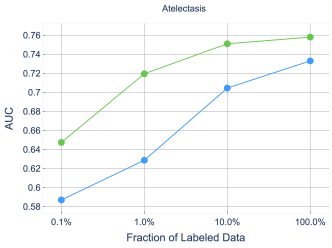


Figure 5.10: Atelectasis

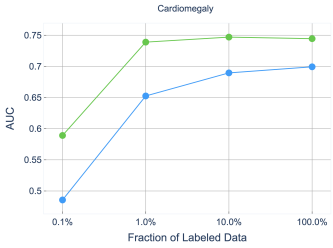


Figure 5.11: Cardiomegaly

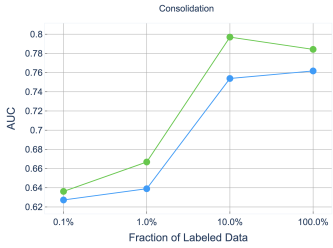


Figure 5.12: Consolidation

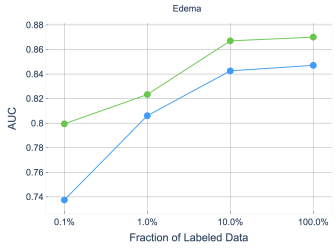


Figure 5.13: Edema

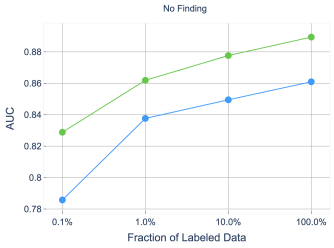


Figure 5.14: No Finding

	Baseline	MoCo-CXR	Improvement	Label Fraction
Atelectasis	0.602(0.571, 0.635)	0.630(0.592, 0.668)	0.030(-0.005, 0.064)	0.1%
	0.612(0.574, 0.650)	0.671(0.623, 0.714)	0.060(0.022, 0.098)	1%
	0.703(0.654, 0.748)	0.751(0.708, 0.796)	0.048(0.011, 0.086)	10%
	0.732(0.685, 0.778)	0.758(0.712, 0.802)	0.025(-0.015, 0.065)	100%
Cardiomegaly	0.485(0.454, 0.516)	0.640(0.605, 0.671)	0.156(0.122, 0.193)	0.1%
	0.634(0.587, 0.674)	0.735(0.690, 0.779)	0.100(0.058, 0.139)	1%
	0.685(0.638, 0.733)	0.745(0.698, 0.788)	0.060(0.019, 0.101)	10%
	0.700(0.652, 0.747)	0.737(0.694, 0.780)	0.038(-0.004, 0.082)	100%
Consolidation	0.633(0.566, 0.694)	0.644(0.554, 0.734)	0.008(-0.085, 0.093)	0.1%
	0.615(0.548, 0.678)	0.699(0.617, 0.771)	0.083(-0.007, 0.166)	1%
	0.752(0.674, 0.821)	0.794(0.699, 0.876)	0.041(-0.059, 0.127)	10%
	0.749(0.665, 0.830)	0.771(0.665, 0.859)	0.019(-0.093, 0.119)	100%
Edema	0.725(0.682, 0.766)	0.781(0.743, 0.814)	0.055(0.016, 0.092)	0.1%
	0.810(0.764, 0.849)	0.847(0.809, 0.883)	0.038(0.005, 0.071)	1%
	0.844(0.801, 0.884)	0.870(0.833, 0.900)	0.024(-0.009, 0.059)	10%
	0.846(0.805, 0.886)	0.867(0.830, 0.898)	0.020(-0.015, 0.052)	100%
Pleural Effusion	0.683(0.644, 0.719)	0.776(0.735, 0.811)	0.096(0.061, 0.130)	0.1%
	0.766(0.716, 0.805)	0.838(0.797, 0.873)	0.070(0.029, 0.112)	1%
	0.799(0.750, 0.840)	0.848(0.807, 0.883)	0.049(0.005, 0.094)	10%
	0.815(0.768, 0.853)	0.850(0.808, 0.884)	0.034(-0.009, 0.080)	100%

Table 5.6: AUC improvements achieved by MoCo-CXR-pretrained linear models against ImageNet-pretrained linear models on CheXpert competition tasks

	Baseline	MoCo-CXR	Improvement	Label Fraction
Atelectasis	0.611(0.582, 0.641)	0.673(0.637, 0.706)	0.061(0.034, 0.089)	0.1%
	0.685(0.651, 0.719)	0.730(0.693, 0.762)	0.043(0.015, 0.069)	1%
	0.732(0.694, 0.770)	0.787(0.750, 0.823)	0.054(0.030, 0.076)	10%
	0.842(0.807, 0.878)	0.821(0.781, 0.860)	-0.021(-0.037, -0.004)	100%
Cardiomegaly	0.593(0.561, 0.624)	0.663(0.628, 0.695)	0.069(0.041, 0.097)	0.1%
	0.728(0.690, 0.765)	0.811(0.777, 0.840)	0.083(0.059, 0.108)	1%
	0.808(0.769, 0.844)	0.820(0.779, 0.856)	0.011(-0.01, 0.029)	10%
	0.857(0.822, 0.890)	0.858(0.824, 0.892)	0.001(-0.014, 0.016)	100%
Consolidation	0.618(0.554, 0.680)	0.646(0.574, 0.713)	0.028(-0.015, 0.072)	0.1%
	0.673(0.614, 0.727)	0.733(0.663, 0.792)	0.061(0.006, 0.120)	1%
	0.809(0.763, 0.853)	0.852(0.790, 0.903)	0.042(0.009, 0.070)	10%
	0.888(0.847, 0.927)	0.904(0.858, 0.939)	0.016(-0.011, 0.046)	100%
Edema	0.771(0.737, 0.804)	0.786(0.752, 0.820)	0.015(-0.014, 0.042)	0.1%
	0.846(0.810, 0.880)	0.850(0.815, 0.882)	0.004(-0.017, 0.024)	1%
	0.865(0.829, 0.898)	0.877(0.840, 0.908)	0.011(-0.008, 0.030)	10%
	0.907(0.876, 0.935)	0.894(0.860, 0.923)	-0.013(-0.027, 0.001)	100%
Pleural Effusion	0.775(0.737, 0.807)	0.813(0.779, 0.842)	0.037(0.015, 0.062)	0.1%
	0.858(0.856, 0.886)	0.885(0.856, 0.909)	0.027(0.006, 0.047)	1%
	0.898(0.872, 0.921)	0.915(0.885, 0.937)	0.017(0.003, 0.031)	10%
	0.942(0.920, 0.961)	0.942(0.921, 0.921)	0.000(-0.009 0.009)	100%

Table 5.7: AUC improvements achieved by MoCo-CXR-pretrained end-to-end models against ImageNet-pretrained end-to-end models on CheXpert competition tasks

Chapter 6

Leveraging Patient Metadata For Contrastive Learning

As we saw in the previous chapter, self-supervised contrastive learning between pairs of multiple views of the same image can successfully leverage unlabeled data to produce meaningful visual representations for both natural and medical images. However, we have not looked at how we can select pairs for medical images when available patient metadata can be leveraged to improve representations. In this work, we develop a method to select positive pairs coming from views of possibly different images through the use of patient metadata. We compare strategies for selecting positive pairs for chest X-ray interpretation including requiring them to be from the same patient, imaging study or laterality. We evaluate downstream task performance by fine-tuning the linear layer on 1% of the labeled dataset for pleural effusion classification. Our best performing positive pair selection strategy, which involves using images from the same patient from the same study across all lateralities, achieves a performance increase of 3.4% and 14.4% in mean AUC from both a previous contrastive method and ImageNet pretrained baseline respectively. Our controlled experiments show that the keys to improving downstream performance on disease classification are (1) using patient metadata to appropriately create positive pairs from different images with the same underlying pathologies, and (2) maximizing the number of different images used in query pairing. In addition, we explore leveraging patient metadata to select hard negative pairs for contrastive learning, but do not find improvement over baselines that do not use metadata. Our method is broadly applicable to medical image interpretation and allows flexibility for incorporating medical insights in choosing pairs for contrastive learning.

This chapter is based on [237].

6.1 Introduction

Self-supervised contrastive learning has recently made significant strides in enabling the learning of meaningful visual representations through unlabeled data [246, 100, 43, 42]. In terms of medical imaging, previous work has found performance improvement when applying contrastive learning on chest X-ray interpretation [214, 217, 11], dermatology classification [11] and MRI segmentation [33]. Despite the early success of these applications, there is only limited work on determining how to improve upon standard contrastive algorithms using medical information [214, 33, 264, 119].

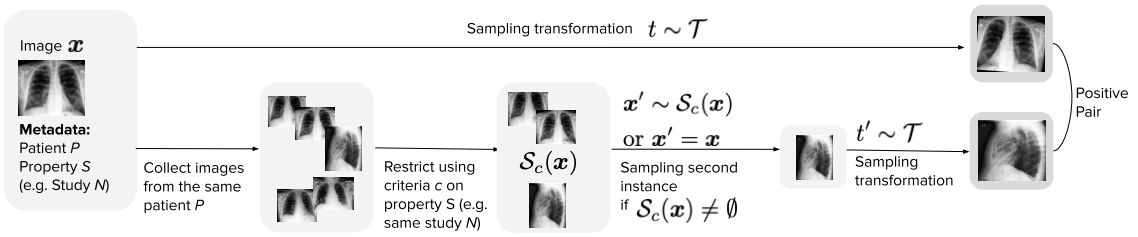
In contrastive learning, the selection of pairs controls the information contained in learned representations, as the loss function dictates that representations of positive pairs are pulled together while those of negative pairs are pushed apart [159]. For natural images where there are no other type of annotations, positive pairs are created using different augmented views of the same image while negative pairs are views of different images [42]. In that setting, [227] argue that good positive pairs are those that contain minimal mutual information apart from common downstream task information, while [224] train a generative model which learns to produce multiple positive views from a single input. However, previous contrastive learning studies on medical imaging have not systematically investigated how to leverage patient metadata available in medical imaging datasets to select positive pairs that go beyond crops of the same image while containing common downstream information.

In this work, we propose a method to treat different images that share common properties found in patient metadata as positive pairs in the context of contrastive learning. We demonstrate the application of this method to a chest X-ray interpretation task. Similar to the concurrent work by [11], we experiment with requiring positive pairs to come from the same patient as these images likely share highly similar visual features. However, our method incorporates these positive pairs with possibly different images directly as part of the view generation scheme in a single contrastive pretraining stage, as opposed to [11], which adds a second pretraining stage where a positive pair must be formed by two distinct images. Further, we go beyond the simple strategy of forming positive pair using any two data points coming from the same patient in [11, 119] and experiment with other metadata such as study number and laterality. Although study number has also been leveraged successfully in [217] to create a sequence of pretrained embeddings representating patient disease progression, our work differs in that we use this information specifically to choose positive pairs during the contrastive pretraining stage.

We conduct MoCo-pretraining [43] using these different criteria and evaluate the quality of pretrained representations by freezing the base model and fine-tuning a linear layer using 1% fraction of the labeled dataset for the task of pleural effusion. Our contributions are:

1. We develop a method, *MedAug*, to use patient metadata to select positive pairs in contrastive learning, and apply this method to chest X-rays for the downstream task of pleural effusion.

Figure 6.1: Selecting positive pairs for contrastive learning with patient metadata



2. Our best pretrained representation achieves a performance increase of 3.4% and 14.4% in mean AUC compared to [214] and the ImageNet pretrained baseline respectively, showing that using patient metadata to select positive pairs from multiple images can significantly improve representations.

3. We perform comparative empirical analysis to show that (1) using positive pairs that share underlying pathologies improves pretrained representations, and (2) increasing the number of distinct images selected to form positive pairs per image query improves the quality of pretrained representations.

4. We perform an exploratory analysis on strategies to select negative pairs using patient metadata, and do not find improvement over the default strategy that does not use metadata.

6.2 Methods

6.2.1 Chest X-ray dataset and task

We use CheXpert, a large collection of de-identified chest X-ray images. The dataset consists of 224,316 images from 65,240 patients labeled for the presence or absence of 14 radiological observations. Following [214], we perform experiments for the pleural effusion classification to provide a head-to-head comparison.

6.2.2 Selecting positive pairs for contrastive learning with patient metadata

Given an input image $\boldsymbol{x}$, encoder g, and a set of augmentations $\mathcal{T}$, most contrastive learning algorithms involve minimizing the InfoNCE loss

$$\mathcal{L}(\boldsymbol{x}) = -\log \frac{\exp[g(\tilde{\boldsymbol{x}}_1) \cdot g(\tilde{\boldsymbol{x}}_2)]}{\exp[g(\tilde{\boldsymbol{x}}_1) \cdot g(\tilde{\boldsymbol{x}}_2)] + \sum_{i=1}^{K} \exp[g(\tilde{\boldsymbol{x}}_1)) \cdot g(\boldsymbol{z}_i))]}. \tag{6.1}$$

Here, the positive pair $(\tilde{\boldsymbol{x}}_1 = t_1(\boldsymbol{x}), \tilde{\boldsymbol{x}}_2 = t_2(\boldsymbol{x}))$ with $t_1, t_2 \in \mathcal{T}$ are augmentations of the input image $\boldsymbol{x}$, while the negative pairs $(\tilde{\boldsymbol{x}}_1, \boldsymbol{z}_i)$, $1 \leq i \leq K$ are pairs of augmentations of different images, with $\boldsymbol{z}_i$ coming from either a queue in the case of MoCo or the minibatch in the case of SimCLR. Recognizing that many augmentation strategies available for natural images are not applicable to medical images, [214] restrict $\mathcal{T}$ to be the set of simple augmentations such as horizontal flipping and random rotation between -10 to 10 degrees. As a result, their method can be thought of as instance discrimination, as $\tilde{\boldsymbol{x}}_1$ and $\tilde{\boldsymbol{x}}_2$ must come from the same image input.

In this work, we propose MedAug, a method to use multiple images as a way to increase the number of positive pair choices. Beyond the disease labels, we can use patient metadata such as patient number, study number, laterality, patient historical record etc. to create appropriate positive pairs. Formally, we can use patient metadata to obtain an enhanced augmentation set dependent on $\boldsymbol{x}$ as follows

$$\mathcal{T}_{\text{enhanced}}(\boldsymbol{x}) = \begin{cases} \{t_i(\boldsymbol{x}') | t_i \in \mathcal{T}, \boldsymbol{x}' \in \mathcal{S}_c(\boldsymbol{x})\} & \text{if } \mathcal{S}_c(\boldsymbol{x}) \neq \emptyset \\ \mathcal{T}(\boldsymbol{x}) & \text{otherwise} \end{cases} \tag{6.2}$$

where $\mathcal{S}_c(\boldsymbol{x})$ is the set of all images satisfying some predefined criteria c in relation to the properties of $\boldsymbol{x}$. The criteria for using the metadata could be informed by clinical insights about the downstream task of interest.

We apply this method on chest X-ray interpretation and pretrain ResNet-18 models using MoCo v2 with hyperparameter choices as in [214]. Since the downstream task is disease classification, we experiment with using $\mathcal{S}_{\text{same patient}}(\boldsymbol{x})$ since images from the same patient are likely to share high amount of visual features. We also experiment with further applying criteria on study numbers as well as laterality. An example application of the method is illustrated in Figure 6.1.

6.2.3 Fine-tuning and evaluation

We evaluate the pretrained representations by (1) training a linear classifier on outputs of the frozen encoder using labeled data and (2) end-to-end fine-tuning. Pretrained checkpoints are selected with k-nearest neighbors algorithm based on Faiss similarity search and clustering library [112]. To simulate label scarcity encountered in medical contexts, we fine-tune using only 1% of the labeled dataset. The fine-tuning experiments are repeated on 5 randomly drawn 1% splits from the labeled dataset to provide an understanding of the model's performance variance. We report the mean AUC and standard deviation over these five 1% fine-tuning splits. Following [214] and [106], we use a

learning rate of 3×10^{-5}, batch size of 16 and 95 epochs for training.

6.3 Experiments

6.3.1 Positive pair selection

Our formulation of using any set of images $\mathcal{S}_c(\boldsymbol{x})$ from the same patient to enhance the set of augmentations for contrastive learning provides the flexibility of experimenting with different criteria c for constraining $\mathcal{S}_c(\boldsymbol{x})$. We experiment with limiting $\mathcal{S}_c(\boldsymbol{x})$ using properties found in the metadata of the query $\boldsymbol{x}$. In particular, we focus on two properties:

Study number. The study number of an image associated with a particular patient reflects the session in which the image was taken. We experiment with three different criteria on study number:

1. All studies: no restriction on $\mathcal{S}_{\text{all studies}}(\boldsymbol{x})$ is dependent on the study number of $\boldsymbol{x}$
2. Same study: only images from the same study with $\boldsymbol{x}$ belong to $\mathcal{S}_{\text{same study}}(\boldsymbol{x})$
3. Distinct studies: only images with different study number from $\boldsymbol{x}$ belong to $\mathcal{S}_{\text{distinct studies}}(\boldsymbol{x})$

Laterality. Chest X-rays can be of either frontal (AP/PA) view or lateral view.

1. All lateralities: no restriction on $\mathcal{S}_{\text{all lateralities}}(\boldsymbol{x})$ is dependent on the laterality of $\boldsymbol{x}$
2. Same laterality: only images from the same laterality with $\boldsymbol{x}$ belong to $\mathcal{S}_{\text{same laterality}}(\boldsymbol{x})$
3. Distinct lateralities: only images with a different laterality from that of $\boldsymbol{x}$ belongs to $\mathcal{S}_{\text{distinct lateralities}}(\boldsymbol{x})$

Table 6.1: Except for criteria c that involve images from different studies, using images from the same patient to select positive pairs result in improved AUC in downstream pleural effusion classification.

Baseline models	**Linear**	**End-to-end**
ImageNet baseline	0.766 ± 0.009	0.858 ± 0.011
MoCo v2 baseline [214]	0.847 ± 0.007	0.881 ± 0.017
MoCo v2 baseline with random crop scale	0.864 ± 0.005	0.890 ± 0.026
Criteria c for creating $\mathcal{S}_c(\boldsymbol{x})$	**Linear**	**End-to-end**
Same patient, same study, same laterality	0.862 ± 0.004	0.894 ± 0.013
Same patient, same study, distinct lateralities	0.865 ± 0.008	0.897 ± 0.008
Same patient, same study	0.876 ± 0.013	0.902 ± 0.007
Same patient, all studies	0.859 ± 0.006	0.877 ± 0.012
Same patient, distinct studies	0.848 ± 0.007	0.874 ± 0.013
Same patient, same study with random crop scale	$\mathbf{0.883 \pm 0.005}$	$\mathbf{0.906 \pm 0.015}$

Results. We report the results of experiments using these criteria in Table 6.1. Except from when $\mathcal{S}_c(\boldsymbol{x})$ includes images with different study numbers from $\boldsymbol{x}$, where there is a drop in performance, we see consistent large improvement from the baseline in [214]. The best result is obtained when using $\mathcal{S}_{\text{same study, all lateralities}}(\boldsymbol{x})$, the set of images from the same patient and same study as that of $\boldsymbol{x}$, regardless of laterality. Incorporating this augmentation strategy while holding other settings from [214] constant results in respective gains of 0.029 (3.4%) and 0.021 (2.4%) in AUC for the linear model and end-to-end model. We also experiment with including random crop augmentation from MoCo v2 [43], where the scaling is modified to be $[0.95, 1.0]$ in order to avoid cropping out areas of interest in the lungs. Adding this augmentation to the same patient, same study strategy, we obtain our best pretrained model, which achieves a linear fine-tuning AUC of 0.883 and an end-to-end fine-tuning AUC of 0.906 on the test set, significantly outperforming previous baselines.

6.3.2 Comparative Empirical Analysis

We perform comparative analysis to understand how different criteria on patient metadata affect downstream performance results seen in Table 6.1.

All studies v.s. same study

We hypothesize that the drop in transfer performance when moving from using images with the same study number to using images regardless of study number is because $\mathcal{S}_{\text{all studies}}(\boldsymbol{x})$ may contain images of a different disease pathology than that seen in $\boldsymbol{x}$. As a result, the model is asked to push the representation of a diseased image close to the representation of a non-diseased image, causing poor downstream performance. To test this hypothesis, we carry out an oracle experiment with $\mathcal{S}_{\text{all studies, same label}}(\boldsymbol{x})$, the set of images from the same patient and with the same downstream label as that of $\boldsymbol{x}$, regardless of study number.

Results. Table 6.2 shows that the model pretrained with $\mathcal{S}_{\text{all studies, same label}}(\boldsymbol{x})$ achieves a respective improvement of 0.034 and 0.022 in AUC over $\mathcal{S}_{\text{all studies}}(\boldsymbol{x})$ strategy for the linear model and end-to-end model. This experiment supports our hypothesis that positive pairs from images with different downstream labels hurt performance.

Table 6.2: Experiment with and without using downstream labels shows that positive pairs with different labels hurt downstream classification performance.

Criteria c for creating $\mathcal{S}_c(\boldsymbol{x})$	**Linear**	**End-to-end**
Same patient, all studies	0.859 ± 0.006	0.877 ± 0.012
Same patient, all studies, same disease label as $\boldsymbol{x}$	$\mathbf{0.893 \pm 0.009}$	$\mathbf{0.899 \pm 0.010}$

All studies v.s. distinct studies

There is a further performance drop when moving from using images across all studies of the same patient to images with a different study number from the current query image (Table 6.1). This finding may also support our hypothesis because there is a larger proportion of positive pairs of different disease pathologies in pairs of images from strictly different studies (see 6.5.1). To make sure this result holds independent of the different number of available images to form pair per query, we repeated these experiments while forcing $|\mathcal{S}_{\text{same study, all lateralities}}(\boldsymbol{x})| = |\mathcal{S}_{\text{same study, same laterality}}(\boldsymbol{x})|$ via random subset pre-selection. Further, we only use distinct images as a pair, i.e. skipping any $\boldsymbol{x}$ with $\mathcal{S}_c(\boldsymbol{x}) = \emptyset$ in (6.2) in order to remove any possible contribution from positive pairs formed from the same image.

Results. Table 6.3 shows the same patient, all studies strategy (AUC = 0.848) outperforms the same patient, distinct studies strategy (AUC = 0.792) even when the size of $\mathcal{S}_c(\boldsymbol{x})$ is controlled. This supports the hypothesis that a higher proportion of positive pairs with different disease pathologies hurts downstream task performance.

Table 6.3: Experiments where we force positive pairs to come from different images and control the size of $\mathcal{S}_c(\boldsymbol{x})$ shows that higher proportion of pairs with different downstream labels contribute to lower downstream performance.

Criteria c for creating $\mathcal{S}(\boldsymbol{x})$	**Linear**	**End-to-end**
Same patient, distinct studies	0.792 ± 0.007	0.841 ± 0.013
Same patient, all studies (size controlled)	$\mathbf{0.848 \pm 0.009}$	$\mathbf{0.863 \pm 0.010}$

All lateralities v.s. distinct lateralities v.s. same laterality

First, we hypothesize that the drop in performance from the all lateralities to the same laterality strategy could be due to $\mathcal{S}_{\text{same study, same laterality}}(\boldsymbol{x})$ having smaller size. To test this, we carry out an experiment in which $\mathcal{S}_{\text{same study, all lateralities}}(\boldsymbol{x})$ is constrained by $|\mathcal{S}_{\text{same study, same laterality}}(\boldsymbol{x})|$, the number of images from the same study and has the same laterality as $\boldsymbol{x}$.

Table 6.4: Experiments with all lateralities where we control the size of $S_{\text{same study, all lateralities}}$ show that the size of $\mathcal{S}_c(\boldsymbol{x})$ affects downstream performance.

Criteria c for creating $\mathcal{S}_c(\boldsymbol{x})$	**Linear**	**End-to-end**
Same patient, same study, same laterality	0.862 ± 0.004	0.894 ± 0.013
Same patient, same study, all lateralities (size controlled)	0.860 ± 0.004	0.899 ± 0.011
Same patient, same study, all lateralities (no control)	$\mathbf{0.876 \pm 0.013}$	$\mathbf{0.902 \pm 0.007}$

Our second hypothesis is that mutual information in images with different lateralities is lower, which benefits retaining only information important to the downstream task, as shown in [227]. We

Table 6.5: Experiments to compare same v.s. distinct lateralities with size restriction on $\mathcal{S}_c(\boldsymbol{x})$ shows no significant difference.

Criteria c for creating $\mathcal{S}_c(\boldsymbol{x})$	**Linear**	**End-to-end**
Same patient, same study, same laterality	0.856 ± 0.016	0.878 ± 0.015
Same patient, same study, distinct lateralities	$\mathbf{0.866 \pm 0.015}$	$\mathbf{0.882 \pm 0.017}$

test this by training two models on images that include at least one counterpart from the other laterality. We pretrain one model with $\mathcal{S}_{\text{same study, same laterality}}(\boldsymbol{x})$ containing only images with the same laterality as $\boldsymbol{x}$, and the other model with $\mathcal{S}_{\text{same study, distinct lateralities}}(\boldsymbol{x})$ containing only images with different laterality from $\boldsymbol{x}$. To prevent the effect of different sizes of $\mathcal{S}_c(\boldsymbol{x})$, we force that $|\mathcal{S}_{\text{same study, same laterality}}(\boldsymbol{x})| = |\mathcal{S}_{\text{same study, distinct lateralities}}(\boldsymbol{x})|$ via random subset pre-selection.

Results. Table 6.4 shows that once we control for the size of $\mathcal{S}_c(\boldsymbol{x})$, there is no significant difference between using images from the same laterality (AUC = 0.862) or from all lateralities (AUC = 0.860). However, the model pretraining with all images from all lateralities achieves much larger downstream AUC of 0.876. Thus, it supports our first hypothesis that the size of $\mathcal{S}_c(\boldsymbol{x})$ influences pretrained representation quality. Table 6.5 shows that once we control for the size of $\mathcal{S}_c(\boldsymbol{x})$, the model pretrained with images from different lateralities only gain 0.010 AUC in linear fine-tuning performance and a non-significant 0.004 in end-to-end performance. This experiment shows that the effect of mutual information from different lateralities on pretrained representation quality is less pronounced.

6.3.3 Negative pair selection

We explore strategies using metadata in the CheXpert dataset to define negative pairs. Similar to our method of defining positive pairs, we take advantage of metadata available in the dataset to select the negative pairs. However, unlike positive pair selection, where only a single pair is required for each image, an image has to pair with the entire queue to select negative pairs. This property makes selecting negative pairs from the same patient as done in selecting positive pairs not suitable because only a small number of images are available for a patient. We instead use a more general property – laterality – across the patients to define negative pairs to retain sufficient negative pairs in the loss function (6.1). Similarly, other metadata such as age and sex may be exploited for the same purpose.

The default negative pair selection strategy is to select all keys from the queue that are not views of the query image. However, we hypothesize that negative pairs with the same laterality are "hard" negative pairs that are more difficult to distinguish and provide more accurate pretrained representations for the downstream task. We describe our four strategies briefly as follows and in more detail in 6.5.2. Our first strategy is to only select images from the queue with the same

laterality as the query to create negative pairs. Our second strategy is to reweight the negative logits based on laterality so in effect queries with each laterality (frontal and lateral) equally contribute to the loss and the queue size remains fixed as in the original MoCo approach. Following a similar idea in [113], our third strategy is to sample a portion of negative pairs with the same laterality for each query and append them to the queue for loss computation. Our fourth strategy is to create synthetic negatives for additional hard negative pairs. Unlike [113], we do not determine hardness of negative pairs based on similarities of representations. Instead, we use existing metadata (image laterality) to approximate hardness of an negative pair. We evaluate the performance of each of these negative pair strategies combined with the positive pair strategy of "same patient, same study, all lateralities".

Results. Results are given in Table 6.6. The default negative pair selection strategy (AUC = 0.876) is not outperformed by any of the metadata-exploiting negative pair selection strategies including same laterality only (AUC = 0.872), same laterality reweighted (AUC = 0.864), same laterality appended (AUC = 0.875) and same laterality synthetic (AUC = 0.870). Thus, our exploratory analysis does not indicate sufficient evidence for performance improvement using strategies that incorporate metadata, but further experiments with other metadata sources may be required to further understand this relationship.

Table 6.6: Experiments with the default negative pair definition (different images) and various negative pair selection strategies.

Negative Pairs Strategy	**Linear**
Default	$\mathbf{0.876 \pm 0.013}$
Same Laterality only	0.872 ± 0.011
Same Laterality (reweighted)	0.864 ± 0.006
Same Laterality (appended)	0.875 ± 0.006
Same Laterality (synthetic)	0.870 ± 0.004

6.4 Discussion

We introduce MedAug, a method to use patient metadata to select positive pairs for contrastive learning, and demonstrate the utility of this method on a chest X-ray interpretation task.

Can we improve performance by leveraging metadata to choose positive pairs? Yes. Our best pretrained strategy with multiple images from the same patient and same study obtains an increase of 3.4% in linear fine-tuning AUC in comparison to the instance discrimination approach implemented in [214]. A similar result has been shown by [119] for ECG signal interpretation. [11] also found improvement in dermatology classification when applying a second contrastive pretraining stage where strictly distinct images from the same patient are selected as positive pairs.

Unlike previous work, our empirical analysis on using images from all studies and distinct studies shows that simply choosing images from the same patient may hurt downstream performance. We show that using appropriate metadata such as study number to select positive pairs that share underlying disease information is needed to obtain the best representation for the downstream task of disease classification. For future studies, it is of interest to experiment with other metadata such as age group, medical history, etc. and how they can inform on tasks other than disease classification.

Our analysis using different criteria on laterality shows that the number of images selected to form positive pairs plays an important role, while the effect of mutual information is less clear. Given time and resources, it would be informative to experiment with how the maximum number of distinct images chosen per query affect downstream performance.

Can we improve performance by leveraging metadata to choose hard negative pairs? Not necessarily. We perform an exploratory analysis of strategies to leverage patient metadata to select negative pairs, and do not find them to outperform the baseline.

In closing, our work demonstrates the potential benefits of incorporating patient metadata into self-supervised contrastive learning for medical images, and can be extended to a broader set of tasks [188, 230].

6.5 Additional Information

6.5.1 Proportions of positive pairs with different disease labels

In 6.3.2, we argue that downstream performance from the $\mathcal{S}_{\text{distinct studies}}(\boldsymbol{x})$ is lower than that of $\mathcal{S}_{\text{all studies}}(\boldsymbol{x})$ because there is a likely a higher proportion of positive pairs with different disease labels in $\mathcal{S}_{\text{distinct studies}}(\boldsymbol{x})$. Figure 6.2 shows that there is almost 9% of $\boldsymbol{x}$ where $\mathcal{S}_{\text{distinct studies}}(\boldsymbol{x})$ contains only images with a different disease label from $\boldsymbol{x}$, whereas this scenario does not appear for $\mathcal{S}_{\text{all studies}}(\boldsymbol{x})$.

6.5.2 Negative Pairs

Following the loss function in equation (1), we denote the exponential sum of the negative pairs by G

$$\mathcal{L}(\boldsymbol{x}) = -\log \frac{\exp[g(\tilde{\boldsymbol{x}}_1) \cdot g(\tilde{\boldsymbol{x}}_2)]}{\exp[g(\tilde{\boldsymbol{x}}_1) \cdot g(\tilde{\boldsymbol{x}}_2)] + G\left(\tilde{\boldsymbol{x}}_1, \boldsymbol{z}_i\right)}. \tag{6.3}$$

where

$$G\left(\tilde{\boldsymbol{x}}_1, \boldsymbol{z}_i\right) = \sum_{\boldsymbol{z}_i \in Q} \exp[g(\tilde{\boldsymbol{x}}_1)) \cdot g(\boldsymbol{z}_i))] \tag{6.4}$$

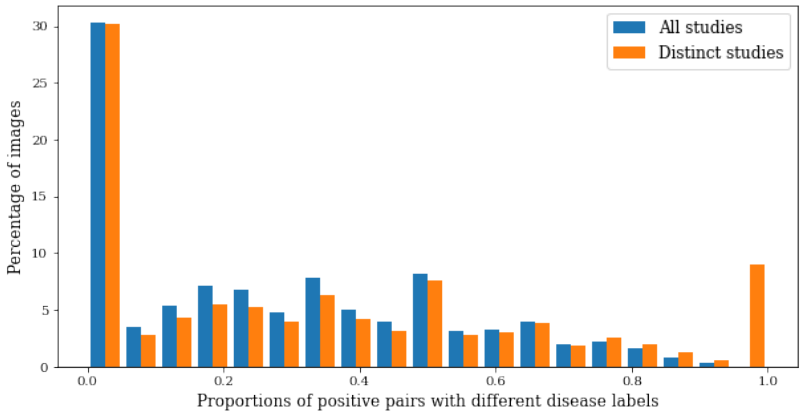


Figure 6.2: Histogram showing the distribution of the proportions of positive pairs with different disease labels in $\mathcal{S}_{\text{distinct studies}}(\boldsymbol{x})$ versus $\mathcal{S}_{\text{all studies}}(\boldsymbol{x})$.

We follow the MoCo setup and denote Q as the queue. Let $\mathcal{S}(\boldsymbol{x})$ be the set of image representations in Q that have the same laterality as $\boldsymbol{x}$. We use the symbol $\|$ to denote list concatenation. We describe each of our negative pair selection strategies as follows:

1. (Same laterality only) For each query, we select keys in the queue that have the same laterality as the query. Specifically, we replace G in equation (4) by G^l

$$G^l\left(\tilde{\boldsymbol{x}}_1, \boldsymbol{z}_i\right) = \sum_{\boldsymbol{z}_i \in \mathcal{S}(\boldsymbol{x})} \exp[g(\tilde{\boldsymbol{x}}_1)) \cdot g(\boldsymbol{z}_i))] \tag{6.5}$$

2. (Same laterality reweighted) The first strategy excluded the keys in the queue that have different laterality from the query. Here we set a target hard negative weight and reweight each *exp* term to achieve the target weight. Let

$$G^w\left(\tilde{\boldsymbol{x}}_1, \boldsymbol{z}_i\right) = \sum_{\boldsymbol{z}_i \in \mathcal{S}(\boldsymbol{x})} w_i^s \exp[g(\tilde{\boldsymbol{x}}_1)) \cdot g(\boldsymbol{z}_i))] + \sum_{\boldsymbol{z}_i \in \mathcal{S}(x)^c} w_i^d \exp[g(\tilde{\boldsymbol{x}}_1)) \cdot g(\boldsymbol{z}_i))] \tag{6.6}$$

where t is the target hard negative weight and $r = \frac{|\mathcal{S}(\boldsymbol{x})|}{|Q|}$ is the proportion of the negative keys in the queue that have the same laterality as $\boldsymbol{x}$. Then $w_i^d = \frac{1-t}{1-r}$ and $w_i^s = \frac{t}{r}$ for all i. In our experiments, we set $t = 0.1$ to allocate 90% of the weight to hard negatives. This allows us to include all negative pairs in the contrastive loss but place emphasis on hard negative pairs with the same laterality.

3. (Same laterality appended) For each query, we select a random sample of the keys that have the same laterality and append them to the existing queue

$$Q = [z_1, z_2, ..., z_K]$$

where K is the queue size. Let

$$A = \{z_{i_1}, z_{i_2}, ..., z_{i_m}\} \subset S(x) \tag{6.7}$$

be the random sample of keys with the same laterality as the query. The new queue is

$$Q^a = Q \parallel A$$

and

$$G^a(\tilde{\boldsymbol{x}}_1, \boldsymbol{z}_i) = \sum_{\boldsymbol{z}_i \in Q^a} \exp[g(\tilde{\boldsymbol{x}}_1)) \cdot g(\boldsymbol{z}_i))]$$

replaces G in equation (4).

4. (Same laterality synthetic) For each query, in addition to appending samples of the keys from $S(\boldsymbol{x})$, we use the samples to generate synthetic keys and append them to the queue. We randomly sample m pairs $(\boldsymbol{s}_i, \boldsymbol{s}_j) \in A = \{\boldsymbol{z}_{i_1}, \boldsymbol{z}_{i_2}, ..., \boldsymbol{z}_{i_m}\}$ and call this set of pairs B.

