Detecting abnormalities on chest X-rays using deep neural networks

Swaroop Kumar M L
Department of Studies in Computer Science
University of Mysore

June 28, 2019

ABSTRACT

Diseases like pneumonia and tuberculosis are leading causes of death worldwide. Although conclusive diagnosis requires other tests such as a sputum culture, chest radiography can be an important diagnostic aid and is routinely recommended since it is fast, affordable and highly sensitive. Moreover, automated detection of abnormalities on the chest X-ray can help in active case finding, screening, and in cases where other tests are not available or are inconclusive. Due to the nature of the domain, it is also important that algorithms not only make inferences but also generate explanations sufficient to convince a human expert.

Inspired by previous work, we develop algorithms that can detect abnormalities on the x-ray. The algorithm explains these detections by generating heatmaps pointing out areas of the image that most influenced it. We establish baselines, benchmark against previous work and show that a) transfer-learning from a large non-TB dataset dramatically improves TB detection, b) models in the domain show inferior performance on external data from a different hospital system but c) recent techniques such as mixup and progressive resizing improve performance and generalization. We achieve performance competitive with previous work in detecting pneumonia-like and other abnormalities on the NIH chestX-ray14 dataset and in detecting tuberculosis on the Shenzhen hospital dataset, and achieve state-of-the-art performance on the Montgomery county tuberculosis dataset. We look for potential sources of bias, test our baseline with respect to gender, age and view-position and evaluate our models on viral and bacterial pneumonia separately.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Problem definition

The lungs are made up of small air-sacks called alveoli. When, for example, air in the alveoli is replaced with pus, blood and other fluids, referred to as consolidation and commonly caused by pneumonia, or when abscesses in the lung rupture forming cavities, indicating a tuberculosis infection, these are visible on the chest x-ray. See figure 1.1 for examples.

Radiologists are trained to look for signs of these abnormalities, use subtle visual features to differentiate among the various types, reason about their causes and help in diagnosis and treatment. An algorithm that can automatically detect these abnormalities can be useful in numerous ways (see section 1.2).

However, building the hardware and software infrastructure for a clinically relevant system that is useful in practice is a problem which presents unique challenges of its own. Our work focuses on the core algorithm. We divide the problem into, and explore, four sub-problems.

1.1.1 Classification

For our primary dataset, a large collection of chest x-rays annotated with multiple abnormalities including pneumonia [1] (see section 2.1), we formulate the problem as a multi-class multi-label classification problem. Given the n-dimensional input feature space $X = \mathbb{R}^n$, and a set of c class labels corresponding to c abnormalities $L = \{ l_1, l_2, l_3 \dots l_c \}$ the task is to learn a function $f : X \to 2^L$ from the training set $D = \{ (x_i, Y_i) \mid 1 \leq i \leq m \}$. For each example $(x_i, Y_i) \in D$, $x_i \in X$ is an n-dimensional feature vector $\{ x_{i1}, x_{i2}, x_{i3} \dots x_{in} \}$, each feature representing the intensity value of a single pixel of the input image, and $Y_i \subseteq L$ is the set of abnormalities associated with x_i. For an unseen

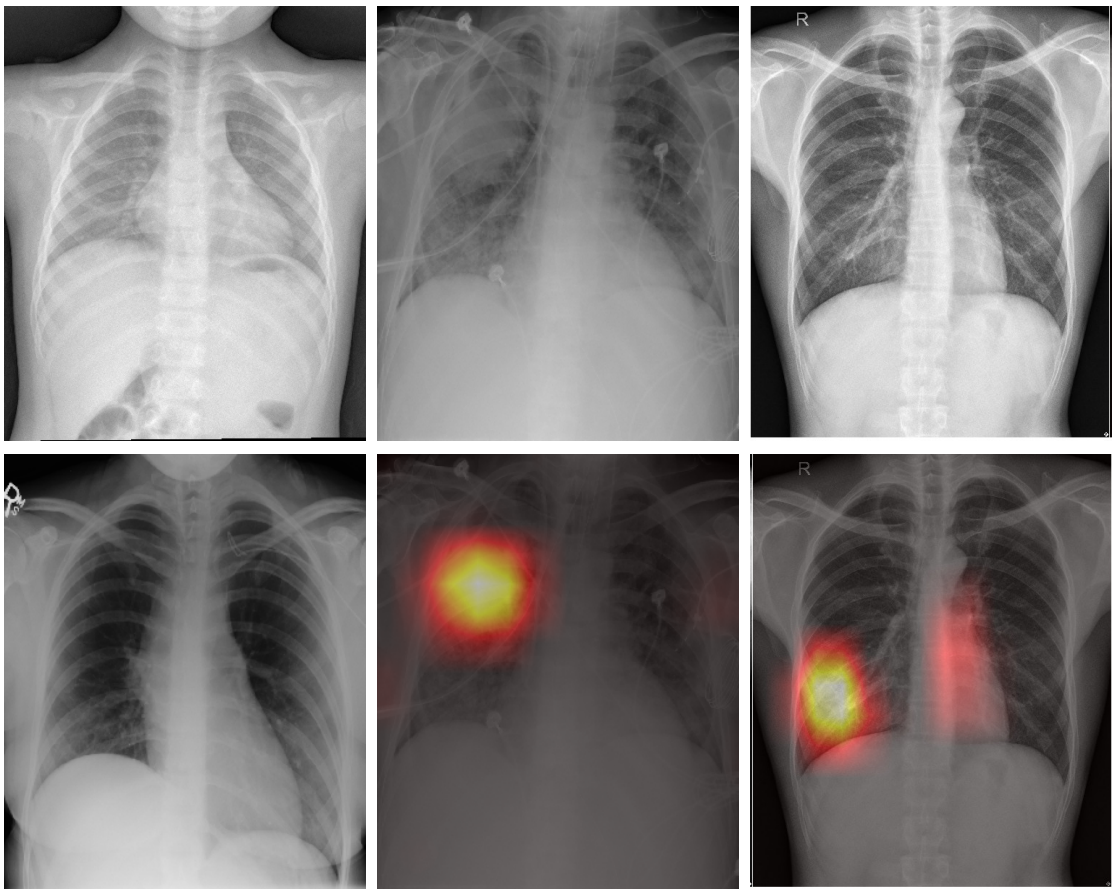

Figure 1.1: From left to right: The first column shows two *Normal* images. Columns 2 and 3 show images with *Pneumonia* and *Tuberculosis* respectively, the first row showing the original images and the second showing the same overlaid with heatmaps localizing the abnormalities, which we call *explanations*

instance x, the classifier $f(\cdot)$ predicts $f(x) \subseteq L$ as the set of abnormalities for x.

For the Shenzhen hospital tuberculosis dataset[2] (see section 2.3), we formulate the problem as a binary classification problem. Again, suppose X is the n-dimensional input feature space. The task is to learn a function $f : X \to \{ 0, 1 \}$ from the training set $D = \{ (x_i, y_i) \mid 1 \leq i \leq m \}$. For each $(x_i, y_i) \in D$, $x_i \in X$ is an n-dimensional feature vector $\{ x_{i1}, x_{i2}, x_{i3} \dots x_{in} \}$, each feature corresponding to the intensity value of a single pixel of the input image, and $y_i \in \{ 0, 1 \}$ is the corresponding label, 0 meaning *normal* and 1 meaning *tuberculosis*. Given an unseen $x \in X$, the classifier $f(\cdot)$ predicts $f(x) \in \{ 0, 1 \}$ as being the label for x.

1.1.2 Explainability

Deep neural networks have outperformed previous methods in several domains. However, they remain black-boxes with millions of parameters, leading to a lack of trust and limiting their use in routine clinical practice. Several methods have been proposed to make these models more interpretable, broadly falling into two categories:

1. Methods that create a proxy model which behaves similarly to the original model, but is simpler and easier to understand. These include methods like LIME[3] and SHAP[4]. While these methods are model-independent, they tend to be very slow.

2. Methods that generate a saliency map which highlights a small portion of the input which is most relevant, in a single forward and backward pass through the network. These include methods like LRP[5], DeepLIFT[6], CAM[7] and Grad-CAM[8].

Given an input $x = \{ x_1, x_2, x_3 \dots x_n \} \in X$, and a set of class labels $L = \{ l_1, l_2, l_3 \dots l_c \}$, the task is to compute the attribution

$$A_j = \{ a_{j1}, a_{j2}, a_{j3} \dots a_{jn} \} \in \mathbb{R}^n \tag{1.1}$$

for each class $l_j \in L$ where $a_{ji} \in A_j$ is a measure of the relevance of the i^{th} feature to the model's inference regarding the j^{th} class.

1.1.3 Generalizability

A test set is considered representative of data that will be encountered in the external world and is used exclusively to evaluate a model. However, true generalization to new datasets may be lower than expected.

1. Since model design choices are based on previous work, methods in an application domain may overfit to one or a few popular datasets. However, [9] shows that this is not the case for CIFAR-10 despite years of methods being tested on this dataset

2. Two datasets may have different distributions. In the context of biomedical imaging, datasets may be collected from different hospital systems and machines. For example, in [10], Zech et al. show that models trained on data from one hospital system showed inferior performance on data from others.

3. The dataset used to train a model may have confounding variables that do not exist in other datasets. For example, in [10], Zech et al. also show that CNNs were able to directly detect the hospital system and department within a hospital system from a chest radiograph where saliency maps showed high activation in image corners. We observed the same phenomenon in networks that were trained to detect abnormalities. Since different departments and machines within a hospital system have different prevalence of a disease, the model may leverage these spurious correlations and fail to generalize.

Therefore, it is important that models are evaluated on external datasets from different hospital systems.

1.1.4 Fairness

Machine learning systems are increasingly being deployed in settings where they may inadvertently learn and leverage biases in the datasets, discriminate based on race, gender, etc. and amplify existing social inequities. For example, in [11], Bolukbasi et al. show that the popular word embedding space Word2Vec encodes gender bias. In [12], Buolamwini et al. show that facial recognition datasets are overwhelmingly composed of light-skinned individuals and that commercial gender classification systems performed worse for dark-skinned people and females, with a difference in accuracy of more than 30% between light-skinned males and dark-skinned females.

There has been substantial work in the research literature on fairness in ML on the development of statistical definitions of fairness [13]–[15] and algorithmic methods to measure and mitigate undesirable biases [15]–[17]. A simple measure against bias in various fields has been to hide variables like gender and race from a model, but complex machine learning models learn to use other correlated variables as proxy for hidden ones (for example, zipcode as correlated with race and the word 'women' in an institution's name as correlated with the gender of its students[18]). Moreover, hiding sensitive variables from researchers exacerbates the problem by limiting their ability to quantify and mitigate these biases. For example, in [19], Estava et al. found that convolutional neural

networks are effective at detecting melanoma from images. However, without labels for skin characteristics such as color, accuracy of the model for different skin-types cannot be measured.

We measure the potential for discrimination by training models with architectures similar to the abnormality detection model, to identify gender and age group from images alone. We then test our baseline model's performance across genders and age groups.

1.2 Motivation

Pneumonia and tuberculosis are leading causes of death worldwide. According to the world health organization, pneumonia disproportionately affects children, accounting for 16% of all deaths of children under the age of 5 years[20]. Tuberculosis is more prevalent in countries where many people live in absolute poverty[21] with limited access to healthcare and in 2017 alone, caused 1.6 million preventable deaths[22].

The global End TB strategy aims for a 95% reduction in deaths due to TB by 2035 compared with 2015. Similarly, the National Strategic Plan (NSP) 2017-2025 sets out to achieve a rapid decline in deaths due to TB and emphasizes the importance of active case finding, that is, detection of TB cases early by seeking out people in targeted groups and scaling up cheap and high sensitivity TB diagnostic tests. The NSP has recommended three tests: sputum smear microscopy, chest x-ray and the new CB-NAAT[1] test.

Conventionally, patients are screened for TB or pneumonia related symptoms, sputum examinations are recommended for those with positive symptoms, and chest x-rays are recommended for those who test negative in the sputum examination.

With automated detection, x-ray tests have the potential to be faster and significantly more affordable. They can be massively scaled up and used

1. For active case finding in high-risk populations, for example, with mobile x-ray vans[23]
2. As an initial screening test before or along with other tests such as a sputum examination
3. To aid a radiologist in her workflow by sorting her queue based on severity, suggesting areas to consider in an image, providing a second opinion, etc.

[1]CB-NAAT or Cartridge Based Nucleic Acid Amplification Test is a molecular test and is known as GeneXpert outside India

Our goal is to make automated abnormality detection systems such as ours more clinically relevant and trust-worthy by a) improving their accuracy and explainability, b) evaluating their ability to generalize to other hospital systems and c) exploring the potential for algorithms in this domain to be unfair.

1.3 Previous work

In [24], Wang et al. collect chest x-ray images and their associated reports from the PAC system of the National Institutes of Health and mine labels from the reports algorithmically. They evaluate the accuracy of these labels and establish a baseline for abnormality detection.

Previous work has explored methods to improve classification performance. In [25]–[28], the authors use attention-guided learning to allow a network to concentrate on abnormal regions of the image. [26] also uses curriculum learning and presents images in increasing order of difficulty. However, [29] uses attention to hide the most salient regions, allowing the network to pay attention to other areas. [27], [30] seek to exploit correlations between abnormalities, [30] by using an LSTM and [27] by extracting saliency maps at an intermediate layer and providing these as input to subsequent layers.

There has also been work on improving localization by combining feature maps from multiple layers of the network [31], [32]. [32] learns a set of *layer relevance weights* for each class, and [31] applies a DenseNet per resolution orthogonal to a standard ResNet followed by upsampling and concatenation.

In [33], Rajpurkar et al. train a variant of DenseNet on the NIH chestX-ray14 dataset relabeled using an ensemble of classifiers and report super-human performance for several abnormalities, comparing board-certified radiologists and the algorithm on a test set labeled by consensus of three cardiothoracic subspeciality radiologists. Both Wang et al. in [24] and Rajpurkar et al. in [33] use weakly supervised localization to explain the network's inference.

For TB detection, previous work such as [34]–[36] have explored various feature extraction techniques, feature selection strategies and classifiers such as logistic regression and SVM. [37]–[40] train deep convolutional neural networks and ensemble these. These methods have also explored the usefulness of segmentation of chest regions. Due to the lack of large publicly available datasets, work in this domain, especially the application of deep learning methods, has been limited. Moreover, results are less relevant to clinical

practice as models trained on small two-class datasets are prone to over-diagnose.

We apply methods proposed by [41], [42]. In [41], van Noord et al. propose a multi-scale CNN which learns both scale-variant and scale-invariant features at an artist attribution task. In [42], Zhang et al. introduce a technique called *mixup* and show that it improves generalization and helps to mitigate the negative effects of label noise.

1.4 Our work

We use a 121-layer dense convolutional network *DenseNet*[43]. The network's connectivity pattern improves the flow of information and gradients and has fewer parameters, making it possible to train very deep networks.