   For each pair $(\boldsymbol{s}_i, \boldsymbol{s}_j) \in B$, we uniformly sample a number u between 0 and 1 and let

$$h = u \cdot \boldsymbol{s}_i + (1 - u) \cdot \boldsymbol{s}_j$$

   A synthetic image representation is defined as the normalized vector $\frac{h}{\|h\|}$. Let H be the set of these m synthetic image representations and

$$Q^h = Q \parallel B \parallel H$$

   is the new queue. G in equation (1) is replaced by

$$G^h(\tilde{\boldsymbol{x}}_1, \boldsymbol{z}_i) = \sum_{z_i \in Q^h} \exp[g(\tilde{\boldsymbol{x}}_1)) \cdot g(\boldsymbol{z}_i))]$$

   Note that unlike [113], we only construct synthetic images once.

Chapter 7

Data Development for Domain Robustness

In previous chapters, we have looked at the development of algorithms for performing medical image interpretation tasks accurately. Now, we will look into the curation of datasets towards the deployment of these algorithms.

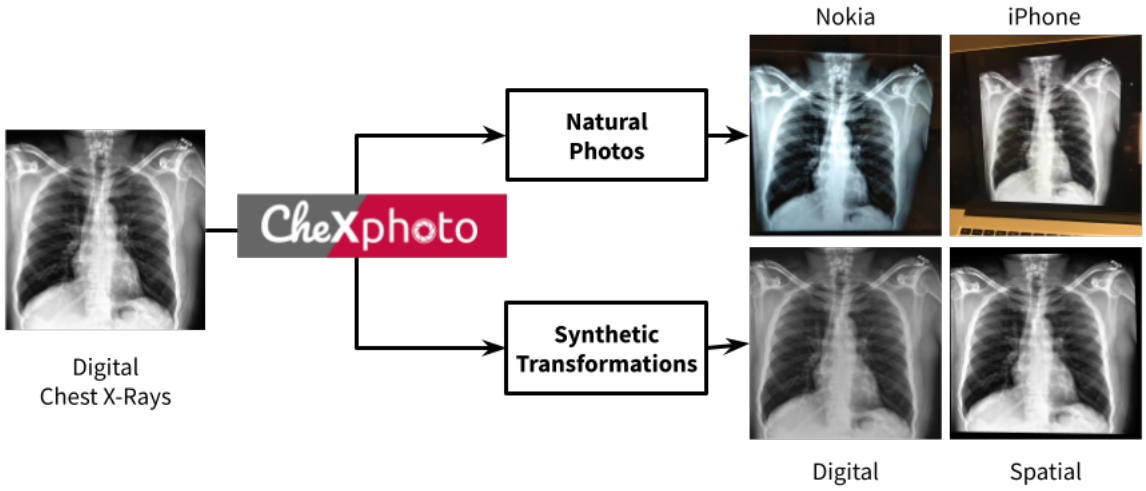
Clinical deployment of deep learning algorithms for chest x-ray interpretation requires a solution that can integrate into the vast spectrum of clinical workflows across the world. An appealing approach to scaled deployment is to leverage the ubiquity of smartphones by capturing photos of x-rays to share with clinicians using messaging services like WhatsApp. However, the application of chest x-ray algorithms to photos of chest x-rays requires reliable classification in the presence of artifacts not typically encountered in digital x-rays used to train machine learning models. We introduce CheXphoto, a dataset of smartphone photos and synthetic photographic transformations of chest x-rays sampled from the CheXpert dataset. To generate CheXphoto we (1) automatically and manually captured photos of digital x-rays under different settings, and (2) generated synthetic transformations of digital x-rays targeted to make them look like photos of digital x-rays and x-ray films. We release this dataset as a resource for testing and improving the robustness of deep learning algorithms for automated chest x-ray interpretation on smartphone photos of chest x-rays.

This chapter is based on [168].

7.1 Background & Summary

One significant obstacle to the adoption of chest x-ray algorithms is that deployment requires a solution that can integrate into the vast spectrum of clinical workflows around the world. Most chest x-ray algorithms are developed and validated on digital x-rays, while the majority of developing

Figure 7.1: Overview of the CheXphoto data generation process.



regions use films [204, 7]. An appealing approach to scaled deployment is to leverage the ubiquity of existing smartphones: automated interpretation of x-ray film through cell phone photography has emerged through a "store-and-forward telemedicine" approach, in which one or more digital photos of chest films are sent as email attachments or instant messages by practitioners to obtain second opinions from specialists as part of clinical care [79, 233]. Furthermore, studies have shown that photographs of films using modern phone cameras are of equivalent diagnostic quality to the films themselves [204], indicating the feasibility of high-quality automated algorithmic interpretation of photos of x-ray films.

Automated interpretation of chest x-ray photos at the same level of performance as digital chest x-rays is challenging because photography introduces visual artifacts not commonly found in digital x-rays, such as altered viewing angles, variable ambient and background lighting conditions, glare, moiré, rotations, translations, and blur [124]. Image classification algorithms have been shown to experience a significant drop in performance when input images are perceived through a camera [124]. Although recent work has demonstrated good generalizability of deep learning algorithms trained on digital x-rays to photographs [184], interpretation performance could be improved through inclusion of x-ray photography in the training process [97, 78]. However, there are currently no large-scale public datasets of photos of chest x-rays.

To meet this need, we developed CheXphoto, a dataset of photos of chest x-rays and synthetic transformations designed to mimic the effects of photography. We believe that CheXphoto will enable researchers to improve and evaluate model performance on photos of x-rays, reducing the barrier to clinical deployment.

7.2 Methods

We introduce CheXphoto, a dataset of photos of chest x-rays and synthetic transformations designed to mimic the effects of photography. Specifically, CheXphoto includes a set of (1) *Natural Photos*: automatically and manually captured photos of x-rays under different settings, including various lighting conditions and locations, and (2) *Synthetic Transformations*: targeted transformations of digital x-rays to simulate the appearance of photos of digital x-rays and x-ray films. The x-rays used in CheXphoto are primarily sampled from CheXpert, a large dataset of 224,316 chest x-rays of 65,240 patients, with associated labels for 14 observations from radiology reports [106].

CheXphoto comprises a training set of natural photos and synthetic transformations of 10,507 x-rays from 3,000 unique patients that were sampled at random from the CheXpert training set, and validation and test sets of natural and synthetic transformations of all 234 x-rays from 200 patients and 668 x-rays from 500 patients in the CheXpert validation and test sets, respectively. In addition, the CheXphoto validation set includes 200 natural photos of physical x-ray films sampled from external data sources, intended to more closely simulate pictures taken by radiologists in developing world clinical settings. As much of the developing world performs x-ray interpretation on film, this distinct set of images enables users to perform additional validation on a novel task that may be encountered in clinical deployment.

7.2.1 Acquiring Natural Photos of Chest X-Rays

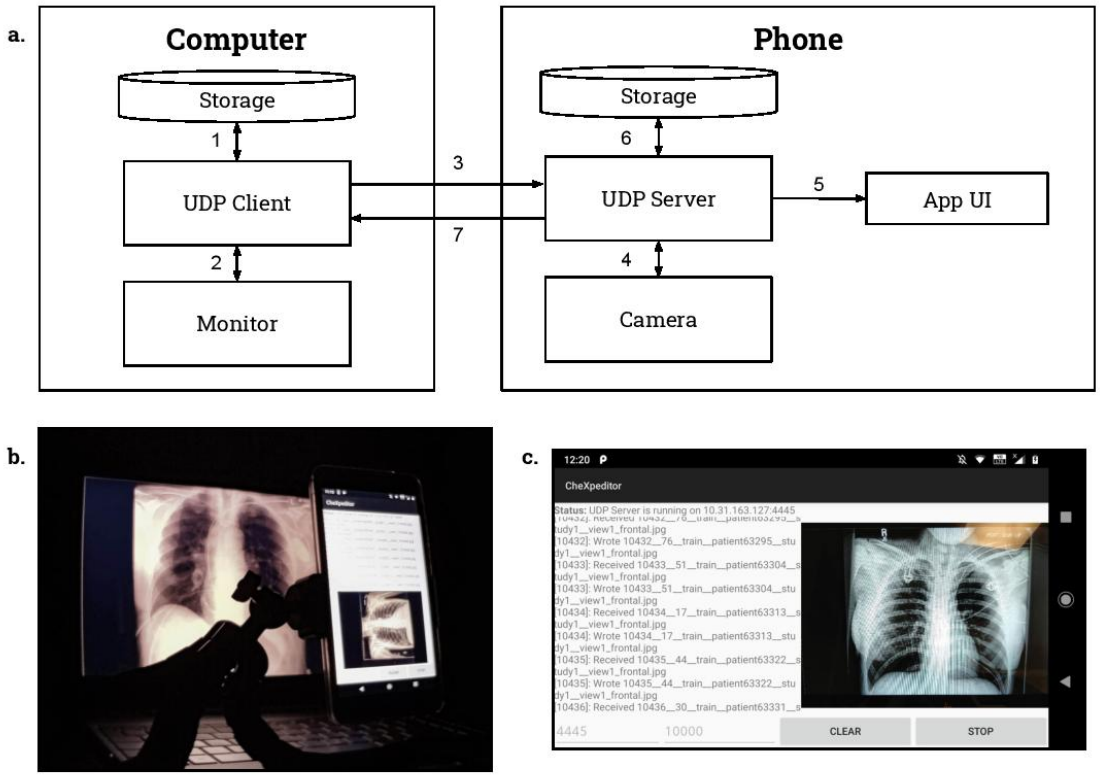
Natural photos consist of x-ray photography using cell phone cameras in various lighting conditions and environments. We developed two sets of natural photos: images captured through an automated process using a Nokia 6.1 cell phone, and images captured manually with an iPhone 8.

Automated Capture of Nokia10k dataset

We developed the 'Nokia10k' dataset by capturing 10,507 images of digital chest x-rays using a tripod-mounted Nokia 6.1 cell phone (16 megapixel camera with a Zeiss sensor) and a custom Android application termed CheXpeditor to fully automate the processes of photography and metadata management. The primary challenge in automation was synchronizing picture-taking on the phone with displaying the chest x-ray on the monitor, to maintain a 1-to-1 correspondence between each chest x-ray and its photographed image. Without a robust synchronization method, photos of chest x-rays might be skipped or duplicated, jeopardizing the data collection process. Thus, bidirectional communication over UDP was established between the phone and the computer driving the monitor to exchange image metadata, take photos, and advance the chest x-ray on the monitor.

The 10,507 x-rays in Nokia10k were indexed deterministically from 1 to N. We selected disjoint subsets of 250 to 500 consecutive indices to be photographed in constant environmental conditions.

Figure 7.2: Acquiring Natural Photos of Chest X-Rays Using Automated Capture **a**. Visual representation of the automated picture-taking process used for Nokia10k. The steps are described: 1. X-ray retrieved from computer storage, 2. X-ray displayed on monitor, 3. X-ray index and metadata sent to phone over UDP, 4. Index verified by phone, and camera triggered, 5. Application UI updated with new picture and filename, 6. Picture saved to phone storage with metadata in filename, 7. Computer notified that imaging was successful. **b**. The physical setup used for Nokia10k, set in an example environment. **c**. Phone application UI, displaying most recent picture and saved filename.



For each subset of indices, photography was conducted as follows:

1) The ith chest x-ray was retrieved from computer storage. 2) The ith chest x-ray was displayed on the monitor. 3) The image metadata m was assembled by the computer, and (i, m) were sent to the phone via UDP. 4) The phone verified that i was one greater than the previous index. If so, its camera was triggered. Else, the computer was notified of an error, and the entire picture-taking process was aborted. 5) The phone application UI, responsible for displaying status and current image, was updated to show the new picture and filename. 6) The picture was saved to phone storage with the metadata m embedded in the filename. 7) The phone notified the computer that the imaging was successful, and the entire process was repeated for the $i + 1$st chest x-ray.

Table 7.1: The distribution of labeled observations for the Nokia10k training dataset.

Pathology	Positive (%)	Uncertain (%)	Negative (%)
No Finding	972 (9.25)	0 (0.00)	9535 (90.75)
Enlarged Cardiomediastinum	518 (4.93)	600 (5.71)	9389 (89.36)
Cardiomegaly	1313 (12.50)	370 (3.52)	8824 (83.98)
Lung Opacity	5184 (49.34)	213 (2.03)	5110 (48.63)
Lung Lesion	415 (3.95)	78 (0.74)	10014 (95.31)
Edema	2553 (24.30)	634 (6.03)	7320 (69.67)
Consolidation	671 (6.39)	1315 (12.52)	8521 (81.10)
Pneumonia	263 (2.50)	885 (8.42)	9359 (89.07)
Atelectasis	1577 (15.01)	1595 (15.18)	7335 (69.81)
Pneumothorax	957 (9.11)	166 (1.58)	9384 (89.31)
Pleural Effusion	4115 (39.16)	607 (5.78)	5785 (55.06)
Pleural Other	170 (1.62)	127 (1.21)	10210 (97.17)
Fracture	391 (3.72)	31 (0.30)	10085 (95.98)
Support Devices	5591 (53.21)	48 (0.46)	4868 (46.33)

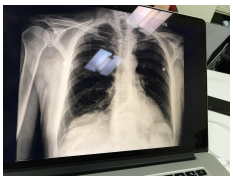
After all images for a Nokia10k subset were taken, they were exported in one batch from the phone to storage. The metadata was parsed from the image filenames and used to automatically assign the correct CheXpert label. Alterations made to the imaging conditions after every subset included moving to a different room, switching the room light on/off, opening/closing the window-blinds, rotating the phone orientation between portrait/landscape, adjusting the position of the tripod, moving the mouse cursor, varying the monitor's color temperature, and switching the monitor's screen finish between matte/glossy. In all conditions, the chest x-ray was centered in the camera view-finder and lung fields were contained within the field of view.

Manual Capture of iPhone1k dataset

We developed the 'iPhone1k dataset' by manually capturing 1,000 images of digital chest x-rays using an iPhone 8 (12 megapixel camera with a Sony Exmor RS sensor). The digital x-rays selected for the iPhone1k dataset are a randomly sampled subset of the x-rays used in the Nokia10k dataset. To produce the iPhone1k dataset, chest x-rays were displayed in full-screen on a computer monitor with 1920 x 1080 screen resolution and a black background. A physician took photos of the chest x-rays with a handheld iPhone 8 using the standard camera app. The physician was advised to change angle and distance from the computer monitor in-between each picture within constraints.

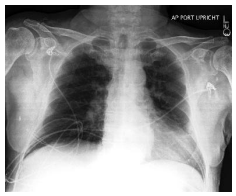
For all images, the chest x-ray was centered in the viewfinder of the camera, and the thoracic and lung fields were contained within the field of view. Conformant to radiological chest x-ray standards, both lung-apices and costodiaphragmatic recesses were included craniocaudally, and the edges of the ribcage were included laterally. Photos were captured in sets of 100 to 200 images at a time; between sets, ambient alterations were made, such as switching the room-lighting on/off, opening or closing of the window-blinds, and physically moving the computer monitor to a different location in the

Table 7.2: Natural Photos (**a-b**) and Synthetic Transformations (Digital (**c-f**) and Spatial (**g-i**)) included in CheXphoto.



(a) iPhone

(b) Nokia



(c) Brightness Up

(d) Brightness Down

(e) Contrast Up

(f) Contrast Down

(g) Glare Matte

(h) Moiré

(i) Tilt

room.

7.2.2 Generating Synthetic Photographic Transformations of Chest X-Rays

Synthetic transformations consist of automated changes to digital x-rays designed to simulate the appearance of photos of digital x-rays and x-ray films. We developed two sets of complementary synthetic transformations: digital transformations to alter contrast and brightness, and spatial transformations to add glare, moiré effects and perspective changes. To ensure that the level of these transformations did not impact the quality of the image for physician diagnosis, the images were verified by a physician. In some cases, the effects may be visually imperceptible, but may still be adversarial for classification models. For both sets, we apply the transformations to the same 10,507 digital x-rays selected for the Nokia10k dataset.

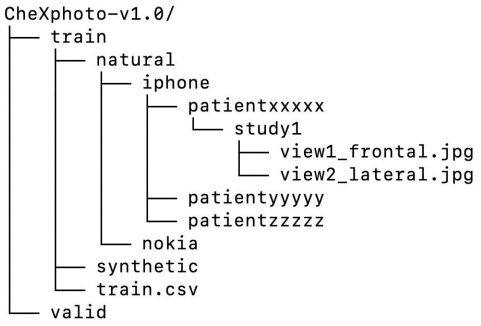
Digital transformations were produced by successive random alterations of contrast and brightness. First, the image was either enhanced for greater or lesser contrast. Setting a contrast factor of 1 for the original image, the contrast up transformation increased contrast by a factor of 1.1 and the contrast down transformation decreased contrast by a factor of 0.83. For both these factors,

random noise between -0.01 and 0.01 was applied. After the contrast modification, the brightness of the image was then transformed randomly up or down using the same numeric factors. Both the brightness and contrast transformations were applied using the Python PIL ImageEnhance class.

Table 7.3: The number of patients, studies, and images in CheXphoto.

Dataset	Patients	Studies	Images
Training			
iPhone	295	829	1,000
Nokia	3,000	8,931	10,507
Synthetic	3,000	8,931	10,507
Validation			
Natural	200	200	234
Synthetic	200	200	234
Film	200	200	200
Test	500	500	668

Figure 7.3: CheXphoto directory structure



Spatial transformations consisted of alterations to add glare, moiré effects and perspective changes. First, we applied a glare matte transformation to simulate the effect of photographing a glossy film which reflects ambient light. This was produced by randomly generating a two-dimensional multivariate normal distribution which describes the location of a circular, white mask. Second, a moiré effect was added to simulate the pattern of parallel lines seen by digital cameras when taking pictures of computer screens. The effect is produced as a result of the difference in rates of shutter speed and LCD screen sweeping refresh rate. The moiré effect was simulated by generating semi-transparent parallel lines, warping them and overlaying them onto each image. Finally, a tilt effect was added to simulate random distortions of perspective that may arise in taking a photo at angle to the screen. The tilt effect was produced by randomly scaling the x and y values of each of the corners by a factor between 0 and 0.05 towards the center. This random movement of corners is used to skew the entire photo.

Both the digital and spatial transformations are provided in CheXphoto. Each transformation may be reproduced individually using the code provided. Additional transformations - glare glossy, blur, motion, rotation, translation - are also included.

7.2.3 Validation and Test

We developed a CheXphoto validation and test set to be used for model validation and evaluation. The validation set comprises natural photos and synthetic transformations of all 234 x-rays in the CheXpert validation set, and is included in the public release, while the test set comprises natural photos of all 668 x-rays in the CheXpert test set, and is withheld for evaluation purposes.

We generated the natural photos of the validation set by manually capturing images of x-rays displayed on a 2560 $\times$ 1080 monitor using a OnePlus 6 cell phone (16 megapixel camera with a Sony

IMX 519 sensor), following a protocol that mirrored the iPhone1k dataset. Synthetic transformations of the validation images were produced using the same protocol as the synthetic training set. The test set was captured using an iPhone 8, following the same protocol as the iPhone1k dataset.

The validation set contains an additional 200 cell phone photos of x-ray films for 200 unique patients. As photos of physical x-ray films, this component of the validation set is distinct from the previously described natural and synthetic transformations of digital x-rays. Films for 119 patients were sampled from the MIMIC-CXR dataset [111], and films for 81 patients were provided by VinBrain, a subsidiary of Vingroup in Vietnam, and originally collected through joint research projects with leading lung hospitals in Vietnam. The film dataset spans 5 observation labels (atelectasis, cardiomegaly, consolidation, edema, pleural effusion), with 40 images supporting each observation. Observation labels for each image were manually verified by a physician. Images were captured using a VinSmart phone with a 12MP camera by positioning the physical x-ray film vertically on a light box in typical clinical lighting conditions, and images were automatically cropped and oriented.

7.2.4 Technical Validation

CheXphoto was developed using images and labels from the CheXpert dataset [106]. Photography of x-rays was conducted in a controlled setting in accordance with the protocols documented in the Methods section, which were developed with physician consultation. Although CheXphoto contains multiple images for some patients, either from the same or different studies, there is no patient overlap between the training, validation, and test sets. Code developed for synthetic transformations is version controlled and made available as an open source resource for review and modification. All images are uniquely identifiable by patient ID, study ID, and view, and the age and sex of each patient is provided in the data description CSV file. The original, unaltered images can be obtained from the CheXpert dataset by the unique identifiers.

The CheXphoto dataset is organized by by transformation; the training and validation sets contain directories corresponding to the method of data generation. Within each directory, the x-ray images are organized in subdirectories by a patient identifier, study ID, and one or more individual views. Images are stored as JPEG files, and image dimensions vary according to the method of generation. Each transformation set has an associated CSV file, which provides observation labels from the CheXpert dataset and relative paths to the corresponding images.

7.2.5 Data Access

The CheXphoto training and validation sets are available for download[1].

The CheXphoto test set is withheld for official evaluation of models. CheXphoto users may submit their executable code, which is then run on the private test set, preserving the integrity of the

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

test results. The testing process is enabled by CodaLab [136], an online platform for collaborative and reproducible computational research. CodaLab Worksheets exposes a simple command-line interface, which enables users to submit a Docker image, dependencies, and the necessary commands to run their models. These features allow us to run arbitrary code submissions on the withheld test set. Once a user has successfully uploaded their code to CodaLab, we will evaluate their performance on the withheld test set and share results on a live leaderboard on the web.

In addition, the code used to prepare the synthetically generated dataset is publicly available[2]. The synthetic transformations can be reproduced by running the `synthesize.py` script with the appropriate `CSV` file containing the paths to the images for which the perturbation is to be applied. Detailed instructions on flags and usage are included in the repository README.

7.3 Conclusion

We believe that CheXphoto will enable greater access to automated chest x-ray interpretation algorithms worldwide, principally in healthcare systems that are presently excluded from the benefits of digital medicine. By facilitating the development, validation, and testing of automated chest x-ray interpretation algorithms with a ubiquitous technology such as smartphone photography, CheXphoto broadens access to interpretation algorithms in underdeveloped regions, where this technology is poised to have the greatest impact on the availability and quality of healthcare.

[2] https://github.com/stanfordmlgroup/cheXphoto

Chapter 8

Data Development for Biomarker Discovery

In this chapter, we will look at the design and curation of a dataset to enable a cancer prediction task. This chapter is based on [236].

Diffuse Large B-Cell Lymphoma (DLBCL) is the most common non-Hodgkin lymphoma. Though histologically DLBCL shows varying morphologies, no morphologic features have been consistently demonstrated to correlate with prognosis. We present a morphologic analysis of histology sections from 209 DLBCL cases with associated clinical and cytogenetic data. Duplicate tissue core sections were arranged in tissue microarrays (TMAs), and replicate sections were stained with H&E and immunohistochemical stains for CD10, BCL6, MUM1, BCL2, and MYC. The TMAs are accompanied by pathologist-annotated regions-of-interest (ROIs) that identify areas of tissue representative of DLBCL. We used a deep learning model to segment all tumor nuclei in the ROIs, and computed several geometric features for each segmented nucleus. We fit a Cox proportional hazards model to demonstrate the utility of these geometric features in predicting survival outcome, and found that it achieved a C-index (95% CI) of 0.635 (0.574,0.691). Our finding suggests that geometric features computed from tumor nuclei are of prognostic importance, and should be validated in prospective studies.

8.1 Background & Summary

Diffuse Large B-Cell Lymphoma (DLBCL) is the most common type of non-Hodgkin lymphoma (NHL), accounting for over a third of cases [175] with more than 20,000 patients diagnosed annually in the United States [102]. DLBCL is fatal without treatment, however approximately 70% of patients can be cured with contemporary therapeutic regimens [131]. Treatment outcomes following

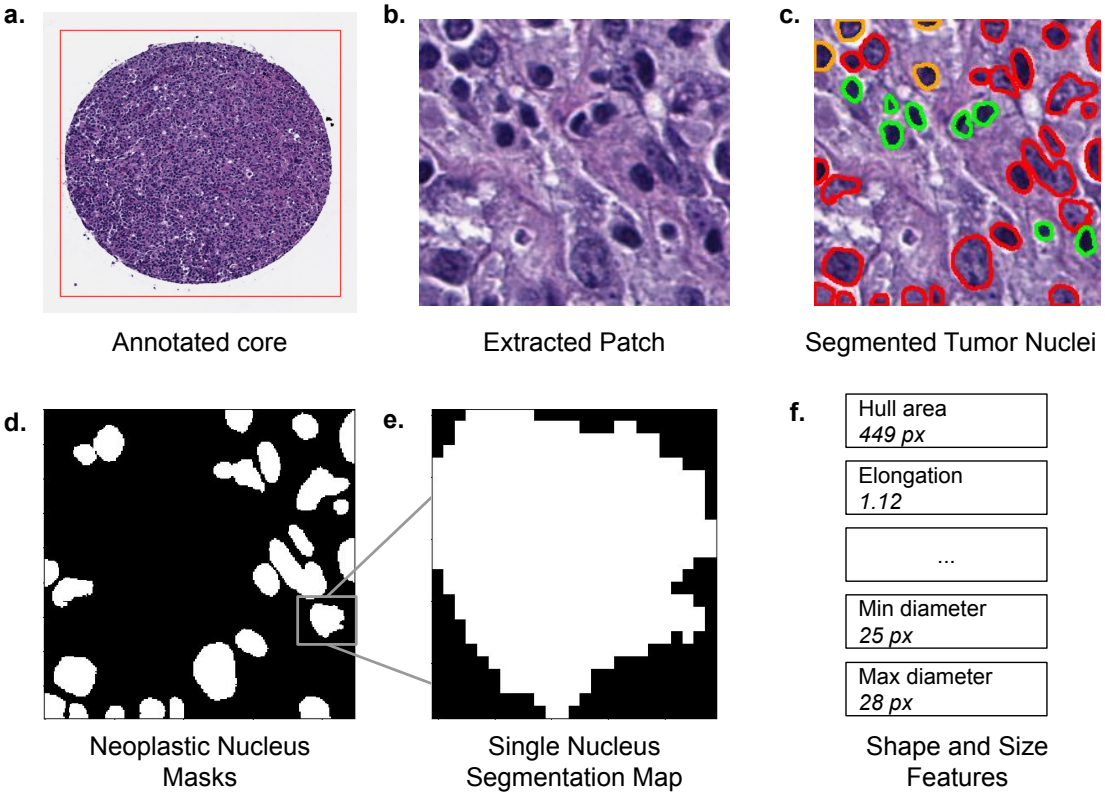


Figure 8.1: Data pipeline for a single core from an H&E stained tissue microarray (TMA). In a) the red rectangle is the pathologist-annotated ROI. In c) red corresponds to cell nuclei classified as "neoplastic" by HoVer-Net. Green corresponds to "inflammatory" and orange corresponds to "non-neoplastic epithelial".

standard R-CHOP (rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone) therapy are highly variable, and depend on a number of clinical, biologic, and genetic factors. Currently, the most effective prognostic classification is the National Comprehensive Cancer Network International Prognostic Index (NCCN-IPI), which incorporates five clinical variables including age, lactate dehydrogenase (LDH), extra-nodal sites of involvement, Ann Arbor stage, and ECOG performance status [265]. The NCCN-IPI model is widely used to risk stratify patients into good, intermediate, and poor-risk categories, however it is insufficient to guide therapeutic decision-making for individual patients.

Gene expression profiling (GEP) studies revealed distinct subtypes of DLBCL that correspond to differences in cell of origin (COO) and show different outcomes in response to R-CHOP therapy [2, 205]. This approach categorizes DLBCL as either germinal center B-cell (GCB), activated B-cell (ABC), or indeterminate, based on the phase of B-cell development it most closely resembles [17]. A practical algorithm employing this approach for immunohistochemically stained, formalin fixed,

paraffin embedded tissue was developed by Hans et al, and despite imperfect concordance with the gold standard GEP method, it is now the most widely used algorithm in the United States for DLBCL [193]. The GCB subtype is associated with more favorable outcomes than the non-GCB subtype [2, 85, 206, 72, 3, 130].

In addition to COO subtyping, double-hit lymphomas with concurrent chromosomal translocations of the MYC and BCL2 genes, or less commonly MYC and BCL6 genes, and double-expressor lymphomas with dual overexpression of MYC and BCL2 proteins have been found to correlate with an aggressive clinical course and poor outcomes when treated with R-CHOP [192]. Determination of these molecular subsets is now standard of care per the World Health Organization (WHO) guidelines and patients harboring dual chromosomal translocations are now formally classified as having high grade B-cell lymphoma, with MYC and BCL2 and/or BCL6 translocations (HGBL) [223].

While COO subtyping by the Hans algorithm corresponds to morphologically distinct benign precursors, germinal center type B-cells and activated B-cells, classification based on the morphologic properties of the tumor itself has historically been challenging due to the significant histomorphologic heterogeneity of DLBCL. Cytologically, DLBCL may resemble centroblasts with multiple peripheral nucleoli and vesicular chromatin or immunoblasts with abundant cytoplasm and a single prominent nucleolus. However, the prognostic significance of these and other recognised cytologic variants, for example anaplastic type DLBCL, is unclear and the subject of continued debate [67, 12, 57, 152, 199, 234].

Though several studies have thus far failed to conclusively demonstrate that morphologic classification can predict outcomes in DLBCL, automated imaging methods could potentially identify novel, prognostically significant morphological or immunohistochemical biomarkers. The ability of automated methods to identify prognostically relevant features on H&E sections that have eluded pathologists has been demonstrated [18, 114, 109]. If successful, automated image analysis could be scaled up into a cost-effective alternative to current classification methods which are typically costly and/or labor intensive. A critical requirement for the development of these models is the availability of datasets containing digitally scanned slides stained to show cell morphology and expression of relevant proteins with accompanying prognostic outcome data.

Here we present DLBCL-Morph, a publicly available dataset containing 42 digitally scanned high-resolution tissue microarrays (TMAs) from 209 DLBCL cases at Stanford Hospital. Each TMA was stained for H&E as well as for CD10, BCL6, MUM1, BCL2, and MYC protein expression. All of the TMAs are accompanied by pathologist-annotated regions-of-interest (ROIs) that indicate areas representative of DLBCL. For each patient in the dataset, we provide survival data, follow-up status, and a wide range of clinical and molecular variables such as age and MYC/BCL2/BCL6 gene translocations. We also segmented out tumor nuclei from ROIs inside the H&E stained TMAs, and provide several geometric features for each tumor nucleus.

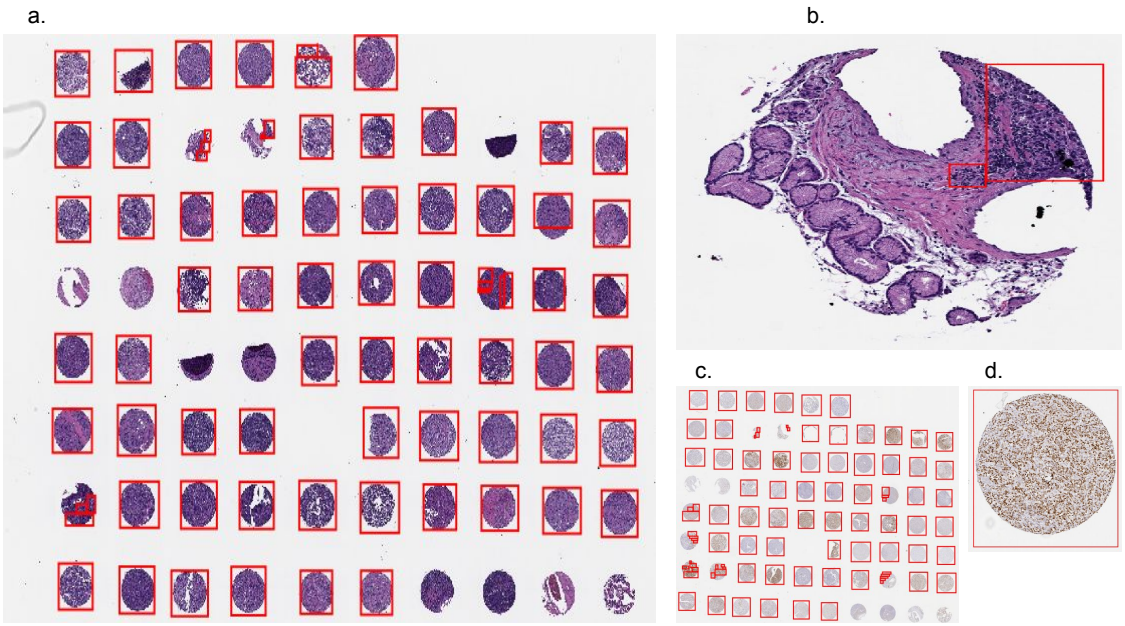


Figure 8.2: Tissue microarrays (TMAs) with region-of-interest (ROI) annotations. a) H&E stained TMA. The red rectangles denote ROIs annotated by a human expert. Some missing or unrepresentative cores have no ROIs. b) A single core from the TMA in a) with ROI that ignores unrepresentative areas of the core. c) BCL6 stained TMA, containing cores from the same patients as a). d) A single annotated core from the TMA in c). Cells stained orange show greater BCL6 expression.

8.2 Methods

Our dataset contains digitally scanned TMAs accompanied by pathologist-annotated ROIs. We extracted patches from the ROIs inside the H&E stained TMAs, and used a deep learning model called HoVer-Net [81] to segment tumor cell nuclei. We then computed several geometric descriptors for each segmented nucleus. Figure 5.2 shows our pipeline for an H&E stained TMA core. Our project was approved by the Institutional Review Board of Stanford University. All protected health information was removed and the project had no impact on clinical care, so the requirement for individual patient consent was waived.

Patient Cohort

The study cohort consists of patients with de novo, CD20+ DLBCL treated with curative intent with R-CHOP or R-CHOP–like immunochemotherapy with available clinical data from the Stanford Cancer Institute, Stanford, California. This patient cohort was included in a prior study with clinicopathologic inclusion criteria are as previously described [195].

Tissue Microarray

Stained tissue microarray (TMA) slides were scanned at 40x magnification (0.25 μm per pixel) on an Aperio AT2 scanner (Leica Biosystems, Nussloch, Germany) in ScanScope Virtual Slide (SVS) format. This high magnification level displays the tissue in very fine detail, which we believe to be beneficial for the development of automated imaging models. Each SVS file also contains a slide label image, a macro camera image, and a thumbnail image. The slide label image is a low-resolution image of the slide's label, which shows the TMA number and the stain (eg: BCL2). The macro camera image is a low-resolution picture of the entire slide. The thumbnail is an image of the whole scanned TMA.

Our dataset includes 7 TMAs, each with a 0.4 micron thick formalin-fixed, paraffin-embedded (FFPE) section of tumors assembled in a grid. Within the microarray each tumor is represented by a 0.6-mm core diameter sample in duplicate. Replicates of each TMA were stained with H&E, which shows cell morphology. They were also stained for the expression of the following 5 oncogenes: CD10, BCL6, MUM1, BCL2, and MYC. We therefore have 6 stains per TMA, resulting in 42 distinct digitally-scanned slides. An example of an H&E stained TMA is shown in Figure 8.2 a) and a BCL6 stained TMA is shown in Figure 8.2 c). Since overexpression of one or more of these proteins is observed in a significant portion of DLBCL cases, automated imaging models can use the immunostained TMAs to potentially identify prognostically significant features related to protein expression.

Pathologist annotations

Although TMA cores were already taken from areas of tissue showing DLBCL, some of the cores were partially or entirely missing. Furthermore, some cores still contained areas of tissue that had very few or no tumor cells. We obtained rectangular ROI annotations from expert pathologists to highlight the core regions which represent DLBCL accurately. The annotations were created for all TMAs and all stains at 40x magnification. The pixel coordinates for the rectangles in ROIs, along with the corresponding deidentified unique patient_id, are included in our dataset. We believe the exclusion of missing or insufficiently representative tissue areas will be beneficial for automated prognostic models which use patches from the TMAs as input. Example ROI annotations are shown in Figure 8.2 b) and d).