For the NIH chestX-ray14 dataset of 112,120 x-ray images of 30,805 unique patients labeled with up-to 14 different abnormalities, we randomly split the dataset into a training set, a validation set and a test set, consisting of roughly 70%, 10% and 20% of the patients respectively. We replace the final fully-connected layer of the network with one that has 14 outputs after which we apply a sigmoid non-linearity. The output of the network is a 14-dimensional vector $P = \{ p_1, p_2, p_3 \dots p_{14} \}$ where $p_k \in P$ corresponds to the k^{th} abnormality and $0 \leq p_k \leq 1$. Using the Adam[44] optimization algorithm, we optimize the sum of binary cross entropy losses

$$l(x, Y) = \sum_{k=1}^{14} [-y_k \log p_k - (1 - y_k) \log(1 - p_k)] \tag{1.2}$$

where (x, Y) is a pair in the training set and y_k is 1 if Y contains the k^{th} abnormality and 0 otherwise. A set of optimal thresholds $T = \{ t_1, t_2, t_3 \dots t_{14} \}$ is determined by maximizing the network's F1-score on the training set and applied to the output of the network. Given an unseen image x, the output of the network is $P = \{ p_1, p_2, p_3 \dots p_{14} \}$ and the final output $Y \subseteq L$ after applying the thresholds is:

$$Y = \{ l_i \in L \mid p_i \in P \wedge t_i \in T \wedge p_i > t_i \} \tag{1.3}$$

Similarly, for the Shenzhen hospital tuberculosis dataset with 662 frontal chest x-ray images, we create 9 folds of the dataset and report the mean and standard deviation of performance. We replace the final fully-connected layer of the network with one that has 2 outputs after which we apply a sigmoid non-linearity. The output of the network is a 2-dimensional vector $P = \{ p_1, p_2 \}$ where $0 \leq p_1 \leq 1$ and $0 \leq p_2 \leq 1$. Using the Adam optimization algorithm, we optimize the binary cross entropy loss

$$l(x, y) = -y \log p_2 - (1 - y) \log(1 - p_1)] \tag{1.4}$$

where (x, y) is a pair in the training set. Given an unseen image x, the output of the network is $P = \{ p_1, p_2 \}$ and the final output y is 1 (tuberculosis) if $p2 > p1$ and 0 (normal) otherwise.

The network consists of a fully convolutional backbone followed by an adaptive-average-pooling layer and a single fully-connected layer. The fully convolutional part of the network results in k w x h feature maps which are averaged along the width and height to form a k dimensional vector, which is fed to the single fully-connected layer with k input nodes and c output nodes. If f_i is the i^{th} feature map and w_i^j is the weight between the i^{th} input node and the j^{th} output node in the fully-connected layer, the saliency map for the j^{th} class M_j is

$$M_j = \sum_i w_i^j f_i \tag{1.5}$$

M_j is a w x h saliency map which is interpolated to the size of the input image and measures the relevance of each pixel to the model's decision regarding the j^{th} class, and can be visualized as a heatmap (see figure 1.1). This serves as an *explanation* of the model's inference, allowing physicians and radiologists to decide how much trust to invest in it.

To test the ability of these networks to generalize to other hospital systems, we use two external datasets, a pediatric pneumonia dataset with 5332 x-ray images[45] and a smaller tuberculosis dataset of 138 frontal chest x-ray images[2].

We evaluate the potential for bias by training similar networks to predict the gender and age group of a patient and the view-position (AP vs. PA) given only the x-ray image and evaluate our baseline for each gender, age group and view position.

We make the following observations:

1. On the NIH chestX-ray14 dataset, using only horizontal flipping for data augmentation led to overfitting but extending this to include rotation, zooming, brightness scaling and perspective warp reduced overfitting and improved performance. Also, a recently proposed data augmentation technique, mixup[42], consistently improved performance on both internal and external datasets.

2. Compared to random initialization, initializing the weights of all but the final layers of the network with those of a similar network trained on a large collection of natural images, ImageNet, significantly improved performance.

3. At the problem of tuberculosis detection, pre-training on the NIH chestX-ray14 dataset significantly improved performance and generalization to the external dataset.

However, this lead to over-diagnosis of TB on the NIH chestX-ray14 dataset.

4. Training a single network on 224 x 224 images until the validation loss plateaued and repeatedly re-training it on progressively higher resolution images improved generalization to external datasets, as did preserving these intermediate models and ensembling them[41].

5. Models trained to detect Tuberculosis on the Shenzhen dataset showed inferior performance on the Montgomery county dataset, and models trained to detect pneumonia and other abnormalities on the NIH chestX-ray14 dataset performed worse on the external dataset than a similar network trained exclusively on the smaller external dataset.

6. For most abnormalities, average saliency maps or explanations weighted by predicted probability, of models trained on the NIH chestX-ray14 dataset showed high activation at image corners, predominantly the top-left and top-right, on both the internal and external datasets. However, cropping of image margins at both training and inference stages, did not negate this effect.

7. Baselines showed similar performance for each gender, age group and view-position. However, networks with architecture similar to the abnormality detection network, after a single epoch of training were able to distinguish between x-ray images of male and female patients and determine the view-position with more than 90% accuracy.

8. Models trained on the NIH CXR-14 dataset were better at detecting viral pneumonia than bacterial pneumonia.

For the Shenzhen tuberculosis dataset, pre-training on the NIH chestX-ray14 dataset resulted in an AUC of 98.4% and an accuracy of 94.6% which are respectively 2.8% and 4.4% better than our baseline. On the external dataset, the same model achieves an AUC of 95.5% and accuracy of 89.4% which are respectively 8.1% and 15.2% better than our baseline. On both the Shenzhen and Montgomery datasets, our model is competitive with previous work and out-performs previous works that use deep neural networks.

On the NIH chestX-ray14 dataset, training a single network on 224 x 224 images, repeatedly re-training on progressively higher resolution images, preserving these intermediate models and ensembling them resulted in an average AUC of 85.6% which is 1.8% better than our baseline and competitive with previous work. This translated to a 3.6% improvement on the pediatric pneumonia dataset.

Chapter 2

Data

We use the NIH ChestX-ray14 dataset[24] and the Shenzhen hospital tuberculosis dataset[2] to train models to detect pneumonia and other abnormalitites, and tuberculosis respectively. We then use two external datasets, the Guangzhou medical center pediatric pneumonia dataset[45] and the Montgomery county tuberculosis dataset[2] as *external* datasets to test the ability of these models to generalize to other hospital systems.

2.1 NIH CXR-14

The NIH chestX-ray14 dataset consists of 112,120 chest x-ray images of 30,805 unique patients which was collected using the PAC system of the National Institutes of Health. Each image was labeled with up-to 14 abnormalities algorithmically using the associated radiology report in natural language.

About half or 60,361 of these are labeled as *No finding* and the rest are labeled with

	Training	**Testing**	**Different Hospital system?**
Pneumonia	NIH CXR-14	NIH CXR-14	No
	NIH CXR-14	Guangzhou	Yes
	Guangzhou	Guangzhou	No
Tuberculosis	Shenzhen	Shenzhen	No
	Shenzhen	Montgomery	Yes
	Montgomery	Montgomery	No

Table 2.1: Datasets used for training and testing

Abnormality	**Number of images**			
	Train	Validation	Test	Total
Atelectasis	7996	1119	2420	11559
Cardiomegaly	1950	240	582	2776
Effusion	9261	1292	2754	13317
Infiltration	13914	2018	3938	19894
Mass	3988	625	1133	5782
Nodule	4375	613	1335	6331
Pneumonia	978	133	242	1431
Pneumothorax	3705	504	1089	5302
Consolidation	3263	447	957	4667
Edema	1690	200	413	2303
Emphysema	1799	208	509	2516
Fibrosis	1158	166	362	1686
Pleural Thickening	2279	372	734	3385
Hernia	144	41	42	227
No Finding	42405	6079	11928	60361
Total	**78468**	**11219**	**22433**	**112120**

Table 2.2: For the NIH CXR-14 dataset, the number of images in the train, validation and test sets per abnormality

one of more of the 14 abnormalities. Some abnormalities are more common than others, the most common being *Infiltration*, which is present in 19,894 images, and the least common being *Hernia*, which is present in only 227 images.

We split the dataset into train, validation and test sets roughly in the ratio 70:10:20. We make sure that there is no patient overlap, that is, all images of a patient are in the same subset since patient overlap may lead to overfitting.

2.1.1 Challenges and issues

The NIH chestX-ray14 dataset is one of the largest publicly accessible chest x-ray datasets. However, it presents a few unique challenges:

Label	**Number of images**			
	Train	Validation	Test	Total
Normal	3199	799	234	4232
Viral pneumonia	1076	269	148	1493
Bacterial pneumonia	2024	506	242	2772
Total	**6299**	**1574**	**624**	**8497**

Table 2.3: For the Guangzhou pediatric pneumonia dataset, the number of images in the train, validation and test sets per label

1. **Noisy labels**
   Labels were extracted using NLP from radiology reports in natural language text. This may lead to label noise. [24] show that these labels are about 90% accurate. Although deep neural networks have been shown to be robust to label noise in general[46], structured noise can be especially detrimental to performance.

2. **Labeling schema**
   Some of these abnormalities are sub-types of others, but labels are provided as a non-hierarchical list. For example, *Pneumonia* on the x-ray is a form of *Consolidation*

3. **Class imbalance**
   The majority of images do not contain abnormalities, and some abnormalities are common while others are rare. For example, *Infiltration* appears in 17% of the images while *Hernia* appears in only 0.2% of the images.

2.2 Guangzhou

The Guangzhou pediatric pneumonia dataset consists of 8,497 chest x-ray images of children under the age of 5 years collected from the Guangzhou Women and Children's Medical Center, Guangzhou. Each image has been labeled as either *Normal* (4232 images) or *Pneumonia* (4265 images). Images labeled as *Pneumonia* have been further tagged as *Bacterial* (2772 images) or *Viral* (1493 images) based on the cause of pneumonia.

We use the standard test set and split the rest of the dataset into train and validation sets in the ratio 80:20. However, when used as an external dataset, we use the entire dataset.

2.3 Shenzhen

The Shenzhen tuberculosis dataset consists of 615 chest x-ray images collected from the Shenzhen No.3 Hospital in Shenzhen, Guangdong providence, China. Each of the images is labeled as either *Normal* or *Tuberculosis*. 340 of these are normal and 275 show manifestations of tuberculosis.

Considering the small size of the dataset, we create 9 folds of the dataset and report average and standard deviation of metrics. Each fold contains all the images split into train, validation and test sets in the ratio 70:10:20.

Label	**Set**	**Number of images**								
		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9
Normal	Train	220	220	220	211	211	211	222	222	222
	Validation	28	39	39	39	38	38	39	31	34
	Test	78	67	67	76	77	77	65	73	70
Tuberculosis	Train	222	222	222	231	231	231	220	220	220
	Validation	45	34	35	33	35	36	34	42	40
	Test	69	80	79	71	70	69	82	74	76

Table 2.4: For the Shenzhen tuberculosis dataset, the number of images in the train, validation and test sets in each fold per label

2.4 Montgomery

The Montgomery county tuberculosis dataset consists of 138 chest x-ray images collected from the tuberculosis control program of the Department of Health and Human Services of Montgomery County, MD, USA. Each of the images is labeled as either *Normal* or *Tuberculosis*. 80 of these are normal and 58 show manifestations of tuberculosis. The dataset also contains manually segmented lung masks for each image.

Similar to the Shenzhen dataset, we create 9 folds and report average and standard deviation of metrics. Each fold contains all the images split into train, validation and test sets in the ratio 70:10:20. However, when used as an external dataset, we ignore these folds and test on the entire dataset.

Label	**Set**	**Number of images**								
		Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9
Normal	Train	55	55	55	53	53	53	52	52	52
	Validation	8	7	10	9	9	9	11	8	9
	Test	17	18	15	18	18	18	17	20	19
Tuberculosis	Train	37	37	37	39	39	39	40	40	40
	Validation	7	8	6	6	6	7	4	7	7
	Test	14	13	15	13	13	12	14	11	11

Table 2.5: For the Montgomery county tuberculosis dataset, the number of images in the train, validation and test sets in each fold per label

Chapter 3

Baselines

3.1 Model architecture

We replace the final fully connected layer of 121-layer dense convolutional neural network with one that has either 14 outputs (for the NIH CXR-14 dataset) or 2 outputs (for all other datasets) after which we apply a sigmoid non-linearity. The fully convolutional backbone of the network results in k w x h feature maps and is followed by a global-average-pooling layer where the k feature maps are averaged along the width and height to form a k dimensional vector. This makes the network independent of input image size and allows us to use the progressive-resizing method (see section 4.7).

The network's connectivity pattern improves the flow of information and gradients and has fewer parameters, making it possible to train very deep networks. The architecture of the model, specifically the fact that the fully convolutional part of the network is followed by a single fully connected layer, forces the model to learn to localize abnormalities given only weak labels (presence or absence of an abnormality).

3.2 Training procedure

We use the Adam optimization algorithm and start the training with an initial learning rate a factor of 10 smaller than the learning rate at which the training loss begins to increase, when the learning rate is increased linearly (using a learning rate finder). We divide the learning rate by 10 if the validation loss plateaus (does not decrease over 5 iterations), and stop training when the validation loss has stopped decreasing. When using k-fold cross validation, we train k different networks on each of the k folds and report average and standard deviation of performance.

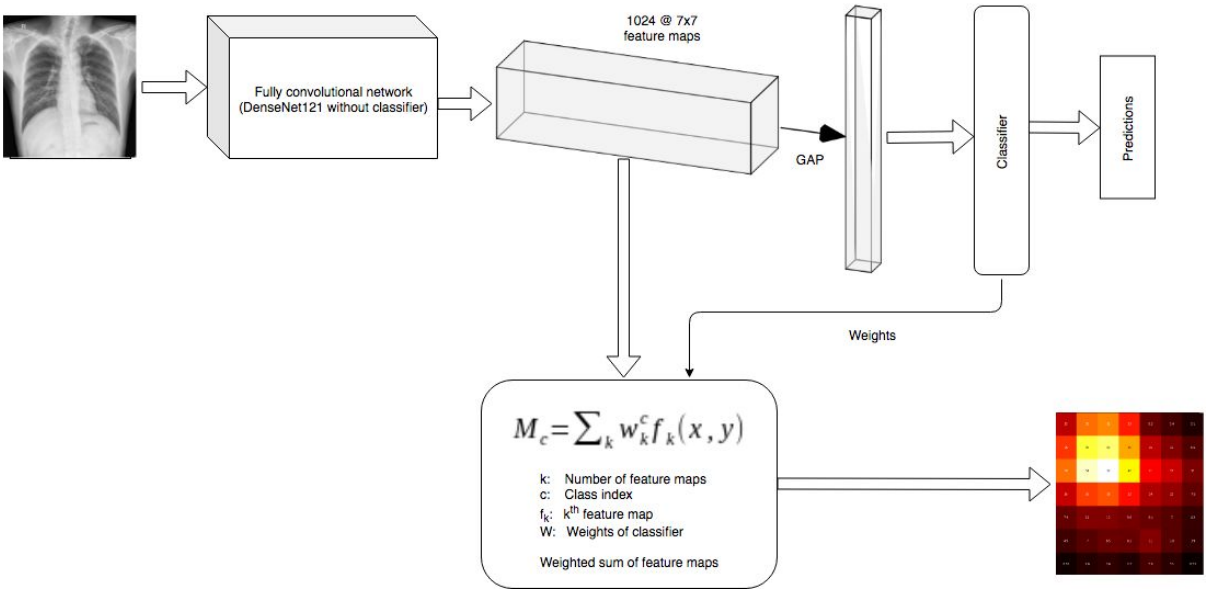

Figure 3.1: Basic architecture of the model

3.3 Inference procedure

At the multi-class multi-label classification task (for the NIH CXR-14 dataset), we merge the train and validation sets after training and use it to compute optimal thresholds for each abnormality by optimizing for the class-specific F1-score. Although it is possible to compute optimal thresholds for the classification task since we have ground truth labels, it is not possible to do so for the localization task since we only have weak labels and not precise locations.

At binary classification tasks, since the network has two output nodes, we do not compute thresholds but simply consider as the proper output the class whose corresponding output node has higher activation.

3.4 Saliency maps and bounding boxes

To compute explanations, we save the feature maps resulting from the final convolutional layer during a forward pass and perform a weighted sum of these feature maps using the weights of the final fully-connected layer between each of the feature maps and the desired output node, as follows.

If f_i is the i^{th} feature map and w_i^j is the weight between the i^{th} input node and the

j^{th} output node in the fully-connected layer, the saliency map for the j^{th} class M_j is

$$M_j = \sum_i w_i^j f_i \tag{3.1}$$

M_j is a w x h saliency map which we interpolate to the size of the input image and visualize as a heatmap (see figure 1.1).

We use a region-growing algorithm to determine bounding boxes given a saliency map. Specifically, we first threshold the saliency map and using the maximum element as a seed point, grow a region around it, including all non-zero neighbours, and repeat the same until all non-zero elements are included in a region, each time choosing as seed the maximum element not included a region. For each region, we determine a bounding box as the smallest rectangle which encloses the entire region. For example, see figure 3.2

3.5 Evaluation metrics

We primarily use the area under the reciever-operator-characteristic curve (AUROC) to measure the performance of a model. AUROC is not affected by the class distribution, does not need thresholds to be set, and is commonly used in the literature. Apart from AUROC and accuracy, we also use

1. **Specificity**
   Specificity is a measure of the model's ability to reject negative examples.

$$Specificity = \frac{|TN|}{|TN| + |FP|} \tag{3.2}$$

2. **Sensitivity**
   Sensitivity is a measure of the model's ability to detect positive examples.

$$Sensitivity = \frac{|TP|}{|TP| + |FN|} \tag{3.3}$$

Where $|TP|$ is the number of true positives, $|TN|$ is the number of true negatives, $|FP|$ is the number of false positives and $|FN|$ is the number of false negatives.

3.6 Baseline performance

Training on 224 x 224 images of the NIH CXR-14 dataset using a batch size of 16, initial learning rate of 0.01, momentum of 0.9, weight decay of 10^{-5} and horizontal flipping

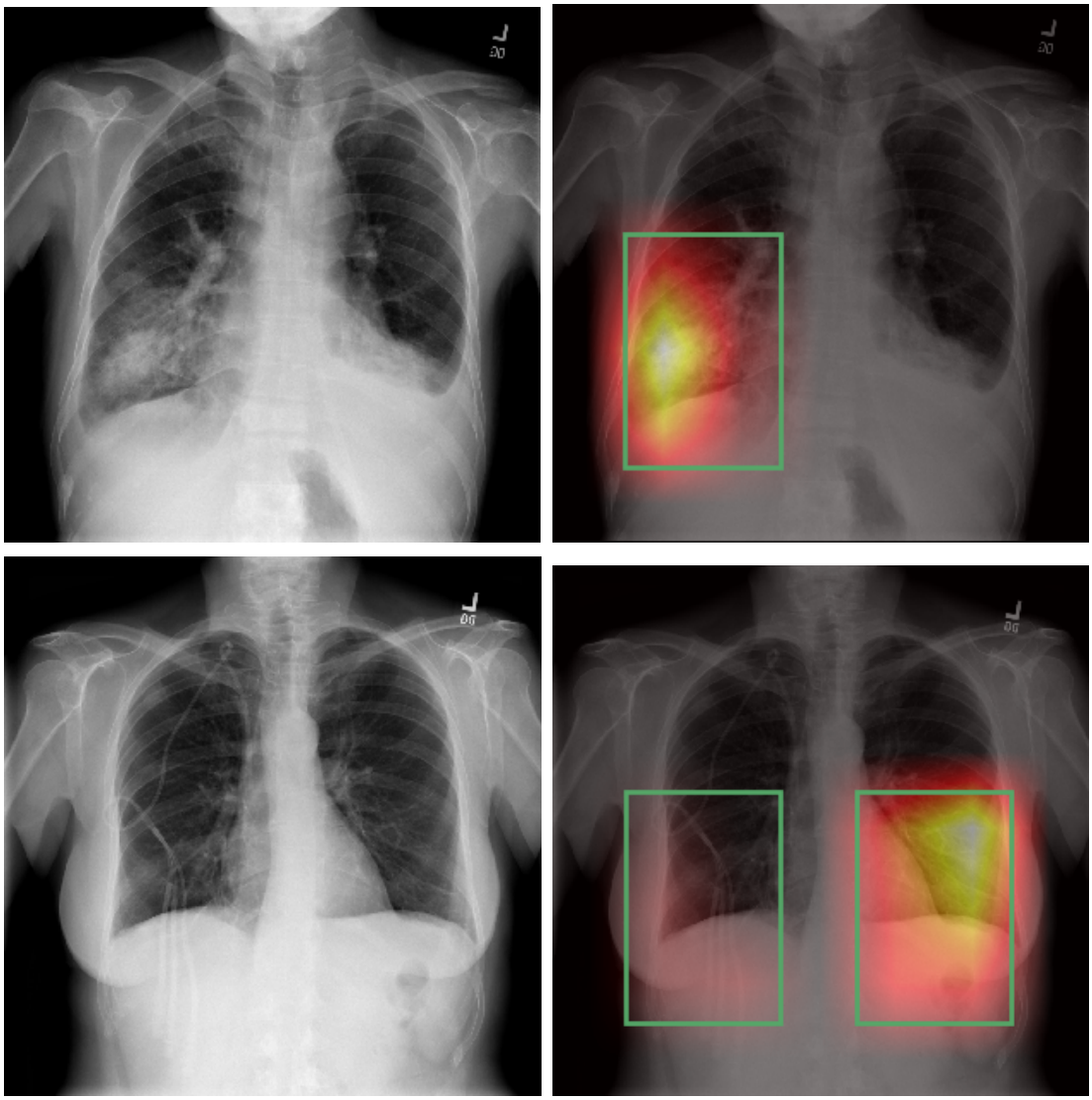

Figure 3.2: Examples of saliency maps with corresponding bounding boxes drawn. The first column shows original x-ray images and the second row shows saliency map and bounding boxes overlaid on the image. Row 1: Pneumonia. Row 2: Atelectasis.

Abnormality	**AUROC**
Atelectasis	0.823
Cardiomegaly	0.905
Effusion	0.879
Infiltration	0.712
Mass	0.839
Nodule	0.778
Pneumonia	0.760
Pneumothorax	0.869
Consolidation	0.806
Edema	0.891
Emphysema	0.923
Fibrosis	0.831
Pleural Thickening	0.785
Hernia	0.929
Average	**0.838**

Table 3.1: Baseline results on the NIH CXR-14 dataset

(with a probability of 0.5) as data augmentation, we obtain average AUROC (of 14 abnormalities) of 0.838 which is competitive with previous work. Results are shown in table 3.1.

Training on 224 x 224 images of the Shenzhen hospital tuberculosis dataset with a batch size of 16, initial learning rate of 0.0001, momentum of 0.9, weight decay of 10^{-5} and using random horizontal flipping (with a probability 0.5), random rotation (0° to 10°), random zoom (1x to 1.1x), random brightness scaling(upto 1.2x with a probability of 0.75) and random perspective warping as data augmentation, we obtain an average AUROC (of 9 folds) of 0.956 with a standard deviation of 0.009, which is competitive with previous work using similar methods. Results are shown in table 3.2.

	AUROC	**Accuracy**	**Specificity**	**Sensitivity**
Mean	0.956	0.902	0.902	0.899
Standard deviation	0.009	0.019	0.050	0.045

Table 3.2: Baseline results on the Shenzhen hospital tuberculosis dataset

Chapter 4

Experiments

4.1 Data augmentation

We observe that the baseline model for the NIH CXR-14 dataset began to overfit after about 20,000 iterations and the validation loss began to increase, as shown in figure 4.1. However, adding more data augmentation, random rotation (0° to 10°), random zoom (1x to 1.1x), random brightness scaling(upto 1.2x with a probability of 0.75) and random perspective warping in addition to horizontal flipping, the model did not overfit upto 70,000 iterations and showed improvement in performance. Data augmentation has the additional benefit of making the model more robust to similar augmentations during inference.

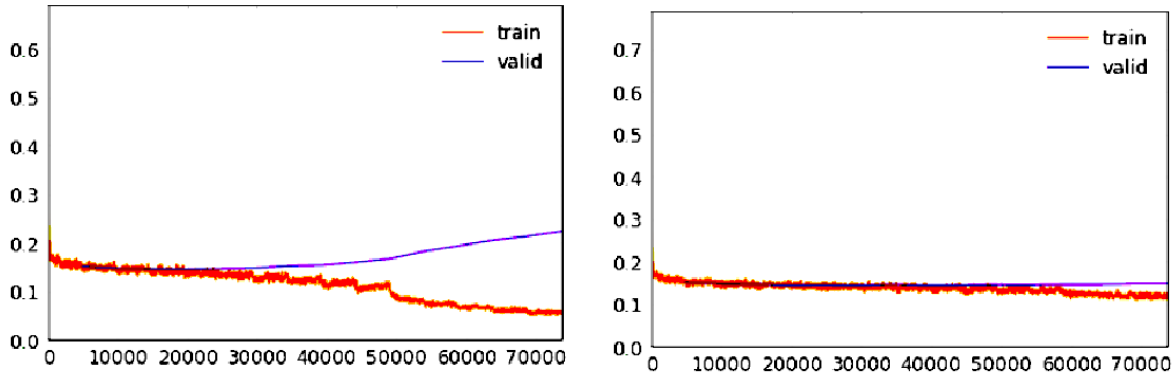

Figure 4.1: Number of iterations (x-axis) vs training and validation loss (y-axis), with only horizontal flipping (left), and more data augmentations (right).

Abnormality	**AUROC**	
	Only horizontal flipping	**More data augmentation**
Atelectasis	0.823	**0.826**
Cardiomegaly	0.905	**0.907**
Effusion	0.879	**0.884**
Infiltration	**0.712**	0.704
Mass	0.839	**0.849**
Nodule	**0.778**	0.773
Pneumonia	0.760	**0.764**
Pneumothorax	0.869	**0.877**
Consolidation	**0.806**	0.799
Edema	0.891	**0.901**
Emphysema	0.923	**0.925**
Fibrosis	**0.831**	0.828
Pleural Thickening	0.785	**0.795**
Hernia	0.929	**0.949**
Average	0.838	**0.841**

Table 4.1: Results on the NIH CXR-14 dataset with and without additional data augmentations

Model	Average AUROC	
	Without TTA	With TTA
Baseline	**0.838**	0.829
More data augmentation	**0.841**	0.838
Higher resolution (512 x 512)	**0.836**	0.834
Progressive resizing (upto 512 x 512)	**0.846**	0.838
Mixup with $\alpha = 1$ (224 x 224) Only horizontal flipping	**0.834**	0.825
Mixup with $\alpha = 1$ (224 x 224)	**0.852**	0.851
Mixup with $\alpha = 0.4$ (224 x 224)	0.849	**0.850**
Mixup with $\alpha = 0.4$ (512 x 512)	0.852	0.852

Table 4.2: Results on the NIH CXR-14 dataset with and without test-time augmentation

4.2 Test-time augmentation

We apply test-time augmentation and perform a weighted average of the model's output on 5 different crops of the image (center and 4 corners) as well as the horizontally flipped versions of each corner. We weight the center-crop by a factor β and each of the other augmentations with a weight $\frac{1-\beta}{8}$.

Using a β of 0.4, we find that performance consistently decreases across abnormalities. See table 4.2 for the results.

4.3 Mixup

Mixup is a recently proposed data augmentation technique [42] that is effective at regularizing models. It has also been shown to combat label noise and improve the model's ability to generalize. Instead of using raw images, we feed the model a linear combination of two images not necessarily from the same class. If I_1 and I_2 are two images, we feed the network a linear combination $M = t \cdot I_1 + (1 - t) \cdot I_2$ where t is drawn from a beta distribution parameterized by some α. The expected output for M is $t \cdot y_1 + (1 - t) \cdot y_2$ where y_1 and y_2 are the targets for I_1 and I_2 respectively.

Mixup improved performance on both the NIH CXR-14 dataset and the Shenzhen tuberculosis dataset. It also improved generalization to external datasets.

Training set	**Testing set**	**Model**	**Average AUROC**	
			Without mixup	With mixup
NIH CXR-14	NIH CXR-14	Baseline	**0.838**	0.834
		More DA (224 x 224)	0.841	**0.852**
		More DA (512 x 512)	0.836	**0.852**
	Guangzhou	More DA (224 x 224)	0.842	**0.873**
		More DA (512 x 512)	0.818	**0.831**
Shenzhen	Shenzhen	(448 x 448)	0.954	**0.956**
		(224 x 224) Pretrained	0.977	**0.979**
		(480 x 480) Pretrained	0.984	**0.985**
	Montgomery	(448 x 448)	0.809	**0.824**
		(224 x 224) Pretrained	0.941	**0.947**
		(480 x 480) Pretrained	**0.957**	0.955

Table 4.3: Results of mixup. Mixup consistently improves performance and generalization to external datasets.

4.4 Transfer-learning from ImageNet

We initialize the weights of a network with those of a similar network trained on a large dataset of millions of natural images, ImageNet and compare this with random initialization (we use the Kaiming He initialization method[47]).

We observe that pre-training on ImageNet significantly improves performance on both internal and external datasets. See table 4.4 for the results.

4.5 Transfer-learning from NIH CXR-14

On the Shenzhen hospital tuberculosis dataset, we compare replacing the weights of a network with those of a similar network a) trained on ImageNet (a large non-x-ray dataset of natural images) and b) trained on ImageNet and then on NIH CXR-14 (a smaller non-tb x-ray dataset).

We observe that pre-training on the NIH CXR-14 dataset both improves performance on both the internal test set and helps to close the generalization gap (see table 4.5).

4.6 Overdiagnosis of TB

The rate of of tuberculosis in the NIH CXR-14 dataset is unknown since tuberculosis is not one of the labels. However, assuming a baseline rate of less than 1%, we observe that models trained to detect tuberculosis on the Shenzhen hospital tuberculosis dataset tend to over-diagnose when tested on the NIH CXR-14 dataset.