Patches from stained TMAs

We extracted patches of size 224x224 from within the ROIs in the stained TMAs, at 40x magnification. The patches were extracted uniformly from inside each annotated rectangle, starting from the top-left corner and proceeding until the bottom-right corner. The patches are non-overlapping, and we omitted patches that are mostly white and contain little tissue. We provide these patches as

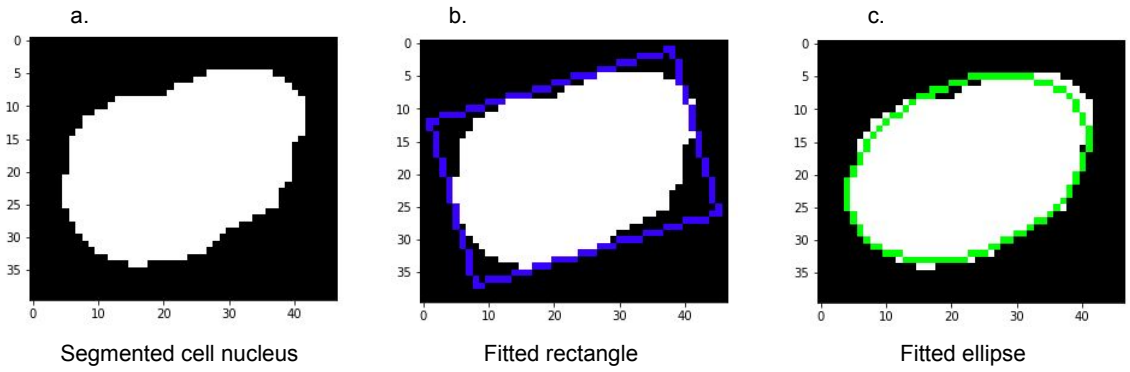


Figure 8.3: Rectangle and ellipse fitted to a single segmented tumor nucleus. a) a binary segmentation image for a tumor cell nucleus. For visual clarity, the image is zero-padded by 5 pixels on each side. b) rotated rectangle fit to the nucleus. Our dataset provides the rectangle's center coordinates, width, height and rotation angle. c) rotated ellipse fit to the nucleus. Our dataset provides the ellipse's center coordinates, perimeter, area, and major and minor axis lengths.

part of our dataset. Due to our ROI annotation process detailed above, our patches exclude missing and unrepresentative areas of cores. Since deep learning based imaging methods typically cannot directly operate on images as large as the 40x magnification image, the patches can instead be used as input. We also used patches from H&E stained TMAs to segment tumor cell nuclei as described below.

Tumor cell nucleus segmentation

We used a deep learning based nucleus segmentation and classification model called HoVer-Net to segment every tumor cell inside each of the patches from H&E stained TMAs. The HoVer-Net operates independently on each patch, and produces an output image segmenting all individual cell nuclei in the patch, and another output image specifying the classification of each segmented nucleus. The HoVer-Net classifies segmented nuclei into 5 categories: neoplastic, non-neoplastic epithelial, inflammatory, connective, dead. HoVer-Net uses a neural network based on a pretrained ResNet-50 architecture to extract image features. These extracted features are then processed in three steps: the nuclear pixel (NP) step, HoVer step, and nuclear classification (NC) step. The NP step determines whether each pixel belongs to a nucleus or the background, and the HoVer step predicts the vertical and horizontal distances of nucleus pixels to their centroid, thereby allowing separation of touching nuclei. Then the NC step classifies each nucleus pixel, and aggregates these across all pixels in a segmented nucleus to classify each nucleus as neoplastic, non-neoplastic epithelial,

inflammatory, connective, or dead. We used the HoVer-Net output to identify each neoplastic cell nucleus in a patch, and saved it as a separate binary image, thereby obtaining one binary image for each tumor cell. Each binary image illustrates the size and shape of the nucleus, and we provide these in our dataset. An example binary image is shown in Figure 5.2 e) and another is shown in Figure 8.3 a). We used these binary images to compute geometric features for each tumor cell nucleus as described below.

Geometric features from tumor nuclei

We used the per-nucleus binary segmentation images to compute several geometric features for each tumor cell nucleus. While end-to-end imaging models may not require such hand-crafted features, prognostic models which use these features can give more explainable results, and can more clearly indicate the prognostic importance of these features.

We fit a (possibly rotated) rectangle of minimum area enclosing the binary mask, and provide the rectangle's top left point coordinates, width and height, and rotation angle. An example rectangle is shown in Figure 8.3 b). The rectangle's top left point is a tuple corresponding to the feature rectCenter. The first element of the tuple corresponds to the x-coordinate, and the second element corresponds to the y-coordinate. The width and height are in a tuple corresponding to the feature rectDimension. The first element of the tuple corresponds to the width, and the second element to the height. The rotation angle corresponds to the feature rotate_angle, which ranges from $-90°$ to $0°$. A value of $-90°$ corresponds to an axis-aligned rectangle. As the rectangle is rotated clockwise, the angle increases toward $0°$, at which point the rectangle is again axis-aligned and the angle resets to $-90°$.

We fit an ellipse around the nucleus in the binary segmentation mask, and provide the ellipse center, major axis, minor axis, perimeter and area of the ellipse. An example ellipse is shown in Figure 8.3 c). The ellip_centroid parameter is a tuple containing the x and y coordinates of the ellipse. The features shortAxis and longAxis correspond to the lengths of the minor and major axes respectively. The feature ellip_perimt corresponds to the ellipse perimeter, and ellip_area corresponds to the ellipse area.

We computed the maximum and minimum Feret diameters for each segmented nucleus, and provide the corresponding angles. Given an object and a fixed direction, the Feret diameter is the distance between two parallel tangents to the object, where the tangents are perpendicular to the fixed direction. The feature maxDiameter contains the Feret diameter maximized over all directions, and maxAngle specifies the angle (between $-180°$ and $180°$) at which the maximum diameter is obtained. The features minDiameter and minAngle are similar but for the minimum Feret diameter. We further computed the convex hull of the segmented nucleus. The feature hull_area corresponds to the area of the convex hull.

Finally we computed a number of geometric features derived from the quantities described above.

These features are esf, csf, sf1, sf2, elongation, and convexity. These are defined below in (1) – (6). The esf, sf1, sf2 and elongation are all simple ratios that can be thought of as measures of how "elongated" the nucleus is. In particular, they are all equal to 1 if the nucleus is perfectly circular. The csf is similar: it is a measure of circularity, and is equal to 1 if the nucleus is perfectly circular. For increasingly elliptical nuclei, the csf decreases towards 0.

$$\text{esf} = \frac{\text{shortAxis}}{\text{longAxis}} \tag{8.1}$$

$$\text{csf} = \frac{4\pi * \text{ellip\_area}}{\text{ellip\_perimt}^2} \tag{8.2}$$

$$\text{sf1} = \frac{\text{shortAxis}}{\text{maxDiameter}} \tag{8.3}$$

$$\text{sf2} = \frac{\text{minDiameter}}{\text{maxDiameter}} \tag{8.4}$$

$$\text{elongation} = \frac{\text{maxDiameter}}{\text{minDiameter}} \tag{8.5}$$

$$\text{convexity} = \sqrt{\frac{\text{ellip\_area}}{\text{hull\_area}}} \tag{8.6}$$

8.3 Data Records

The DLBCL-morph dataset [235] is organized into three folders, *TMA*, *Patches*, and *Cells* as is shown by Figure 8.4. The clinical data of the patients together with the outcome is stored in *clinical_data.xlsx* and *clinical_data_cleaned.csv* where the latter contains all the patients for which the outcome is recorded and all categorical variables are converted to numerical values, e.g. 'neg', 'pos', and 'no data' were converted to 0, 1, and NaN, respectively for the variable CD10 IHC. Each patient has a unique identifier. There are 209 patients recorded in *clinical_data_cleaned.csv*. The column *OS* records the overall survival which is the length of time (in years) from the end of treatment until death or last follow-up. The column *Follow-up Status* (FUS) is 1 if the patient was deceased at the time of last follow-up, else 0. A description of each column can be found in Supplementary File 1.

TMA

The *TMA* folder contains a total of 42 digitally-scanned TMAs, which are organized within subfolders for each stain. The filename of each TMA is a TMA id which is the same across all stains, i.e.

DLBCL-Morph/TMA/HE/TMA255 and *DLBCL-Morph/TMA/BCL2/TMA255* contains cores of the same set of patients. The TMA id together with the row and column number of each core, starting with 0 and 0, respectively in the upper left corner, can be linked to the patient id through *core.csv*, each patient has two cores. The *annotations.csv* contains coordinates of ROIs annotated by human experts. For each annotation there is a patient id, TMA id, and stain where the TMA id and the stain is used to locate the TMA file that the annotation belongs to. The annotations are rectangular and the coordinates record the upper left and lower right corners based on the 40x magnification level of the TMAs.

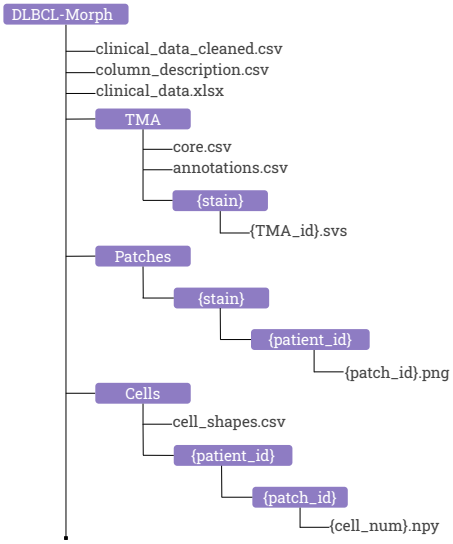


Figure 8.4: The directory structure of DLBCL-Morph

Patches

The Patches folder contains subfolders of stains which contains subfolders of patients that has at least one ROI. The patches are localized in the folders of patient ids with a patch id as the filename and are stored in PNG format. There are 195 patients that have at least one patch from at least one stained TMA. However, some patients do not contain patches for all 6 stains, which can occur if the core for a particular stain was missing or not covered by any ROIs.

Cells

The *Cells* folder contains subfolders of patient ids which contains subfolders of patch ids. The binary segmentation images for tumor cell nucleus are localized in the folders of patch ids with the cell number as the filename and stored in NPY format. The NPY format is used by the Numpy

package for Python to save arrays, in this case we are storing 2-dimensional arrays with binary values as segmentations of tumor cell nucleus. The cell numbers are non-consecutive since all non-tumor cells are discarded in each patch. All the geometric features computed from tumor nuclei are stored in *cell_shapes.csv* and can be linked to the nucleus segmentation images through the patch id and the cell number.

8.4 Technical Validation

Segmentation Output Validation

We validated the quality of the tumor nucleus segmentation masks output by the HoVer-Net by comparing these against segmentations provided by an expert pathologist. We randomly sampled 20 H&E-stained patches, each from a distinct patient. An expert pathologist segmented all tumor nuclei visible in each patch. Using the pathologist's segmentation masks as ground truth, we measured the agreement with the HoVer-Net's tumor nuclei segmentations by computing the mean Intersection over Union (mIOU) over all patches. We also measured the mean precision and recall over all patches, defined as follows:

$$\text{precision} = \frac{|\text{ground\_truth\_mask} \ \wedge \ \text{output\_mask}|}{|\text{output\_mask}|}$$

$$\text{recall} = \frac{|\text{ground\_truth\_mask} \ \wedge \ \text{output\_mask}|}{|\text{ground\_truth\_mask}|}$$

where ground_truth_mask is the binary segmentation mask provided by the pathologist, output_mask is the binary segmentation mask for tumor nuclei provided by the HoVer-Net, and $\wedge$ is the logical AND operator.

We obtained an mIOU of 0.372, a mean precision of 0.393, and a mean recall of 0.906. The high mean recall and the relatively low mean precision suggest that the HoVer-Net segmentations covered the ground-truth tumor nuclei segmentations, but they also classified additional nuclei as neoplastic. These additional nuclei were not classified as tumor in the ground-truth segmentation mask provided by the pathologist. These other nuclei, annotated as tumor nuclei by HoVer-Net, were further evaluated by a pathologist and appeared to in part include plasma cells, endothelial cells, and nuclear material compatible with dying lymphoma cells (pyknotic nuclei).

Survival Regression

We performed survival regression using the geometric and clinical features in our dataset to measure the utility of these features in predicting prognostic outcome. This analysis was performed on the 170 patients for whom patches from H&E stained TMAs were available. For each of the geometric

features computed per tumor nucleus, we computed the mean and standard deviation across all nuclei for each patient. We then fit Cox Proportional Hazards models using the binary Follow-up Status (FUS) feature as an indicator of censoring, and the overall survival (OS) feature as the time to event or censoring (in years). We evaluated our models using Harrel's C-index [89]. Random prediction would give a C-index of 0.5. Specifically we fit three models: i) using both clinical and geometric features ii) using only clinical features iii) using only geometric features.

We used the bootstrap method to obtain an "optimism-corrected" C-index [88]. We sampled 1000 bootstrap replicates with replacement and fit the model on each bootstrap replicate. We then evaluated the model on both the original data and the bootstrap replicate. We recorded the performance decrease between evaluating on the bootstrap replicate and evaluating on the original data. This decrease, averaged over all bootstrap replicates, was subtracted from the original C-index to obtain the optimism-corrected C-index. We also generated the corresponding 95% two-sided confidence intervals (CI) for the optimism-corrected C-indices using the non-parametric percentile bootstrap method [65] with 1000 bootstrap replicates.

The resulting optimism-corrected C-indices with 95% CIs for our models were: i) 0.700 $(0.651, 0.744)$ using clinical and geometric features, ii) 0.674 $(0.602, 0.737)$ using only clinical features and iii) 0.635 $(0.574, 0.691)$ using only geometric features. Thus, use of the geometric features alone allowed significantly better than random survival prediction. Use of both clinical and geometric features led to a higher performance than the use of clinical features alone, although this performance difference was not statistically significant. While prognostic classification based on the morphologic properties of the tumor has proved to be challenging and the subject of continued debate, [67, 12, 57, 152, 199, 234] our results suggest that geometric features computed from H&E-stained tumor nuclei can provide a significant signal to predict survival outcome. This finding should be further validated by using improved tumor nuclei segmentation models, which may have better agreement with expert pathologists, and evaluating on external and prospective datasets.

8.5 Usage Notes

The iPython notebooks used for processing the data are publicly available [1]. The data is organized as shown in Figure 8.4. We have provided publicly available Jupyter Notebooks [166, 120] to illustrate computation of geometrical features as well as usage of the data. One notebook uses the clinical and geometric variables in the dataset to reproduce the survival regression results described in the Technical Validation section. Another notebook visualizes and reproduces the computation of several geometric features for any segmented tumor nucleus in our dataset. Finally, we provide another notebook to extract patches uniformly from inside any of the ROIs in the dataset. These patches are already included as part of the dataset, but we believe this notebook will be beneficial

[1] https://github.com/stanfordmlgroup/DLBCL-Morph

for researchers who work with the SVS files in our dataset.

8.6 Code availability

The code to compute all geometric features from all tumor nuclei in our dataset, along with notebooks to illustrate usage of our data and reproduce all survival regression results, is publicly available at https://github.com/stanfordmlgroup/DLBCL-Morph.

Chapter 9

Accurate Radiology Report Labeling

In previous chapters, we have looked at the curation of datasets that enable large-scale training of medical imaging models. In this chapter, we will look specifically at the process of labeling images by extracting labels from radiology text reports.

Existing approaches to report labeling typically rely either on sophisticated feature engineering based on medical domain knowledge or manual annotations by experts. In this work, we introduce a BERT-based approach to medical image report labeling that exploits both the scale of available rule-based systems and the quality of expert annotations. We demonstrate superior performance of a biomedically pretrained BERT model first trained on annotations of a rule-based labeler and then fine-tuned on a small set of expert annotations augmented with automated backtranslation. We find that our final model, CheXbert, is able to outperform the previous best rule-based labeler with statistical significance, setting a new SOTA for report labeling on one of the largest datasets of chest x-rays.

This chapter is based on [213].

9.1 Introduction

The extraction of labels from radiology text reports enables important clinical applications, including large-scale training of medical imaging models [239]. Many natural language processing systems have been designed to label reports using sophisticated feature engineering of medical domain knowledge [172]. On chest x-rays, the most common radiological exam, rule-based methods have been engineered to label some of the largest available datasets [111]. While these methods have generated considerable advances, they have been unable to capture the full diversity of complexity, ambiguity

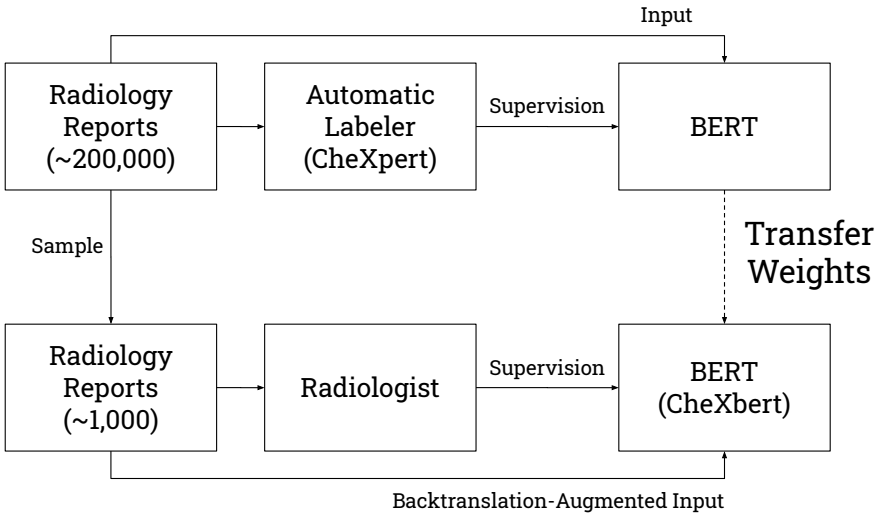


Figure 9.1: We introduce a method for radiology report labeling, in which a biomedically pretrained BERT model is first trained on annotations of a rule-based labeler, and then fine-tuned on a small set of expert annotations augmented with automated backtranslation.

and subtlety of natural language in the context of radiology reporting.

More recently, Transformers have demonstrated success in end-to-end radiology report labeling [60, 244]. However, these methods have shifted the burden from feature engineering to manual annotation, requiring considerable time and expertise for high quality. Moreover, these methods do not take advantage of existing feature-engineered labelers, which represent state-of-the-art on many medical tasks.

We introduce a simple method for gaining the benefits of both existing radiology report labelers and expert annotations to achieve highly accurate automated radiology report labeling. This approach begins with a biomedically pretrained BERT model [56, 163] trained on the outputs of an existing labeler, and performs further fine-tuning on a small corpus of expert annotations augmented with automated backtranslation. We apply this approach, shown in Figure 9.1, to the task of radiology report labeling of chest x-rays, and call our resulting model *CheXbert*.

CheXbert outperforms the previous best reported labeler [106] on an external dataset, MIMIC-CXR [111], with an improvement of 0.055 (95% CI 0.039, 0.070) on the F1 metric, and is only 0.007 F1 away from a radiologist performance benchmark. We expect this method of training medical report labelers is broadly useful for natural language processing within the medical domain, where collection of expert labels is expensive, and feature engineered labelers already exist for many tasks.

9.2 Related Work

Many natural language processing systems have been developed to extract structured labels from free-text radiology reports [172, 250, 90, 8, 202, 241, 39, 25]. In many cases, these methods have relied on heavy feature engineering that include controlled vocabulary and grammatical rules to find and classify properties of radiological findings. NegEx [34], a popular component of rule-based methods, uses simple regular expressions for detecting negation of findings and is often used in combination with ontologies such as the Unified Medical Language System (UMLS) [24]. NegBio [162], an extension to NegEx, utilizes universal dependencies for pattern definition and subgraph matching for graph traversal search, includes uncertainty detection in addition to negation detection for multiple pathologies in chest x-ray reports, and is used to generate labels for the ChestX-Ray14 dataset [239].

The CheXpert labeler [106] improves upon NegBio on chest x-ray report classification through more controlled extraction of mentions and an improved NLP pipeline and rule set for uncertainty and negation extraction. The CheXpert labeler has been applied to generate labels for the CheXpert dataset and MIMIC-CXR [111], which are amongst the largest chest x-ray datasets publicly available.

Deep learning approaches have also been trained using expert-annotated sets of radiology reports [249]. In these cases, training set size, often driving the performance of deep learning approaches, is limited by radiologist time and expertise. [38] trained CNNs with GloVe embeddings [165] on 1000 radiologist-labeled reports for classification of pulmonary embolism in chest CT reports and improved upon the previous rule-based SOTA, peFinder [29] trained both recurrent and convolutional networks in combination with attention mechanisms on 27,593 physician-labeled radiology reports and apply their labeler to generate labels. More recently, Transformer-based models have also been applied to the task of radiology report labeling. [60] trained classifiers using BERT [56] and XLNet [251] on 3,856 radiologist labeled reports to detect normal and abnormal labels. [244] developed ALARM, an MRI head report classifier on head MRI data using BioBERT [129] models trained on 1500 radiologist-labeled reports, and demonstrate improvement over simpler fixed embedding and word2vec-based [145] models [260].

Our work is closely related to approaches to reduce the number of expert annotations required for training medical report labelers [31, 191, 16]. A method of weak supervision known as data programming [190] has seen successful application to medical report labeling: in this method, users write heuristic labelling functions that programmatically label training data. [197] used data programming to incorporate labeling functions consisting of regular expressions that look for phrases in radiology reports, developed with the help of a clinical expert in a limited time window, to label for intracranial hemorrhage in head CTs. [63] demonstrated that in under 8 hours of cumulative clinician time, a data programming method can approach the efficacy of large hand-labeled training sets annotated over months or years for training medical imaging models, including chest x-ray classifiers on the task of normal / abnormal detection. Beyond data programming approaches, [60] developed

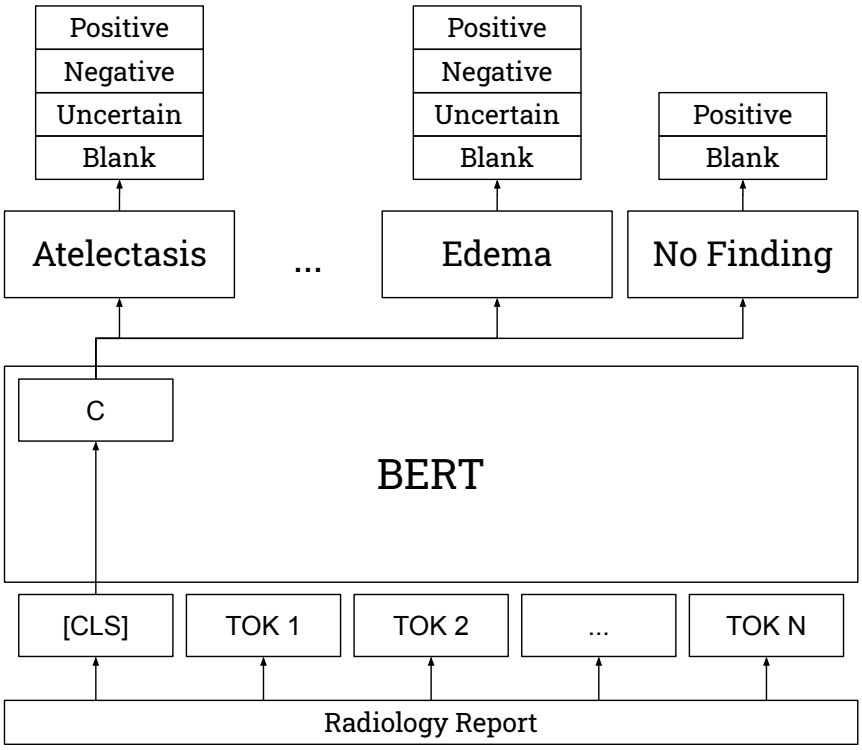


Figure 9.2: Model architecture. The model contains 14 linear heads, one for each medical observation, but only 3 heads are shown here.

a fully unsupervised approach utilizing a Siamese Neural Network and Gaussian Mixture Models, reporting performance similar to the CheXpert labeler without requiring any radiologist-labeled reports on the simplified task of normal / abnormal detection. Concurrently developed to our work is the CheXpert++ labeler [143], which was trained on the outputs of the rule-based CheXpert labeler and showed improved performance after a single additional epoch of fine-tuning on expert-labeled report sentences.

9.3 Methods

9.3.1 Task

The report labeling task is to extract the presence of one or more clinically important observations (e.g. consolidation, edema) from a free-text radiology report. More formally, a labeler takes in as inputs sentences from a radiology report and outputs for 13 observations one of the following classes: blank, positive, negative, and uncertain. For the 14th observation corresponding to *No Finding*, the labeler only outputs one of the two following classes: blank or positive.

9.3.2 Data

Two existing large datasets of chest x-rays, CheXpert [106] (consisting of 224,316 images), and MIMIC-CXR [111] (consisting of 377,110 images) are used in this study. Both datasets have corresponding radiology reports that have been labeled for the same set of 14 observations using the CheXpert labeler [106], from the *Impression* section, or other parts of the radiology report.

A subset of both datasets also contain manual annotations by expert radiologists. On CheXpert, a total of 1000 reports (*CheXpert manual set*) were reviewed by 2 board certified radiologists with disagreement resolution through consensus. On MIMIC-CXR, a total of 687 reports (*MIMIC-CXR test set*) were reviewed by 2 board certified radiologists and manually labeled for the same 14 medical observations as in CheXpert. In this study, CheXpert is used for the development of models, and the MIMIC-CXR test set is used for evaluation.

Some reports from the same patient appear multiple times in the CheXpert dataset. Removing duplicate reports as well as the CheXpert manual set from the CheXpert dataset results in 190,460 reports, the class prevalences for which are shown in Table 9.5 of the Appendix. We remove excess spaces and newlines from all reports.

9.3.3 Model Architecture

All models use a modification of the BERT-base architecture [56] with 14 linear heads (as shown in Figure 9.2): 12 heads correspond to various medical abnormalities, 1 to medical support devices, and 1 to 'No Finding'. Each radiology report text is tokenized, and the maximum number of tokens in each input sequence is capped at 512. The final-layer's hidden state corresponding to the CLS token is then fed as input to each of the linear heads.

9.3.4 Training Details

For all our models, unless otherwise specified, we fine-tune all layers of the BERT model, including the embeddings, and feed the CLS token into the 14 linear heads to generate class scores for each medical observation. BERT-Base contains ~ 110 million parameters, and the linear heads contain $\sim 40,000$ parameters.

All models are trained using cross-entropy loss and Adam optimization with a learning rate of 2×10^{-5}, as used in [56] for fine-tuning tasks. The cross-entropy losses for each of the 14 observations are added to produce the final loss. During training, we periodically evaluate our model on the dev set and save the checkpoint with the highest performance averaged over all 14 observations. All models are trained using 3 TITAN-XP GPUs with a batch size of 18.

	Model	F1 (95% CI)
Training Strategy	T-rad	0.705 (0.680, 0.725)
	T.cls-rad	0.286 (0.265, 0.305)
	T.token-rad	0.396 (0.374, 0.416)
	T-auto	0.755 (0.731, 0.774)
	T-hybrid	0.775 (0.753, 0.795)
Biomedical Representations	Tbio-rad	0.616 (0.587, 0.639)
	Tclinical-rad	0.677 (0.651, 0.699)
	Tblue-rad	0.741 (0.714, 0.763)
With Backtranslation Augmentation	T-rad-bt	0.729 (0.702, 0.749)
	T-hybrid-bt	0.795 (0.772, 0.815)
	Tblue-rad-bt	0.770 (0.747, 0.790)
	Tblue-hybrid-bt (CheXbert)	0.798 (0.775, 0.816)
Previous SOTA	CheXpert	0.743 (0.719, 0.764)
Benchmark	Radiologist	0.805 (0.784, 0.823)

Table 9.1: Average F1 score with 95% confidence intervals for all our models, with comparisons to CheXpert labeler and radiologist benchmark.

9.3.5 Evaluation

Models are evaluated on their average performance on three retrieval tasks: positive extraction, negative extraction, and uncertainty extraction. For each of the tasks, the class of interest (e.g. negative for the negative extraction and uncertain for the uncertainty extraction) is treated as the positive class, and the other classes are considered negative. For each of the 14 observations, we compute a weighted average of the F1 scores on each of the above three tasks, weighted by the support for each class of interest, which we call the *weighted-F1* metric, henceforth simply abbreviated to F1.

We report the simple average of the F1 across all of the observations. We include the 95% two-sided confidence intervals of the F1 using the nonparametric percentile bootstrap method with 1000 bootstrap replicates [65].

9.4 Experiments

9.4.1 Supervision Strategies

We investigate models trained using three strategies: trained only on radiologist-labeled reports, trained only on labels generated automatically by the CheXpert labeler [106], and trained on a combination of the two.

Radiologist Labels **T-rad** is obtained by training the model on the CheXpert manual set, fine-tuning all weights. As baselines, we also train models that freeze all weights in the BERT layers, and only update the weights in the linear heads: **T.cls-rad** is identical to T-rad in architecture, while

Category	CheXbert	Improvement over CheXpert
Pneumonia	0.835 (0.789, 0.881)	0.151 (0.093, 0.206)
Fracture	0.791 (0.665, 0.895)	0.120 (0.019, 0.236)
Consolidation	0.877 (0.810, 0.935)	0.105 (0.029, 0.192)
Enlarged Cardiom.	0.713 (0.623, 0.783)	0.100 (0.038, 0.166)
No Finding	0.640 (0.482, 0.759)	0.097 (0.007, 0.182)
Pleural Other	0.534 (0.372, 0.671)	0.056 (0.008, 0.124)
Cardiomegaly	0.815 (0.759, 0.860)	0.051 (0.018, 0.086)
Pneumothorax	0.928 (0.892, 0.960)	0.046 (0.015, 0.076)
Atelectasis	0.940 (0.910, 0.971)	0.023 (-0.001, 0.051)
Support Devices	0.888 (0.856, 0.919)	0.021 (0.004, 0.040)
Edema	0.881 (0.843, 0.916)	0.017 (-0.007, 0.042)
Pleural Effusion	0.919 (0.892, 0.947)	0.014 (-0.005, 0.034)
Lung Lesion	0.664 (0.550, 0.771)	-0.019 (-0.098, 0.056)
Lung Opacity	0.741 (0.684, 0.792)	-0.021 (-0.056, 0.006)
Average	0.798 (0.775, 0.816)	0.055 (0.039, 0.070)

Table 9.2: The F1 scores for CheXbert as well as improvements over the CheXpert labeler on the MIMIC-CXR test set, in descending order of improvement, and reported with 95% confidence intervals.

T.token-rad averages the non-padding output tokens as the input into the linear heads rather than using the CLS token output. All models are trained using a random 75%-25% train-dev split on the CheXpert manual set, and are trained until convergence.

Automatic Labels **T-auto** is obtained using labels generated by the rule-based CheXpert labeler, described in [106]. T-auto is trained using a random 85%-15% train-dev split of the CheXpert dataset, different from the models trained on radiologist labels. T-auto is trained for 8 epochs, since slightly higher dev performance is observed compared to the typical 2-4 epochs for BERT fine-tuning tasks.

Hybrid Labels **T-hybrid** is obtained by initializing a model with the weights of T-auto, and then fine-tuning it on radiologist-labeled reports, as for T-rad.

Results As shown in Table 9.1, T-rad achieves an F1 of 0.705 $(0.680, 0.725)$, significantly higher than the performance of the baselines with T.cls-rad at 0.286 $(0.265, 0.305)$, and T.token-rad at 0.396 $(0.374, 0.416)$. T-auto achieves a higher F1 of 0.755 $(0.731, 0.774)$. Superior performance is obtained by T-hybrid, with an F1 of 0.775 $(0.753, 0.795)$.

9.4.2 Biomedical Language Representations

We investigate the effect of having models pre-trained on biomedical data. For the following models, we use an identical training procedure to T-rad, but initialize the weights differently. **Tbio-rad**

is obtained by using BioBERT weight initializations [129]. BioBERT was obtained by further pretraining the BERT weights on a large biomedical corpus comprising PubMed abstracts (4.5 billion words) and PMC full-text articles (13.5 billion words). **Tclinical-rad** is obtained by using Clinical BioBERT weight initializations [4], which were obtained by further pretraining the BioBERT weights on 2 million clinical notes from the MIMIC-III database. Finally, **Tblue-rad** is obtained by using BlueBERT, a BERT model pretrained on PubMed abstracts and clinical notes (MIMIC-III) [163].

Results As shown in Table 9.1, Tbio-rad achieves an F1 of 0.616 $(0.587, 0.639)$ and Tclinical-rad achieves an F1 of 0.677 $(0.651, 0.699)$, lower than T-rad. However, Tblue-rad achieves an F1 of 0.741 $(0.714, 0.763)$, higher than T-rad. The drop in performance with Tbio-rad and Tclinical-rad may possibly be attributed to using different vocabulary, sequence length, and other configurations (stopping procedure, embedding dimensions) than those used by Tblue-rad, which uses the configurations provided in [56].

9.4.3 Data Augmentation using Backtranslation

We investigate the use of backtranslation to improve the performance of the models. Backtranslation is designed to generate alternate formulations of sentences by translating them to another language and back. Although backtranslation has been successfully used to augment text data in a variety of NLP tasks [257, 171], to our knowledge, the technique is yet to be applied to a medical report extraction task. In this experiment, we augment the CheXpert manual set using Facebook-FAIR's winning submission to the WMT'19 news translation task [154] to generate backtranslations. Although this submission includes models that produce German/English and Russian/English translations, initial experiments with Russian did not demonstrate semantically correct translations, so we only continued experiments with German. We use beam search with a beam size of 1 to select the single most likely translation. We perform this experiment using our best models: **Tblue-rad-bt** is obtained by using an identical training procedure to Tblue-rad on the augmented dataset (which is twice the size of the CheXpert manual set). **Tblue-hybrid-bt** is obtained by first training a BlueBERT-based labeler on automatically generated CheXpert labels, and then fine-tuning on radiologist-labeled reports of the CheXpert manual set, augmented by backtranslation. We also report the performance of T-rad-bt and T-hybrid-bt.

Results As shown in Table 9.1, T-rad-bt achieves an F1 score of 0.729 $(0.702, 0.749)$, higher than that of T-rad. Similarly, T-hybrid-bt achieves an F1 of 0.795 $(0.772, 0.815)$. Tblue-rad-bt achieves an F1 of 0.770 $(0.747, 0.790)$, higher than that of the CheXpert labeler. Tblue-hybrid-bt achieves a superior F1 score of 0.798 $(0.775, 0.816)$.

9.4.4 Comparison to previous SOTA and radiologist benchmark

We compare the performance of our best model to the previous best reported labeler, the CheXpert labeler [106], and to a radiologist benchmark. CheXpert is an automated rule-based labeler that extracts mentions of conditions like pneumonia by searching against a large manually curated list of words associated with the condition and then classifies mentions as uncertain, negative, or positive using rules on a universal dependency parse of the report. For the radiologist benchmark, the annotations by one of the 2 radiologists on the MIMIC-CXR test set is used, while the other is used as ground truth. We report the improvement of our best model, Tblue-hybrid-bt, which we also call **CheXbert**, over the CheXpert labeler by computing the paired differences in F1 scores on 1000 bootstrap replicates and provide the mean difference along with a 95% two-sided confidence interval.

Results We observe that CheXbert has a statistically significant improvement ($p < 0.001$) over the existing SOTA, CheXpert, which achieves a score of 0.743 $(0.719, 0.764)$. Notably, we also find that Tblue-rad-bt, the best model trained only on manually labeled radiology reports, performs at least as well as the CheXpert labeler.

Table 9.2 shows the F1 per class (along with 95% confidence intervals) for CheXbert and for the improvements over CheXpert. CheXbert records an improvement in all but 2 medical conditions, and a statistically significant improvement in 9 of the 14 conditions. The largest improvements are observed for Pneumonia [0.151 $(0.093, 0.206)$], Fracture [0.120 $(0.019, 0.236)$], Consolidation [0.105 $(0.029, 0.192)$], Enlarged Cardiomediastinum [0.100 $(0.038, 0.166)$], and No Finding [0.097 $(0.007, 0.182)$]. Further significant improvements are observed for Pleural Other [0.056 $(0.008, 0.124)$], Cardiomegaly [0.051 $(0.018, 0.086)$], Pneumothorax [0.046 $(0.015, 0.076)$] and Support Devices [0.021 $(0.004, 0.040)$]. Overall, CheXbert achieves a statistically significant improvement on F1 of 0.055 $(0.039, 0.070)$. The board-certified radiologist achieves an F1 of 0.805 $(0.784, 0.823)$, which is 0.007 F1 points higher than the performance of CheXbert.