The rate of over-diagnosis is especially high when models are pre-trained on the NIH CXR-14 dataset. However, models trained using progressive resizing without pre-training on the NIH CXR-14 dataset have the lowest rate of over-diagnosis (see table 4.6).

4.7 Progressive resizing

For the NIH CXR-14 dataset dataset, we first train on 224 x 224 images until the validation loss plateaus. We then re-train the same network on 256 x 256 images, 288 x 288 images, etc. upto 512 x 512 with a smaller learning rate and for fewer epochs. The intuition behind progressive resizing is that first training on lower resolution images is equivalent to pre-training and is better than training on high resolution images from scratch. It may also make networks more robust to scale variation and behave similar to

Training set	Testing set	Model	Average AUROC	
			Without pre-training on ImageNet	With pre-training on ImageNet
NIH CXR-14	NIH CXR-14	More DA (224 x 224)	0.794	**0.841**
		More DA (512 x 512)	0.791	**0.836**
	Guangzhou	More DA (224 x 224)	0.817	**0.842**
		More DA (512 x 512)	0.768	**0.818**
Shenzhen	Shenzhen	More DA (224 x 224)	0.894	**0.956**
		More DA (672 x 672)	0.876	**0.960**
		Progressive resizing up-to 672 x 672	0.902	**0.954**
	Montgomery	More DA (224 x 224)	0.596	**0.871**
		More DA (672 x 672)	0.583	**0.813**
		Progressive resizing up-to 672 x 672	0.617	**0.829**

Table 4.4: Results of pre-training on ImageNet. Pre-training on Imagenet, a large collection of natural images, significantly improves both performance and generalization

Test set	**Model**	**AUROC**		**Accuracy**	
		Mean	Standard deviation	Mean	Standard deviation
Shenzhen	Baseline	0.956	0.009	0.902	0.019
	Pre-trained on NIH CXR-14 (224 x 224)	0.977	0.006	0.934	0.017
	Pre-trained on NIH CXR-14 (480 x 480)	**0.984**	**0.006**	**0.946**	**0.015**
Montgomery	Baseline	0.871	0.029	0.755	0.038
	Pre-trained on NIH CXR-14 (224 x 224)	0.941	0.014	0.860	0.030
	Pre-trained on NIH CXR-14 (480 x 480)	**0.957**	**0.012**	**0.890**	**0.019**

Table 4.5: Results of pre-training on NIH CXR-14. Pre-training networks on the NIH CXR-14 14 improves performance on both the internal test set and the external dataset

Model	**Over-diagnosis (%)**
Baseline	8.5
Pre-trained on ImageNet (448 x 448)	7.9
Pre-trained on ImageNet (672 x 672)	14.5
Pre-trained on NIH CXR-14 (224 x 224)	**45.5**
Pre-trained on NIH CXR-14 (480 x 480)	**36**
Progressive resizing up-to (448 x 448)	1.1
Progressive resizing up-to (672 x 672)	3.7

Table 4.6: Results of over-dianosis. Models trained on the Shenzhen dataset to detect tuberculosis tend to over-diagnose TB on the NIH CXR-14 dataset

Internal dataset	External dataset	AUROC	
		Without PR	With PR
NIH CXR-14	NIH CXR-14	0.836	**0.846**
	Gaungzhou	0.818	**0.871**
Shenzhen	Shenzhen	0.96	0.954
	Montgomery	0.813	**0.829**

Table 4.7: Results for progressive resizing.

Internal dataset	External dataset	AUROC	
		Baseline	Ensemble
NIH CXR-14	NIH CXR-14	0.838	**0.856**
	Gaungzhou	0.842	**0.878**
Shenzhen	Shenzhen	0.956	**0.963**
	Montgomery	**0.871**	0.663

Table 4.8: Results for ensembling of predictions of multiple models trained using progressive resizing on different resolutions

data augmentation preventing overfitting.

For the Shenzhen hospital tuberculosis dataset, we first train on 224 x 224 images, until the validation loss plateaus, and retrain the same network on 448 x 448 images and then on 672 x 672 images.

We observe that progressive resizing consistently improves performance and generalization to external datasets. See table 4.7 for the results.

4.8 Ensembling predictions

We ensemble the predictions from multiple models trained on different resolutions using progressive resizing. This is similar to [41], the intuition being that scale-invariance emerges from ensembling.

We observe that ensembling improves performance across abnormalitites on the NIH CXR-14 dataset and improves generalization to the Guangzhou dataset. It also improves performance on the Shenzhen dataset for tuberculosis detection but fails to generalize to the Montgomery dataset.

4.9 Ensembling saliency maps

At the problem of tuberculosis detection, we experiment with ensembling saliency maps from multiple models trained on different resolutions after interpolating these to the size of the largest saliency map. Predictions are derived from the saliency maps by using averaging across the width and height.

We observe that this method results in performance similar to ensembling final predictions.

4.10 Cropping of image margins

For most models, saliency maps showed high activation at image corners when detecting abnormalities, perhaps due to the presence of tokens on the x-ray which differ between hospital systems and departments, and which may be correlated with abnormalities. [48] makes the same observation with models trained to identify the hospital system from x-ray images. We experiment with cropping out the top, bottom, left and right margins of the image. However, average saliency maps still showed high activation in image corners. Figure 4.2 shows average saliency maps of pneumonia weighted by predicted probabilities for various models.

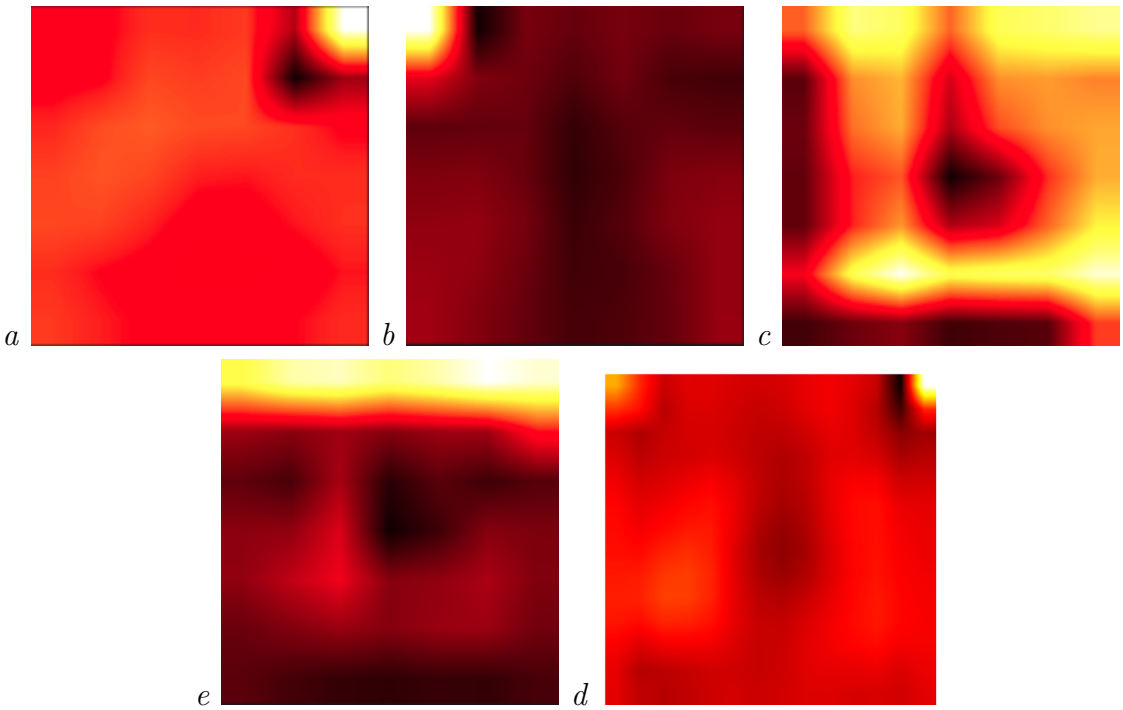

Figure 4.2: Average saliency maps for *pneumonia*. Clockwise from the top-left: a) Baseline, b) Baseline with more data augmentation, c) Trained with margins cropped, d) Trained on NIH CXR-14, tested on Guangzhou and e) Trained without pre-training on ImageNet

4.11 Fairness

We look for potential sources of bias such as gender, age and view-position by a) looking at the distribution of ground truth by gender, age and view-position and variable rates of abnormalities, b) training models with architecture similar to our baseline model to predict these from images alone. We then measure variable performance of our baseline model across gender, age-group and view-position.

4.11.1 Age

Age follows a rougly gaussian distribution with a mean of 46.17 years and standard deviation of 16.73. On dividing patients into 10 age groups (0-9 years, 10-19 years ...90-99 years), abnormality rate increases with age, with a maximum rate of 57.4% for the age group 80-89 years and a minimum rate of 38.7% for the age group 0-9 years. *No finding* is negatively correlated with age, with a Pearson product-moment correlation coefficient (PMCC) of -0.07. When broken down into 3 age groups (less than 25 years, between 25 and 65 years, and greater than 65 years) and by specific abnormalities, *Hernia* is 2.8 times more likely if the patient is old-aged (more than 65 years old) and *Pneumonia* is

	Number of images			**Rate of abnormality**
	Total	Abnormal	Normal	
Male	56804	26368	30436	0.464
Female	44097	20200	23897	0.458
0 to 9 years	1214	470	744	0.387
10 to 19 years	4883	2036	2847	0.417
20 to 29 years	11575	4952	6623	0.428
30 to 39 years	14515	5971	8544	0.411
40 to 49 years	19543	8635	10908	0.442
50 to 59 years	24949	11976	12973	0.480
60 to 69 years	17500	8996	8504	0.514
70 to 79 years	5691	2946	2745	0.518
80 to 89 years	978	562	416	0.575
90 to 99 years	39	17	22	0.436
PA	60463	25179	35284	0.416
AP	40438	21389	19049	0.529

Table 4.9: Distribution of *normal* and *abnormal* images by gender, age group and view-position.

1.4 times more likely if the patient is young (less than 25 years old).

However, a network with a similar architecture with 3 output nodes trained to detect the age group from x-ray images predicted the most common 2^{nd} age group (25 to 65 years) for every image. A network with a similar architecture with a single output node trained to predict age as a continuous variable achieved a mean absolute error of 10.9 years, which is not significantly better than the mean absolute deviation of a gaussian distribution with the same mean and standard deviation as that of patient-age, 13.3, meaning that the model's predictions are not much better than a naive algorithm which predicts the mean age of 46 years for every image.

We evaluated our baseline for variable performance for each age group and found that the model showed similar performance (in terms of AUROC) for each.

Abnormality	**Number of images**				**Prior**	**Posterior**			**Posterior / Prior**			**AUROC**		
	Total	0-25	25-65	65-99		0-25	25-65	65-99	0-25	25-65	65-99	0-25	25-65	65-99
Atelectasis	10416	910	7846	1660	0.103	0.077	0.107	0.107	0.74	1.03	1.04	0.943	0.956	0.958
Cardiomegaly	2532	275	1907	350	0.025	0.023	0.026	0.023	0.92	1.03	0.90	0.832	0.837	0.840
Effusion	12015	1067	9050	1898	0.119	0.090	0.123	0.123	0.75	1.03	1.03	0.916	0.923	0.921
Infiltration	17852	2501	13110	2241	0.177	0.210	0.178	0.145	1.19	1.01	0.82	0.900	0.896	0.895
Mass	5121	503	3953	665	0.051	0.042	0.054	0.043	0.83	1.06	0.85	0.732	0.730	0.719
Nodule	5710	532	4440	738	0.057	0.045	0.060	0.048	0.79	1.07	0.84	0.874	0.880	0.866
Pneumonia	1220	210	858	152	0.012	0.018	0.012	0.010	1.46	0.96	0.81	0.786	0.802	0.796
Pneumothorax	4794	767	3467	560	0.048	0.065	0.047	0.036	1.36	0.99	0.76	0.784	0.783	0.794
Consolidation	4220	576	3083	561	0.042	0.048	0.042	0.036	1.16	1.00	0.87	0.894	0.906	0.902
Edema	2103	278	1635	190	0.021	0.023	0.022	0.012	1.12	1.07	0.59	0.808	0.823	0.832
Emphysema	2308	314	1597	397	0.023	0.026	0.022	0.026	1.16	0.95	1.12	0.906	0.918	0.916
Fibrosis	1520	80	1131	309	0.015	0.007	0.015	0.020	0.45	1.02	1.33	0.935	0.936	0.939
Pleural Thickening	3013	273	2213	527	0.030	0.023	0.030	0.034	0.77	1.01	1.14	0.827	0.845	0.831
Hernia	186	1	105	80	0.002	0.000	0.001	0.005	0.05	0.77	2.81	0.841	0.824	0.833

Table 4.10: Distribution by age-group, and prior and posterior probabilities of each disease given the age-group.

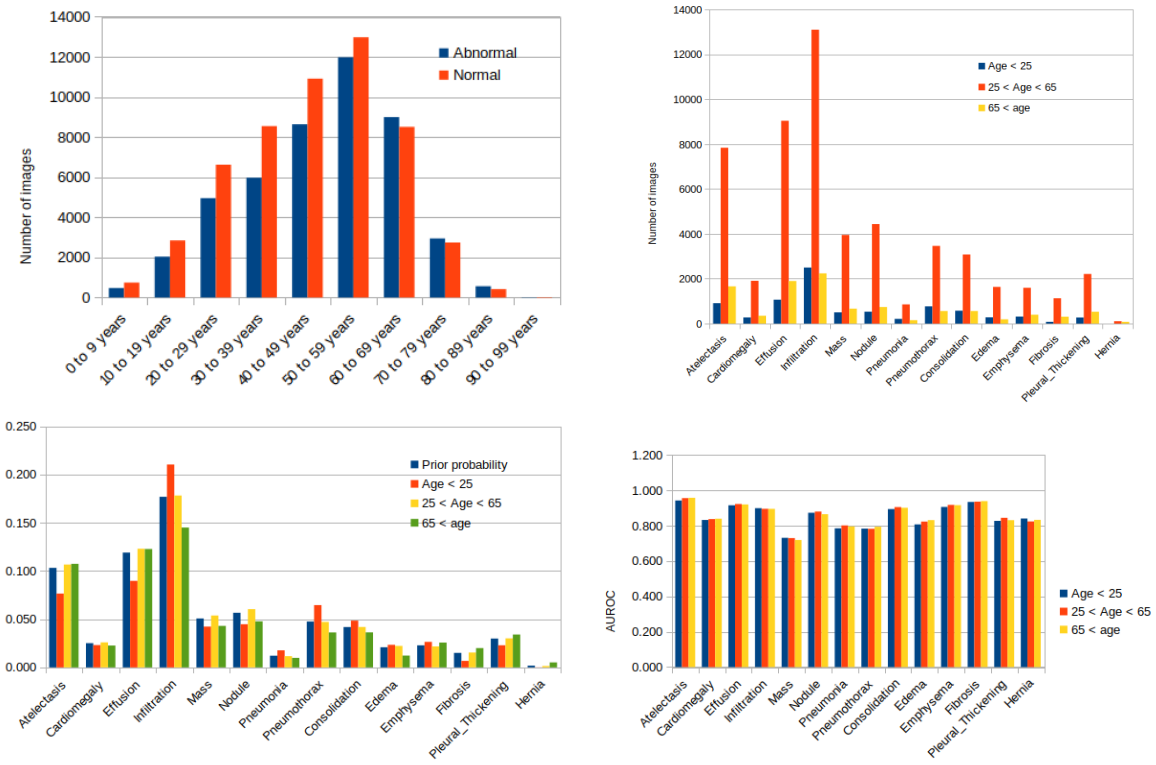

Figure 4.3: Age bias. From top-left clockwise, a) distribution by age, b) distribution by age broken down by abnormality, c) baseline model's AUROC for each abnormality and d) prior and posterior probabilities of each disease given age-group.

4.11.2 Gender

In the combined training and test sets of the NIH CXR-14 dataset (90%), the male to female ratio is approximately 1.28, with similar rates of abnormality, 53.5% and 54.1% for males and females respectively, and *No finding* is only weakly correlated with *female*, with a Pearson product-moment correlation coefficient (PMCC) of 0.006. However, when broken down by specific abnormalities, the ratio of the posterior probability of an abnormality given the gender to its prior probability shows significant variation, with *Hernia* becoming 1.3 times more likely and *Cardiomegaly* becoming 1.2 times more likely if the patient is female (pregnancy is a common cause of *Cardiomegaly*).

Moreover, a similar network (with the same architecture, the only difference being 2 output nodes in the final layer instead of the 14) trained to identify gender from x-ray images on this dataset acheived an accuracy of 93.8% (AUROC of 98.9%) on this task when trained for a single epoch. Saliency maps showed high activations at and around regions of the image containing female breasts (as shown in figure 4.4).

Although this does not necessarily mean that our abnormality-detection models are

biased, the two findings above show that some bias exists in the dataset and that these models are capable of exploiting these. We evaluated our baseline model for variable performance for males and females and found that the model showed similar performance (in terms of AUROC) for both genders.

Abnormality	**Number of images**			**Prior**	**Posterior**		**Posterior / Prior**		**AUROC**	
	Total	Female	Male		Female	Male	Female	Male	Female	Male
Atelectasis	10416	4202	6214	0.103	0.095	0.109	0.92	1.06	0.951	0.942
Cardiomegaly	2532	1347	1185	0.025	0.031	0.021	1.22	0.83	0.835	0.840
Effusion	12015	5311	6704	0.119	0.120	0.118	1.01	0.99	0.921	0.923
Infiltration	17852	7622	10230	0.177	0.173	0.180	0.98	1.02	0.896	0.896
Mass	5121	2035	3086	0.051	0.046	0.054	0.91	1.07	0.732	0.727
Nodule	5710	2417	3293	0.057	0.055	0.058	0.97	1.02	0.877	0.877
Pneumonia	1220	517	703	0.012	0.012	0.012	0.97	1.02	0.797	0.803
Pneumothorax	4794	2351	2443	0.048	0.053	0.043	1.12	0.91	0.771	0.783
Consolidation	4220	1810	2410	0.042	0.041	0.042	0.98	1.01	0.907	0.901
Edema	2103	997	1106	0.021	0.023	0.019	1.08	0.93	0.815	0.826
Emphysema	2308	847	1461	0.023	0.019	0.026	0.84	1.12	0.919	0.912
Fibrosis	1520	702	818	0.015	0.016	0.014	1.06	0.96	0.935	0.936
Pleural Thickening	3013	1205	1808	0.030	0.027	0.032	0.92	1.07	0.837	0.841
Hernia	186	106	80	0.002	0.002	0.001	1.30	0.76	0.828	0.824

Table 4.11: Distribution by gender, and prior and posterior probabilities of each disease given the gender.

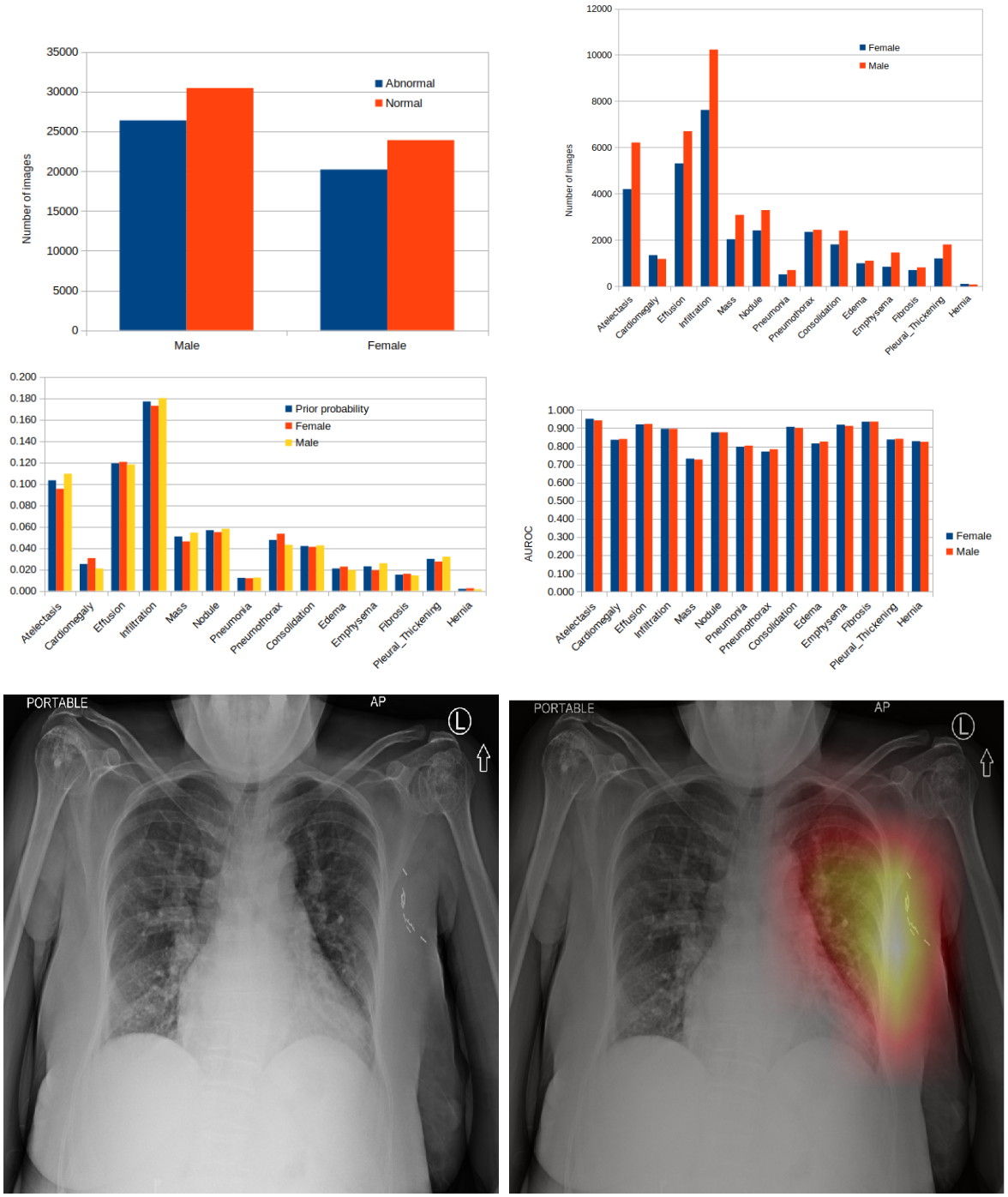

Figure 4.4: Gender bias. First two rows: From top-left clockwise, a) distribution by gender, b) distribution by gender broken down by abnormality, c) baseline model's AUROC for each abnormality and d) prior and posterior probabilities of each disease given gender. The 3^{rd} row shows saliency map for a *female* prediction showing high activation around regions of the image containing female breasts.

4.11.3 View-position

The PA (posterioanterior) view is preferred over the AP (anterioposterior) view. However, the AP view is usually chosen over the PA view for younger children and is necessitated for very ill patients who cannot stand erect. The PA view is more common in the NIH CXR-14 dataset with 60% of the images showing the PA view. Abnormality rate is higher for the AP view (52.8%) compared to the PA view (41.6%), and *No finding* is positively correlated with PA with a Pearson product-moment correlation coefficient (PMCC) of 0.11. When broken down by specific abnormalities, the ratio of the posterior probability of an abnormality given the view-position to its prior probability shows significant variation, with *Edema* and *Consolidation* becoming 2.2 times and 1.7 times more likely respectively if the x-ray image shows AP view.

Moreover, a network with a similar architecture with 2 output nodes trained to identify the view-position from x-ray images acheived an accuracy of 98.7% (AUROC of 99.7%) on this task when trained for a single epoch. Saliency maps showed high activations at and around the anterior aspect of the ribs and around shadows of tokens on x-ray that identified the machine as being a *portable* machine (as shown in figure 4.5).

We evaluated our baseline for variable performace for each view-position and found that the model showed similar performance (in terms of AUROC, sensitivity and specificity).

Abnormality	**Number of images**			**Prior**	**Posterior**		**Posterior / Prior**		**AUROC**	
	Total	PA	AP		PA	AP	PA	AP	PA	AP
Atelectasis	10416	5161	5255	0.103	0.085	0.130	0.83	1.26	0.942437	0.95283
Cardiomegaly	2532	1435	1097	0.025	0.024	0.027	0.95	1.08	0.836189	0.837878
Effusion	12015	5977	6038	0.119	0.099	0.149	0.83	1.25	0.917303	0.928475
Infiltration	17852	8342	9510	0.177	0.138	0.235	0.78	1.33	0.894947	0.898967
Mass	5121	3171	1950	0.051	0.052	0.048	1.03	0.95	0.729112	0.726252
Nodule	5710	3794	1916	0.057	0.063	0.047	1.11	0.84	0.873476	0.881735
Pneumonia	1220	537	683	0.012	0.009	0.017	0.73	1.40	0.800406	0.803007
Pneumothorax	4794	3056	1738	0.048	0.051	0.043	1.06	0.90	0.780797	0.797678
Consolidation	4220	1378	2842	0.042	0.023	0.070	0.54	1.68	0.903083	0.904439
Edema	2103	256	1847	0.021	0.004	0.046	0.20	2.19	0.824132	0.820317
Emphysema	2308	1359	949	0.023	0.022	0.023	0.98	1.03	0.914292	0.91912
Fibrosis	1520	1268	252	0.015	0.021	0.006	1.39	0.41	0.933933	0.934433
Pleural Thickening	3013	2156	857	0.030	0.036	0.021	1.19	0.71	0.843553	0.829741
Hernia	186	159	27	0.002	0.003	0.001	1.43	0.36	0.828903	0.82359

Table 4.12: Distribution by view-position, and prior and posterior probabilities of each disease given the view-position.

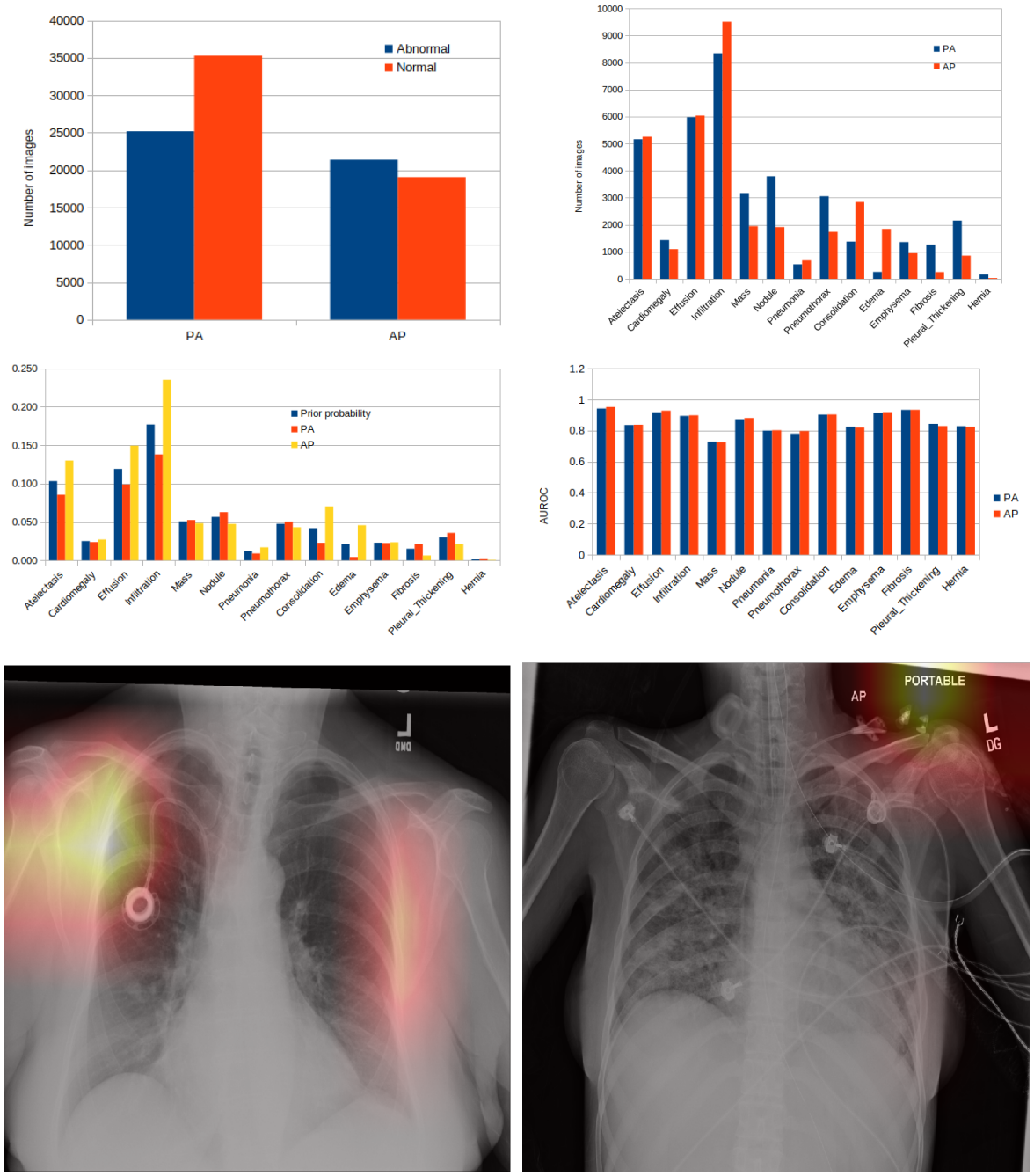

Figure 4.5: View bias. First two rows: From top-left clockwise, a) distribution by view-position, b) distribution by view-position broken down by abnormality, c) baseline model's AUROC for each abnormality and d) prior and posterior probabilities of each disease given view-position. The 3^{rd} from left to right shows saliency maps for *PA* and *AP* showing high activation at and around the anterior aspect of the ribs and around shadows of tokens on the x-ray identifying the machine as being a *portable* machine.