Training times For all our models except the baselines, training on radiologist-labeled reports takes $\sim$ 30 minutes, training on the radiologist-labeled reports augmented via backtranslation takes $\sim$ 50 minutes. Training on the larger automatically labeled report set takes $\sim$ 7 hours.

Inference times We benchmark the time taken by CheXbert and CheXpert to label all 190,460 report impressions in the CheXpert dataset. On a system with 32GB RAM and 1 CPU core, the CheXbert model takes $\sim$ 3.7 hours. This is an order of magnitude faster than the 36 hours required for CheXpert. With a single TITAN-XP GPU, the CheXbert model's inference time reduces to $\sim$ 18 minutes.

Report Segment and Labels	Reasoning
...two views of chest demonstrate *cariomegaly* with no focal consolidation... Cardiomegaly CheXpert: Blank ✗ T-auto: Positive ✓	T-auto, in contrast to CheXpert, recognizes conditions with misspellings in the report like "cariomegaly" in place of "cardiomegaly".
...consistent with acute and/or chronic pulmonary edema.... Edema CheXpert: Positive ✓ T-auto: Uncertain ✗	T-auto incorrectly detects uncertainty in the edema label, likely from the "and/or"; CheXpert correctly classifies this example as positive.
...Normal heart size, mediastinal and hilar contours are unchanged in appearance... Enlarged Cardiomediastinum CheXpert: Negative ✗ T-auto: Negative ✗ CheXbert: Uncertain ✓	T-auto and CheXpert both incorrectly label this example as negative for enlarged cardiomediastinum; CheXbert correctly classifies it as uncertain, likely recognizing that "unchanged" is associated with uncertainty of the condition. The condition cannot be labeled positive or negative without more information.

Table 9.3: Phrases from reports where CheXpert, T-auto, and CheXbert provide different labels. The correct label is indicated by a checkmark in the first column. The CheXpert versus T-auto comparisons are conducted on the CheXpert manual set. The CheXbert versus T-auto/CheXpert comparison is conducted on the MIMIC-CXR test set.

9.5 Analysis

9.5.1 T-auto versus CheXpert

We analyze whether T-auto, which is trained exclusively on labels from CheXpert (a rules-based labeler), can generalize beyond those rules.

We analyze specific examples in the CheXpert manual test set which T-auto correctly labels but CheXpert mislabels. On one example, T-auto is able to correctly detect uncertainty expressed in the phrase "cannot be entirely excluded," which CheXpert is not able to detect because the phrase does not match any rule in its ruleset. Similarly, on another example containing "no evidence of pneumothorax or bony fracture," T-auto correctly labels fracture as negative, while CheXpert labels fracture as positive since the phrasing does match any negation construct part of its ruleset. T-auto, in contrast to CheXpert, also recognizes conditions with misspellings in the report like "cariomegaly" in place of "cardiomegaly" and "mediastnium" in place of "mediastinum". Examples of T-auto correctly labeling conditions mislabeled by CheXpert are provided in Table 9.8 of the Appendix. Table 9.9 of the Appendix contains examples of CheXpert correctly labeling conditions mislabeled by T-auto. An example of each case is shown in the top two rows of Table 9.3.

9.5.2 CheXbert versus T-auto and CheXpert

We analyze how CheXbert improves on T-auto and CheXpert using examples which CheXbert correctly labels but T-auto and CheXpert incorrectly label.

CheXbert is able to correctly detect conditions that CheXpert and T-auto are not able to. On one example, T-auto and CheXpert both mislabel a "mildly enlarged heart" as blank for cardiomegaly, while CheXbert correctly labels it positive. On another containing "Right hilum appears slightly more prominent" (an indicator for enlarged cardiomediastinum), CheXbert correctly classifies enlarged cardiomediastinum as positive, while T-auto and CheXpert do not detect the condition.

Furthermore, CheXbert correctly labels nuanced expressions of negation that both CheXpert and T-auto mislabel. On the example containing "heart size is slightly larger but still within normal range," CheXpert and T-auto mistakenly label cardiomegaly as positive, while CheXbert correctly labels cardiomegaly as negative. On another example containing the phrase "interval removal of PICC lines", CheXpert and T-auto detect "PICC lines" as an indication of a support device but are unable to detect the negation indicated by "removal", which CheXbert correctly does.

Additionally, CheXbert is able to correctly detect expressions of uncertainty that both CheXpert and T-auto mislabel. On an example containing "new bibasilar opacities, which given the clinical history are suspicious for aspiration," CheXbert correctly identifies lung opacity as positive while CheXpert and T-auto incorrectly detect uncertainty (associating "suspicious" as a descriptor of "opacities"). More examples which CheXbert correctly labels but CheXpert and T-auto mislabel can be found in Table 9.10 of the Appendix. A selected example is shown in the last row of Table 9.3.

9.5.3 Report Changes with Backtranslation

We analyze the phrasing and vocabulary changes that backtranslation introduces into the reports. Backtranslation frequently rephrases text. For instance, the sentence "redemonstration of multiple right-sided rib fractures" is backtranslated to "redemonstration of several rib fractures of the right side". Backtranslation also introduces some error: the phrase "left costophrenic angle" is backtranslated to "left costophrine angle" ("costophrine" is not a word), and the phrase "left anterior chest wall pacer in place" is backtranslated to "pacemaker on the left front of the chest wall", which omits the critical attribute of being in place. In many examples, the backtranslated text paraphrases medical vocabulary into possible semantic equivalents: "cutaneous" becomes "skin", "left clavicle" becomes "left collarbone", "osseous" becomes "bone" or "bony", "anterior" becomes "front", and "rib fracture" becomes "broken ribs". More backtranslations with analyses are provided in Table 9.11 of the Appendix. Additionally, a physician validated that the backtranslation outputs used correct radiology language and maintained the semantics of the original report. The results are provided in Table 9.4 of the Appendix.

9.6 Limitations

Our study has several limitations. First, our hybrid/auto approaches require an already-existing labeler. Second, our report labeler has a maximum input token size of 512 tokens, but this may be easily extended to work with longer medical/radiology reports. In the CheXpert dataset, we found that only 3 of the 190,460 report impressions were longer than 512 tokens. Third, our task is limited to the 14 observations labeled for, and we do not test for the model's ability to label rarer conditions. However, CheXbert can mark No Finding as blank, which can indicate the presence of another condition if the other 13 conditions are also blank. Fourth, the ground truth labels for the MIMIC-CXR test set were determined by a single board-certified radiologist, and the use of more radiologists could demonstrate a truer comparison to the radiologist benchmark. Fifth, while we do test performance on a dataset from an institution unseen in training, additional datasets across institutions could be useful in further establishing the model's ability to generalize.

9.7 Conclusion

In this study, we propose a simple method for combining existing report labelers with hand-annotations for accurate radiology report labeling. In this method, a biomedically pretrained BERT model is first trained on the outputs of a labeler, and then further fine-tuned on the manual annotations, the set of which is augmented using backtranslation. We report five findings on our resulting model, CheXbert. First, we find that CheXbert outperforms models trained only on radiologist-labeled reports, or only on the existing labeler's outputs. Second, we find that CheXbert outperforms the BERT-based model not pretrained on biomedical data. Third, we find that CheXbert outperforms models which do not use backtranslation. Fourth, we find that CheXbert outperforms the previous best labeler, CheXpert (which was rules-based), with an improvement of 0.055 (95% CI 0.039, 0.070) on the F1 metric; we also find that the best model trained only on manually labeled radiology reports (Tblue-rad-bt) performs at least as well as the CheXpert labeler. Fifth, we find that CheXbert is 0.007 F1 points from the radiologist performance benchmark, suggesting that the gap to ceiling performance is narrow.

We expect this method of training medical report labelers is broadly useful within the medical domain, where collection of expert labels can produce a small set of high quality labels, and existing feature engineered labelers can produce labels at scale. Extracting highly accurate labels from medical reports by taking advantage of both sources can enable many important downstream tasks, including the development of more accurate and robust medical imaging models required for clinical deployment.

9.8 Appendix

9.8.1 Physician validation of backtranslation quality

Table 9.4: Physician validation of backtranslation output quality on a set of 100 randomly sampled reports from the CheXpert manual set and their backtranslations.

Score	Valid radiology language	Preserves semantic information
1	6	14
2	48	26
3	46	60

Although the CheXbert model shows empirical improvements using backtranslated reports, backtranslation can introduce additional noise into the reports. A physician validated the quality of the backtranslation outputs. For this experiment, we randomly sampled 100 reports from the CheXpert manual set. The physician read each original report and its backtranslation, and assigned a score for whether the backtranslation a) used valid radiology language, and b) maintained the semantics of the report. For each of tasks a) and b), the expert assigned a score of 1 (worst), 2 or 3 (highest).

For task a), a score of 3 means the backtranslation contained near-perfect radiology language, a 2 means the backtranslation had only minor deviations from valid radiology language, and 1 means the backtranslation had a major deviation from valid radiology language. For task b), a score of 3 means the backtranslation fully preserved the semantics of the original, a 2 means the backtranslation contained minor semantic errors, and a 1 means the backtranslation had a major change or loss of semantic information compared to the original report.

9.8.2 Additional results

Table 9.5: After removing duplicate reports for the same patient from the CheXpert dataset (excluding the CheXpert manual set), we are left with a total of 190,460 reports. Labels for these reports are provided by the CheXpert labeler. The class prevalences of this set are displayed for each medical condition.

Condition	**Positive**	**Negative**	**Uncertain**	**Blank**
Atelectasis	29,818 (15.66%)	1,018 (0.53%)	29,832 (15.66%)	129,792 (68.15%)
Cardiomegaly	23,302 (12.23%)	7,809 (4.10%)	6,682 (3.51%)	152,667 (80.16%)
Consolidation	12,977 (6.81%)	19,397 (10.18%)	24,345 (12.78%)	133,741 (70.22%)
Edema	49,725 (26.11%)	15,867 (8.33%)	11,746 (6.17%)	113,122 (59.39%)
Enlarged Cardiomed.	9,129 (4.79%)	15,165 (7.96%)	10,278 (5.40%)	155,888 (81.85%)
Fracture	7,364 (3.87%)	1,960 (1.03%)	488 (0.26%)	180,648 (94.85%)
Lung Lesion	6,955 (3.65%)	758 (0.40%)	1,084 (0.57%)	181,663 (95.38%)
Lung Opacity	94,156 (49.44%)	5,006 (2.63%)	4,404 (2.31%)	86,894 (45.62%)
No Finding	16,795 (8.82%)	NA	NA	173,665 (91.18%)
Pleural Effusion	77,028 (40.44%)	25,097 (13.18%)	9,565 (5.02%)	78,770 (41.36%)
Pleural Other	2,481 (1.30%)	210 (0.11%)	1,801 (0.95%)	185,968 (97.64%)
Pneumonia	4,647 (2.44%)	1,851 (0.97%)	15,907 (8.35%)	168,055 (88.24%)
Pneumothorax	17,688 (9.29%)	47,566 (24.97%)	2,704 (1.42%)	122,502 (64.32%)
Support Devices	107,601 (56.50%)	5,319 (2.79%)	910 (0.48%)	76,630 (40.23%)

Table 9.6: Dev set F1 scores for all our models. The dev set for all rad models and T-hybrid consists of 250 randomly sampled reports from the CheXpert manual set. The dev set for T-auto is a random 15% split of the CheXpert dataset. The dev set for all models using backtranslation is obtained by augmenting the 250 randomly sampled reports from the CheXpert manual set by backtranslation. Tblue-hybrid-bt is first trained on labels generated by the CheXpert labeler, and then fine-tuned on radiologist labels augmented by backtranslation. Before fine-tuning on radologist labels, it obtains an F1 of 0.977 on the 15% dev split of the CheXpert dataset.

	Model	F1
Training Strategy	T-rad	0.848
	T.cls-rad	0.411
	T.token-rad	0.518
	T-auto	0.977
	T-hybrid	0.904
Biomedical Representations	Tbio-rad	0.760
	Tclinical-rad	0.802
	Tblue-rad	0.866
With Backtranslation Augmentation	T-rad-bt	0.846
	T-hybrid-bt	0.905
	Tblue-rad-bt	0.865
	Tblue-hybrid-bt (CheXbert)	0.912

Table 9.7: The differences in the number of times labels were correctly assigned by one model versus another model. For example, in the first column named "T-auto > CheXpert," we report the difference between the number of times T-auto correctly classifies a label and the number of times CheXpert correctly classifies a label. We record the differences between a pair of models by category (blank, positive, negative, uncertain) and by total. These occurrences are obtained on the MIMIC-CXR test set.

	T-auto > CheXpert	**CheXbert > CheXpert**	**CheXbert > Radiologist**
Blank	0	29	56
Positive	-22	11	56
Negative	14	45	9
Uncertain	16	46	-3
Total	8	131	118

Table 9.8: Examples where T-auto correctly assigns a label while CheXpert misassigns that label on the CheXpert manual set. We include speculative reasoning for the classifications.

Example & Labels	**Reasoning**
...redemonstration of diffuse nodular air space opacities which are unchanged from prior examination **which may represent air space pulmonary edema versus infection**, as clinically correlated... Edema CheXpert: Positive ✗ T-auto: Uncertain ✓	T-auto appears to detect uncertainties indicated by words like "may" and "versus" on conditions. In this case, this phrase did not match an uncertainty detection rule in the CheXpert classifier.
... there has been interval development of left basilar patchy airspace opacity, which likely represents atelectasis, **although consolidation cannot be entirely excluded**... Consolidation CheXpert: Positive ✗ T-auto: Uncertain ✓	Unlike CheXpert, T-auto correctly detects uncertainty conveyed in the phrase "cannot be entirely excluded".
1. no radiographic evidence of acute cardiopulmonary disease. 2. **no evidence of pneumothorax or bony fracture**. Fracture CheXpert: Positive ✗ T-auto: Negative ✓	In this example, T-auto is able to detect a negation indicated by "no evidence of". CheXpert is not able to pick up this negation construction as part of its ruleset.

Table 9.9: Examples where CheXpert correctly assigns a label while T-auto misassigns that label on the CheXpert manual set. We include speculative reasoning for the classifications.

Example & Labels	**Reasoning**
...2.**mild cardiomegaly**. persistent small bilateral pleural effusions, left greater than right... Cardiomegaly CheXpert: Positive ✓ T-auto: Uncertain ✗	T-auto mistakenly labels "mild cardiomegaly" as uncertain for cardiomegaly.
...2.there are diffuse increased interstitial markings and prominence of the central vasculature, **consistent with acute and/or chronic pulmonary edema**... Edema CheXpert: Positive ✓ T-auto: Uncertain ✗	T-auto may have incorrectly detected uncertainty from "and/or," which is a conjunction between "acute" and "chronic".

Table 9.10: Examples where CheXbert correctly assigns a label while both T-auto and CheXpert misassign that label on the MIMIC-CXR test set. We include speculative reasoning for the classifications.

Example & Labels	**Reasoning**
New bibasilar opacities, which given the clinical history are suspicious for aspiration, possibly developing pneumonia. Lung Opacity CheXpert: Uncertain ✗ T-auto: Uncertain ✗ CheXbert: Positive ✓	The word "suspicious" does not modify "opacities" in this sentence. Although CheXbert correctly identifies this, CheXpert and T-auto misclassify the "opacities" as uncertain.
...Coalescent areas in the left upper and lower zones **could well reflect regions of consolidation**. The right lung is essentially clear... Consolidation CheXpert: Positive ✗ T-auto: Positive ✗ CheXbert: Uncertain ✓	CheXbert correctly detects that consolidation is uncertain, as indicated by the phrase "could well reflect".
Removal of dialysis catheter with no evidence of pneumothorax. **Heart is mildly enlarged** and is accompanied by vascular engorgement and new septal lines consistent with interstitial edema... Cardiomegaly CheXpert: Blank ✗ T-auto: Blank ✗ CheXbert: Positive ✓	Due to a ruleset limitation, CheXpert only looks at "the heart" or "heart size" but not "heart" independently when checking for mentions of cardiomegaly. However, CheXbert recognizes mentions of cardiomegaly implied by phrases like "heart is mildly enlarged".
No previous images. There is hyperexpansion of the lungs suggestive of chronic pulmonary disease. Prominence of engorged and ill-defined pulmonary vessels is consistent with the clinical diagnosis of pulmonary vascular congestion, though in the absence of previous images it is difficult to determine whether any this appearance could reflect underlying chronic pulmonary disease. **The possibility of supervening consolidation would be impossible to exclude** on this single study, especially without a lateral view. No evidence of pneumothorax. Consolidation CheXpert: Positive ✗ T-auto: Positive ✗ CheXbert: Uncertain ✓	CheXbert correctly detects uncertainty for consolidation indicated by the word "possibility". Both T-auto and CheXpert misclassify consolidation.

Example (cont.) & Labels (cont.)	**Reasoning (cont.)**
1. Left suprahilar opacity and fiducial seeds are again seen, although appears slightly less prominent/small in size, although as mentioned on the prior study, could be further evaluated by chest CT or PET-CT. 2. **Right hilum appears slightly more prominent** as compared to the prior study, which may be due to patient positioning, although increased right hilar lymphadenopathy is not excluded. Enlarged Cardiomediast. CheXpert: Blank ✗ T-auto: Blank ✗ CheXbert: Positive ✓	The right hilum appearing more prominent is an indicator of enlarged cardiomediastinum, which is clinically understood. If the hilum is growing, then the entire mediastinum is growing. Although both CheXpert and T-auto mislabeled this report impression, CheXbert successfully labeled it positive for enlarged cardiomediastinum.
As compared to the previous radiograph, there is no relevant change. The reduced volume of the right hemithorax with areas of lateral pleural thickening. The areas of pleural thickening are constant, size and morphology. Unchanged perihilar areas of fibrosis. **Unchanged size and aspect of the cardiac silhouette**, no pathologic changes in the left lung. Cardiomegaly CheXpert: Positive ✗ T-auto: Positive ✗ CheXbert: Uncertain ✓	CheXbert correctly identifies uncertainty, as the cardiac silhouette is "unchanged," which means that it cannot be labeled positive or negative without additional information regarding the previous state. Both CheXpert and T-auto incorrectly label this example as positive for cardiomegaly.
AP chest compared to ___: Small-to-moderate left pleural effusion has increased slightly over the past several days. Moderate enlargement of the cardiac silhouette accompanied by mediastinal vascular engorgement is also slightly more pronounced. Pulmonary vasculature is engorged but there is no edema. Consolidation has been present without appreciable change in the left lower lobe since at least ___. **Mediastinum widened** at the thoracic inlet by a combination of tortuous vessels and mediastinal fat deposition. Right jugular introducer ends just above the junction with left brachiocephalic vein. Enlarged Cardiomediast. CheXpert: Blank ✗ T-auto: Blank ✗ CheXbert: Positive ✓	CheXbert correctly identifies enlarged cardiomediastinum from the phrase "mediastinum widened," which is a slightly different way of describing enlarged cardiomediastinum that CheXpert and T-auto both miss.

Example (cont.) & Labels (cont.)	**Reasoning (cont.)**
Moderately enlarged heart size, stable since ___. **No findings concerning for pulmonary edema** or pneumonia. Edema CheXpert: Uncertain ✗ T-auto: Uncertain ✗ CheXbert: Negative ✓	Unlike T-auto and CheXpert, CheXbert correctly labels edema as negative, presumably understanding that the initial phrase "no findings" applies to both edema and pneumonia.
AP chest compared to ___ and ___: As far as I can tell, given the severe anatomic distortion of the chest cage and its contents, lungs were clear on ___. Small region of opacification may have been developing lateral to the left hilus on ___, and today there is a suggestion of some new opacification at the base of the lung, but these observations are far from certain. I am not even confident that conventional radiographs, should the patient be able to cooperate for them, would clarify the issue. CT scanning, if feasible, would certainly confirm if the lungs are clear, but in the absence of a baseline study **it might be difficult to distinguish atelectasis from pneumonia**. Pleural effusion is minimal if any. Heart is probably not enlarged. Nasogastric tube is looped in the stomach. Right PIC line ends in the mid SVC. No pneumothorax. Atelectasis CheXpert: Positive ✗ T-auto: Positive ✗ CheXbert: Uncertain ✓	The report states that "it might be difficult to distinguish atelectasis from pneumonia" which indicates uncertainty, and this is correctly identified by CheXbert. CheXpert and T-auto simply label atelectasis as positive.
Two frontal views of the chest show new mild interstitial pulmonary edema. Interval increase in mediastinal caliber therefore is probably due to distention of mediastinal veins. **Heart size is slightly larger but still within normal range**. Pleural effusions are minimal, if any. No focal pulmonary abnormality. No pneumothorax. ET tube is in standard placement and a nasogastric tube passes below the diaphragm and out of view. Cardiomegaly CheXpert: Positive ✗ T-auto: Positive ✗ CheXbert: Negative ✓	Although CheXpert and T-auto mistakenly label cardiomegaly as positive given the phrase the "heart is slightly larger," the following phrase "but still within normal range" implies that cardiomegaly is negative. CheXbert correctly classifies this example as negative for cardiomegaly.

Example (cont.) & Labels (cont.)	**Reasoning (cont.)**
As compared to the previous radiograph, the pre-existing right upper lobe **pneumonia is completely resolved**. The pre-existing signs of mild fluid overload, however, are still present. The pre-existing cardiomegaly is unchanged. Several calcified lung nodules are also unchanged. Unchanged alignment of the sternal wires. No acute pneumonia, no pleural effusions. Pneumonia CheXpert: Positive ✗ T-auto: Positive ✗ CheXbert: Negative ✓	CheXbert correctly labels pneumonia as negative, as implied by the phrase " pneumonia is completely resolved," while CheXpert and T-auto both mislabel pneumonia as positive.
Subsegmental right lung base atelectasis. Increasing loss of vertebral body height at T11. Stable L1 compression fracture. Right shoulder humeral DJD. **Interval removal of PICC lines**. Support Devices CheXpert: Positive ✗ T-auto: Positive ✗ CheXbert: Negative ✓	CheXbert, presumably using a semantic understanding of the word "removal", correctly labels support devices as negative. CheXpert and T-auto pick up on "PICC lines" but do not detect the negation. Both incorrectly label support devices as positive.
AP chest compared to ___: Small-to-moderate left pleural effusion has increased slightly over the past several days. Moderate enlargement of the cardiac silhouette accompanied by mediastinal vascular engorgement is also slightly more pronounced. Pulmonary vasculature is engorged but there is no edema. Consolidation has been present without appreciable change in the left lower lobe since at least ___. Mediastinum widened at the thoracic inlet by a combination of tortuous vessels and mediastinal fat deposition. **Right jugular introducer ends just above the junction with left brachiocephalic vein**. Support Devices CheXpert: Blank ✗ T-auto: Blank ✗ CheXbert: Positive ✓	A jugular introducer is a support device that wasn't included in CheXpert's list of mentions for support devices. Consequently CheXpert and T-auto, which trains on CheXpert labels, both incorrectly label support devices as blank. CheXbert, however, correctly labels support devices as positive.
1. Interval removal of the sternal wires with placement of new sternal closure devices, mediastinal staples and tubes. Lungs are well inflated with linear streaky opacities seen at the left base likely representing scarring and/or subsegmental atelectasis. No evidence of pulmonary edema, pneumothorax, pleural effusions or focal airspace consolidation to suggest pneumonia. **Slight lucency at the left apex is felt to be related to underlying emphysema rather than representing a pneumothorax**. Pneumothorax CheXpert: Positive ✗ T-auto: Positive ✗ CheXbert: Negative ✓	CheXbert correctly labels pneumothorax as negative, as the radiologist notes that the observation is related to emphysema rather than pneumothorax. In this complex negation, T-auto and CheXpert incorrectly label pneumothorax as positive.

Table 9.11: Examples of additional data samples generated using backtranslation on radiologist-annotated reports from the CheXpert manual set. Augmenting our relatively small set of radiologist-annotated reports with backtranslation proved useful in improving performance of our labeler on the MIMIC-CXR test set.

Original Report	**Backtranslation**	**Changes**
1. marked cardiomegaly with a configuration that raises concern for a pericardial effusion. possible mild edema. 2. healed left-sided rib fractures.	1. pronounced cardiomegaly with a configuration that raises concerns about a pericardial effusion. possible mild edema. 2. healed left-sided rib fractures.	"marked" is changed to the synonym "pronounced", and "raises concern for" is rephrased as "raises concerns about".
1. redemonstration of right side pleural effusion and bibasilar atelectasis unchanged from comparison. 2. redemonstration of multiple right-sided rib fractures.	1. redemonstration of the pleural effusion of the right side and the bibasilar atelectasia unchanged compared to the comparison. 2. redemonstration of several rib fractures of the right side.	"right side pleural effusion" is rephrased as "pleural effusion of the right side", "unchanged from comparison" is rephrased to "compared to the comparison" and "multiple right-sided rib fractures" is rephrased as "several rib fractures of the right side". However, "atelectasis" is incorrectly changed to "atelectasia".
1. single ap portable semiupright view of the chest demonstrates no change in medical support devices. 2. persistent dense retrocardiac opacity and small to moderate left pleural effusion present. right perihilar opacity appears resolved. 3. stable cardiomediastinal silhouette. no pulmonary edema. 4. multilevel degenerative changes of the spine.	1. single ap portable semiupright view of the breast showing no change in medical aids. 2. persistent dense retrocardiac opacity and small to moderate left pleural effusion presented. right perihilar opacity appears resolved. 3. stable cardiomediastinal silhouette. no pulmonary edema. 4. multi-level degenerative changes of the spine.	"semiupright" becomes "semiupright", "medical support devices" is changed to "medical aids", "present" is changed to "presented" and "multilevel" is changed to "multi-level". However, "chest" is incorrectly changed to "breast".

Original Report (cont.)	**Backtranslation (cont.)**	**Changes (cont.)**
1. single frontal view of the chest demonstrates a surgical drain projecting over the neck, a tracheostomy tube, a feeding tube which extends below the diaphragm and beyond the inferior margin of the film. cutaneous staples project over the left clavicle, and surgical clips are seen within the left neck. no evidence of pneumothorax. 2. a dense retrocardiac opacity may represent atelectasis versus consolidation.,small bilateral pleural effusions are present. A convex opacity at the right paratracheal region is of uncertain significance; recommend upright pa and lateral for further evaluation when the patient is able. 3. the cardiomediastinal silhouette and pulmonary vasculature are unremarkable.	1. a single frontal view of the breast shows a surgical drain extending over the neck, a tracheostolic tube, a feeding tube extending under the diaphragm and over the lower edge of the film. skin clamps protrude over the left collarbone, and surgical clips are visible in the left cervical area. no indication of pneumothorax. 2. dense retrocardiac opacity may represent ateltasia versus consolidation. small bilateral pleural effusions are present. convex opacity in the right paratracheal area is of uncertain importance; recommend upright pa and lateral for further assessment if the patient is able to do so. 3. the cardiastinal silhouette and pulmonary vasculature are unobtrusive.	"demonstrates a surgical drain projecting over" rephrased to "shows a surgical drain extending over", "a feeding tube which extends below the diaphragm and beyond the inferior margin of the film" rephrased to "a feeding tube extending under the diaphragm and over the lower edge of the film", "surgical clips are seen within the left neck" changed to the semantically equivalent "surgical clips are visible in the left cervical area", "region is of uncertain significance" rephrased as "area is of uncertain importance", "further evaluation when the patient is able" is rephrased as "further assessment if the patient is able to do so", and "pulmonary vasculature are unremarkable" is changed to the semantically close "pulmonary vasculature are unobtrusive". However "chest" incorrectly changed to "breast, "tracheostomy tube" incorrectly changed to "tracheostolic tube", "cutaneous staples project over the left clavicle" changed to the semantically similar "skin clamps protrude over the left collarbone", but "skin clamps" is suboptimal, "atelectasis" incorrectly changed to "ateltasia", "cardiomediastinal" is incorrectly changed to "cardiastinal".

Original Report (cont.)	**Backtranslation (cont.)**	**Changes (cont.)**
1. single ap view of the chest demonstrates hyperinflation of the lungs. 2. there are prominent interstitial opacities which are stable. there is a residual tiny left apical pneumothorax without interval change. 3. cardiomediastinal silhouette is stable. 4. there is nonvisualization of the left costophrenic angle limiting its evaluation and if concerned, repeat study can be performed.	1. a single view of the breast shows hyperinflation of the lungs. 2. there are prominent interstitial opacities that are stable. there is a remaining tiny left apical pneumothorax without interval change. 3. the cardiomediastinal silhouette is stable 4. there is no visualization of the left costophrine angle that restricts its assessment, and if affected, a repeat study can be conducted.	"demonstrates hyperinflation" is rephrased as "shows hyperinflation", "residual" is changed to the synonym "remaining", and "angle limiting its evaluation and if concerned, repeat study can be performed" is rephrased to "angle that restricts its assessment, and if affected, a repeat study can be conducted". The replacement of "concerned" with "affected" appears suboptimal. However, "ap" is incorrectly removed from the phrase "single ap view of the chest", "chest" is incorrectly changed to "breast", and "costophrenic angle" is incorrectly changed to "cortophrine angle".

Chapter 10

Improving Label Quality By Addressing Distributional Shift

As seen in the previous chapter, automatic extraction of medical conditions from free-text radiology reports is critical for supervising computer vision models to interpret medical images. In this chapter, we show that radiologists labeling reports significantly disagree with radiologists labeling corresponding chest X-ray images, which reduces the quality of report labels as proxies for image labels. We develop and evaluate methods to produce labels from radiology reports that have better agreement with radiologists labeling images. Our best performing method, called VisualCheXbert, uses a biomedically-pretrained BERT model to directly map from a radiology report to the image labels, with a supervisory signal determined by a computer vision model trained to detect medical conditions from chest X-ray images. We find that VisualCheXbert outperforms an approach using an existing radiology report labeler by an average F1 score of 0.14 (95% CI 0.12, 0.17). We also find that VisualCheXbert better agrees with radiologists labeling chest X-ray images than do radiologists labeling the corresponding radiology reports by an average F1 score across several medical conditions of between 0.12 (95% CI 0.09, 0.15) and 0.21 (95% CI 0.18, 0.24).

This chapter is based on [110].

10.1 Introduction

Because manually annotating a large number of medical images is costly [155, 29, 1, 84, 211, 238, 54], an appealing solution is the use of automatic labelers to extract labels from medical text reports that accompany the images. On the task of chest X-ray interpretation, high-performing vision models have been successfully trained [184, 167, 255, 185, 226, 187] on large, publicly available chest X-ray datasets [106, 111, 239, 168] labeled by automated radiology report labelers [106, 162, 143, 213].

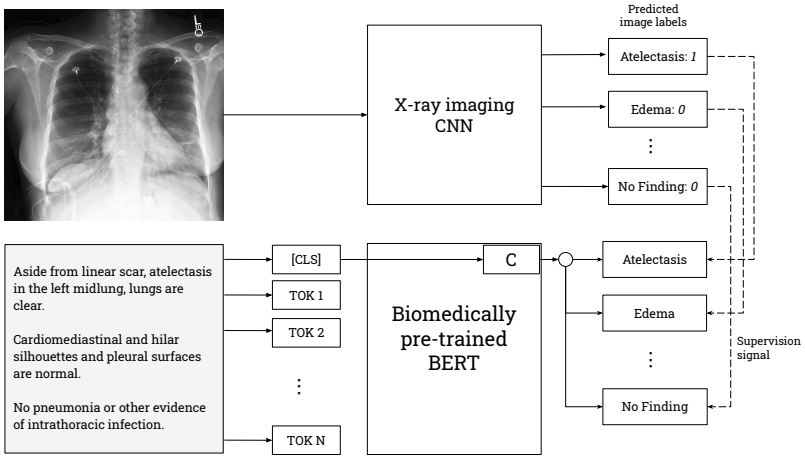


Figure 10.1: The VisualCheXbert training procedure. VisualCheXbert uses a biomedically-pretrained BERT model to directly map from a radiology report to the labels obtained by a radiologist interpreting the associated X-ray image. The training procedure for VisualCheXbert is supervised by a computer vision model trained to detect medical conditions from chest X-ray images.

However, training these vision models on labels obtained from reports assumes that the report labels are good proxies for image labels. Prior work has found that report labels may not accurately reflect the visual content of medical images [157, 158, 225].

We investigate this assumption in the setting of automated chest X-ray labeling and develop methods to produce labels from radiology reports that better agree with radiologists labeling the corresponding X-ray images. Our primary contributions are:

1. We quantify the agreement between radiologists labeling reports and radiologists labeling images across several medical conditions. We find that there is significant disagreement between board-certified radiologists when labeling a chest X-ray image and when labeling the corresponding radiology report.

2. Upon board-certified radiologist review of examples of disagreements between radiologists labeling reports and radiologists labeling images, we find various reasons for disagreement related to (a) label hierarchy relationships, (b) access to clinical history, (c) the use of the *Impression* and *Findings* section of radiology reports, and (d) the inherent noise of the labeling task.

3. We find many significant relationships between presence of conditions labeled using reports and presence of conditions labeled using images. We report and clinically interpret various radiology report labels that increase (or decrease) the odds of particular conditions in an image with statistical significance.

4. We learn to map textual radiology reports directly to the X-ray image labels. Our best performing method, called *VisualCheXbert*, uses a biomedically-pretrained BERT model to directly map from a radiology report to the image labels. We find that VisualCheXbert better agrees with radiologists labeling chest X-ray images than do radiologists labeling the corresponding radiology reports by an average F1 score across several medical conditions of between 0.12 (95% CI 0.09, 0.15) and 0.21 (95% CI 0.18, 0.24). We also find that VisualCheXbert outperforms an approach using the CheXpert radiology report labeler [106] by an average F1 score of 0.14 (95% CI 0.12, 0.17).

We expect that our methods of addressing the discrepancy between medical report labels and image labels are broadly useful across the medical domain and may facilitate the development of improved medical imaging models.

10.2 Data

We made use of two large publicly available datasets of chest X-rays: CheXpert [106] and MIMIC-CXR [111]. For both datasets, we use the *Impression* section of the radiology reports, which summarizes the key findings in the radiographic study. Each of the X-rays in these datasets was labeled for 14 commonly occurring medical conditions. CheXpert consists of 224,316 chest radiographs, with labels generated from the corresponding radiology report impression by the automatic, rules-based CheXpert labeler. Given a radiology report impression as input, the CheXpert labeler labels each medical condition (except "No Finding") as "positive", "negative", "uncertain" or "blank". A "blank" label is produced by the CheXpert labeler if the condition was not mentioned at all in the report impression. If the condition was mentioned but its presence was negated, a "negative" label is produced. If the condition was mentioned but its presence was uncertain, an "uncertain" label is produced. For "No Finding", the CheXpert labeler only produces "positive" or "blank" labels. "No Finding" is only labeled as "positive" if no medical abnormality whatsoever was mentioned in the report impression. The MIMIC-CXR dataset consists of 377,110 chest X-rays and their corresponding radiology reports, and it has also been labeled by the CheXpert labeler.

The CheXpert dataset contains a separate set of 200 chest X-ray studies called the "CheXpert validation set" and another set of 500 chest X-ray studies called the "CheXpert test set". The CheXpert validation set is labeled by the majority vote of 3 board-certified radiologists examining the X-ray images and labeling each of the 14 conditions as "positive" or "negative", similar to the image ground truth on the CheXpert test set, which is described below. No radiologist report labels are obtained for the validation set.

The CheXpert test set, which was collected by [106], is labeled by radiologists in two distinct ways:

Table 10.1: Agreement between radiologists looking at reports and radiologists looking at the corresponding X-ray images. The high and low scores are obtained by mapping uncertain labels in the radiologist report labels to the image ground truth labels and the opposite of the image ground truth labels respectively.