	Training	Testing	AUROC
Pneumonia	NIH CXR-14	NIH CXR-14	0.760
	NIH CXR-14	Guangzhou	0.842
	Guangzhou	Guangzhou	0.9
Tuberculosis	Shenzhen	Shenzhen	0.956
	Shenzhen	Montgomery	0.871
	Montgomery	Montgomery	0.631

Table 4.13: Evaluation results for our baseline models' ability to generalize to external datasets

4.12 Generalization

We evaluate the ability of our baselines to generalize to other hospital systems. At the problem of pneumonia detection, we train on the NIH CXR-14 dataset and test on both the internal test set and the external Guangzhou dataset, a pediatric pneumonia dataset from a different hospital system.

At the problem of tuberculosis detection, we train on the Shenzhen tuberculosis and and test on both the internal test set and the external Montgomery tuberculosis dataset from a different hospital system.

We find that the baseline model trained to detect pneumonia on the NIH CXR-14 dataset performs better on the Guangzhou dataset than the internal test set but worse that a model trained exclusively on the external dataset, and the baseline model trained to detect tuberculosis on the Shenzhen dataset shows inferior performace on the external dataset, but which is better than a model trained exclusively on the external dataset.

4.13 Viral and bacterial pneumonia

Images in the Guangzhou pediatric pneumonia dataset that show manifestations of pneumonia are further categorized as *Viral* and *bacterial*. Since viral and bacterial pneumonia present different levels of emergency and warrant different courses of treatment, we evaluate our models for variable performance on these categories to ensure that they are not biased toward one or the other type.

We found that the models trained on the NIH CXR-14 dataset were better at detecting

Model	**AUROC**	
	Bacterial pneumonia	Viral pneumonia
Baseline	0.835	**0.855**
Higher resolution (512 x 512)	0.816	**0.821**
Progressive resizing (upto 512 x 512)	0.860	**0.891**
Mixup, with $\alpha = 0.4$	0.860	**0.895**

Table 4.14: Variable performance on viral and bacterial pneumonia of models trained on the NIH CXR-14 dataset.

Model	**Average AUROC**		
	Un-segmented	Segmented	Segmented and cropped
Baseline (224 x 224)	0.871	0.867	0.874
(320 x 320)	0.846	0.854	0.859
(448 x 448)	0.809	0.811	0.795
(540 x 540)	0.828	0.826	0.812
(672 x 672)	0.813	0.814	0.828
(224 x 224) Pretrained	0.941	0.940	0.942
(480 x 480) Pretrained	0.957	0.957	0.955

Table 4.15: Performance of models trained on the Shenzhen dataset and evaluated on the Montgomery dataset without segmentation, with segmentation and with segmentation and cropping.

viral pneumonia than bacterial pneumonia.

4.14 Segmentation and centering

The Montgomery dataset includes hand-annotated segmentation masks of both the left and right lungs for each image. We evaluate models trained on the Shenzhen hospital tuberculosis dataset, on the Montgmomery dataset after a) segmenting the lung regions and b) segmeting the lung regions and cropping to the smallest rectangle which encloses both the lungs.

Across multiple models and trials and averaged across 9 folds, we failed to see a significant increase or decrease in performance among models trained on un-segmented images, segmented images and segmented and cropped images.

Chapter 5

Results

5.1 Comparison to previous work and human radiologists

For the NIH CXR-14 dataset, we achieve performance competitive with previous work and show improvements over our baseline. See table 5.1 for the results. Rajpurkar et al. in [49] measured human performance in terms of AUROC for each disease, using the majority vote of 3 independent board-certified cardiothoracic specialist radiologists (average experience 15 years) as ground truth, and measure the the performance of 6 BC radiologists from 3 academic institutions (average experience 12 years) and 3 senior radiology residents by fitting a curve to these 9 radiologists' operating points and calculating the area under it. We compare our models with human radiologist performance and find that the model's performance is on average within 2% of that of human radiologists (see table 5.2 for the comparison).

On the Shenzhen and Montgomery datasets, we achieve performance comparable to previous work and show improvement over our baseline. See tables 5.3 and 5.4 for the results.

Authors	**Average AUROC**
Wang et al. (2017)	0.738
Y. Shen et al.	0.775
H. Wang et al. (ChestNet)	0.781
P. Kumar et al.	0.792
Yao et al. (2017)	0.803
Y. Tang et al.	0.805
S. Guendel et al.	0.807
Yan et al.	0.83
X. Xu et al. (DeepCXray)	0.832
Rajpurkar et al. (CheXNet)	0.841
B. Zhou et al.	0.842
Rajpurkar et al. (ChexNext)	0.849
Our model	**0.856**
Q. Guan et al.	0.871

Table 5.1: Comparison to previous work on the NIH CXR-14 dataset

Abnormality	**AUROC**			
	Baseline	Ensemble	Radiologist	Difference (%)
Atelectasis	0.823	0.839	0.808	-3.06
Cardiomegaly	0.899	0.916	0.888	-2.79
Effusion	0.881	0.89	0.9	0.96
Infiltration	0.705	0.72	0.734	1.39
Mass	0.857	0.868	0.886	1.76
Nodule	0.779	0.817	0.899	8.17
Pneumonia	0.767	0.765	0.823	5.83
Pneumothorax	0.881	0.895	0.94	4.46
Consolidation	0.822	0.819	0.841	2.22
Edema	0.911	0.902	0.91	0.83
Emphysema	0.913	0.944	0.911	-3.33
Fibrosis	0.824	0.854	0.897	4.31
Pleural Thickening	0.81	0.805	0.779	-2.59
Hernia	0.906	0.944	0.985	4.1
Average	**0.841**	**0.856**	**0.8715**	**1.59**

Table 5.2: Comparison to human radiologists on the NIH CXR-14 dataset

Authors	**AUROC**	**Accuracy**
Jaeger et al	0.9	0.841
Hwang et al	0.93	0.837
Lopez and Valiati	0.926	0.846
MT Islam et al	0.94	0.9
Haloi et al	0.949	
Liu et al (ResNet-152)	0.967	0.923
Liu et al (Inception-ResNet-v2)	0.983	0.917
Vajda et al	0.99	0.957
Our baseline	**0.956**	**0.902**
Our best model Pretrained on NIH CXR-14 with mixup $\alpha = 0.4$	**0.985**	**0.949**

Table 5.3: Comparison to previous work on the Shenzhen tuberculosis dataset

Authors	**AUROC**	**Accuracy**
Jaeger et al	0.869	0.783
Lopez and Valiati	0.926	0.826
Liu et al (Inception-ResNet-v2)	0.957	0.844
Liu et al (ResNet-152)	0.951	0.890
Vajda et al	0.870	0.783
Our baseline	**0.871**	**0.755**
Our best model Pre-trained on NIH CXR-14 (480 x 480)	**0.957**	**0.89**

Table 5.4: Comparison to previous work on the Montgomery tuberculosis dataset

5.2 Examples

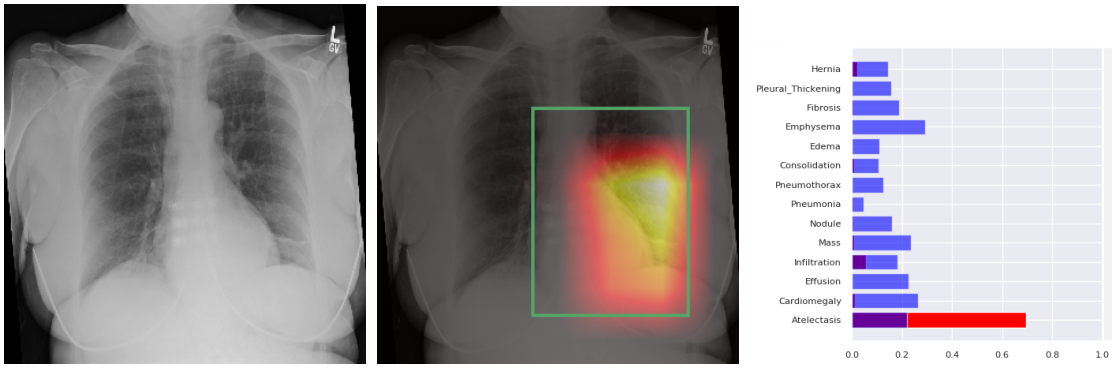

Figure 5.1: Original image, image overlaid with saliency map and bounding boxes for *Atelectasis*, and predicted probabilities for an x-ray image.

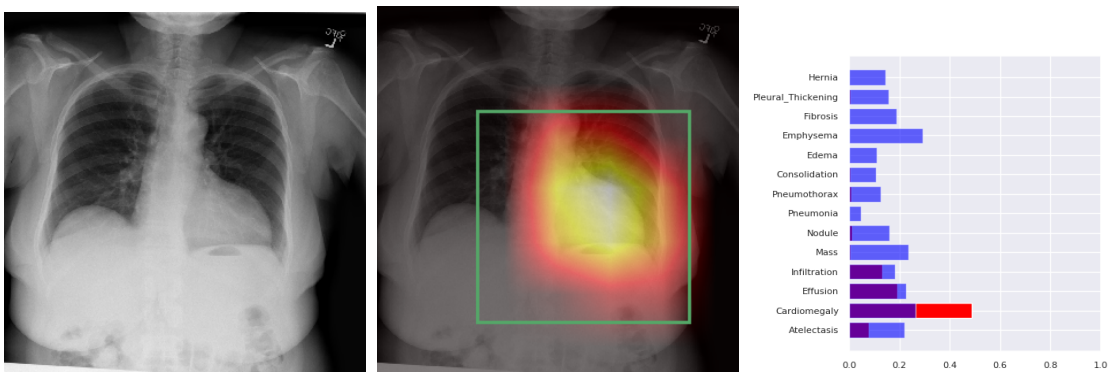

Figure 5.2: Original image, image overlaid with saliency map and bounding boxes for *Cardiomegaly*, and predicted probabilities for an x-ray image.

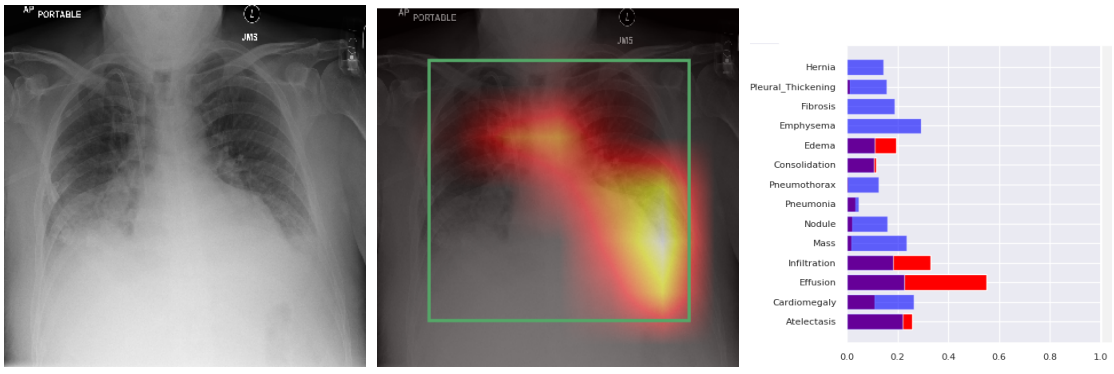

Figure 5.3: Original image, image overlaid with saliency map and bounding boxes for *Effusion*, and predicted probabilities for an x-ray image.

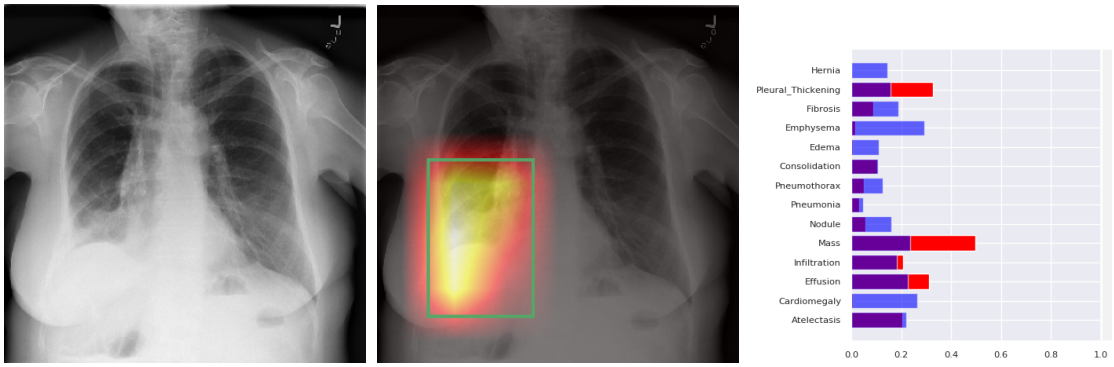

Figure 5.4: Original image, image overlaid with saliency map and bounding boxes for *Infiltration*, and predicted probabilities for an x-ray image.

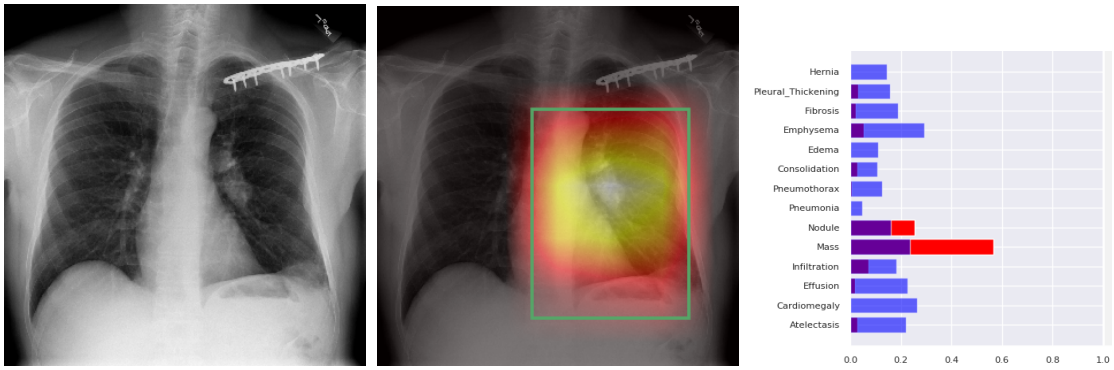

Figure 5.5: Original image, image overlaid with saliency map and bounding boxes for *Mass*, and predicted probabilities for an x-ray image.

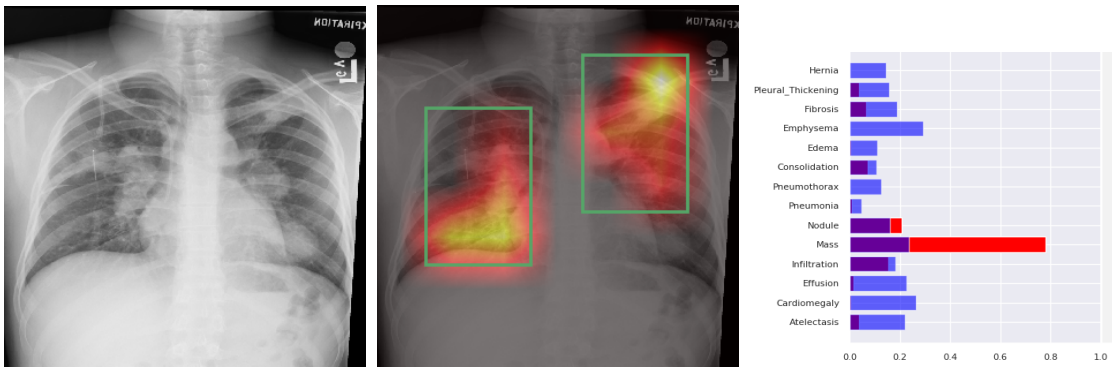

Figure 5.6: Original image, image overlaid with saliency map and bounding boxes for *Nodule*, and predicted probabilities for an x-ray image.

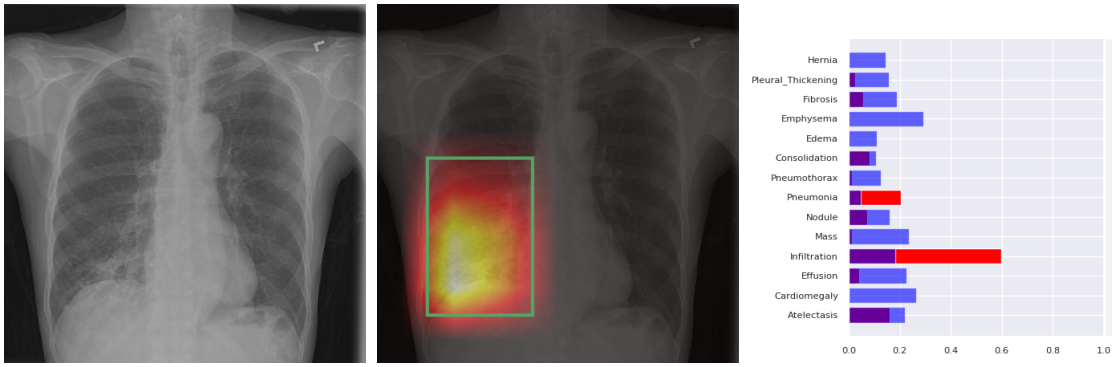

Figure 5.7: Original image, image overlaid with saliency map and bounding boxes for *Pneumonia*, and predicted probabilities for an x-ray image.

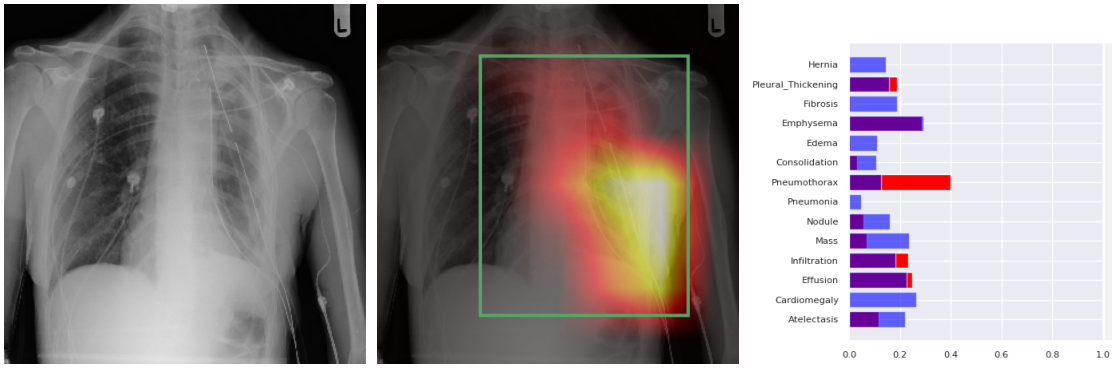

Figure 5.8: Original image, image overlaid with saliency map and bounding boxes for *Pneumothorax*, and predicted probabilities for an x-ray image.

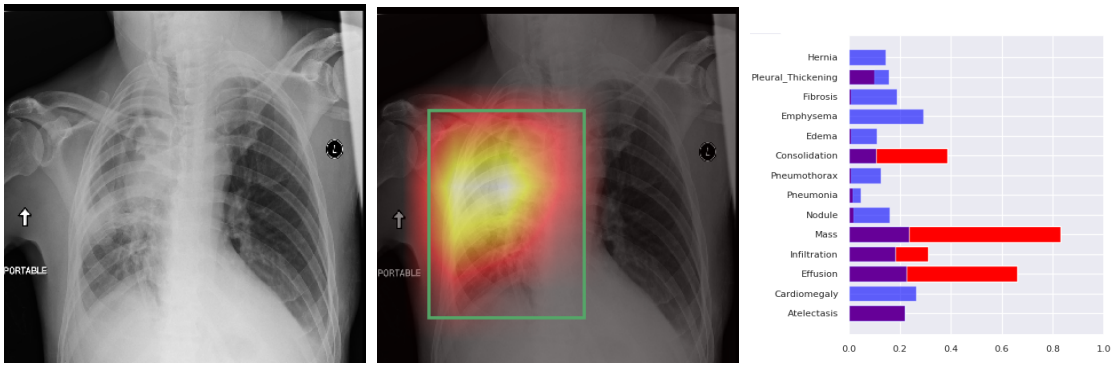

Figure 5.9: Original image, image overlaid with saliency map and bounding boxes for *Consolidation*, and predicted probabilities for an x-ray image.

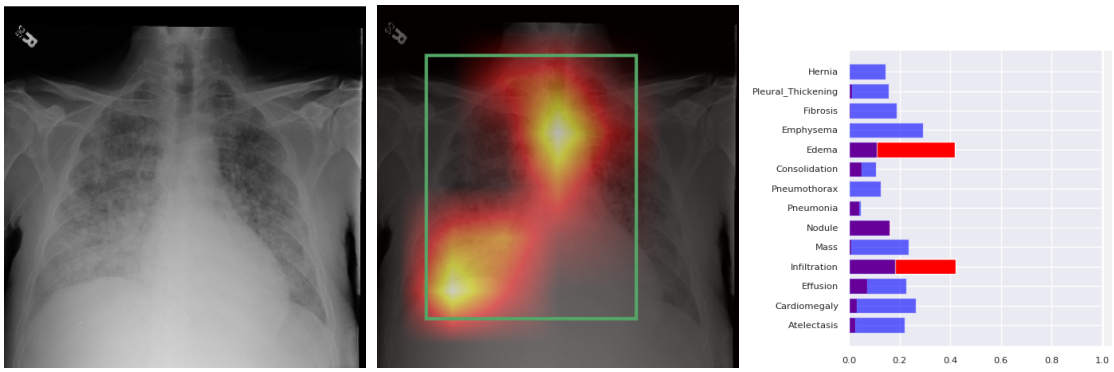

Figure 5.10: Original image, image overlaid with saliency map and bounding boxes for *Edema*, and predicted probabilities for an x-ray image.

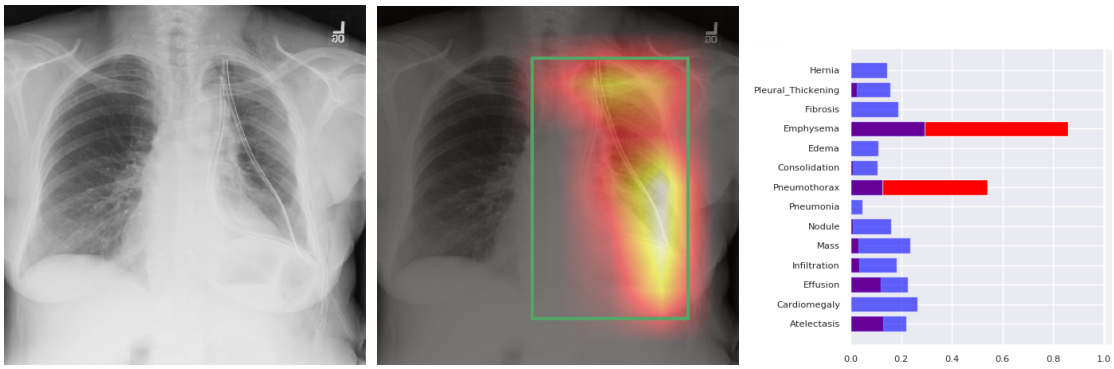

Figure 5.11: Original image, image overlaid with saliency map and bounding boxes for *Emphysema*, and predicted probabilities for an x-ray image.

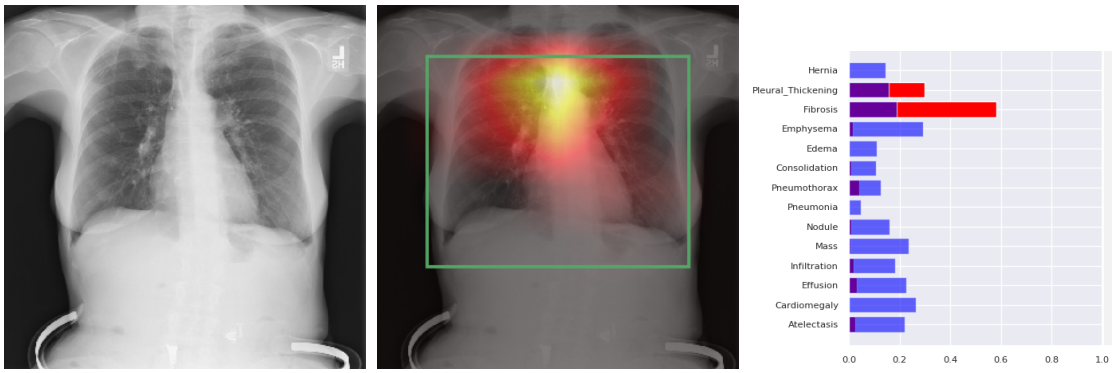

Figure 5.12: Original image, image overlaid with saliency map and bounding boxes for *Fibrosis*, and predicted probabilities for an x-ray image.

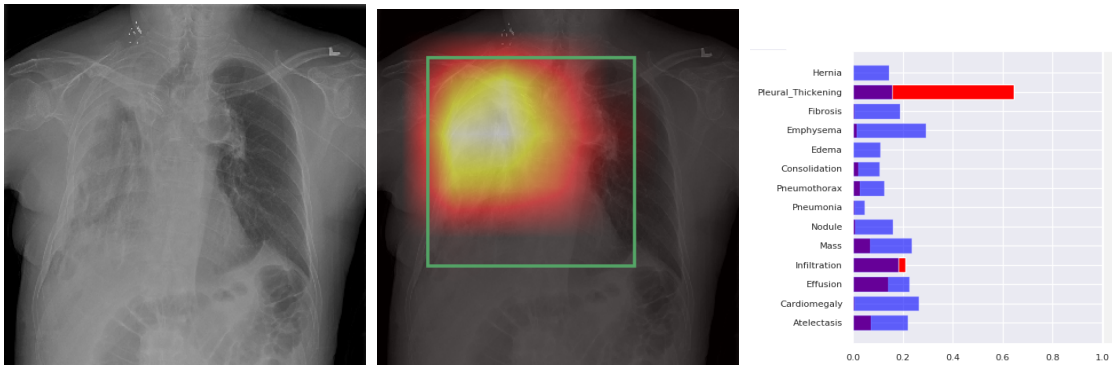

Figure 5.13: Original image, image overlaid with saliency map and bounding boxes for *Pleural Thickening*, and predicted probabilities for an x-ray image.

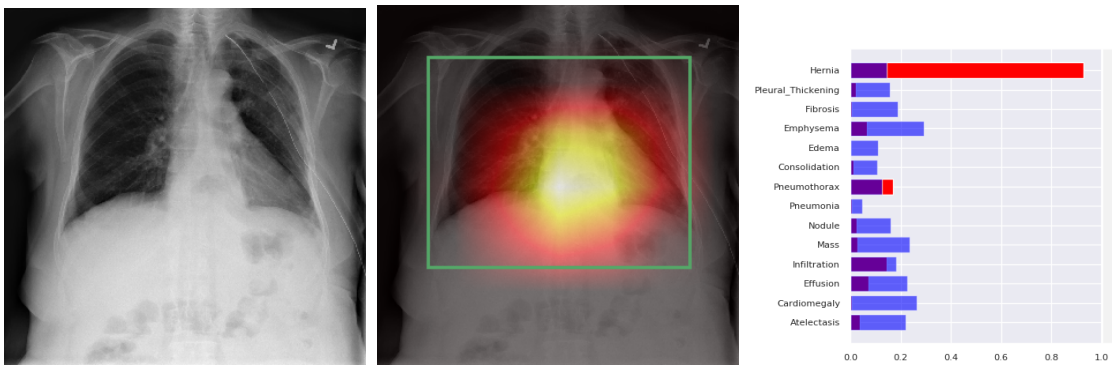

Figure 5.14: Original image, image overlaid with saliency map and bounding boxes for *Hernia*, and predicted probabilities for an x-ray image.

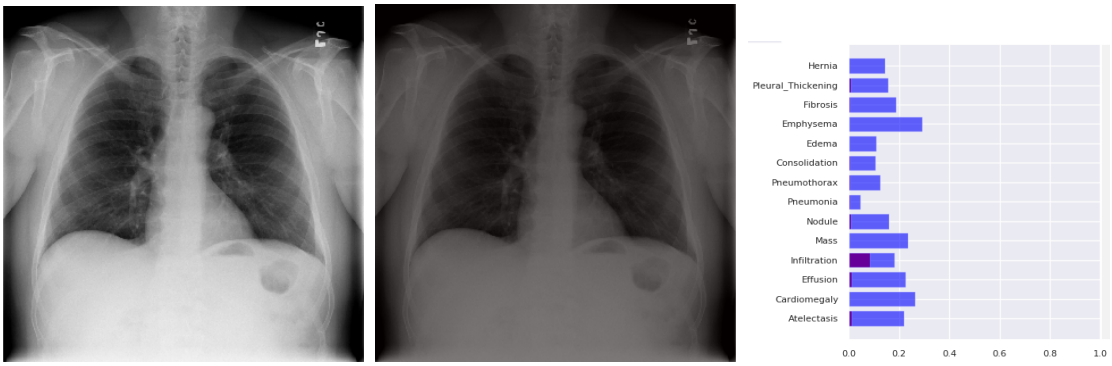

Figure 5.15: Original image, image overlaid with saliency map and bounding boxes, and predicted probabilities for an x-ray image showing no abnormalities.