Condition (n = # positive)	**Low F1**	**High F1**	**Low Kappa**	**High Kappa**
Atelectasis (n=153)	0.230	0.595	-0.014	0.457
Cardiomegaly (n=151)	0.422	0.463	0.290	0.344
Edema (n=78)	0.453	0.581	0.335	0.492
Pleural Effusion (n=104)	0.638	0.710	0.511	0.613
Enlarged Cardiom. (n=253)	0.089	0.208	-0.053	0.097
Lung Opacity (n=264)	0.683	0.686	0.401	0.405
Support Devices (n=261)	0.863	0.863	0.737	0.737
No Finding (n=62)	0.381	0.381	0.292	0.292
Average	0.470	0.561	0.312	0.430
Weighted Average	0.492	0.575	0.320	0.427

Image ground truth 5 board-certified radiologists looked at each X-ray image and labeled each of the 14 conditions as "positive" or "negative". The final label is their majority vote. These radiologists only observed the X-ray images and did not have access to the radiology report or patients' historical records at the time of image labeling.

Radiologist report labels A board-certified radiologist looked at each radiology report impression corresponding to the X-rays and labeled each of the 14 conditions as being "positive", "negative", "uncertain", or "blank". This radiologist did not observe any X-ray images. A condition was labeled as "blank" if it was not at all mentioned in the report impression. If the condition was mentioned but its presence in the chest X-ray was negated, then the condition was labeled as "negative". If the condition was mentioned but its presence was uncertain, it was labeled as "uncertain".

10.3 Evaluation

We only evaluate our models on medical conditions for which at least 50 out of the 500 chest X-ray studies in the CheXpert test set were marked positive by the radiologists labeling the X-ray images (image ground truth). These conditions, which we refer to as the *evaluation conditions*, are: Atelectasis, Cardiomegaly, Edema, Pleural Effusion, Enlarged Cardiomediastinum, Lung Opacity, Support Devices, and No Finding. We evaluate models using the average and weighted average of the F1 score across conditions on the CheXpert test set with the image ground truth. To compute the weighted average, each condition is weighted by the portion of positive labels for that condition in the CheXpert test set.

Table 10.2: Clinical explanations of disagreements between radiologists looking at reports and radiologists looking at images on the CheXpert test set. Given access to the X-ray image, the full radiology report, the radiology report impression, the radiology report labels, and the image ground truth, a board-certified radiologist explained disagreements between radiologist report labels and the image ground truth. We show select examples with explanations in this table.

Report Impression and Labels	**Clinical Explanation**
1. single ap upright view of the chest showing a mildly increased opacity at the left lung base that could represent atelectasis versus consolidation. *Cardiomegaly* Radiologist Report Label: Negative Image Ground Truth: Positive	The radiologist looking at the report marks Cardiomegaly as negative as it is not mentioned in the report. Since the image is a Intensive Care Unit (ICU) film and cardiomegaly is not a clinically relevant condition for the population selected for in ICU films, the presence of cardiomegaly was never mentioned in the report, resulting in the discrepancy between radiologists looking at the report and radiologists looking at the image.
1. pulmonary vascular congestion. left lower lobe opacity comp--atible with atelectasis and/or consolidation. *Cardiomegaly* Radiologist Report Label: Negative Image Ground Truth: Positive	Although cardiomegaly was mentioned in the radiology report "Findings" section, cardiomegaly was not mentioned in the report "Impression". Since the radiologist looking at the report only had access to the "Impression" section, they labeled Cardiomegaly as negative when it was actually present in the image.
1. decreased pulmonary edema. stable bilateral pleural effusions and bibasilar atelectasis. *Edema* Radiologist Report Label: Positive Image Ground Truth: Negative	The phrase "decreased pulmonary edema" shows that the radiologist writing the report had relevant clinical context, as the edema has "decreased" compared to a previous report or image. However, the radiologist looking at the image does not have this clinical context, resulting in a discrepancy.
1. single frontal radiograph of the chest is limited secondary to poor inspiration and rotation. 2. cardiac silhouette is partially obscured secondary to rotation. lungs demonstrate bibasilar opacities, likely reflecting atelectasis. possible small right pleural effusion. no pneumotho--rax. 3. visualized osseous structures and soft tissues unremarkable. *Pleural Effusion* Radiologist Report Label: Positive Image Ground Truth: Negative	The phrase "possible small right pleural effusion" indicates the uncertainty regarding the presence of pleural effusion. This natural uncertainty may explain the disagreement between radiologists looking at the image and radiologists looking at the report. On review, it was noted that pleural effusion was borderline in this example.
1. crowding of the pulmonary vasculature. cannot exclude mild interstitial pulmonary edema. 2. no focal air space consolidation. the cardiomediastinal silhou--ette appears grossly within normal limits. *Pleural Effusion* Radiologist Report Label: Negative Image Ground Truth: Positive	Upon review by a board-certified radiologist, there was an error in the radiology report, which did not mention the presence of pleural effusion. The error in the report itself may explain the disagreement between the image and report labels.

10.4 Experiments

10.4.1 Radiologist Report / Image Labeling Agreement

We first investigate the extent of the disagreement between board-certified radiologists when labeling a chest X-ray image and when labeling the corresponding radiology report.

Method

We compute the level of agreement between radiologists labeling X-ray images and radiologists labeling the corresponding radiology reports on the CheXpert test set. The CheXpert test set contains a set of labels from radiologists labeling X-ray images as well as another set of labels from

radiologists labeling the corresponding radiology reports. Using the labels from X-ray images as the ground truth, we compute Cohen's Kappa [50] as well as the F1 score to measure the agreement between these two sets of labels. To compare the radiologist report labels to the image ground truth labels, we convert the radiologist report labels to binary labels as follows. We map the blank labels produced for the radiology report to negative labels. We map uncertain labels to either the image ground truth label or the opposite of the image ground truth label, and we record the results for both these strategies to obtain "Low F1", "High F1", "Low Kappa", and "High Kappa" scores. The low and high scores represent the most pessimistic and optimistic mapping of the uncertainty labels.

Results

We find that there is significant disagreement, which is indicated by low Kappa and F1 scores for almost all conditions evaluated. For example, Enlarged Cardiomediastinum and No Finding have a relatively small "High Kappa" score of 0.097 and 0.292 and a "High F1" score of 0.208 and 0.381, indicating high levels of disagreement even when assuming the most optimistic mapping of the uncertainty labels. Atelectasis, Cardiomegaly, Edema, Pleural Effusion, and Lung Opacity also have a low "High Kappa" score of 0.457, 0.344, 0.492, 0.613, and 0.405 respectively and a "High F1" score of 0.595, 0.463, 0.581, 0.710, and 0.686 respectively. Support Devices has the highest Kappa score, with a "High Kappa" of 0.737, and the highest F1 score, with a "High F1" of 0.863. The average Kappa score is between 0.312 and 0.430, and the average F1 score is between 0.470 and 0.561. The low and high F1 / Kappa scores for the evaluation conditions are shown in Table 10.1.

10.4.2 Disagreement Reasons

We investigate why there is disagreement between board-certified radiologists when labeling a chest X-ray image and when labeling the corresponding radiology report.

Method

A board-certified radiologist was given access to the chest X-ray image, the full radiology report, the radiology report impression section, the image ground truth across all conditions, and the radiologist report labels across all conditions for each of the 500 examples in the CheXpert test set. The radiologist then explained examples where radiologists labeling reports disagree with radiologists labeling X-ray images. We also calculated the counts of disagreements between radiologists labeling reports and radiologists labeling X-ray images for each condition on the CheXpert test set. A board-certified radiologist explained why there were large numbers of disagreements on certain conditions.

Results

We find various reasons why radiologists labeling reports might disagree with radiologists labeling images. First, there is a difference between the setup of the report labeling and image labeling tasks related to the label hierarchy. On the report labeling task on the CheXpert test set, radiologists were instructed to label only the most specific condition as positive and leave parent conditions blank. For example, although Lung Opacity is a parent condition of Edema, a radiologist marking a report as positive for Edema would leave Lung Opacity blank. Blank report labels are typically mapped to negative image labels. However, radiologists labeling images label each condition as positive or negative independent of the presence of other conditions. Second, radiologists labeling reports have access to clinical report history, which biases radiologists towards reporting certain conditions in reports while a radiologist labeling the image may not observe the condition on the image. [28] explain biases from clinical history in terms of framing bias, where the presentation of the clinical history can lead to different diagnostic conclusions, and attribution bias, where information in the clinical history can lead to different diagnostic conclusions. Third, radiologists labeling reports were only given access to the report impression section when labeling the CheXpert test set. Sometimes, conditions are mentioned in the *Findings* section of the report but not mentioned in the *Impression* section. This results in more negative labels when radiologists looked at reports. For chest CT scan reports, [75] also find that a condition mentioned in the *Findings* section is not always mentioned in the *Impression* section of the report. Fourth, labeling images and reports is inherently noisy to a certain extent, resulting in disagreement. Drivers of noise include mistakes on the part of radiologists labeling reports and radiologists labeling images, uncertainty regarding the presence of a condition based on an image or report, and different thresholds for diagnosing conditions as positive among radiologists. [26] describe additional factors that contribute to discrepancies in radiologist interpretations, including radiologist specific causes of error like under reading as well as system issues like excess workload. [243], in their analysis on MRI neuroradiology reports, also note factors, such as a difference in observations within reports depending on the referrer, that likewise result in discrepancies.

Next, we explain the counts of the largest disagreements between radiologists labeling reports and radiologists labeling images. Out of the 500 examples on the CheXpert test set, there were 223 examples where the image was labeled positive while the report was labeled negative for Enlarged Cardiomediastinum. We hypothesize that this results from the difference in the task setup related to the label hierarchy. Since Enlarged Cardiomediastinum is a parent condition of Cardiomegaly, radiologists labeling reports were instructed to leave Enlarged Cardiomediastinum blank if they labeled Cardiomegaly positive. There were 101 examples where the image was labeled positive while the report was labeled negative for Cardiomegaly. Diagnosis of cardiomegaly on chest radiographs can depend on patient positioning and clinical history. Further, particularly in the ICU setting in which multiple consecutive radiographs are taken, cardiomegaly is not consistently described in the

report even when present unless a clinically significant change is observed (i.e. pericardial effusion). There were 100 examples where the image was labeled positive while the report was labeled negative for Lung Opacity. We hypothesize that this results from the difference in task setup related to label hierarchy, as Lung Opacity is a parent condition. Further, particularly in the setting of atelectasis, lung opacity may not have risen to clinical relevance for the reporting radiologist despite being seen on the isolated imaging task. There were 65 examples where the image was labeled negative while the report was labeled positive for Pleural Effusion. We hypothesize that this partially results from both the variant thresholds for diagnosis of pleural effusion among radiologists and the clinical setting in which the reporting radiologist has access to prior films. It was common to see the report state "decreased" or "trace residual" effusion due to the context of prior imaging on that patient. However, in the isolated image labeling task, the perceived likelihood of the condition fell below the threshold of a board-certified radiologist. There were 49 examples where the image was labeled negative, while the report was labeled positive for Edema. Similar to the effusion example, clinical context and prior imaging played a role in these discrepancies as, again, diagnoses were carried forward from prior studies and language such as "some residual" or "nearly resolved" in the report were used to indicate the presence of edema based on the clinical context. However, when labeling the corresponding image in isolation, the presence of edema fell below the threshold of a board-certified radiologist. Table 10.2 contains specific examples of these disagreements with clinical explanations. Table 10.3 shows the counts of disagreements between radiologists labeling reports and radiologists labeling images by condition.

10.4.3 Relationships between reports labeled and image labeled conditions

To determine whether there are significant relationships between conditions labeled from reports and conditions labeled from images, we learn a mapping from the output of radiologists labeling reports to the output of radiologists labeling images. We then analyze the significant relationships implied by this mapping from a clinical perspective.

Method

We train logistic regression models to map the radiologist report labels for all conditions to the image ground truth for each of the evaluation conditions. We quantitatively measure the relationship between the radiologist report labels and the image ground truth by obtaining odds ratios from the coefficients of these logistic regression models. We review the odds ratios from these models with a board-certified radiologist to understand how particular radiologist report labels might clinically change the odds of image labels.

Table 10.3: Counts of disagreements by condition between radiologists labeling reports and radiologists labeling the corresponding X-ray images on the CheXpert test set. The first column reports the number of times the image ground truth was positive, while the radiologist report label was negative. The second column reports the number of times the image ground truth was negative, while the radiologist report label was positive.

Condition	**Positive on Image Negative on Report**	**Negative on Image Positive on Report**
No Finding	38	40
Enlarged Cardiom.	223	5
Cardiomegaly	101	15
Lung Opacity	100	50
Lung Lesion	2	12
Edema	26	49
Consolidation	16	17
Pneumonia	6	5
Atelectasis	75	31
Pneumothorax	1	13
Pleural Effusion	11	65
Pleural Other	3	15
Fracture	3	21
Support Devices	53	13

Training details

We one-hot encode the radiologist report labels and provide these binary variables as inputs to a logistic regression model. For example, the "Atelectasis Positive" variable is 1 if the radiologist labels Atelectasis as positive on the report and 0 otherwise. Similarly, the "Atelectasis Negative" variable is 1 if the radiologist labels Atelectasis as negative on the report and 0 otherwise. The same logic applies to the "Atelectasis Uncertain" variable as well as the other variables for each condition. We then train the logistic regression model with L1 regularization ($\alpha = 0.5$) on the CheXpert test set using the one-hot encoded radiologist report labels (for all conditions) as input and the image ground truth for a condition as output. In total, we train different logistic regression models to map the radiologist report labels to binary image labels for each of the 8 evaluation conditions. We compute odds ratios by exponentiating the coefficients of the logistic regression models.

Results

After training the logistic regression models, we find that particular radiology report labels increased (or decreased) the odds of particular conditions in an image with statistical significance (P ¡ 0.05). As expected, we find that radiology report labels associated with a condition increase the odds of that same condition in the image; for example, a Cardiomegaly positive report label increases the odds of Cardiomegaly in the image. We also find that the regression model corrects for label

hierarchy. A Cardiomegaly positive report label increases the odds of Enlarged Cardiomediastinum (the parent of Cardiomegaly) on the image by 9.6 times. We similarly observe the model correcting for the label hierarchy of Lung Opacity. Radiology report labels of Edema positive, Consolidation positive, and Atelectasis positive, which all correspond to child conditions of Lung Opacity, increase the odds of Lung Opacity. We also find that the model maps particular uncertainties in report labels to the presence of a condition in the image. For example, Atelectasis uncertain report labels and Edema uncertain report labels increase the odds of Lung Opacity by 2.9 and 7.9 times respectively.

Next, we find that the model maps positive report labels to the presence of other conditions in the image. A Pleural Effusion positive report label increases the odds of Lung Opacity by 4.4 times. We hypothesize that this results from co-occurrence between Pleural Effusion and child conditions of Lung Opacity such as Atelectasis and Edema. Pleural effusion physiologically often leads to adjacent lung collapse, atelectasis, and is often seen in physiologic fluid overload conditions, edema. We find that an Atelectasis positive report label decreases the odds of Support Devices in the image by 0.28 times. On the patient population who have support devices, many of whom are in an Intensive Care Unit (ICU) setting, it is not clinically useful for radiologists to comment on the presence of atelectasis on reports, as they would rather focus on more clinically relevant changes. This may explain the mechanism by which the presence of atelectasis in a report signals that there are no support devices in the image. We find that a Fracture positive report label decreases the odds of Support Devices by 0.17 times. We hypothesize that this results from a negative co-occurrence between Fractures and Support Devices, as the two observations select for different patient populations: X-rays for fractures are often done in the Emergency Department (ED) or other outpatient settings rather than the ICU setting. We find that an Edema positive report label increases the odds of Enlarged Cardiomediastinum on the image by 2.1 times. This may be explained by the fact that Edema and Enlarged Cardiomediastinum often co-occur in a clinical setting, as they can both be caused by congestive heart failure. Lastly, we find that a Support Devices positive report label decreases the odds of No Finding in the image by 0.03 times. This may be explained by the fact that patients with support devices are usually in the ICU setting and sick with other pathologies. We visualize these statistically significant odds ratios for each type of radiologist report label (such as "Atelectasis Negative") as a factor for the presence of an evaluation condition in the X-ray image in Figure 10.2.

10.4.4 Naive mapping from labels obtained from reports to X-ray image labels

We map the output of an automated radiology report labeler to X-ray image labels using simple uncertainty handling strategies.

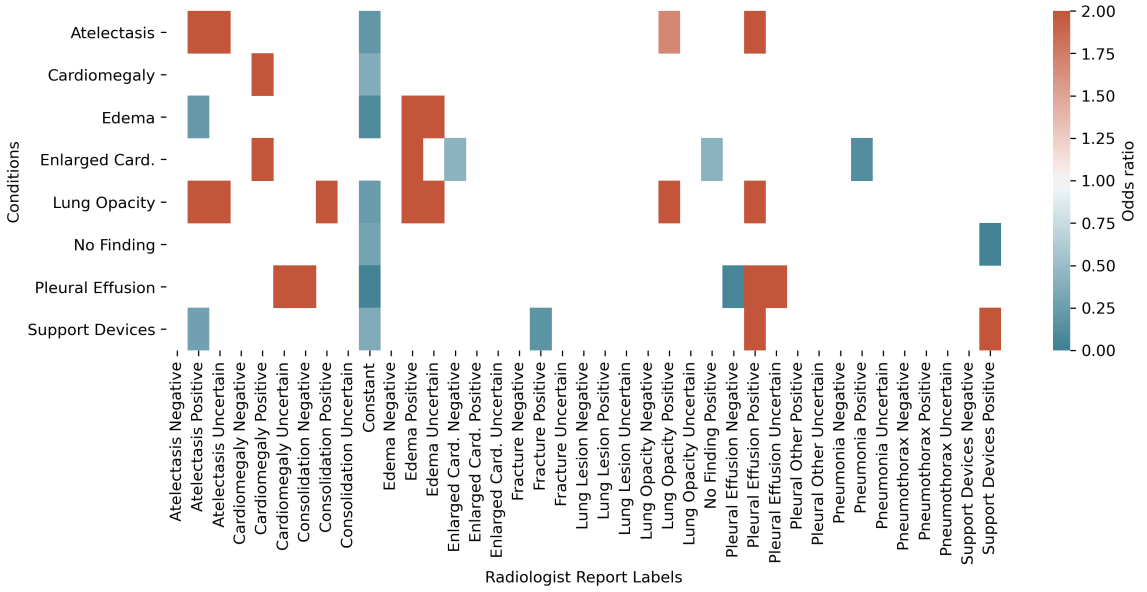


Figure 10.2: Odds ratios for radiologist report labels as factors for the presence of a condition in the X-ray image. We map the radiologist report labels across all conditions to the image ground truth using a logistic regression model. We obtain odds ratios for the input variables, which are the one-hot encoded radiologist report labels, and only display odds ratios for which the corresponding P value (two-sided t test) is less than 0.05.

Table 10.4: F1 scores obtained by the Zero-One and LogReg baselines, evaluated on the CheXpert test set. The weighted average is weighted by prevalence (n = # positive).

Condition (n = # positive)	**Zero-One Baseline**	**LogReg Baseline**
Atelectasis (n=153)	0.52	**0.63**
Cardiomegaly (n=151)	0.46	**0.56**
Edema (n=78)	**0.53**	0.47
Pleural Effusion (n=104)	**0.65**	**0.65**
Enlarged Cardiom. (n=253)	0.20	**0.67**
Lung Opacity (n=264)	0.69	**0.81**
Support Devices (n=261)	**0.85**	0.84
No Finding (n=62)	0.39	**0.55**
Average	0.54	**0.65**
Weighted Average	0.56	**0.70**

Method

For a baseline approach, we naively map labels obtained from running the CheXpert labeler on the radiology report impressions to X-ray image labels. The CheXpert labeler is an automatic, rules-based radiology report labeler [106]. The labels produced by the CheXpert labeler include 4 classes

Table 10.5: F1 scores for BERT+Thresholding and BERT+LogReg trained on the MIMIC-CXR and CheXpert datasets. We refer to the BERT+Thresholding method on the MIMIC-CXR dataset as VisualCheXbert. The models here are evaluated on the CheXpert test set.

Condition (n = # positive)	**BERT+Thresholding on MIMIC-CXR, DenseNet Labels**	**BERT+LogReg on MIMIC-CXR, DenseNet Labels (VisualCheXbert)**	**BERT+Thresholding on CheXpert, DenseNet Labels**	**BERT+LogReg on CheXpert, DenseNet Labels**
Atelectasis (n=153)	**0.65**	0.64	**0.67**	0.66
Cardiomegaly (n=151)	0.53	**0.62**	**0.61**	**0.61**
Edema (n=78)	**0.55**	0.54	0.49	**0.53**
Pleural Effusion (n=104)	0.64	**0.65**	0.57	**0.67**
Enlarged Cardiom. (n=253)	0.44	**0.73**	0.60	**0.70**
Lung Opacity (n=264)	0.81	**0.83**	0.70	**0.83**
Support Devices (n=261)	0.85	**0.87**	0.80	**0.84**
No Finding (n=62)	0.44	**0.54**	0.46	**0.52**
Average	0.61	**0.68**	0.61	**0.67**
Weighted Average	0.65	**0.73**	0.65	**0.72**

per medical condition (positive, negative, uncertain, and blank). Since the image ground truth only has positive or negative labels per condition, we must map the labels produced by the CheXpert labeler to binary labels. We map the blank labels produced by the CheXpert labeler to negative labels. We do not change the positive and negative labels produced by the CheXpert labeler. To handle the uncertain labels, we use the two common uncertainty handling strategies in [106]: we map the uncertain labels to either all negative labels (zeros-uncertainty handling strategy) or all positive labels (ones-uncertainty handling strategy). We record the F1 score from the better performing strategy on the CheXpert test set, using as ground truth the labels provided by radiologists labeling X-ray images (image ground truth). We refer to this method as the *Zero-One Baseline*. Since we only report the maximum of the zeros-uncertainty handling strategy and the ones-uncertainty handling strategy, the F1 scores for the Zero-One Baseline represent the most optimistic global mapping of the uncertainty labels for this method.

Results

We find that the average and weighted average F1 scores across the evaluation conditions for the Zero-One Baseline are 0.54 and 0.56 respectively, which are in between the average / weighted average "Low F1" and "High F1" scores for radiologists labeling reports (see Table 10.1). This indicates that the Zero-One Baseline is not strictly better or worse than radiologists labeling reports, who we previously show to have poor agreement with radiologists labeling images. The Zero-One Baseline F1 scores for Atelectasis, Cardiomegaly, Edema, Pleural Effusion, and Enlarged Cardiomediastinum are 0.52, 0.46, 0.53, 0.65, and 0.20 respectively, which are all between the respective "Low F1" and "High F1" scores for radiologists labeling reports. The Zero-One Baseline F1 scores for Lung Opacity and No Finding are 0.69 and 0.39 respectively, which are slightly higher ($\sim$ 0.01 difference) than the respective "High F1" scores for radiologists labeling reports. Similarly, the Zero-One Baseline F1 score for Support Devices is 0.39, which is slightly lower ($\sim$ 0.01 difference) than the "Low F1" Support Devices score for radiologists labeling reports. The F1 scores for the Zero-One Baseline

across the evaluation conditions are shown in Table 10.4.

10.4.5 Mapping labels obtained from reports to X-ray image labels

We map the output of an automated radiology report labeler to X-ray image labels, similarly to how we previously map the output of radiologists labeling reports to the output of radiologists labeling images. Previous work by [64] showed that labels obtained from noisy labeling functions on radiology reports can be mapped to labels that are of similar quality to image labels produced by radiologists for the simpler task of classifying X-rays as normal or abnormal.

Method

This approach, motivated by a prior experiment in which we map radiologist report labels to image labels, improves upon the naive uncertainty mapping strategy used in the Zero-One Baseline. As before, we obtain report labels by running the CheXpert labeler on radiology report impressions. For each of the evaluation conditions, we train a logistic regression model that maps the CheXpert labeler's output on a radiology report impression to a positive or negative label for the target condition. This approach makes use of the automated report labels for all 14 conditions to predict the label for each target condition. We refer to this approach as the *LogReg Baseline*.

Training details

We one-hot encode the report labels outputted by the CheXpert labeler and provide these binary variables as inputs to a logistic regression model. We train a logistic regression model with $L2$ regularization ($C = 1.0$) and a max iteration of 500 using the one-hot encoded report labels (for all conditions) as input and the image ground truth for a condition as output. The class weights are the inverse prevalence of the respective class in the training set. We use a leave-one-out cross-validation strategy to train and validate the logistic regression model on the CheXpert test dataset. For each of the 8 evaluation conditions, we train different logistic regression models to map the labels produced by the CheXpert labeler to binary image labels.

Results

We find that the LogReg Baseline approach improves upon the Zero-One Baseline for most conditions. Compared to the Zero-One Baseline, the LogReg Baseline increases the average F1 score from 0.54 to 0.65 and the weighted average F1 score from 0.56 to 0.70. The LogReg Baseline increases the F1 score compared to the Zero-One Baseline from 0.52 to 0.63 for Atelectasis, 0.46 to 0.56 for Cardiomegaly, 0.20 to 0.67 for Enlarged Cardiomediastinum, 0.69 to 0.81 for Lung Opacity, and 0.39 to 0.55 for No Finding. However, the LogReg Baseline decreases the F1 scores compared to the Zero-One Baseline from 0.53 to 0.47 for Edema and 0.85 to 0.84 for Support Devices. For Pleural

Table 10.6: Improvement in F1 score obtained by VisualCheXbert, evaluated on the CheXpert test set and reported with 95% confidence intervals. The left-most column shows the improvement over the Zero-One Baseline. The middle column shows the improvement over the radiologist report labels with uncertains mapped to the image ground truth label. The right-most column shows the improvement over the radiologist report labels with uncertains mapped to the opposite of image ground truth label.

Condition (n = # positive)	**Improvement over Zero-One Baseline**	**Improvement over Higher Radiologist Score**	**Improvement over Lower Radiologist Score**
Atelectasis (n=153)	0.12 (0.04, 0.20)	0.04 (-0.04, 0.12)	0.41 (0.32, 0.49)
Cardiomegaly (n=151)	0.16 (0.07, 0.25)	0.15 (0.07, 0.25)	0.20 (0.11, 0.28)
Edema (n=78)	0.01 (-0.05, 0.07)	-0.04 (-0.09, 0.02)	0.09 (0.02, 0.17)
Pleural Effusion (n=104)	-0.01 (-0.04, 0.03)	-0.06 (-0.10, -0.02)	0.01 (-0.03, 0.05)
Enlarged Cardiom. (n=253)	0.53 (0.46, 0.60)	0.52 (0.44, 0.60)	0.64 (0.57, 0.71)
Lung Opacity (n=264)	0.14 (0.09, 0.20)	0.15 (0.09, 0.20)	0.15 (0.10, 0.20)
Support Devices (n=261)	0.02 (-0.01, 0.06)	0.01 (-0.02, 0.04)	0.01 (-0.02, 0.04)
No Finding (n=62)	0.15 (0.05, 0.26)	0.16 (0.05, 0.28)	0.16 (0.05, 0.28)
Average	0.14 (0.12, 0.17)	0.12 (0.09, 0.15)	0.21 (0.18, 0.24)
Weighted Average	0.17 (0.15, 0.20)	0.15 (0.13, 0.18)	0.24 (0.21, 0.26)

Effusion, both the LogReg Baseline and the Zero-One Baseline have an F1 score of 0.65. Although the LogReg Baseline is not better than the Zero-One Baseline for all conditions, these results suggest that a learned mapping from radiologist report labels to X-ray image labels can outperform naively mapping all uncertain labels to positive or negative for most conditions. The F1 scores obtained by the LogReg Baseline, along with head-to-head comparisons to the Zero-One Baseline, are shown in Table 10.4.

10.4.6 Mapping textual radiology reports directly to the X-ray image labels

Previously, we mapped the output of an existing automated report labeler, which takes text reports as input, to X-ray image labels. We now map the textual radiology report directly to the X-ray image labels.

Method

We develop a deep learning model that maps a radiology report directly to the corresponding X-ray image labels.

Since it is too expensive to obtain labels from radiologists for hundreds of thousands of X-ray images to supervise our model, we instead train a single DenseNet model [103] to detect medical conditions from chest X-ray images, as is described by [106], and we use this computer vision model as a proxy for a radiologist labeling an X-ray image. We use the DenseNet model to output probabilities

for each of the 14 conditions for all X-rays in the MIMIC-CXR dataset and the CheXpert training dataset. To obtain the output of the vision model on the MIMIC-CXR dataset, we train the DenseNet on the CheXpert training dataset. Similarly, to obtain the output of the vision model on the CheXpert training dataset, we train the DenseNet on the MIMIC-CXR dataset. We find that the DenseNet trained on the CheXpert training set has an AUROC of 0.875 on the CheXpert test set across all conditions, and the DenseNet trained on the MIMIC-CXR dataset has an AUROC of 0.883 on the CheXpert test set across all conditions.

We then use the probabilities outputted from these computer vision models as ground truth to fine-tune a BERT-base model. We train one BERT model using the MIMIC-CXR dataset and one using the CheXpert training dataset. The BERT model takes a tokenized radiology report impression from the MIMIC-CXR or CheXpert dataset as input and is trained to output the labels produced by the DenseNet model. We feed the BERT model's output corresponding to the *[CLS]* token into linear heads (one head for each medical condition) to produce scores for each medical condition. We use the cross-entropy loss to fine-tune BERT. The BERT model is initialized with biomedically pretrained weights produced by [164]. This model training process is shown in Figure 10.1.

After training the BERT model, we map the outputs of BERT, which are probabilities, to positive or negative labels for each condition. To do so, we try two different methods. Our first method uses optimal probability thresholds to convert the BERT outputs to binary labels. We calculate optimal thresholds by finding the threshold for each condition that maximizes Youden's index [256] (the sum of sensitivity and specificity minus one) on the CheXpert validation dataset. We refer to this approach as *BERT+Thresholding*. Our second method trains a logistic regression model to map the output of BERT across all 14 conditions to a positive or negative label for the target condition. We refer to this approach as *BERT+LogReg*. Ultimately, we develop four different models by using both methods on outputs from a BERT model trained on the MIMIC-CXR dataset and a BERT model trained on the CheXpert training dataset. The four resulting models are called BERT+Thresholding on MIMIC-CXR, BERT+LogReg on MIMIC-CXR, BERT+Thresholding on CheXpert, and BERT+LogReg on CheXpert. We refer to the BERT+LogReg model trained on the MIMIC-CXR dataset with labels provided by the DenseNet model, which is our best performing approach, as *VisualCheXbert*.

Training details

We train the BERT model on 3 TITAN-XP GPUs using the Adam optimizer [118] with a learning rate of 2×10^{-5}, following [56] for fine-tuning tasks. We use a random 85%-15% training-validation split, as in [213]. The BERT model is trained until convergence. We use a batch size of 18 radiology report impressions. For the BERT+LogReg approach, the logistic regression model uses $L2$ regularization ($C = 1.0$) and a max iteration of 500. Similar to the LogReg Baseline, the class weights are

the inverse prevalence of the respective class in the training set, and we use a leave-one-out cross-validation strategy to train and test the logistic regression model on the CheXpert test dataset. We train different logistic regression models to map the probabilities outputted by the BERT model to the binary image labels for each of the 8 evaluation conditions.

Results

We compare the performance of the different BERT approaches on the CheXpert test set. First, we find that on most conditions, BERT+LogReg outperforms BERT+Thresholding. This finding holds true on both the CheXpert and MIMIC-CXR datasets. Second, we find that despite being trained on datasets from different institutions, the models trained on MIMIC-CXR and CheXpert datasets perform similarly. This indicates that the BERT model trained on radiology report impressions from the MIMIC-CXR distribution (Beth Israel Deaconess Medical Center Emergency Department between 2011–2016) [111] can perform as well as a model trained on radiology report impressions from the CheXpert distribution (Stanford Hospital between 2002-2017) [106], even when both models are evaluated on a test set from the CheXpert distribution. Since we obtain a slightly higher average and weighted average F1 using the MIMIC-CXR dataset, we use BERT trained on MIMIC-CXR in our final approach called VisualCheXbert. The performance of the BERT approaches is shown in Table 10.5.

Next, we compare VisualCheXbert to the Zero-One Baseline. When comparing VisualCheXbert to the Zero-One Baseline as well as the higher and lower scores of radiologists labeling reports described below, we report the improvements by computing the paired differences in F1 scores on 1000 bootstrap replicates and providing the mean difference along with a 95% two-sided confidence interval [66]. Overall, VisualCheXbert improves the average F1 and weighted average F1 over the Zero-One Baseline with statistical significance, increasing the average F1 score by 0.14 (95% CI 0.12, 0.17) and the weighted average F1 score by 0.17 (95% CI 0.15, 0.20). We find that VisualCheXbert obtains a statistically significant improvement over the Zero-One Baseline on most conditions. VisualCheXbert increases the F1 score on Enlarged Cardiomediastinum, Cardiomegaly, No Finding, Lung Opacity, and Atelectasis compared to the Zero-One Baseline by 0.53 (95% CI 0.46, 0.60), 0.16 (95% CI 0.07, 0.25), 0.15 (95% CI 0.05, 0.26), 0.14 (95% CI 0.09, 0.20), and 0.12 (95% CI 0.04, 0.20), respectively. VisualCheXbert obtains similar performance (no statistically significant difference) to the Zero-One Baseline on the rest of the conditions, which are Edema, Pleural Effusion, and Support Devices, with improvements of 0.01 (95% CI -0.05, 0.07), -0.01 (95% CI -0.04, 0.03), and 0.02 (95% CI -0.01, 0.06), respectively.

Lastly, we compare the F1 scores for VisualCheXbert to the higher and lower scores of radiologists labeling reports. The higher scores for radiologists labeling reports are obtained by mapping the uncertain radiologist report labels to the image ground truth label, while the lower scores for radiologists labeling reports are obtained by mapping the uncertain radiologist report labels to the

opposite of the ground truth. Overall, VisualCheXbert obtains a statistically significant improvement over both the higher and lower radiologist scores, increasing the average F1 score by 0.12 (95% CI 0.09, 0.15) over the higher radiologist score and 0.21 (95% CI 0.18, 0.24) over the lower radiologist score and increasing the weighted average F1 score by 0.15 (95% CI 0.13, 0.18) over the higher radiologist score and 0.24 (95% CI 0.21, 0.26) over the lower radiologist score. Statistically significant improvements over the higher radiologist score are observed for Cardiomegaly (0.15 [95% CI 0.07, 0.25]), Enlarged Cardiomediastinum (0.52 [95% CI 0.44, 0.60]), Lung Opacity (0.15 [95% CI 0.09, 0.20]), and No Finding (0.16 [95% CI 0.05, 0.28]). VisualCheXbert performs similarly (no statistically significant difference) to the higher radiologist score on Atelectasis (0.04 [95% CI -0.04, 0.12]), Edema (-0.04 [95% CI -0.09, 0.02]), and Support Devices (0.01 [95% CI -0.02, 0.04]). VisualCheXbert performs slightly worse than the higher radiologist score on one condition, which is Pleural Effusion (-0.06 [95% CI -0.10, -0.02]). VisualCheXbert observes considerable, statistically significant improvements compared to the lower radiologist score on all but two conditions. There is no statistically significant difference between VisualCheXbert and the lower radiologist score on these two conditions, which are Pleural Effusion (0.01 [95% CI -0.03, 0.05]) and Support Devices (0.01 [95% CI -0.02, 0.04]). We show the improvements obtained by VisualCheXbert over the Zero-One Baseline and the improvements over radiologists labeling reports in Table 10.6.

10.5 Limitations

Our work has the following limitations. First, our study only made use of the *Impression* section of the radiology reports, which is a summary of the radiology report. Prior work regarding automated chest X-ray labeling has also extensively used the impression section in radiology reports [106, 111, 239]. However, conditions are sometimes mentioned in the *Findings* section of the report but not in the *Impression* section. As a result, negative and blank labels are more frequent when using the *Impression* section, and this could increase the disparity between labels extracted from the impression and the corresponding chest X-ray image labels. Second, the VisualCheXbert model has a maximum input size of 512 tokens. In practice, only 3 of the report impressions in the entire CheXpert dataset were longer than this limit. Third, the CheXpert test set, on which we evaluated our models, consists of 500 radiology studies and is therefore limited in size. As a result, some of the medical conditions contained very few positive examples; we only evaluated our models on conditions for which at least 10% of the examples in the CheXpert test set were positive. Using a larger test set would allow evaluation on rarer conditions. Fourth, our models are evaluated on chest X-rays from a single institution. Further evaluation on data from other institutions could be used to evaluate the generalizability of our models.