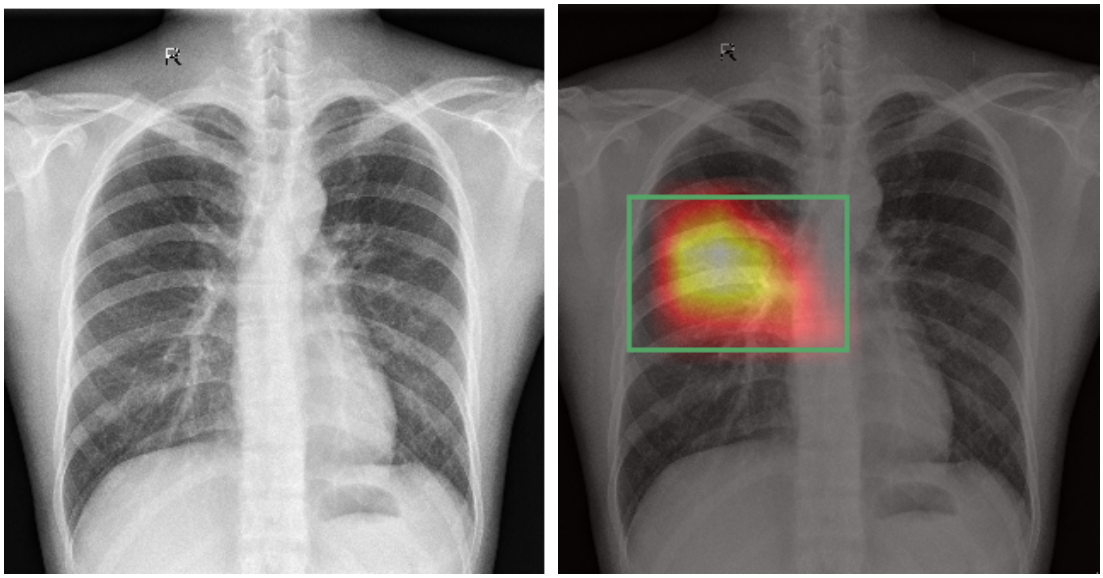

Figure 5.16: Original image, image overlaid with saliency map and bounding boxes for *Tuberculosis*

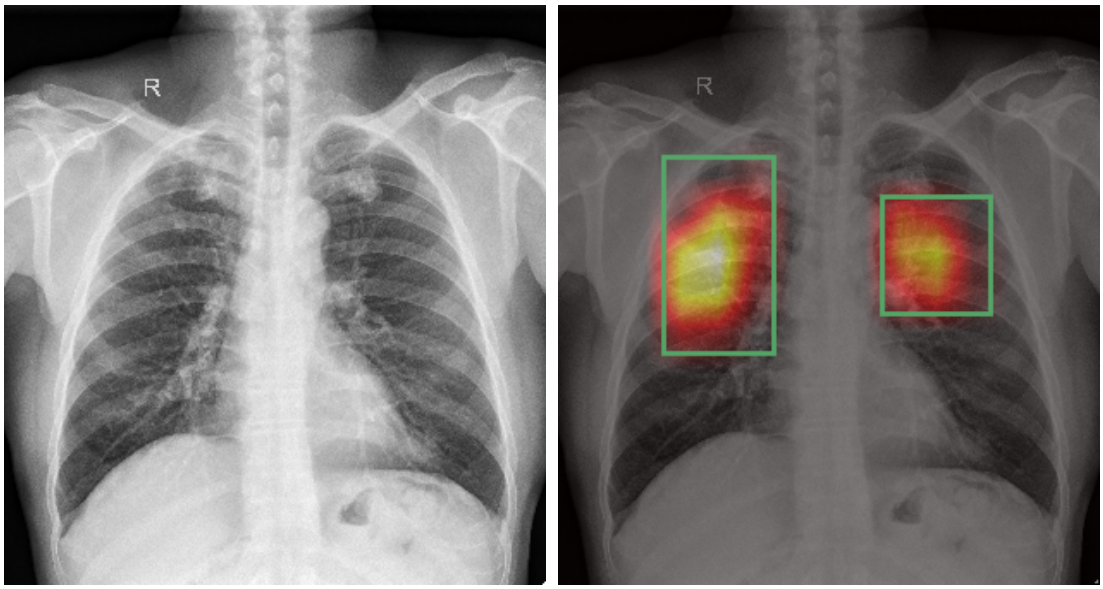

Figure 5.17: Original image, image overlaid with saliency map and bounding boxes for *Tuberculosis*

Chapter 6

Conclusion

We developed algorithms that can detect abnormalities on the x-ray and explain these detections by generating heatmaps pointing out areas of the image that most influenced it. We established baselines, benchmarked against previous work and showed that a) transfer-learning from a large non-TB dataset dramatically improves TB detection, b) models in the domain show inferior performance on external data from a different hospital system but c) recent techniques such as mixup and progressive resizing improve performance and generalization. We achieved performance competitive with previous work in detecting pneumonia-like and other abnormalities on the NIH chestX-ray14 dataset and in detecting tuberculosis on the Shenzhen hospital dataset, and achieved state-of-the-art performance on the Montgomery county tuberculosis dataset.

Avenues for future research include:

1. Training on images from multiple hospital systems to improve the model's ability to generalize to other hospital systems and machine types, perhaps using techiniques in domain adaptation or deep domain confusion [50] to prevent the model from learning features that are necessary to identify the hospital system.

2. Exploring ways to infuse domain knowledge into the algorithm to exploit correlations between the abnormalities.

3. Using attention to allow the network to focus on pathological areas.

4. Training models end-to-end to generate radiology reports in natural language from x-ray images.

5. Segmentation of lung regions and bone shadow suppression to reduce the number of false-positives.

6. Using recently released datasets such as CheXPert[51] and PadChest[52] which have more images and a hierarchical labeling schema.

Bibliography

[1] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, *ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases*, 2017.

[2] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, p. 475, 2014.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, ACM, 2016, pp. 1135–1144.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: `http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf`.

[5] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, e0130140, 2015.

[6] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, JMLR. org, 2017, pp. 3145–3153.

[7] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[9] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do cifar-10 classifiers generalize to cifar-10?" *arXiv preprint arXiv:1806.00451*, 2018.

[10] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLoS medicine*, vol. 15, no. 11, e1002683, 2018.

[11] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," in *Advances in neural information processing systems*, 2016, pp. 4349–4357.

[12] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 77–91.

[13] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[14] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, ACM, 2012, pp. 214–226.

[15] M. Hardt, E. Price, N. Srebro, *et al.*, "Equality of opportunity in supervised learning," in *Advances in neural information processing systems*, 2016, pp. 3315–3323.

[16] A. Agarwal, A. Beygelzimer, M. Dudık, J. Langford, and H. Wallach, "A reductions approach to fair classification," *arXiv preprint arXiv:1803.02453*, 2018.

[17] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, 2017, pp. 4066–4076.

[18] J. Dastin, *Insight - amazon scraps secret ai recruiting tool that showed bias...* Oct. 2018. [Online]. Available: `https://in.reuters.com/article/amazon-com-jobs-automation/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idINKCN1MK0AH`.

[19] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, p. 115, 2017.

[20] *Pneumonia*, Nov. 2016. [Online]. Available: `https://www.who.int/news-room/fact-sheets/detail/pneumonia`.

[21] *Tb and poverty*. [Online]. Available: `https://www.tbalert.org/about-tb/global-tb-challenges/tb-poverty/`.

[22] *Tuberculosis (tb)*, Sep. 2018. [Online]. Available: `https://www.who.int/news-room/fact-sheets/detail/tuberculosis`.

[23] A. Modi and R. Suresh, *Scaling up tb screening with ai: Deploying automated x-ray screening in remote regions*, Apr. 2019. [Online]. Available: http://blog.qure.ai/notes/scaling-up-tb-screening-with-ai.

[24] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," May 2017. DOI: 10.1109/CVPR.2017.369. arXiv: 1705.02315. [Online]. Available: http://arxiv.org/abs/1705.02315%20http://dx.doi.org/10.1109/CVPR.2017.369.

[25] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Diagnose like a Radiologist: Attention Guided Convolutional Neural Network for Thorax Disease Classification," Jan. 2018. arXiv: 1801.09927. [Online]. Available: http://arxiv.org/abs/1801.09927.

[26] Y. Tang, X. Wang, A. P. Harrison, L. Lu, J. Xiao, and R. M. Summers, "Attention-Guided Curriculum Learning for Weakly Supervised Classification and Localization of Thoracic Diseases on Chest Radiographs," in, Springer, Cham, Sep. 2018, pp. 249–258. DOI: 10.1007/978-3-030-00919-9_29. [Online]. Available: http://link.springer.com/10.1007/978-3-030-00919-9%7B%5C_%7D29.

[27] H. Wang and Y. Xia, "ChestNet: A Deep Neural Network for Classification of Thoracic Diseases on Chest Radiography," Jul. 2018. arXiv: 1807.03058. [Online]. Available: http://arxiv.org/abs/1807.03058.

[28] E. Pesce, P.-P. Ypsilantis, S. Withey, R. Bakewell, V. Goh, and G. Montana, "Learning to detect chest radiographs containing lung nodules using visual attention networks," Dec. 2017. DOI: 10.1016/j.media.2018.12.007. arXiv: 1712.00996. [Online]. Available: http://arxiv.org/abs/1712.00996%20http://dx.doi.org/10.1016/j.media.2018.12.007.

[29] J. Cai, L. Lu, A. P. Harrison, X. Shi, P. Chen, and L. Yang, "Iterative Attention Mining for Weakly Supervised Thoracic Disease Pattern Localization in Chest X-Rays," in, Springer, Cham, Sep. 2018, pp. 589–598. DOI: 10.1007/978-3-030-00934-2_66. [Online]. Available: http://link.springer.com/10.1007/978-3-030-00934-2%7B%5C_%7D66.

[30] L. Yao, E. Poblenz, D. Dagunts, arXiv preprint arXiv ..., and undefined 2017, "Learning to diagnose from scratch by exploiting dependencies among labels," *arxiv.org*, [Online]. Available: https://arxiv.org/abs/1710.10501.

[31] L. Yao, J. Prosky, E. Poblenz, B. Covington, and K. Lyman, "Weakly Supervised Medical Diagnosis and Localization from Multiple Resolutions," Mar. 2018. arXiv: 1803.07703. [Online]. Available: http://arxiv.org/abs/1803.07703.

[32] S. Sedai, D. Mahapatra, Z. Ge, R. Chakravorty, and R. Garnavi, "Deep Multi-scale Convolutional Feature Learning for Weakly Supervised Localization of Chest Pathologies in X-ray Images," in, Springer, Cham, Sep. 2018, pp. 267–275. DOI: 10.1007/978-3-030-00919-9_31. [Online]. Available: http://link.springer.com/10.1007/978-3-030-00919-9%7B%5C_%7D31.

[33] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, B. N. Patel, K. W. Yeom, K. Shpanskaya, F. G. Blankenberg, J. Seekins, T. J. Amrhein, D. A. Mong, S. S. Halabi, E. J. Zucker, A. Y. Ng, and M. P. Lungren, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLOS Medicine*, vol. 15, no. 11, A. Sheikh, Ed., e1002686, Nov. 2018, ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1002686. [Online]. Available: http://dx.plos.org/10.1371/journal.pmed.1002686.

[34] S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, Zhiyun Xue, K. Palaniappan, R. K. Singh, S. Antani, G. Thoma, Yi-Xiang Wang, Pu-Xuan Lu, and C. J. McDonald, "Automatic Tuberculosis Screening Using Chest Radiographs," *IEEE Transactions on Medical Imaging*, vol. 33, no. 2, pp. 233–245, Feb. 2014, ISSN: 0278-0062. DOI: 10.1109/TMI.2013.2284099. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24108713%20http://ieeexplore.ieee.org/document/6616679/.

[35] U. Lopes and J. Valiati, "Pre-trained convolutional neural networks as feature extractors for tuberculosis detection," *Computers in Biology and Medicine*, vol. 89, pp. 135–143, Oct. 2017, ISSN: 0010-4825. DOI: 10.1016/J.COMPBIOMED.2017.08.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482517302548.

[36] S. Vajda, A. Karargyris, S. Jaeger, K. Santosh, S. Candemir, Z. Xue, S. Antani, and G. Thoma, "Feature Selection for Automatic Tuberculosis Screening in Frontal Chest Radiographs," *Journal of Medical Systems*, vol. 42, no. 8, p. 146, Aug. 2018, ISSN: 0148-5598. DOI: 10.1007/s10916-018-0991-9. [Online]. Available: http://link.springer.com/10.1007/s10916-018-0991-9.

[37] S. Hwang, H.-E. Kim, J. Jeong, and H.-J. Kim, "A novel approach for tuberculosis screening based on deep convolutional neural networks," G. D. Tourassi and S. G. Armato, Eds., vol. 9785, International Society for Optics and Photonics, Mar. 2016, 97852W. DOI: 10.1117/12.2216198. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2216198.

[38] M. T. Islam, M. A. Aowal, A. T. Minhaz, and K. Ashraf, "Abnormality Detection and Localization in Chest X-Rays using Deep Convolutional Neural Networks,"

May 2017. arXiv: 1705.09850. [Online]. Available: http://arxiv.org/abs/1705.09850.

[39] M. Haloi, K. R. Rajalakshmi, and P. Walia, "Towards Radiologist-Level Accurate Deep Learning System for Pulmonary Screening," Jun. 2018. arXiv: 1807.03120. [Online]. Available: http://arxiv.org/abs/1807.03120.

[40] J. Liu, Y. Liu, C. Wang, A. Li, B. Meng, X. Chai, and P. Zuo, "An Original Neural Network for Pulmonary Tuberculosis Diagnosis in Radiographs," in, Springer, Cham, Oct. 2018, pp. 158–166. DOI: 10.1007/978-3-030-01421-6_16. [Online]. Available: http://link.springer.com/10.1007/978-3-030-01421-6%7B%5C_%7D16.

[41] N. van Noord and E. Postma, "Learning scale-variant and scale-invariant features for deep image classification," *Pattern Recognition*, vol. 61, pp. 583–592, Jan. 2017, ISSN: 0031-3203. DOI: 10.1016/J.PATCOG.2016.06.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320316301224.

[42] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," Oct. 2017. arXiv: 1710.09412. [Online]. Available: http://arxiv.org/abs/1710.09412.

[43] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[45] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[46] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," *arXiv preprint arXiv:1705.10694*, 2017.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[48] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Medicine*, vol. 15, no. 11, A. Sheikh, Ed., e1002683, Nov. 2018, ISSN: 1549-1676. DOI: 10.1371/journal.pmed.

1002683. [Online]. Available: http://dx.plos.org/10.1371/journal.pmed.1002683.

[49] P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, B. N. Patel, K. W. Yeom, K. Shpanskaya, F. G. Blankenberg, J. Seekins, T. J. Amrhein, D. A. Mong, S. S. Halabi, E. J. Zucker, A. Y. Ng, and M. P. Lungren, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS Medicine*, vol. 15, no. 11, A. Sheikh, Ed., e1002686, Nov. 2018, ISSN: 15491676. DOI: 10.1371/journal.pmed.1002686. [Online]. Available: http://dx.plos.org/10.1371/journal.pmed.1002686.

[50] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[51] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," Jan. 2019. arXiv: 1901.07031. [Online]. Available: http://arxiv.org/abs/1901.07031.

[52] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "PadChest: A large chest x-ray image dataset with multi-label annotated reports," Jan. 2019. arXiv: 1901.07441. [Online]. Available: http://arxiv.org/abs/1901.07441.