10.6 Conclusion

We investigate the discrepancy between labels extracted from radiology reports and the X-ray image ground truth labels. We then develop and evaluate methods to address this discrepancy. In our work, we aim to answer the following questions.

Do radiologists labeling reports agree with radiologists labeling X-ray images? We find that there is significant disagreement between radiologists labeling reports and radiologists labeling images. On the CheXpert test set, we observe low Kappa scores for almost all conditions evaluated. The average Kappa across the evaluation conditions is between 0.312 and 0.430. These bounds are based on the most pessimistic mapping and most optimistic mapping of uncertain radiology report labels.

Why do radiologists labeling reports disagree with radiologists labeling X-ray images? Upon a board-certified radiologist review of examples of disagreements between radiologists labeling reports and radiologists labeling images, we find four main reasons for disagreement. First, on the CheXpert test set, radiologists labeling reports typically do not mark a parent condition as positive if a child condition is positive. An example of a parent and child condition would be Lung Opacity and Edema, respectively. Second, radiologists labeling reports have access to clinical report history, which biases their diagnoses compared to radiologists labeling images who do not have access to this information. Third, conditions are sometimes reported in the *Findings* section of radiology reports but not the *Impression* section of radiology reports. However, the *Impression* section of radiology reports is commonly used to label reports. This discrepancy can cause radiologists labeling reports to miss pathologies present on the X-ray image. Fourth, labeling images and reports is noisy to a certain extent due to factors such as human mistakes, uncertainty in both reports and images, and different thresholds for diagnosing conditions as positive among radiologists.

Are there significant relationships between conditions labeled from reports and conditions labeled from images? We find many significant relationships between conditions labeled from reports and conditions labeled from images. We report and clinically interpret various radiology report labels that increase (or decrease) the odds of particular conditions in an image with statistical significance (P¡ 0.05). As expected, we find that positive report labels for a condition increase the odds of that condition in an image. We find that positive report labels for children of a condition increase the odds of the parent condition in an image, a phenomenon that is correcting for the label hierarchy. We find that particular uncertain report labels for a condition increase the odds of the condition (and/or its parent condition). We also find that positive report labels for certain conditions increase (or decrease) the odds of other conditions in the image. One example is that a positive Atelectasis report label decreases the odds of Support Devices in the X-ray image by 0.28 times. We explain potential mechanisms by which the presence of a condition in a report signals the presence (or absence) of another condition in the image.

Can we learn to map radiology reports directly to the X-ray image labels? We learn to map a textual radiology report directly to the X-ray image labels. We use a computer vision model

trained to detect diseases from chest X-rays as a proxy for a radiologist labeling an X-ray image. Our final model, VisualCheXbert, uses a biomedically-pretrained BERT model that is supervised by the computer vision model. When evaluated on radiologist image labels on the CheXpert test set, VisualCheXbert increases the average F1 score across the evaluation conditions by between 0.12 (95% CI 0.09, 0.15) and 0.21 (95% CI 0.18, 0.24) compared to radiologists labeling reports. VisualCheXbert also increases the average F1 score by 0.14 (95% CI 0.12, 0.17) compared to a common approach that uses a previous rules-based radiology report labeler.

Given the considerable, statistically significant improvement obtained by VisualCheXbert over the approach using an existing radiology report labeler [106] when evaluated on the image ground truth, we hypothesize that VisualCheXbert's labels could be used to train better computer vision models for automated chest X-ray diagnosis.

Chapter 11

Generalization to Rare and Unseen Diseases

In previous chapters, we have seen that medical image interpretation algorithms are often evaluated on the same dataset distributions on which they were trained. However, the deployment of these algorithms requires understanding of their performance under clinically relevant distribution shifts.

In this chapter, we systematically evaluate the performance of deep learning models in the presence of diseases not labeled for or present during training. First, we evaluate whether deep learning models trained on a subset of diseases (seen diseases) can detect the presence of any one of a larger set of diseases. We find that models tend to falsely classify diseases outside of the subset (unseen diseases) as "no disease". Second, we evaluate whether models trained on seen diseases can detect seen diseases when co-occurring with diseases outside the subset (unseen diseases). We find that models are still able to detect seen diseases even when co-occurring with unseen diseases. Third, we evaluate whether feature representations learned by models may be used to detect the presence of unseen diseases given a small labeled set of unseen diseases. We find that the penultimate layer of the deep neural network provides useful features for unseen disease detection. Our results can inform the safe clinical deployment of deep learning models trained on a non-exhaustive set of disease classes.

This chapter is based on [210].

11.1 Introduction

Safe clinical deployment of deep learning models for disease diagnosis would require models to not only diagnose diseases that they have been trained to detect, but also recognize the presence of diseases they have not been trained to detect for possible deferral to a human expert [150, 184].

Figure 11.1: Overview of the experimental setup. /

Medical imaging datasets used to train models typically only provide labels for a limited number of common diseases because of the challenge and costs associated with labeling for all possible diseases. For example, some serious diseases, including pneumomediastinum, are not part of any commonly used chest X-ray databases [106, 111, 239]. However, it is unknown whether deep learning models for chest x-ray interpretation can maintain performance in presence of diseases not seen during training, or whether they can detect the presence of such diseases.

In this study, we provide a systematic evaluation of deep learning models in the presence of diseases not labeled for or present during training. Specifically, we first evaluate whether deep learning models trained on a subset of diseases (seen diseases) can detect the presence of any one of a larger set of diseases. We find that models tend to falsely classify diseases outside of the subset (unseen diseases) as "no disease". Second, we evaluate whether models trained on seen diseases can detect seen diseases when co-occurring with diseases outside the subset (unseen diseases). We find that models are still able to detect seen diseases even when co-occurring with unseen diseases. Third, we conduct an initial exploration of unseen disease detection methods, focused on evaluation of feature representations. We evaluate whether feature representations learned by models may be used to detect the presence of unseen diseases given a small labeled set of unseen diseases. We find that the penultimate layer provides useful features for unseen disease detection. Our results

can inform the safe clinical deployment of deep learning models trained on a non-exhaustive set of disease classes.

11.2 Related Work

Traditional machine learning frameworks assume a "closed world" assumption, where no new classes exist in the test set. However, in real world applications, trained models could encounter new classes. Deep learning models for image recognition are known to suffer in performance when applied to a test distribution that differs from their training distribution [98, 178, 201]. Several methodologies have been explored in computer vision for novelty or abnormality detection, including reconstruction-based methods, self-representation, statistical modeling, and deep adversarial learning [169, 248, 140]. Several methodologies were explored for open set clinical decision making showing no clear superior techniques [174]. However, most methodologies are designed for multiclass problems but not multi-label problems [74, 19]. Specifically for healthcare applications, there have been limited studies on out-of-distribution medical imaging [32, 151], with no previous studies investigating the performance of medical imaging classification models when facing unseen diseases.

11.3 Methods

11.3.1 Data

We form a dataset which has disease labels split into two categories: "seen diseases" and "unseen diseases" (Figure 11.1 A). We modify the CheXpert dataset, consisting of 224,316 chest radiographs from 65,240 patients labeled for the presence of 14 observations [106]. To be able to split disease labels without overlap, we remove children and parent label classes shared by diseases, specifically the Enlarged Cardiomediastinum, Airspace Opacity and Pneumonia label classes. We also remove the Support Devices label, as it is a clinically insignificant observation. We divide the remaining labels into four seen labels (No Disease, Consolidation, Pleural Effusion, Cardiomegaly), and six unseen labels (Pleural Other, Edema, Lung Lesion, Atelectasis, Fracture, and Pneumothorax). This division is based on each disease's prevalence in the dataset in order to evaluate model performance when trained only on commonly occurring diseases (Figure 1A). We use the CheXpert validation set to select models and to train unseen disease classifiers, which is a set of 200 labeled studies, where ground truth was set by annotation from a consensus of 3 radiologists. We use the CheXpert test set, which consists of 500 chest x-ray studies annotated with a radiologist majority vote, to evaluate the performance of models.

(a) Detection of "no disease" vs "any disease" overall, and in three subgroups: images with only unseen diseases, images with only seen diseases, images with both unseen and seen diseases co-occurring in one image.

(b) Performance of models in detecting seen diseases overall and for individual seen diseases: consolidation, pleural effusion and cardiomegaly. The overall evaluation strategy considers the average AUC over all seen diseases.

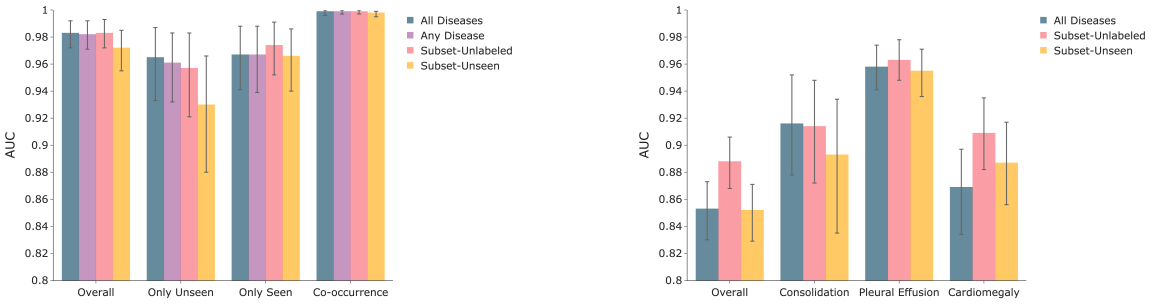


Figure 11.2: Performance of multi-label models under various setups.

11.3.2 Multi-Label Models

We train four multi-label models with different sets of images and labels, and evaluate the multi-label models on their disease detection performance. An overview of the models' setup is outlined in Figure 11.1 B.

All Diseases (Control) The All Diseases model is trained with all images and all ten disease labels ("no disease" and both seen and unseen), and serves as a comparison to models trained on a subset of diseases.

Any Disease (Control) The Any Disease model is trained with all images as a binary model with the "no disease" label (signifying a normal chest X-ray without any disease) and serves as another control comparison.

Subset-Unlabeled The Subset-Unlabeled model is trained with all images, but with the labels for unseen diseases removed. In the Subset-Unlabeled model, all image studies are included in training, while removing the six unseen disease labels.

Subset-Unseen The Subset-Unseen model is trained with images that have either no disease or have only seen diseases. Image studies with one or more unseen labels are removed, while also removing the six unseen disease labels.

11.3.3 Model Training

During training, the uncertain label is treated as a different class, resulting in a 3-class classifier for each label (negative, uncertain, positive). We use DenseNet121 for all experiments [103]. Images are fed into the network with size 320 × 320 pixels. We use the Adam optimizer with default β-parameters of $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a fixed learning rate 1×10^{-4} [118]. Batches are sampled using a fixed batch size of 16 images. We train for 3 epochs, saving checkpoints every 4800 iterations. Model training and inference code is written using PyTorch 1.5 on a Python 3.6.5 environment and run on 2 Nvidia GTX 1070 GPUs.

11.3.4 Forming Ensembles for Evaluation

For evaluating the performance of the multi-label models, we formed an ensemble of each model by running the model three times with different random initializations. Each run produced 10 top checkpoints. We created an ensemble of the 30 generated checkpoints on the validation set by computing the mean of the output probabilities over the 30 checkpoints for each task.

11.3.5 Visualizations of feature representations

To demonstrate the effectiveness of feature representations of multi-label models as inputs to unseen classifiers, we plot the 2D t-SNE [231] clusters in Figure 11.3 of the feature representations for each multi-label model. The t-SNE was run using a perplexity of 30 with 1000 iterations and a learning rate of 200. The clusters are color coded with the disease subset label from the validation set (seen/unseen) that we describe in Section 11.3.1.

11.4 Statistical analysis

To determine statistical significance between 2 models, we use the 95% confidence intervals of the difference between bootstrap samples. To generate confidence intervals, we used the non-parametric bootstrap with 1000 bootstrap replicates. Statistically significant differences between models were established using the non-parametric bootstrap on the mean AUC difference on the test set. We calculate p-values from the confidence interval using the method described in [5] with a threshold of 0.05 for hypothesis testing. This method was chosen to evaluate whether 2 models were similar in performance with respect to their average AUC over the bootstrap sample, and to test statistically significant performance differences in either direction using the 95% confidence intervals. We use the Benjamini-Hochberg method to correct for multiple hypothesis testing between various models.

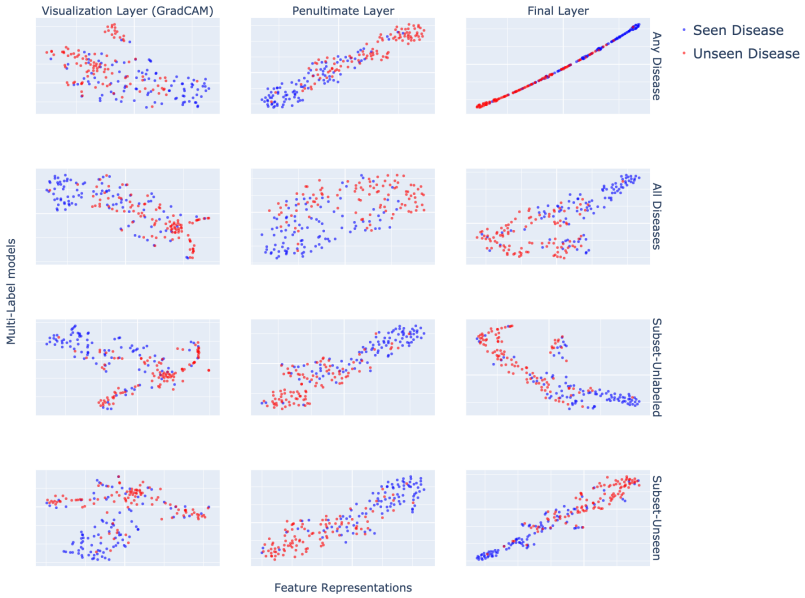


Figure 11.3: t-SNE plots of feature representations of each multi-label model

11.5 Detection of any disease vs no disease

We evaluate the performance of the multi-label models on detecting the presence of any disease (vs "no disease") on a test set containing both seen and unseen diseases. Results are summarized in Figure 11.2a, and Tables 11.1 and 11.2.

Subset-Unlabeled vs Controls The Subset-Unlabeled model is not statistically significantly different from the Any Disease model (mean AUC difference 0.001, [95% CI -0.004, 0.005]), and the All Diseases model (mean AUC difference 0.000, [95% CI -0.003, 0.003]).

Subset-Unseen vs Controls The Subset-Unseen model performs statistically significantly lower than the Any Disease model overall (mean AUC difference -0.010, [95% CI -0.019,-0.004]), but is not statistically significantly different to the Any Disease model when evaluating examples with only seen diseases (mean AUC difference -0.002, [95% CI -0.015,0.015]). We find that the Subset-Unseen model performs statistically significantly lower than the All Diseases model overall (mean AUC difference -0.010, [95% CI -0.018, -0.003]), but is not statistically significantly different in evaluating examples with co-occurring seen and unseen diseases (mean AUC difference -0.004, [95% CI -0.010, 0.000]).

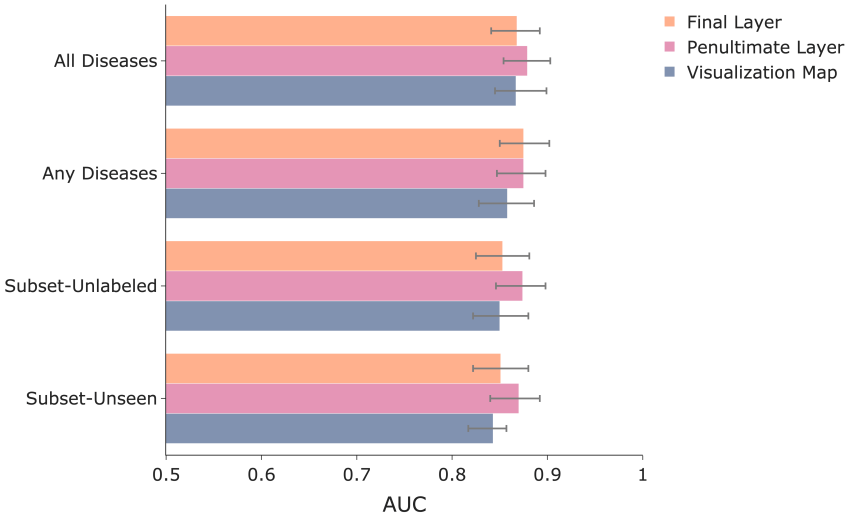


Figure 11.4: Performance on the task of unseen disease detection using unseen scores. Unseen scores were outputted by random forest classifiers trained using three different feature representations to detect the presence of unseen disease(s): the final prediction layer, penultimate layer and visualization map of the trained multi-label classifiers.

Subset-Unlabeled vs Subset-Unseen The Subset-Unlabeled model performs statistically significantly higher than the Subset-Unseen model in detecting "no disease" vs "any disease" in the presence of only unseen diseases (mean AUC difference 0.028 , [95% CI 0.011, 0.047]), and in the presence of co-occurring seen and unseen diseases (mean AUC difference 0.004, [95% CI 0.001, 0.009]). The Subset-Unlabeled model is not statistically significantly different from the Subset-Unseen model for only seen diseases (mean AUC difference -0.008, [95% CI -0.019, 0.001]).

11.6 Detection of seen diseases in the presence of seen and unseen diseases

We evaluate whether a multi-label model trained on seen diseases can successfully detect seen diseases on a test set containing both seen and unseen diseases. Results are summarized in Figure 11.2b, Table. 11.3 and 11.4.

Subset-Unseen vs All Diseases The Subset-Unseen model is not statistically significantly different from the All Diseases model overall in detecting seen diseases (mean AUC difference -0.011, [95% CI -0.020, 0.000]).

Subset-Unlabeled vs All Diseases The Subset-Unlabeled model has statistically significantly higher performance when compared to the All Diseases model in detecting seen diseases overall (mean AUC difference 0.033, [95% 0.025, 0.042]) and in detecting cardiomegaly (mean AUC difference 0.032, [95% CI 0.019, 0.046]).

Subset-Unseen vs Subset-Unlabeled The Subset-Unseen model has a statistically significantly lower performance when compared to the Subset-Unlabeled model in detecting seen diseases overall (mean AUC difference -0.044, [95% CI -0.054, -0.033]), pleural effusion (mean AUC difference -0.014, [95% CI -0.025, -0.004]), and cardiomegaly (mean AUC difference -0.019, [95% CI -0.033, -0.005]), and is not statistically significantly different from the Subset-Unlabeled model in detecting consolidation (mean AUC difference -0.023, [95% CI -0.067, 0.020]).

11.7 Unseen disease detection

We develop classifiers to detect the presence of any unseen disease given an X-ray image. Applying the four trained multi-label models to the validation set, we collect the outputs from the final prediction layer, the penultimate layer, and the visualization map (generated using the Grad-CAM method [208]). The output of the visualization map using GradCAM is used as a matrix directly in the following steps. The feature representations are extracted from running the validation set on the trained classification models. The three sets of outputs are then used to train a random forest classifier and a logistic regression classifier, with a binary outcome on whether the chest X-ray has an unseen disease or not, to produce an "unseen score" using unseen disease labels on the validation set (shown in Figure 11.1C). Logistic regression classifier is a commonly used standard for binary classification. Random forest classifiers, compared to logistic regression, are able to create nonlinear decision boundaries. Unseen scores are the output of the random forest classifier or the logistic regression classifier, and a numeric number between 0 and 1, signifying how likely the chest X-ray image has an unseen disease. The performance of these classifiers is reported on the test set. Results are summarized in Figure 11.4 and Table 11.5.

Comparing feature representations Unseen scores derived from the penultimate layer have the best average performance (AUC 0.873, [95% CI 0.848, 0.897]), followed by those from the final prediction layer (AUC 0.860, [95% CI 0.833, 0.889]) and the visualization map (AUC 0.851 [95% CI 0.832, 0.879]). The performance of unseen scores derived from the penultimate layer is statistically significantly higher than those from the final prediction layer (mean AUC difference 0.013 [95% CI 0.009, 0.017]), which is higher than those from the visualization map (mean AUC difference 0.009 [95% CI 0.007, 0.011]).

Comparing classifiers Over all of the representations and the multi-label models, the random forest classifier has a high average performance of AUC 0.862 [95% CI 0.837, 0.892], but this is not statistically significantly higher than the performance of the logistic regression classifier (mean AUC difference 0.002 [95% CI 0.000, 0.003]).

Comparing multi-label models Using the random forest classifier trained on the penultimate layer from the four multi-label models, the unseen score derived from the Any Disease model has the best performance (mean AUC 0.879, [95% CI 0.849, 0.901]) at predicting the presence of unseen disease, followed by the unseen score from the All Diseases model (mean AUC 0.875, [95% CI 0.850, 0.899]). The performance of the unseen score from the Any Diseases model is statistically significantly higher than that of the unseen score derived from the All Diseases model (mean AUC difference 0.004, [95% CI 0.003, 0.006]). The unseen scores derived from the Subset-Unlabeled and the Subset-Unseen models have the lowest performance among the unseen scores of the four models (AUC 0.874, [95% CI 0.846, 0.897]) and (AUC 0.870, [95% CI 0.842, 0.894]) respectively. Finally, the performance of the unseen scores derived from the Subset-Unlabeled model is statistically significantly higher than that of the Subset-Unseen model (mean AUC difference 0.005, [95% CI 0.003, 0.007]).

	Overall	Only Unseen Diseases	Only Seen Diseases	Co-occurring Seen and Unseen Diseases
All Diseases	0.980 (0.966,0.990)	0.957 (0.922,0.982)	0.962 (0.932,0.984)	0.999 (0.996,1.000)
Any Disease	0.982 (0.971,0.992)	0.961 (0.932,0.983)	0.967 (0.939,0.988)	0.999 (0.997,1.000)
Subset-Unlabeled	0.983 (0.972,0.993)	0.957 (0.921,0.983)	0.974 (0.952,0.991)	0.999 (0.997,1.000)
Subset-Unseen	0.975 (0.959,0.988)	0.937 (0.891,0.971)	0.968 (0.941,0.987)	0.997 (0.991,0.999)

Table 11.1: Performance in detecting "no disease" vs "any disease" overall and by each subgroup [mean area under curve (AUC), (95% confidence interval)].

	Overall	Only Unseen Diseases	Only Seen Diseases	Co-occurring Seen and Unseen Diseases
Any Diseases	-0.001 (-0.007,0.004) [p: 0.791]	-0.004 (-0.018,0.007) [p: 0.583]	0.000 (-0.011,0.010) [p: 1.0]	0.000 (-0.001,0.002) [p: 1.0]
Subset-Unlabeled	0.000 (-0.003, 0.003) [p: 1.000]	-0.007 (-0.017,0.000) [p: 0.124]	0.006 (-0.006,0.017) [p: 0.396]	0.000 (-0.001,0.001) [p: 1.000]
Subset-Unseen	-0.010 (-0.018, -0.003) [p: 0.006]	-0.035 (-0.059,-0.016) [p: 0.002]	-0.001 (-0.015,0.012) [p: 0.961]	-0.004 (-0.010,0.000) [p: 0.149]

Table 11.2: Differences in performance in detecting "no disease" vs "any disease" overall and by each subgroup, compared to the All Diseases model [mean area under curve (AUC), (95% confidence interval)] and p-values with $\alpha \leq 0.05$.

11.8 Limitations

There are three main limitations of our study. First, the dataset has a limited number of diseases, with six unseen diseases in this study. Ideally, a wider variety of unseen diseases would be evaluated

	Overall	Consolidation	Pleural Effusion	Cardiomegaly
All Diseases	0.851 (0.828,0.871)	0.910 (0.870,0.948)	0.956 (0.938,0.971)	0.863 (0.829,0.894)
Subset-Unlabeled	0.888 (0.868,0.906)	0.914 (0.872,0.948)	0.963 (0.948,0.978)	0.909 (0.882,0.935)
Subset-Unseen	0.861 (0.839,0.881)	0.909 (0.863,0.947)	0.958 (0.942,0.974)	0.886 (0.856,0.913)

Table 11.3: Performance in detecting seen diseases overall and by each disease [mean area under curve (AUC), (95% confidence interval)].

	Overall	Consolidation	Pleural Effusion	Cardiomegaly
Subset-Unlabeled	0.033 (0.025, 0.042) [p: 1.662e-13]	-0.003 (-0.030,0.022) [p: 0.776]	0.005 (-0.003, 0.015) [p: 0.307]	0.032 (0.019, 0.046) [p: 2.520e-06]
Subset-Unseen	-0.011 (-0.020, 0.000) [p: 0.839]	-0.027 (-0.069,0.021) [p: 0.352]	-0.009 (-0.019, 0.002) [p: 0.104]	0.012 (-0.001, 0.025) [p: 0.076]

Table 11.4: Differences in performance in detecting "no disease" vs "any disease" overall and by each subgroup, compared to the All Diseases model [mean area under curve (AUC), (95% confidence interval)] and p-values with $\alpha \leq 0.05$.

to minimize the impact of disease correlations on performance evaluation. Moreover, the ability to expand towards more diseases while maintaining performance is important for a useful model in the actual clinical setting. Second, our study is limited to an internal validation set without an external test set including different unseen diseases. Third, our study did not explore training strategies for multi-label models that could mitigate the performance drop with the All Diseases model compared to the Seen-Unlabeled model.

11.9 Discussion

In this study, we evaluate the performance of deep learning models in the presence of diseases not labeled for or present during training.

Can models detect seen diseases in the presence of unseen diseases? Our results show that the Subset-Unlabeled model, which is trained with unseen disease examples but not unseen disease labels, and the Subset-Unseen model, trained without unseen disease examples or labels, are able to detect "any disease" vs "no disease" in images with co-occurring unseen and seen diseases. However, their performance decreases when facing images with only unseen diseases (Figure 11.2a). These results show that in a real-world clinical setting, deep learning models may succeed in identifying "no disease" vs "any disease" when an unseen disease co-occurs with a seen disease, but may likely falsely report "no disease" if an unseen disease appears alone. Such mistake can result in delays in correct diagnosis and treatment, and therefore can be life-threatening in some medical conditions [15]. This result re-emphasizes the necessity for unseen disease detection to avoid misclassification of unseen diseases as "no disease."

Our results also show that the Subset-Unlabeled model and the Subset-Unseen model are able to detect seen diseases, even in the presence of unseen diseases, at a level comparable to the All Diseases model. We find that the Subset-Unlabeled model performs better than the All Diseases model, which

	All Diseases	Any Disease	Subset-Unlabeled	Subset-Unseen
Final Prediction layer				
Logistic Regression	0.863 (0.841,0.898)	0.871 (0.845,0.897)	0.850 (0.822,0.878)	0.848 (0.822,0.877)
Random Forest	0.868 (0.839,0.897)	0.875 (0.848,0.899)	0.853 (0.828,0.880)	0.851 (0.823,0.879)
Penultimate Layer				
Logistic Regression	0.874 (0.848,0.899)	0.875 (0.847,0.899)	0.872 (0.845,0.898)	0.870 (0.843,0.895)
Random Forest	0.875 (0.850,0.899)	0.879 (0.849,0.901)	0.874 (0.846,0.897)	0.870 (0.842,0.894)
Visualization Map				
Logistic Regression	0.850 (0.823,0.879)	0.856 (0.828,0.882)	0.844 (0.816,0.871)	0.843 (0.813,0.871)
Random Forest	0.858 (0.826,0.883)	0.867 (0.841,0.873)	0.850 (0.820,0.878)	0.843 (0.815,0.873)

Table 11.5: Performance in detecting unseen diseases [mean area under curve (AUC), (95% confidence interval)]. We used three different representations to predict the presence of unseen disease(s): the final prediction layers, penultimate layers and visualization maps from the trained classifiers. For each representation, we trained a logistic regression model and a random forest model.

may be because of the multi-task nature of the problem, where the optimization landscape may cause detrimental gradient interference between the different tasks and impede learning [258]. We find that the Subset-Unlabeled model has a statistically significantly higher performance compared to the Subset-Unseen model, likely because the Subset-Unlabeled model is exposed to additional training examples.

Can unseen diseases be detected without explicit training? On unseen disease detection, we conduct an initial exploration of unseen disease detection methods, borrowing philosophy from few shot learning, while focusing on evaluating feature representations extracted from classification models. We find that the unseen scores from the Subset-Unlabeled model has higher performance than those from the Subset-Unseen model, likely because the Subset-Unlabeled model learns representations of the unlabeled diseases during training. We find that unseen scores from the penultimate layer are the best for unseen disease detection, followed by the final layer and the visualization map. A possible explanation is that the penultimate layer contains information representing the unseen diseases, whereas the final prediction layer discards this information to reduce training loss. We find that the visualization map is outperformed by both the penultimate and the final prediction layer, perhaps because some diseases in our dataset, including lung lesion, pneumothorax, fracture, atelectasis, can occur in different locations in the chest X-ray than the seen diseases. Overall, our results demonstrate that using feature representations of multi-label models trained on diseases form suitable baselines for unseen disease detection. Exploration of the optimal model for training the unseen disease classifiers using the feature representations evaluated in this work would be an important future research direction.

Chapter 12

Generalization to Clinically Different Distributions

As we have seen in the previous chapter, poor generalization due to data distribution shifts in clinical settings is a key barrier to implementation. In this chapter, we look at the diagnostic performance for eight different chest X-ray models when applied to (1) smartphone photos of chest X-rays and (2) external datasets without any finetuning. All models were developed by different groups and submitted to the CheXpert challenge, and re-applied to test datasets without further tuning. We found that (1) on photos of chest X-rays, all eight models experienced a statistically significant drop in task performance, but only 3 performed significantly worse than radiologists on average, and (2) on the external set, none of the models performed statistically significantly worse than radiologists, and five models performed statistically significantly better than radiologists. Our results demonstrate that some chest X-ray models, under clinically relevant distribution shifts, were comparable to radiologists while other models were not. Future work should investigate aspects of model training procedures and dataset collection that influence generalization in the presence of data distribution shifts.

This chapter is based on [186].

12.1 Introduction

Automating cognitive tasks in medical imaging interpretation with deep learning models could improve access, efficiency, and augment existing workflows [182, 153, 212, 177]. However, poor generalization due to data distribution shifts in clinical settings is a key barrier to implementation.

First, though leveraging deep learning models in automated interpretation of photos of medical imaging examinations may serve as an infrastructure agnostic approach to deployment, particularly

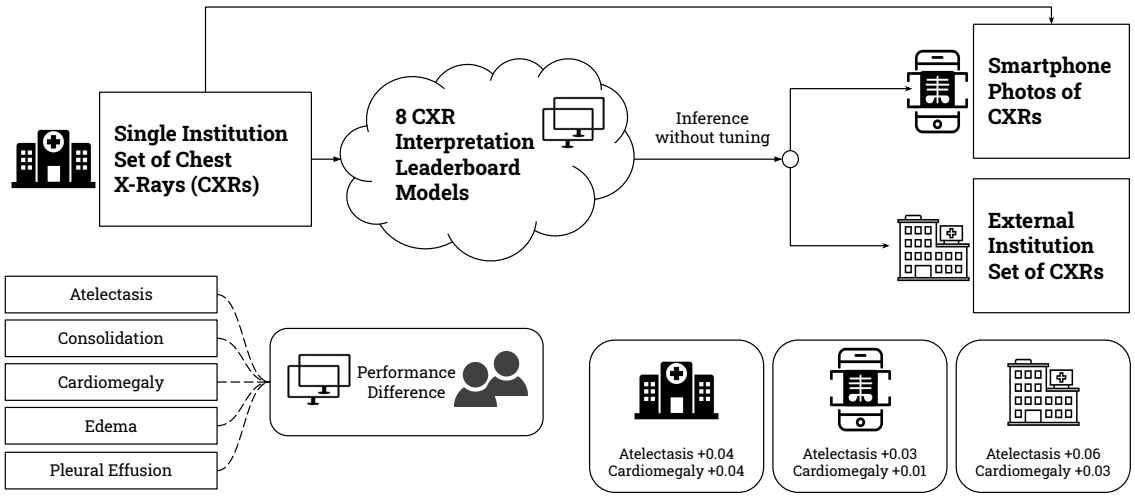


Figure 12.1: We measured the diagnostic performance for 8 different chest X-ray models when applied to (1) smartphone photos of chest X-rays and (2) external datasets without any finetuning. All models were developed by different groups and submitted to the CheXpert challenge, and re-applied to test datasets without further tuning.

in resource limited settings, significant technical barriers exist in automated interpretation of photos of chest X-rays. Photographs of X-rays introduce visual artifacts which are not commonly found in digital X-rays, such as altered viewing angles, variable lighting conditions, glare, moiré, rotations, translations, and blur [168]. These artifacts have been shown to reduce algorithm performance when input images are perceived through a camera [125]. The extent to which such artifacts reduces the performance of chest X-ray models has not been well investigated.

A second major obstacle to clinical adoption of chest X-ray models is that clinical deployment requires models trained on data from one institution to generalize to data from another institution [116, 36]. Early work has shown that chest X-ray models may not generalize well when externally validated on data from a different institution and are possibly vulnerable to distribution shift stemming from change in patient population or rely on non-medically relevant cues between institutions [261]. However, the difference in diagnostic performance of more recent chest X-ray models to external datasets has not been investigated.

We measured the diagnostic performance for 8 different chest X-ray models when applied to (1) photos of chest X-rays, and (2) chest X-rays obtained at a different institution. Specifically, we applied these models to a dataset of smartphone photos of 668 X-rays from 500 patients, and a set of 420 frontal chest X-rays from the ChestXray-14 dataset collected at the National Institutes of Health Clinical Center [239]. All models were developed by different groups and submitted to the CheXpert challenge, a large public competition for digital chest X-ray analysis [106]. Models were evaluated on their diagnostic performance in binary classification, as measured by Matthew's Correlation

Coefficient (MCC) [46], on the following pathologies selected in [106]: atelectasis, cardiomegaly, consolidation, edema, and pleural effusion [106].

We found that:

1. In comparison of model performance on digital chest X-rays to photos, all 8 models experienced a statistically significant drop in task performance on photos with an average drop of 0.036 MCC. In comparison of performance of models on photos compared to radiologist performance, three out of eight models performed significantly worse than radiologists on average, and the other five had no significant difference.

2. On the external set (NIH), none of the models performed statistically significantly worse than radiologists. On average over the pathologies, five models performed significantly better than radiologists. On specific pathologies (consolidation, cardiomegaly, edema, and atelectasis), there were some models that achieved significantly better performance than radiologists.

Our systematic examination of the generalization capabilities of existing models can be extended to other tasks in medical AI, and provide a framework for tracking technical readiness towards clinical translation.

Metric	Comparison	Average	Pleural Effusion	Edema	Atelectasis	Consolidation	Cardiomegaly
AUC	Photos	0.856 (0.840,0.869)	0.950 (0.932,0.968)	0.917 (0.884,0.943)	0.882 (0.856,0.912)	0.914 (0.865,0.946)	0.921 (0.900,0.940)
	Standard	0.871 (0.855,0.883)	0.960 (0.944,0.975)	0.926 (0.892,0.950)	0.885 (0.858,0.910)	0.918 (0.879,0.948)	0.934 (0.914,0.951)
	Standard - Photos	0.016 (0.012,0.019)	0.011 (0.004,0.019)	0.009 (0.001,0.018)	0.003 (-0.006,0.013)	0.005 (-0.009,0.016)	0.013 (0.006,0.023)
MCC	Photos	0.534 (0.507,0.559)	0.571 (0.526,0.631)	0.556 (0.481,0.639)	0.574 (0.505,0.634)	0.316 (0.246,0.386)	0.580 (0.522,0.630)
	Standard	0.570 (0.543,0.599)	0.621 (0.575,0.670)	0.550 (0.474,0.637)	0.587 (0.529,0.640)	0.336 (0.264,0.418)	0.643 (0.584,0.695)
	Standard - Photos	0.036 (0.024,0.048)	0.049 (0.020,0.070)	-0.006 (-0.039,0.033)	0.012 (-0.016,0.041)	0.020 (-0.011,0.047)	0.063 (0.036,0.084)

Table 12.1: AUC and MCC performance of models and radiologists on the standard X-rays and the photos of chest X-rays, with 95% confidence intervals.

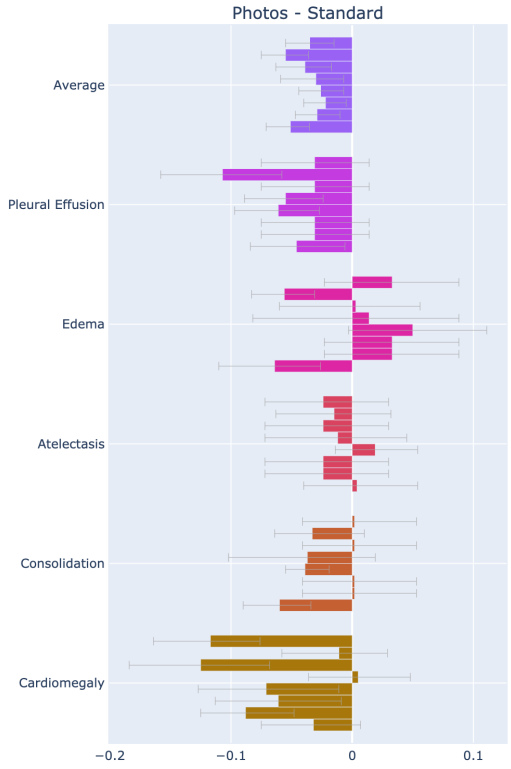
Comparison	Average	Pleural Effusion	Edema	Atelectasis	Consolidation	Cardiomegaly
Photos	0.534 (0.507,0.559)	0.571 (0.526,0.631)	0.556 (0.481,0.639)	0.574 (0.505,0.634)	0.316 (0.246,0.386)	0.580 (0.522,0.630)
Radiologists	0.568 (0.542,0.597)	0.671 (0.618,0.727)	0.507 (0.431,0.570)	0.548 (0.496,0.606)	0.359 (0.262,0.444)	0.566 (0.511,0.620)
Radiologists - Photos	0.035 (0.009,0.065)	0.099 (0.056,0.145)	-0.049 (-0.136,0.029)	-0.027 (-0.086,0.050)	0.042 (-0.056,0.124)	-0.014 (-0.069,0.029)

Table 12.2: MCC performance of models on the photos of chest X-rays, radiologist performance, and their difference, with 95% confidence intervals.

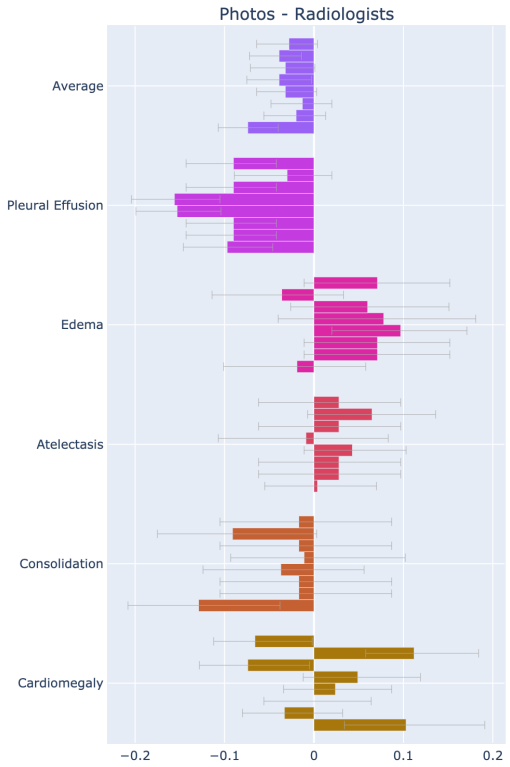
12.2 Methods

12.2.1 Photos of Chest X-rays

We collected a test set of photos of chest x-rays, described in [168]. In this set, chest X-rays from each CheXpert test study were displayed on a non-diagnostic computer monitor. Chest X-rays



(a) MCC differences of 8 chest X-ray models on different pathologies between photos of the X-rays and the original X-rays with 95% confidence intervals.



(b) MCC differences of the same models on photos of chest X-rays compared to radiologist performance with 95% confidence intervals.

were displayed in full screen on a computer monitor with 1920×1080 screen resolution and a black background. A physician was instructed to capture the photos, keeping the mobile camera stable and center the lung fields in the camera view. A time-restriction of 5 seconds per image was imposed to simulate a busy healthcare environment. Subsequent inspection of photos showed that they were taken with slightly varying angles; some photos included artefacts such as Moiré patterns and subtle screen glare. Photos were labeled using the ground truth for the corresponding digital X-ray image. The reference standard on this set was determined using a majority vote of 5 board-certified radiologists. Three separate board-certified radiologists were used for the comparison against the models and all radiologists used the original chest X-ray images for making their diagnoses, rather than the photos.

12.2.2 Running Models on New Test Sets

CheXpert used a hidden test set for official evaluation of models. Teams submitted their executable code, which was then run on a test set that was not publicly readable to preserve the integrity

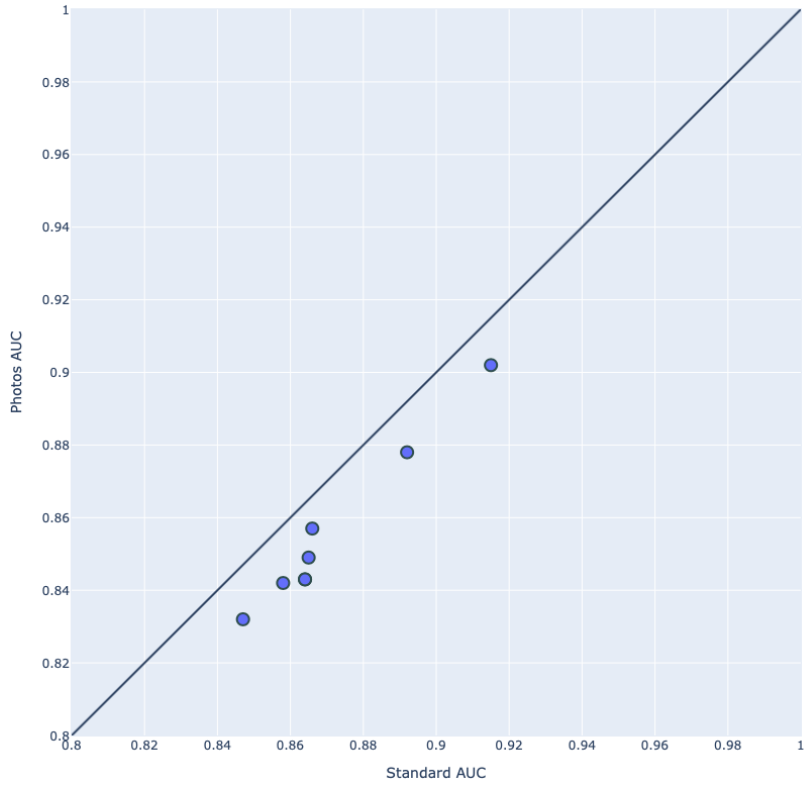


Figure 12.3: Comparison of the average AUC of 8 individual models on photos of chest X-rays compared to on standard images

of the test results. We made use of the CodaLab platform to re-run these chest X-ray models by substituting the hidden CheXpert test set with the datasets used in this study.

12.2.3 Evaluation Metrics

Our primary evaluation metric was Matthew's Correlation Coefficient (MCC), a statistical rate which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (true positives, false negatives, true negatives, and false positives); MCC is proportionally both to the size of positive elements and the size of negative elements in the dataset [46].

We reported the average MCC of 8 models for five pathologies, namely atelectasis, cardiomegaly, consolidation, edema, and pleural effusion. Additionally, in experiments comparing the models on standard chest X-rays to photos of chest X-rays, we reported the AUC and MCC of the models. In experiments comparing models to board-certified radiologists, we reported the difference in MCC for each of the five pathologies.

12.3 Results

12.3.1 Model Performance on Photos of Chest X-rays vs Original X-rays.

Performance Drop In Application To Photos

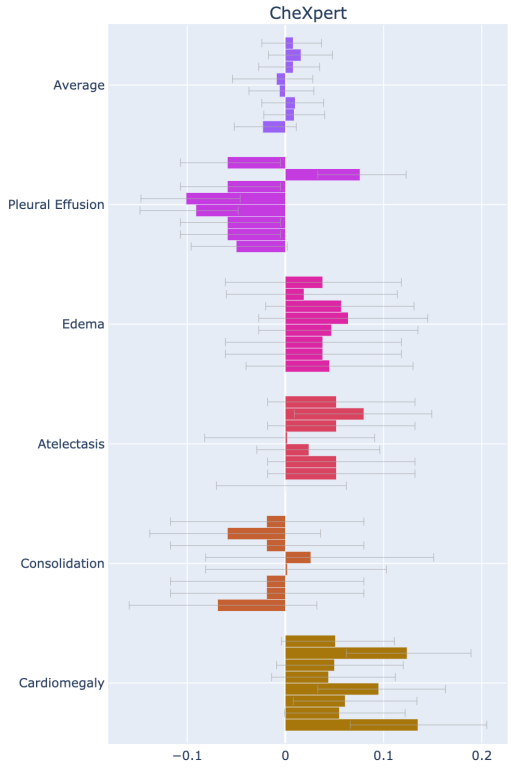
In comparison of model performance on digital chest X-rays to photos, all eight models experienced a statistically significant drop in task performance on photos with an average drop of 0.036 MCC (95% CI 0.024, 0.048) (See Figure 12.2a, Table 12.1). All models had a statistically significant drop on at least one of the pathologies between native digital image to photos. One model had a statistically significant drop in performance on three pathologies: pleural effusion, edema, and consolidation. Two models had a significant drop on two pathologies: one on pleural effusion and edema, and the other on pleural effusion and cardiomegaly. The cardiomegaly and pleural effusion tasks led to decreased performance in five and four models respectively.

Performance on Photos In Comparison to Radiologist Performance on Standard Images

In comparison of performance of models on photos compared to radiologist performance, three out of eight models performed significantly worse than radiologists on average, and the other five had no significant difference (see Figure 12.2b). On specific pathologies, there were some models that had a significantly higher performance than radiologists: two models on cardiomegaly, and one model on edema. Conversely, there were some models that had a significantly lower performance than radiologists: two models on cardiomegaly, and one model on consolidation. The pathology with the greatest number of models that had a significantly lower performance than radiologists was pleural effusion (seven models).

Performance drop in context of radiologist performance

Our results demonstrated that while most models experienced a significant drop in performance when applied to photos of chest X-rays compared to the native digital image, their performance was nonetheless largely equivalent to radiologist performance. We found that although there were thirteen times that models had a statistically significant drop in performance on photos on the different pathologies, the models had significantly lower performance than radiologists only 6 of those 13 times. Comparison to radiologist performance provides context in regard to clinical applicability: several models remained comparable to radiologist performance standard despite decreased performance on photos. Further investigation could be directed towards understanding how different model training procedures may affect model generalization to photos of chest X-rays, and understanding etiologies behind trends for changes in performance for specific pathologies or specific artifacts.



(a) MCC differences in performance of models on the CheXpert test set, with 95% confidence intervals (higher than 0 is in favor of the models being better).

(b) MCC differences in performance of the same models compared to another set of radiologists across the same pathologies on an external institution's (NIH) data.

Implication

While using photos of chest X-rays to input into chest X-ray algorithms could enable any physician with a smartphone to get instant AI algorithm assistance, the performance of chest X-ray algorithms on photos of chest X-rays has not been thoroughly investigated. Several studies have highlighted the importance of generalizability of computer vision models with noise in [97]. [58] demonstrated that deep neural networks perform poorly compared to humans on image classification on distorted images. [73], [203] have found that convolutional neural networks trained on specific image corruptions did not generalize, and the error patterns of network and human predictions were not similar on noisy and elastically deformed images.

12.3.2 Comparison of Models and Radiologists on External Institution

We measured the change in diagnostic performance of the same eight chest X-ray models on chest X-rays obtained at a different institution. We applied these models, trained on the CheXpert dataset

Institution	Comparison	Average	Pleural Effusion	Edema	Atelectasis	Consolidation	Cardiomegaly
CheXpert	Radiologists	0.568 (0.542,0.597)	0.671 (0.618,0.727)	0.507 (0.431,0.570)	0.548 (0.496,0.606)	0.359 (0.262,0.444)	0.566 (0.511,0.620)
	Models	0.570 (0.543,0.599)	0.621 (0.575,0.670)	0.550 (0.474,0.637)	0.587 (0.529,0.640)	0.336 (0.264,0.418)	0.643 (0.584,0.695)
	Models - Radiologists	0.002 (-0.028,0.030)	-0.05 (-0.092,-0.007)	0.043 (-0.033,0.114)	0.039 (-0.029,0.106)	-0.022 (-0.104,0.076)	0.077 (0.040,0.135)
NIH	Radiologists	0.537 (0.515,0.555)	0.642 (0.590,0.690)	0.618 (0.549,0.669)	0.469 (0.423,0.515)	0.455 (0.385,0.509)	0.492 (0.443,0.530)
	Models	0.578 (0.551,0.601)	0.673 (0.605,0.734)	0.662 (0.582,0.742)	0.529 (0.454,0.595)	0.551 (0.499,0.623)	0.517 (0.466,0.567)
	Models - Radiologists	0.041 (0.010,0.072)	0.032 (-0.019,0.078)	0.044 (-0.028,0.124)	0.060 (-0.003,0.126)	0.096 (0.027,0.155)	0.025 (-0.028,0.078)

Table 12.3: MCC performance of models and radiologists on the CheXpert and NIH sets of chest X-rays, and their difference, with 95% confidence intervals.

from the Stanford Hospital, to a set of 420 frontal chest X-rays labeled as part of [182]. These X-rays are sourced from the ChestXray-14 dataset collected at the National Institutes of Health Clinical Center [239], and sampled to contain at least 50 cases of each pathology according to the original labels provided in the dataset. The reference standard on this set (NIH) was determined using a majority vote of three cardiothoracic subspecialty radiologists; six board-certified radiologists were used for comparison against the models.

Performance on external institution in comparison to radiologists

On the external set (NIH), none of the models performed statistically significantly worse than radiologists (see Figure 12.4b). On average over the pathologies, five models performed significantly better than radiologists. On specific pathologies, there were some models that achieved significantly better performance than radiologists: six models on consolidation, three models on cardiomegaly, four on edema, and two on atelectasis, one on pleural effusion.

Implication

Our finding that these models perform comparably to or at a level exceeding radiologists differs from a previous study which reported that a chest X-ray model failed to generalize to new populations or institutions separate from the training data, relying on institution specific and/or confounding cues to infer the label of interest [261]. Our findings may be attributed to the improvement in the generalizability of chest X-ray models owing to larger and higher-quality datasets that have been publicly released [106, 111] Future work should investigate specific aspects of model training and dataset quality and size that lend to these differences, and whether self-supervised training procedures [214] increase generalizability across institutions.

Performance change in context of radiologist performance

Comparing performances on the CheXpert and NIH test sets, we found that on the NIH data set, in 16 instances models had a significantly better performance than radiologists; on the internal CheXpert test set, we observed that in 6 instances, models had a significantly higher performance than radiologists (see Figure 12.4a). This difference may be attributed to a variety of factors including

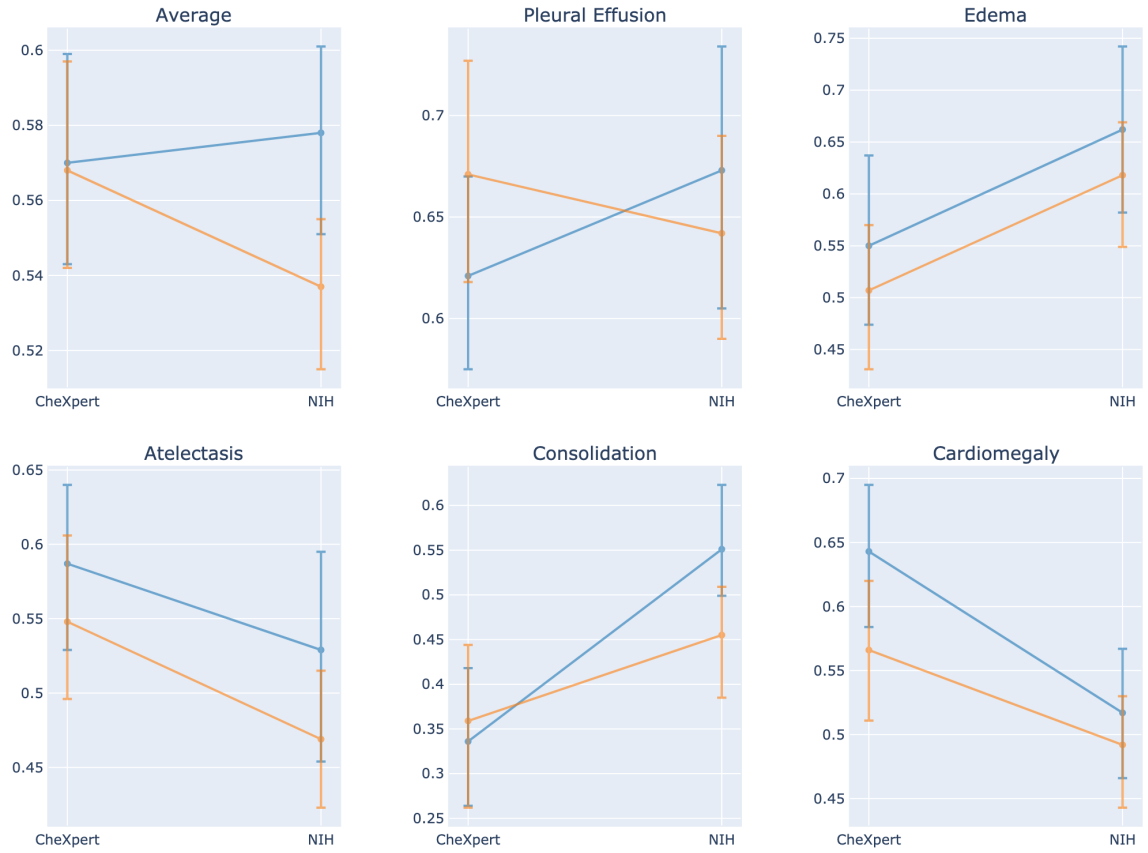


Figure 12.5: Overall change in performance of models (blue) and radiologists (orange) across CheXpert and the external institution dataset (NIH).

the difference in prevalence of pathologies or the difficulty in identifying them in the external test set compared to the internal set. We are able to contextualize the generalization ability of models to external institutions by comparing their differences to a radiologist performance benchmark, rather than provide a comparison of their absolute performances, which would not control for these possible differences. For instance, when considering cardiomegaly (see Figure 12.5), we observe a drop in model performance, which in isolation would indicate poor generalizability. However, in light of a similar drop in radiologist performance, we may be able to attribute the difference to differences in difficulties between the two datasets.

12.4 Discussion

The purpose of this work was to systematically address the key translation challenges for chest X-ray models in clinical application to common real-world scenarios. We found that several chest X-ray

models had a drop in performance when applied to smartphone photos of chest X-rays, but even with this drop, some models still performed comparably to radiologists. We also found that when models were tested on an external institution's data, they performed comparably to radiologists. In both forms of clinical distribution shifts we found that high-performance chest X-ray interpretation models trained on CheXpert produced clinically useful diagnostic performance.

Our work makes significant contributions over another investigation of chest X-ray models [184]. While their study considered the differences in AUC of models when applied to photos of X-rays, they did not (1) compare the resulting performances against radiologists, (2) investigate the drop in performances on specific tasks, or (3) analyze drops in performances of individual models across tasks. Finally, while they compared the performance of models to radiologists on an external dataset, they did not investigate the change in performance of models between the internal dataset and the external dataset.

Strengths of our study include our systematic investigation of generalization performance of several chest X-ray models developed by different teams. Limitations of our work include that our study is still retrospective in nature, and prospective studies would further advance understanding of generalization under distribution shifts. Our systematic examination of the generalization capabilities of existing models can be extended to other tasks in medical AI [61, 104, 117, 232, 228], and provide a framework for tracking technical readiness towards clinical translation.

Chapter 13

Conclusion

Overall, the thesis demonstrates progress in deep learning for medical image interpretation using the combination of developments in (1) algorithmic advancements in the context of large and small labeled datasets, (2) dataset advancements through clinically-informed curation and labeling, (3) and advancements in studies systematically analyzing the performance of algorithms under clinically relevant distribution shifts.

The future of AI technologies for medicine is in their seamless integration into clinical workflows. A key prerequisite for usage of AI technologies in clinical workflows is understanding how to combine AI predictions with human decisions. One straightforward model of collaboration is to simply share the AI system's prediction with the clinician. Even in this model, though, there are important consideration points, including the form the prediction should take and how the prediction should be communicated to the clinician (especially important given the "black-box"" criticism of deep learning models) [200]. More complex models for collaboration include mechanisms that adapt to the individual clinician or clinician-team, incorporating levels of expertise in the context of the decision being made. I believe it will be important to focus on the design and development of methods for AI-clinician collaboration in real hospital workflows. Designing for this interface would not only inform the development of learning algorithms for clinical medicine, but also would inform the workflow management of hospitals, accelerating the adoption and evolution of such technologies.

A major obstacle to the translation of AI technologies to clinical settings is the lack of frameworks for reasoning about failures and mitigating their potentially catastrophic effects in deployment settings. The design and management of clinical workflows and operations that can incorporate fail-safe mechanisms for AI technologies remains an open avenue for research. As an example, before medical image interpretation algorithms can be deployed in a hospital workflow, it will be important to identify when they are prone to errors [37]. For example, algorithms may be prone to errors when deployed in an environment in which the distribution of patients is different from that used to train the algorithm. As another example, models may fail when applied to medical imaging

exams with certain comorbidities and/or clinical features. When such failures can be identified in advance, clinical workflows can be designed with human-monitoring checks. The investigation and improvement of reliability of AI technologies can thus inform the design of clinical workflows with fail-safe mechanisms for the use of these technologies.

Finally, the labor-intensive process of constructing biomedical image datasets with accurate labels is expensive to scale, limiting limits the scope of medical applications which can be effectively tackled using data-intensive deep learning methods. Self-supervised contrastive learning approaches have demonstrated promising results on natural image classification tasks - these may enable a paradigm shift for training models for medicine where a relatively small number of annotations enables training of highly-accurate models. However, these methods are still in their infancy, with few works exploiting properties of data unique to medical settings. For instance, natural sources of supervision for medical images, including associated free-text reports and structured clinical data, are starting to be incorporated into contrastive learning mechanisms. I believe that development of methodologies which work across biomedical image, language, structured and multi-modal data will be important for medical settings where only a small set of expert annotations can be acquired.

As we tackle these key challenges over the next years, we will enable a world in which AI will transform medicine and consequently patient lives.

Bibliography

[1] Michael David Abràmoff, Yiyue Lou, Ali Erginay, Warren Clarida, Ryan Amelon, James C Folk, and Meindert Niemeijer. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. *Invest Ophthalmol Vis Sci*, 57(13):5200–5206, Oct 2016.

[2] Ash A. Alizadeh, Michael B. Eisen, R. Eric Davis, Chi Ma, Izidore S. Lossos, Andreas Rosenwald, Jennifer C. Boldrick, Hajeer Sabet, Truc Tran, Xin Yu, John I. Powell, Liming Yang, Gerald E. Marti, Troy Moore, James Hudson, Lisheng Lu, David B. Lewis, Robert Tibshirani, Gavin Sherlock, Wing C. Chan, Timothy C. Greiner, Dennis D. Weisenburger, James O. Armitage, Roger Warnke, Ronald Levy, Wyndham Wilson, Michael R. Grever, John C. Byrd, David Botstein, Patrick O. Brown, and Louis M. Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, February 2000.

[3] Ash A. Alizadeh, Andrew J. Gentles, Alvaro J. Alencar, Chih Long Liu, Holbrook E. Kohrt, Roch Houot, Matthew J. Goldstein, Shuchun Zhao, Yasodha Natkunam, Ranjana H. Advani, Randy D. Gascoyne, Javier Briones, Robert J. Tibshirani, June H. Myklebust, Sylvia K. Plevritis, Izidore S. Lossos, and Ronald Levy. Prediction of survival in diffuse large b-cell lymphoma based on the expression of 2 genes reflecting tumor and microenvironment. *Blood*, 118(5):1350–1358, August 2011.

[4] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[5] Douglas G Altman and J Martin Bland. How to obtain the p value from a confidence interval. *BMJ*, 343, 2011.

[6] Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro,

JingDong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 173–182, 2016.

[7] Savvas Andronikou, Kieran McHugh, Nuraan Abdurahman, Bryan Khoury, Victor Mngomezulu, William E. Brant, Ian Cowan, Mignon McCulloch, and Nathan Ford. Paediatric radiology seen from africa. part i: providing diagnostic imaging to a young population. *Pediatric Radiology*, 41(7):811–825, Jul 2011.

[8] Mauro Annarumma, Samuel J. Withey, Robert J. Bakewell, Emanuele Pesce, Vicky Goh, and Giovanni Montana. Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. *Radiology*, 291(1):196–202, January 2019. Publisher: Radiological Society of North America.

[9] Ioannis D Apostolopoulos and Tzani A Mpesiana. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine*, page 1, 2020.

[10] Shane G Artis, RG Mark, and GB Moody. Detection of atrial fibrillation using artificial neural networks. In *Computers in Cardiology 1991, Proceedings.*, pages 173–176. IEEE, 1991.

[11] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. *arXiv preprint arXiv:2101.05224*, 2021.

[12] J W Baars, D de Jong, E M Willemse, L Gras, O Dalesio, P v Heerde, P C Huygens, H v d Lelie, and A E G Kr v.d. Borne. Diffuse large b-cell non-hodgkin lymphomas: the clinical relevance of histological subclassification. *British Journal of Cancer*, 79(11-12):1770–1776, March 1999.

[13] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15535–15545, 2019.

[14] Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E. Petersen, Yike Guo, Paul M. Matthews, and Daniel Rueckert. Self-supervised learning for cardiac MR image segmentation by anatomical position prediction. *CoRR*, abs/1907.02757, 2019.

[15] Ioana Baiu and David Spain. Rib fractures. *The Journal of the American Medical Association*, 321(18):1836–1836, 2019. Publisher: American Medical Association.

[16] Imon Banerjee, Matthew C. Chen, Matthew P. Lungren, and Daniel L. Rubin. Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort. *Journal of Biomedical Informatics*, 77:11–20, 2018.

[17] Katia Basso and Riccardo Dalla-Favera. Germinal centres and b cell lymphomagenesis. *Nature Reviews Immunology*, 15(3):172–184, February 2015.

[18] Andrew H. Beck, Ankur R. Sangoi, Samuel Leung, Robert J. Marinelli, Torsten O. Nielsen, Marc J. van de Vijver, Robert B. West, Matt van de Rijn, and Daphne Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational Medicine*, 3(108):108ra113–108ra113, 2011.

[19] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Towards open set deep networks*, pages 1563–1572, 2016.

[20] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. *NPJ digital medicine*, 3(1):1–8, 2020.

[21] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[22] K Berbaum, Jr EA Franken, and WL Smith. The effect of comparison films upon resident interpretation of pediatric chest radiographs. *Investigative radiology*, 20:124–128, 1985.

[23] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[24] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, January 2004.

[25] Selen Bozkurt, Emel Alkim, Imon Banerjee, and Daniel L. Rubin. Automated Detection of Measurements and Their Descriptors in Radiology Reports Using a Hybrid Natural Language Processing Algorithm. *Journal of Digital Imaging*, 32(4):544–553, August 2019.

[26] Adrian Brady, Risteárd Ó Laoide, Peter McCarthy, and Ronan McDermott. Discrepancy and error in radiology: concepts, causes and consequences. *The Ulster medical journal*, 81(1):3, 2012.

[27] Keno K. Bressem, Lisa Adams, Christoph Erxleben, Bernd Hamm, Stefan Niehues, and Janis Vahldiek. Comparing different deep learning architectures for classification of chest radiographs, 2020.

[28] Lindsay P Busby, Jesse L Courtier, and Christine M Glastonbury. Bias in radiology: the how and why of misses and misinterpretations. *Radiographics*, 38(1):236–247, 2018.

[29] 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, 2019.

[30] Remi Cadene. pretrainedmodels 0.7.4. https://pypi.org/project/pretrainedmodels/, 2018.

[31] Alison Callahan, Jason A. Fries, Christopher Ré, James I. Huddleston, Nicholas J. Giori, Scott Delp, and Nigam H. Shah. Medical device surveillance with electronic health records. *npj Digital Medicine*, 2(1):1–10, September 2019. Number: 1 Publisher: Nature Publishing Group.

[32] Tianshi Cao, Chinwei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. A benchmark of medical out of distribution detection. *arXiv preprint*, 2020.

[33] Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations, 2020.

[34] Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34(5):301–310, October 2001.

[35] Eugene Charniak and Mark Johnson. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 173–180. Association for Computational Linguistics, 2005.

[36] David Chen, Sijia Liu, Paul Kingsbury, Sunghwan Sohn, Curtis B. Storlie, Elizabeth B. Habermann, James M. Naessens, David W. Larson, and Hongfang Liu. Deep learning and alternative learning strategies for retrospective real-world clinical data. *npj Digital Medicine*, 2(1):43, December 2019.

[37] Emma Chen, Andy Kim, Rayan Krishnan, Jin Long, Andrew Y Ng, and Pranav Rajpurkar. Chexbreak: Misclassification identification for deep learning models interpreting chest x-rays. *arXiv preprint arXiv:2103.09957*, 2021.

[38] Matthew C. Chen, Robyn L. Ball, Lingyao Yang, Nathaniel Moradzadeh, Brian E. Chapman, David B. Larson, Curtis P. Langlotz, Timothy J. Amrhein, and Matthew P. Lungren. Deep Learning to Classify Radiology Free-Text Reports. *Radiology*, 286(3):845–852, November 2017. Publisher: Radiological Society of North America.

[39] Po-Hao Chen, Hanna Zafar, Maya Galperin-Aizenberg, and Tessa Cook. Integrating Natural Language Processing and Machine Learning Algorithms to Categorize Oncologic Response in Radiology Reports. *Journal of Digital Imaging*, 31(2):178–184, April 2018.

[40] S. Chen and Q. Zhao. Shallowing deep networks: Layer-wise pruning based on feature representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):3048–3056, 2019.

[41] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[42] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*, 2020.

[43] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[44] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A survey of model compression and acceleration for deep neural networks. *CoRR*, abs/1710.09282, 2017.

[45] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019.

[46] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6, 2020.

[47] François Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016.

[48] Gari D Clifford, Chengyu Liu, Benjamin Moody, H Lehman Li-wei, Ikaro Silva, Qiao Li, AE Johnson, and Roger G Mark. Af classification from a short single lead ecg recording: the physionet/computing in cardiology challenge 2017. In *2017 Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2017.

[49] Douglas A Coast, Richard M Stern, Gerald G Cano, and Stanley A Briller. An approach to cardiac arrhythmia analysis using hidden markov models. *IEEE Transactions on biomedical Engineering*, 37(9):826–836, 1990.

[50] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 2021/01/13 1960.

[51] Jeffrey De Fauw, Joseph R. Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, George van den Driessche, Balaji Lakshminarayanan, Clemens Meyer, Faith

Mackinder, Simon Bouton, Kareem Ayoub, Reena Chopra, Dominic King, Alan Karthikesalingam, Cían O. Hughes, Rosalind Raine, Julian Hughes, Dawn A. Sim, Catherine Egan, Adnan Tufail, Hugh Montgomery, Demis Hassabis, Geraint Rees, Trevor Back, Peng T. Khaw, Mustafa Suleyman, Julien Cornebise, Pearse A. Keane, and Olaf Ronneberger. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, Sep 2018.

[52] Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. Universal stanford dependencies: A cross-linguistic typology. In *LREC*, volume 14, pages 4585–4592, 2014.

[53] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.

[54] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2015.

[55] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[56] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*, May 2019. arXiv: 1810.04805.

[57] J. Diebold, J.R. Anderson, J.O. Armitage, J.M. Connors, K.A. Maclennan, H.K. Müller-Hermelink, B.N. Nathwani, F. Ullrich, and D.D. Weisenburger. Diffuse large b-cell lymphoma: A clinicopathologic analysis of 444 cases classified according to the updated kiel classification. *Leukemia & Lymphoma*, 43(1):97–104, January 2002.

[58] Samuel Dodge and Lina Karam. A Study and Comparison of Human and Deep Learning Recognition Performance under Visual Distortions. In *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–7, July 2017.

[59] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[60] Ignat Drozdov, Daniel Forbes, Benjamin Szubert, Mark Hall, Chris Carlin, and David J. Lowe. Supervised and unsupervised language modelling in Chest X-Ray radiological reports. *PLOS ONE*, 15(3):e0229963, March 2020. Publisher: Public Library of Science.

[61] Tony Duan, Pranav Rajpurkar, Dillon Laird, Andrew Y. Ng, and Sanjay Basu. Clinical Value of Predicting Individual Treatment Effects for Intensive Blood Pressure Therapy: A Machine Learning Experiment to Estimate Treatment Effects from Randomized Trial Data. *Circulation: Cardiovascular Quality and Outcomes*, 12(3), March 2019.

[62] Dale Dubin. *Rapid Interpretation of EKG's*. USA: Cover Publishing Company, 1996, 1996.

[63] Jared Dunnmon, Alexander Ratner, Nishith Khandwala, Khaled Saab, Matthew Markert, Hersh Sagreiya, Roger Goldman, Christopher Lee-Messer, Matthew Lungren, Daniel Rubin, and Christopher Ré. Cross-Modal Data Programming Enables Rapid Medical Machine Learning. *arXiv:1903.11101 [cs, eess, stat]*, March 2019. arXiv: 1903.11101.

[64] Jared Dunnmon, Alexander Ratner, Nishith Khandwala, Khaled Saab, Matthew Markert, Hersh Sagreiya, Roger Goldman, Christopher Lee-Messer, Matthew Lungren, Daniel Rubin, and Christopher Ré. Cross-modal data programming enables rapid medical machine learning, 2019.

[65] B. Efron and R. Tibshirani. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, 1(1):54–75, February 1986. Publisher: Institute of Mathematical Statistics.

[66] Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75, 1986.

[67] Marianne Engelhard, Gunter Brittinger, Dieter Huhn, Heinrich H. Gerhartz, Peter Meusers, Wolfgang Siegert, Eckhard Thiel, Wolfgang Wilmanns, Ulker Aydemir, Stefan Bierwolf, Henrik Griesser, Markus Tiemann, and Karl Lennert. Subclassification of Diffuse Large B-Cell Lymphomas According to the Kiel Classification: Distinction of Centroblastic and Immunoblastic Lymphomas Is a Significant Prognostic Risk Factor. *Blood*, 89(7):2291–2297, 04 1997.

[68] Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? pages 201–208, 2010.

[69] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):1–9, 2021.

[70] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[71] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, February 2017.

[72] Kai Fu, Dennis D. Weisenburger, William W.L. Choi, Kyle D. Perry, Lynette M. Smith, Xinlan Shi, Christine P. Hans, Timothy C. Greiner, Philip J. Bierman, R. Gregory Bociek, James O. Armitage, Wing C. Chan, and Julie M. Vose. Addition of rituximab to standard chemotherapy improves the survival of both the germinal center b-cell–like and non–germinal center b-cell–like subtypes of diffuse large b-cell lymphoma. *Journal of Clinical Oncology*, 26(28):4587–4594, October 2008.

[73] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv:1811.12231 [cs, q-bio, stat]*, January 2019.

[74] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. Publisher: IEEE.

[75] Esteban F Gershanik, Ronilda Lacson, and Ramin Khorasani. Critical finding capture in the impression section of radiology reports. In *AMIA Annual Symposium Proceedings*, volume 2011, page 465. American Medical Informatics Association, 2011.

[76] John K Gohagan, Philip C Prorok, Richard B Hayes, and Barnett-S Kramer. The prostate, lung, colorectal and ovarian (plco) cancer screening trial of the national cancer institute: history, organization, and status. *Controlled clinical trials*, 21(6):251S–272S, 2000.

[77] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[78] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014.

[79] Hans Goost, Johannes Witten, Andreas Heck, Dariusch Hadizadeh, Oliver Weber, Ingo Gräff, Christof Burger, Mareen Montag, Felix Koerfer, and Koroush Kabir. Image and diagnosis quality of x-ray image transmission via cell phone camera: A project study evaluating quality and reliability. *PloS one*, 7:e43402, 10 2012.

[80] John W Graham. Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60:549–576, 2009.

[81] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images, 2018.

[82] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. *arXiv preprint arXiv:1801.09927*, 2018.

[83] Maya E Guglin and Deepak Thatai. Common errors in computer electrocardiogram interpretation. *International journal of cardiology*, 106(2):232–237, 2006.

[84] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, 2016.

[85] Gonzalo Gutiérrez-García, Teresa Cardesa-Salzmann, Fina Climent, Eva González-Barca, Santiago Mercadal, José L. Mate, Juan M. Sancho, Leonor Arenillas, Sergi Serrano, Lourdes Escoda, Salomé Martínez, Alexandra Valera, Antonio Martínez, Pedro Jares, Magdalena Pinyol, Adriana García-Herrera, Alejandra Martínez-Trillos, Eva Giné, Neus Villamor, Elías Campo, Luis Colomo, and Armando López-Guillermo and. Gene-expression profiling and not immunophenotypic algorithms predicts prognosis in patients with diffuse large b-cell lymphoma treated with immunochemotherapy. *Blood*, 117(18):4836–4843, May 2011.

[86] Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1):65–69, January 2019.

[87] David M Hansell, Alexander A Bankier, Heber MacMahon, Theresa C McLoud, Nestor L Muller, and Jacques Remy. Fleischner society: glossary of terms for thoracic imaging. *Radiology*, 246(3):697–722, 2008.

[88] Frank E. Harrell, Kerry L. Lee, and Daniel B. Mark. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, February 1996.

[89] Jr Harrell, Frank E., Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546, 05 1982.

[90] Saeed Hassanpour, Curtis P. Langlotz, Timothy J. Amrhein, Nicholas T. Befera, and Matthew P. Lungren. Performance of a Machine Learning Classifier of Knee MRI Reports

in Two Large Academic Radiology Practices: A Tool to Estimate Diagnostic Yield. *American Journal of Roentgenology*, 208(4):750–753, January 2017. Publisher: American Roentgen Ray Society.

[91] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[92] Kaiming He, Ross B. Girshick, and Piotr Dollár. Rethinking imagenet pre-training. *CoRR*, abs/1811.08883, 2018.

[93] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[94] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[95] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016.

[96] Bo Hedén, Mattias Ohlsson, Holger Holst, Mattias Mjöman, Ralf Rittner, Olle Pahlm, Carsten Peterson, and Lars Edenbrandt. Detection of frequently overlooked electrocardiographic lead reversals using artificial neural networks. *The American journal of cardiology*, 78(5):600–604, 1996.

[97] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *CoRR*, abs/1903.12261, 2019.

[98] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[99] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[100] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[101] Olle G. Holmberg, Niklas D. Köhler, Thiago Martins, Jakob Siedlecki, Tina Herold, Leonie Keidel, Ben Asani, Johannes Schiefelbein, Siegfried Priglinger, Karsten U. Kortuem, and Fabian J. Theis. Self-supervised retinal thickness prediction enables deep learning from unlabeled data to boost classification of diabetic retinopathy. *bioRxiv*, 2019.

[102] Matej Horvat, Vesna Zadnik, Tanja Južnič Šetina, Lučka Boltežar, Jana Pahole Goličnik, Srdjan Novaković, and Barbara Jezeršek Novaković. Diffuse large b-cell lymphoma: 10 years' real-world clinical experience with rituximab plus cyclophosphamide, doxorubicin, vincristine and prednisolone. *Oncology Letters*, January 2018.

[103] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, volume 1, page 3, 2017.

[104] Shih-Cheng Huang, Tanay Kothari, Imon Banerjee, Chris Chute, Robyn L Ball, Norah Borus, Andrew Huang, Bhavik N Patel, Pranav Rajpurkar, Jeremy Irvin, et al. Penet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric ct imaging. *NPJ digital medicine*, 3(1):1–9, 2020.

[105] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[106] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

[107] Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. Speeding up convolutional neural networks with low rank expansions. *CoRR*, abs/1405.3866, 2014.

[108] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

[109] Mika S Jain and Tarik F Massoud. Predicting tumour mutational burden from histopathological images using multiscale deep learning. *bioRxiv*, 2020.

[110] Saahil Jain, Akshay Smit, Steven QH Truong, Chanh DT Nguyen, Minh-Thanh Huynh, Mudit Jain, Victoria A Young, Andrew Y Ng, Matthew P Lungren, and Pranav Rajpurkar. Visualchexbert: addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 105–115, 2021.

[111] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv:1901.07042 [cs, eess]*, November 2019. arXiv: 1901.07042.

[112] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017.

[113] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning, 2020.

[114] Jakob Nikolas Kather, Alexander T. Pearson, Niels Halama, Dirk Jäger, Jeremias Krause, Sven H. Loosen, Alexander Marx, Peter Boor, Frank Tacke, Ulf Peter Neumann, Heike I. Grabsch, Takaki Yoshikawa, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, Christian Trautwein, and Tom Luedde. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature Medicine*, 25(7):1054–1056, 2019.

[115] Alexander Ke, William Ellsworth, Oishi Banerjee, Andrew Y Ng, and Pranav Rajpurkar. Chextransfer: performance and parameter efficiency of imagenet models for chest x-ray interpretation. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 116–124, 2021.

[116] Christopher J. Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1):195, December 2019.

[117] Amirhossein Kiani, Bora Uyumazturk, Pranav Rajpurkar, Alex Wang, Rebecca Gao, Erik Jones, Yifan Yu, Curtis P. Langlotz, Robyn L. Ball, Thomas J. Montine, Brock A. Martin, Gerald J. Berry, Michael G. Ozawa, Florette K. Hazard, Ryanne A. Brown, Simon B. Chen, Mona Wood, Libby S. Allard, Lourdes Ylagan, Andrew Y. Ng, and Jeanne Shen. Impact of a deep learning assistant on the histopathologic classification of liver cancer. *npj Digital Medicine*, 3(1):23, December 2020.

[118] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[119] Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals. *arXiv preprint arXiv:2005.13249*, 2020.

[120] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, Carol Willing, and Jupyter development team. Jupyter notebooks - a publishing format for reproducible computational workflows. In Fernando Loizides and Birgit Scmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87–90. IOS Press, 2016.

[121] Anton Kolesov, Dmitry Kamyshenkov, Maria Litovchenko, Elena Smekalova, Alexey Golovizin, and Alex Zhavoronkov. On multilabel classification methods of incompletely labeled biomedical text data. *Computational and mathematical methods in medicine*, 2014, 2014.

[122] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[123] Pulkit Kumar, Monika Grewal, and Muktabh Mayank Srivastava. Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs. In *International Conference Image Analysis and Recognition*, pages 546–552. Springer, 2018.

[124] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *ICLR Workshop*, 2017.

[125] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *CoRR*, abs/1607.02533, 2016. _eprint: 1607.02533.

[126] Pablo Laguna, Roger G Mark, A Goldberg, and George B Moody. A database for evaluation of algorithms for measurement of qt and other waveform intervals in the ecg. In *Computers in Cardiology 1997*, pages 673–676. IEEE, 1997.

[127] Jonathan Laserson, Christine Dan Lantsman, Michal Cohen-Sfady, Itamar Tamir, Eli Goz, Chen Brestel, Shir Bar, Maya Atar, and Eldad Elnekave. Textray: Mining clinical reports to gain a broad understanding of chest x-rays. *arXiv preprint arXiv:1806.02121*, 2018.

[128] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[129] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, page btz682, September 2019. arXiv: 1901.08746.

[130] G. Lenz, G. W. Wright, N. C. T. Emre, H. Kohlhammer, S. S. Dave, R. E. Davis, S. Carty, L. T. Lam, A. L. Shaffer, W. Xiao, J. Powell, A. Rosenwald, G. Ott, H. K. Muller-Hermelink, R. D. Gascoyne, J. M. Connors, E. Campo, E. S. Jaffe, J. Delabie, E. B. Smeland, L. M. Rimsza, R. I. Fisher, D. D. Weisenburger, W. C. Chan, and L. M. Staudt. Molecular subtypes of diffuse large b-cell lymphoma arise by distinct genetic pathways. *Proceedings of the National Academy of Sciences*, 105(36):13520–13525, September 2008.

[131] John P. Leonard, Kathryn S. Kolibaba, James A. Reeves, Anil Tulpule, Ian W. Flinn, Tatjana Kolevska, Robert Robles, Christopher R. Flowers, Robert Collins, Nicholas J. DiBella, Steven W. Papish, Parameswaran Venugopal, Andrew Horodner, Amir Tabatabai, Julio Hajdenberg, Jaehong Park, Rachel Neuwirth, George Mulligan, Kaveri Suryanarayan, Dixie-Lee Esseltine, and Sven de Vos. Randomized phase II study of r-CHOP with or without bortezomib in previously untreated patients with non–germinal center b-cell–like diffuse large b-cell lymphoma. *Journal of Clinical Oncology*, 35(31):3538–3546, November 2017.

[132] Cuiwei Li, Chongxun Zheng, and Changfeng Tai. Detection of ECG characteristic points using wavelet transforms. *IEEE Transactions on biomedical Engineering*, 42(1):21–28, 1995.

[133] Feng Li, Zheng Liu, Hua Chen, Minshan Jiang, Xuedian Zhang, and Zhizheng Wu. Automatic Detection of Diabetic Retinopathy in Retinal Fundus Photographs Based on Deep Learning Algorithm. *Translational Vision Science & Technology*, 8(6):4–4, 11 2019.

[134] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *ICCV*, pages 1928–1936, 2017.

[135] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Fei-Fei Li. Thoracic disease identification and localization with limited supervision. *arXiv preprint arXiv:1711.06373*, 2017.

[136] Percy Liang, Isabelle Guyon, Evelyne Viegas, Sergio Escalera, Xavier Baró Solé, and Eric Carmichael. *CodaLab*, 2020. https://codalab.org.

[137] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[138] Jingyu Liu, Gangming Zhao, Yu Fei, Ming Zhang, Yizhou Wang, and Yizhou Yu. Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10632–10641, 2019.

[139] Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The lancet digital health*, 1(6):e271–e297, 2019.

[140] Markos Markou and Sameer Singh. Novelty detection: a review—part 1: statistical approaches. *Signal processing*, 83(12):2481–2497, 2003. Publisher: Elsevier.

[141] Juan Pablo Martínez, Rute Almeida, Salvador Olmos, Ana Paula Rocha, and Pablo Laguna. A wavelet-based ECG delineator: evaluation on standard databases. *IEEE Transactions on biomedical engineering*, 51(4):570–581, 2004.

[142] David McClosky. Any domain parsing: automatic domain adaptation for natural language parsing. 2010.

[143] Matthew B. A. McDermott, Tzu Ming Harry Hsu, Wei-Hung Weng, Marzyeh Ghassemi, and Peter Szolovits. Chexpert++: Approximating the chexpert labeler for speed,differentiability, and probabilistic output, 2020.

[144] SL Melo, LP Caloba, and J Nadal. Arrhythmia analysis using artificial neural network and decimated electrocardiographic data. In *Computers in Cardiology 2000*, pages 73–76. IEEE, 2000.

[145] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[146] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020.

[147] Akinori Mitani, Abigail Huang, Subhashini Venugopalan, Greg S. Corrado, Lily Peng, Dale R. Webster, Naama Hammel, Yun Liu, and Avinash V. Varadarajan. Detection of anaemia from retinal fundus images via deep learning. *Nature Biomedical Engineering*, 4(1):18–27, Jan 2020.

[148] George B Moody and Roger G Mark. A new method for detecting atrial fibrillation using RR intervals. *Computers in Cardiology*, 10(1):227–230, 1983.

[149] George B Moody and Roger G Mark. The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001.

[150] Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert, 2021.

[151] Gustav Mårtensson, Daniel Ferreira, Tobias Granberg, Lena Cavallin, Ketil Oppedal, Alessandro Padovani, Irena Rektorova, Laura Bonanni, Matteo Pardini, and Milica G Kramberger. The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. *Medical Image Analysis*, page 101714, 2020. Publisher: Elsevier.

[152] Hirokazu Nakamine, Dennis D. Weisenburger, Robert G. Bagin, Julie M. Vose, Martin A. Bast, Philip J. Bierman, and James O. Armitage. Prognostic significance of clinical and pathologic features in diffuse large b-cell lymphoma. *Cancer*, 71(10):3130–3137, May 1993.

[153] Ju Gang Nam, Sunggyun Park, Eui Jin Hwang, Jong Hyuk Lee, Kwang-Nam Jin, Kun Young Lim, Thienkai Huy Vu, Jae Ho Sohn, Sangheum Hwang, Jin Mo Goo, and others. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary

nodules on chest radiographs. *Radiology*, 290(1):218–228, 2018. Publisher: Radiological Society of North America.

[154] Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. Facebook FAIR's WMT19 News Translation Task Submission. *arXiv:1907.06616 [cs]*, July 2019. arXiv: 1907.06616.

[155] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, 2021.

[156] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

[157] Luke Oakden-Rayner. Exploring large scale public medical image datasets, 2019.

[158] Tobi Olatunji, Li Yao, Ben Covington, Alexander Rhodes, and Anthony Upton. Caveats in generating medical imaging labels from radiology reports, 2019.

[159] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[160] Jiapu Pan and Willis J Tompkins. A real-time QRS detection algorithm. *IEEE transactions on biomedical engineering*, (3):230–236, 1985.

[161] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[162] Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports, 2017.

[163] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. *arXiv:1906.05474 [cs]*, June 2019. arXiv: 1906.05474.

[164] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets, 2019.

[165] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[166] F. Perez and B. E. Granger. Ipython: A system for interactive scientific computing. *Computing in Science Engineering*, 9(3):21–29, 2007.

[167] Hieu H Pham, Tung T Le, Dat T Ngo, Dat Q Tran, and Ha Q Nguyen. Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. *arXiv preprint arXiv:2005.12734*, 2020.

[168] Nick A Phillips, Pranav Rajpurkar, Mark Sabini, Rayan Krishnan, Sharon Zhou, Anuj Pareek, Nguyet Minh Phu, Chris Wang, Andrew Y Ng, and Matthew P Lungren. Chexphoto: 10,000+ smartphone photos and synthetic photographic transformations of chest x-rays for benchmarking deep learning robustness. *arXiv preprint arXiv:2007.06199*, 2020.

[169] Marco AF Pimentel, David A Clifton, Lei Clifton, and Lionel Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014. Publisher: Elsevier.

[170] John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[171] Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. Investigating Backtranslation in Neural Machine Translation. *arXiv:1804.06189 [cs]*, April 2018. arXiv: 1804.06189.

[172] Ewoud Pons, Loes M. M. Braun, M. G. Myriam Hunink, and Jan A. Kors. Natural Language Processing in Radiology: A Systematic Review. *Radiology*, 279(2):329–343, April 2016. Publisher: Radiological Society of North America.

[173] EJ Potchen, JW Gard, P Lazar, P Lahaie, and M Andary. Effect of clinical history data on chest film interpretation-direction or distraction. In *Investigative Radiology*, volume 14, pages 404–404. LIPPINCOTT-RAVEN PUBL 227 EAST WASHINGTON SQ, PHILADELPHIA, PA 19106, 1979.

[174] Viraj Prabhu, Anitha Kannan, Geoffrey J Tso, Namit Katariya, Manish Chablani, David Sontag, and Xavier Amatriain. Open set medical diagnosis. *arXiv preprint arXiv:1910.02830*, 2019.

[175] The Non-Hodgkin's Lymphoma Classification Project. A Clinical Evaluation of the International Lymphoma Study Group Classification of Non-Hodgkin's Lymphoma. *Blood*, 89(11):3909–3918, 06 1997.

[176] Preetham Putha, Manoj Tadepalli, Bhargava Reddy, Tarun Raj, Justy Antony Chiramal, Shalini Govil, Namita Sinha, Manjunath KS, Sundeep Reddivari, Pooja Rao, et al. Can artificial intelligence reliably report chest x-rays?: Radiologist validation of an algorithm trained on 1.2 million x-rays. *arXiv preprint arXiv:1807.07455*, 2018.

[177] Chunli Qin, Demin Yao, Yonghong Shi, and Zhijian Song. Computer-aided detection in chest radiography based on artificial intelligence: a survey. *BioMedical Engineering OnLine*, 17(1):113, August 2018.

[178] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009.

[179] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. *arXiv preprint arXiv:1712.04440*, 2017.

[180] Maithra Raghu, Chiyuan Zhang, Jon M. Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning with applications to medical imaging. *CoRR*, abs/1902.07208, 2019.

[181] Pranav Rajpurkar, Awni Y Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y Ng. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836*, 2017.

[182] Pranav Rajpurkar, Jeremy Irvin, Robyn L. Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P. Langlotz, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Francis G. Blankenberg, Jayne Seekins, Timothy J. Amrhein, David A. Mong, Safwan S. Halabi, Evan J. Zucker, Andrew Y. Ng, and Matthew P. Lungren. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLOS Medicine*, 15(11):e1002686, November 2018.

[183] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv:1711.05225 [cs, stat]*, November 2017. arXiv: 1711.05225.

[184] Pranav Rajpurkar, Anirudh Joshi, Anuj Pareek, Phil Chen, Amirhossein Kiani, Jeremy Irvin, Andrew Y Ng, and Matthew P Lungren. Chexpedition: investigating generalization challenges for translation of chest x-ray algorithms to the clinical setting. *arXiv preprint arXiv:2002.11379*, 2020.

[185] Pranav Rajpurkar, Anirudh Joshi, Anuj Pareek, Jeremy Irvin, Andrew Y. Ng, and Matthew Lungren. Chexphotogenic: Generalization of deep learning models for chest x-ray interpretation to photos of chest x-rays, 2020.

[186] Pranav Rajpurkar, Anirudh Joshi, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Chexternal: Generalization of deep learning models for chest x-ray interpretation to photos of chest x-rays and external clinical settings, 2021.

[187] Pranav Rajpurkar, Chloe O'Connell, Amit Schechter, Nishit Asnani, Jason Li, Amirhossein Kiani, Robyn L Ball, Marc Mendelson, Gary Maartens, Daniël J van Hoving, et al. Chexaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with hiv. *NPJ digital medicine*, 3(1):1–8, 2020.

[188] Pranav Rajpurkar, Allison Park, Jeremy Irvin, Chris Chute, Michael Bereket, Domenico Mastrodicasa, Curtis P Langlotz, Matthew P Lungren, Andrew Y Ng, and Bhavik N Patel. Appendixnet: Deep learning for diagnosis of appendicitis from a small dataset of ct exams using video pretraining. *Scientific reports*, 10(1):1–7, 2020.

[189] Suhail Raoof, David Feigin, Arthur Sung, Sabiha Raoof, Lavanya Irugulpati, and Edward C Rosenow III. Interpretation of plain chest roentgenogram. *Chest*, 141(2):545–558, 2012.

[190] Alex Ratner, Braden Hancock, Jared Dunnmon, Roger Goldman, and Christopher Ré. Snorkel MeTaL: Weak Supervision for Multi-Task Learning. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, DEEM'18, pages 1–4, Houston, TX, USA, June 2018. Association for Computing Machinery.

[191] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: rapid training data creation with weak supervision. *The VLDB Journal*, 29(2):709–730, May 2020.

[192] Peter A. Riedell and Sonali M. Smith. Double hit and double expressors in lymphoma: Definition and treatment. *Cancer*, 124(24):4622–4632, September 2018.

[193] Peter A. Riedell and Sonali M. Smith. Should we use cell of origin and dual-protein expression in treating DLBCL? *Clinical Lymphoma Myeloma and Leukemia*, 18(2):91–97, February 2018.

[194] Youngmin Ro and Jin Young Choi. Layer-wise pruning and auto-tuning of layer-wise learning rates in fine-tuning of deep networks, 2020.

[195] Andreas Rosenwald, Susanne Bens, Ranjana Advani, Sharon Barrans, Christiane Copie-Bergman, Mad-Helenie Elsensohn, Yaso Natkunam, Maria Calaminici, Birgitta Sander, Maryse Baia, Alexandra Smith, Daniel Painter, Luu Pham, Shuchun Zhao, Marita Ziepert, Ekaterina S. Jordanova, Thierry J. Molina, Marie José Kersten, Eva Kimby, Wolfram Klapper,

John Raemaekers, Norbert Schmitz, Fabrice Jardin, Wendy B.C. Stevens, Eva Hoster, Anton Hagenbeek, John G. Gribben, Reiner Siebert, Randy D. Gascoyne, David W. Scott, Philippe Gaulard, Gilles Salles, Catherine Burton, Daphne de Jong, Laurie H. Sehn, and Delphine Maucort-Boulch. Prognostic significance of myc rearrangement and translocation partner in diffuse large b-cell lymphoma: A study by the lunenburg lymphoma biomarker consortium. *Journal of Clinical Oncology*, 37(35):3359–3368, 2019. PMID: 31498031.

[196] Jonathan Rubin, Deepan Sanghavi, Claire Zhao, Kathy Lee, Ashequl Qadir, and Minnan Xu-Wilson. Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks. *arXiv preprint arXiv:1804.07839*, 2018.

[197] Khaled Saab, Jared Dunnmon, Roger Goldman, Alex Ratner, Hersh Sagreiya, Christopher Ré, and Daniel Rubin. Doubly Weak Supervision of Deep Learning Models for Head CT. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Lecture Notes in Computer Science, pages 811–819, Cham, 2019. Springer International Publishing.

[198] Tara N Sainath, Abdel-rahman Mohamed, Brian Kingsbury, and Bhuvana Ramabhadran. Deep convolutional neural networks for lvcsr. In *Acoustics, speech and signal processing (ICASSP), 2013 IEEE international conference on*, pages 8614–8618. IEEE, 2013.

[199] A. Salar, A. Fernández Sevilla, V. Romagosa, A. Domingo-Claros, E. González-Barca, J. Pera, J. Climent, and A. Grañena. Diffuse large b-cell lymphoma: is morphologic subdivision useful in clinical management? *European Journal of Haematology*, 60(3):202–208, April 2009.

[200] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Ng, et al. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *medRxiv*, 2021.

[201] Seelwan Sathitratanacheewin and Krit Pongpirul. Deep learning for automated classification of tuberculosis-related chest x-ray: Dataset specificity limits diagnostic performance generalizability. 2018.

[202] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association : JAMIA*, 17(5):507–513, 2010.

[203] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially Robust Generalization Requires More Data. In S. Bengio, H. Wallach,

H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5014–5026. Curran Associates, Inc., 2018.

[204] Adam B Schwartz, Gina Siddiqui, John S Barbieri, Amana L Akhtar, Woojin Kim, Ryan Littman-Quinn, Emily F Conant, Narainder K Gupta, Bryan A Pukenas, Parvati Ramchandani, Anna S Lev-Toaff, Jennifer D Tobey, Drew A Torigian, Amy H Praestgaard, and Carrie L Kovarik. The accuracy of mobile teleradiology in the evaluation of chest x-rays. *Journal of Telemedicine and Telecare*, 20(8):460–463, 2014. PMID: 25322696.

[205] David W. Scott. Cell-of-origin in diffuse large b-cell lymphoma: Are the assays ready for the clinic? *American Society of Clinical Oncology Educational Book*, (35):e458–e466, May 2015.

[206] David W. Scott, Anja Mottok, Daisuke Ennishi, George W. Wright, Pedro Farinha, Susana Ben-Neriah, Robert Kridel, Garrett S. Barry, Christoffer Hother, Pau Abrisqueta, Merrill Boyle, Barbara Meissner, Adele Telenius, Kerry J. Savage, Laurie H. Sehn, Graham W. Slack, Christian Steidl, Louis M. Staudt, Joseph M. Connors, Lisa M. Rimsza, and Randy D. Gascoyne. Prognostic significance of diffuse large b-cell lymphoma cell of origin determined by digital gene expression in formalin-fixed paraffin-embedded tissue biopsies. *Journal of Clinical Oncology*, 33(26):2848–2856, September 2015.

[207] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR, abs/1610.02391*, 7, 2016.

[208] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016.

[209] Atman P Shah and Stanley A Rubin. Errors in the computerized electrocardiogram interpretation of cardiac rhythm. *Journal of electrocardiology*, 40(5):385–390, 2007.

[210] Siyu Shi, Ishaan Malhi, Kevin Tran, Andrew Y Ng, and Pranav Rajpurkar. Chexseen: Unseen disease detection for deep learning interpretation of chest x-rays. *arXiv preprint arXiv:2103.04590*, 2021.

[211] George Shih, Carol C. Wu, Safwan S. Halabi, Marc D. Kohli, Luciano M. Prevedello, Tessa S. Cook, Arjun Sharma, Judith K. Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, Ritu R. Gill, Myrna C.B. Godoy, Stephen Hobbs, Jean Jeudy, Archana Laroia, Palmi N. Shah, Dharshan Vummidi, Kavitha Yaddanapudi, and Anouk Stein. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

[212] Ramandeep Singh, Mannudeep K. Kalra, Chayanin Nitiwarangkul, John A. Patti, Fatemeh Homayounieh, Atul Padole, Pooja Rao, Preetham Putha, Victorine V. Muse, Amita Sharma, and Subba R. Digumarthy. Deep learning in chest radiography: Detection of findings and presence of change. *PLoS ONE*, 13(10), October 2018.

[213] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert, 2020.

[214] Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. *arXiv preprint arXiv:2010.05352*, 2020.

[215] Hannah Spitzer, Kai Kiwitz, Katrin Amunts, Stefan Harmeling, and Timo Dickscheid. Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks. *CoRR*, abs/1806.05104, 2018.

[216] Suraj Srinivas and R. Venkatesh Babu. Data-free parameter pruning for deep neural networks. *CoRR*, abs/1507.06149, 2015.

[217] A Sriram, M Muckley, K Sinha, F Shamout, J Pineau, KJ Geras, L Azour, Y Aphinyanaphongs, N Yakubova, and W Moore. Covid-19 prognosis via self-supervised representation learning and multi-image prediction. 2021.

[218] Anuroop Sriram, Matthew Muckley, Koustuv Sinha, Farah Shamout, Joelle Pineau, Krzysztof J. Geras, Lea Azour, Yindalon Aphinyanaphongs, Nafissa Yakubova, and William Moore. Covid-19 deterioration prediction via self-supervised representation learning and multi-image prediction, 2021.

[219] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[220] Ewout W Steyerberg. *Clinical prediction models: a practical approach to development, validation, and updating*. Springer Science & Business Media, 2008.

[221] Xu Sun and Weichao Xu. Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11):1389–1393, 2014.

[222] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019.

[223] Steven H. Swerdlow, Elias Campo, Stefano A. Pileri, Nancy Lee Harris, Harald Stein, Reiner Siebert, Ranjana Advani, Michele Ghielmini, Gilles A. Salles, Andrew D. Zelenetz, and Elaine S. Jaffe. The 2016 revision of the World Health Organization classification of lymphoid neoplasms. *Blood*, 127(20):2375–2390, 05 2016.

[224] Alex Tamkin, Mike Wu, and Noah Goodman. Viewmaker networks: Learning views for unsupervised representation learning, 2020.

[225] Siyi Tang, Amirata Ghorbani, Rikiya Yamashita, Sameer Rehman, Jared A. Dunnmon, James Zou, and Daniel L. Rubin. Data valuation for medical imaging using shapley value: Application on a large-scale chest x-ray dataset, 2020.

[226] Yu-Xing Tang, You-Bao Tang, Yifan Peng, Ke Yan, Mohammadhadi Bagheri, Bernadette A. Redd, Catherine J. Brandon, Zhiyong Lu, Mei Han, Jing Xiao, and Ronald M. Summers. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *npj Digital Medicine*, 3(1):70, 2020.

[227] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243*, 2020.

[228] Eric J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, January 2019.

[229] Mintu P Turakhia, Donald D Hoang, Peter Zimetbaum, Jared D Miller, Victor F Froelicher, Uday N Kumar, Xiangyan Xu, Felix Yang, and Paul A Heidenreich. Diagnostic utility of a novel leadless arrhythmia monitoring device. *The American journal of cardiology*, 112(4):520–524, 2013.

[230] Bora Uyumazturk, Amirhossein Kiani, Pranav Rajpurkar, Alex Wang, Robyn L. Ball, Rebecca Gao, Yifan Yu, Erik Jones, Curtis P. Langlotz, Brock Martin, Gerald J. Berry, Michael G. Ozawa, Florette K. Hazard, Ryanne A. Brown, Simon B. Chen, Mona Wood, Libby S. Allard, Lourdes Ylagan, Andrew Y. Ng, and Jeanne Shen. Deep learning for the digital pathologic diagnosis of cholangiocarcinoma and hepatocellular carcinoma: Evaluating the impact of a web-based diagnostic assistant, 2019.

[231] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[232] Maya Varma, Mandy Lu, Rachel Gardner, Jared Dunnmon, Nishith Khandwala, Pranav Rajpurkar, Jin Long, Christopher Beaulieu, Katie Shpanskaya, Li Fei-Fei, Matthew P. Lungren, and Bhavik N. Patel. Automated abnormality detection in lower extremity radiographs using deep learning. *Nature Machine Intelligence*, 1(12):578–583, December 2019.

[233] DJ Vassallo, PJ Buxton, JH Kilbey, and M Trasler. The first telemedicine link for the british forces. *BMJ Military Health*, 144(3):125–130, 1998.

[234] Luis Villela, Armando Lopez-Guillermo, Silvia Montoto, Susana Rives, Francesc Bosch, Maria Perales, Ana Ferrer, Jordi Esteve, Lluis Colomo, Elias Campo, and Emilio Montserrat. Prognostic features and outcome in patients with diffuse large b-cell lymphoma who do not achieve a complete response to first-line regimens. *Cancer*, 91(8):1557–1562, 2001.

[235] Damir Vrabac, Akshay Smit, Rebecca Rojansky, Yaso Natkunam, Ranjana Hira Advani, Andrew Y. Ng, Sebastian Fernandez-Pol, and Pranav Rajpurkar. Morphological feature annotation using deep learning for a clinically, histologically, and cytogenetically annotated digital image set for dlbcl. *figshare*, 2021.

[236] Damir Vrabac, Akshay Smit, Rebecca Rojansky, Yasodha Natkunam, Ranjana H Advani, Andrew Y Ng, Sebastian Fernandez-Pol, and Pranav Rajpurkar. Dlbcl-morph: Morphological features computed using deep learning for an annotated digital dlbcl image set. *Scientific Data*, 8(1):1–8, 2021.

[237] Yen Nhi Truong Vu, Richard Wang, Niranjan Balachandar, Can Liu, Andrew Y Ng, and Pranav Rajpurkar. Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation. *arXiv preprint arXiv:2102.10663*, 2021.

[238] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):19549, 2020.

[239] 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. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, Honolulu, HI, July 2017. IEEE.

[240] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9049–9058, 2018.

[241] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49, January 2018.

[242] Ross Wightman. timm 0.2.1. https://pypi.org/project/timm/, 2020.

[243] David A Wood, Sina Kafiabadi, Aisha Al Busaidi, Emily Guilhem, Jeremy Lynch, Matthew Townend, Antanas Montvila, Juveria Siddiqui, Naveen Gadapa, Matthew Benger, et al. Labelling imaging datasets on the basis of neuroradiology reports: a validation study. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 254–265. Springer, 2020.

[244] David A. Wood, Jeremy Lynch, Sina Kafiabadi, Emily Guilhem, Aisha Al Busaidi, Antanas Montvila, Thomas Varsavsky, Juveria Siddiqui, Naveen Gadapa, Matthew Townend, Martin Kiik, Keena Patel, Gareth Barker, Sebastian Ourselin, James H. Cole, and Thomas C. Booth. Automated Labelling using an Attention model for Radiology reports of MRI scans (ALARM). *arXiv:2002.06588 [cs]*, February 2020. arXiv: 2002.06588.

[245] Baoyuan Wu, Siwei Lyu, Bao-Gang Hu, and Qiang Ji. Multi-label learning with missing labels for image annotation and facial action unit recognition. *Pattern Recognition*, 48(7):2279–2289, 2015.

[246] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[247] Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*, 2016.

[248] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015.

[249] Kui Xue, Yangming Zhou, Zhiyuan Ma, Tong Ruan, Huanhuan Zhang, and Ping He. Fine-tuning BERT for Joint Entity and Relation Extraction in Chinese Medical Text. August 2019.

[250] Kabir Yadav, Efsun Sarioglu, Hyeong Ah Choi, Walter B. Cartwright, Pamela S. Hinds, and James M. Chamberlain. Automated Outcome Classification of Computed Tomography Imaging Reports for Pediatric Traumatic Brain Injury. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine*, 23(2):171–178, February 2016.

[251] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv:1906.08237 [cs]*, January 2020. arXiv: 1906.08237.

[252] Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv preprint arXiv:1710.10501*, 2017.

[253] Li Yao, Jordan Prosky, Eric Poblenz, Ben Covington, and Kevin Lyman. Weakly supervised medical diagnosis and localization from multiple resolutions. *arXiv preprint arXiv:1803.07703*, 2018.

[254] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics, 1995.

[255] Wenwu Ye, Jin Yao, Hui Xue, and Yi Li. Weakly supervised lesion localization with probabilistic-cam pooling, 2020.

[256] William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

[257] Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. *arXiv:1804.09541 [cs]*, April 2018. arXiv: 1804.09541.

[258] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning, 2020.

[259] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 694–699. ACM, 2002.

[260] John Zech, Margaret Pain, Joseph Titano, Marcus Badgeley, Javin Schefflein, Andres Su, Anthony Costa, Joshua Bederson, Joseph Lehar, and Eric Karl Oermann. Natural Language–based Machine Learning Models for the Annotation of Clinical Radiology Reports. *Radiology*, 287(2):570–580, January 2018. Publisher: Radiological Society of North America.

[261] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, November 2018.

[262] Li Zhang, Mengya Yuan, Zhen An, Xiangmei Zhao, Hui Wu, Haibin Li, Ya Wang, Beibei Sun, Huijun Li, Shibin Ding, Xiang Zeng, Ling Chao, Pan Li, and Weidong Wu. Prediction of hypertension, hyperglycemia and dyslipidemia from retinal fundus photographs via deep learning: A cross-sectional study of chronic diseases in central china. *PLOS ONE*, 15(5):1–11, 05 2020.

[263] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016.

[264] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

[265] Zheng Zhou, Laurie H. Sehn, Alfred W. Rademaker, Leo I. Gordon, Ann S. LaCasce, Allison Crosby-Thompson, Ann Vanderplas, Andrew D. Zelenetz, Gregory A. Abel, Maria A. Rodriguez, Auayporn Nademanee, Mark S. Kaminski, Myron S. Czuczman, Michael Millenson, Joyce Niland, Randy D. Gascoyne, Joseph M. Connors, Jonathan W. Friedberg, and Jane N. Winter. An enhanced international prognostic index (NCCN-IPI) for patients with diffuse large b-cell lymphoma treated in the rituximab era. *Blood*, 123(6):837–842, February 2014.

[266] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 384–393. Springer, 2019.

[267] Jiuwen Zhu, Yuexiang Li, Yifan Hu, Kai Ma, S. Kevin Zhou, and Yefeng Zheng. Rubik's cube+: A self-supervised feature learning framework for 3d medical image analysis. *Medical Image Analysis*, 64:101746, 2020.

[268] Xiaojin Zhu. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison*, 2(3):4, 2006.

[269] Xinrui Zhuang, Yuexiang Li, Yifan Hu, Kai Ma, Yujiu Yang, and Yefeng Zheng. Self-supervised feature learning for 3d medical images by playing a rubik's cube. *CoRR*, abs/1910.02241, 2019.