CS230

Diagnosing cardiomegaly through cardiothoracic ratio calculation from chest x-ray images

Ignacio Berdinas
Stanford University
berdinas@stanford.edu

Abstract

Developing models that are easy to understand is key for the adoption of ML models in medical practice. We propose an approach to calculate the cardiothoracic ratio, a measurement used by radiologists to predict cardiomegaly, by using two models to predict lung masks and heart bounding boxes. We train our models with public datasets and, after that, fine-tune them on CheXpert to improve cross-dataset generalization and validate our approach. We achieve an AUROC score of 0.909 comparable to state-of-the-art end-to-end cardiomegaly prediction approaches.

1 Introduction

Automated chest radiograph interpretation can provide benefits in medical settings. While there has been a lot of progress using novel datasets, there are still improvements to be made in the classification of edge cases, as explained in the results analysis from the CheXpert paper (1). Cardiomegaly can be diagnosed by measuring the cardiothoracic ratio (CTR)(2) (see figure 1), which has a range considered as normal, but radiologists have uncertainty near the edges of that range.

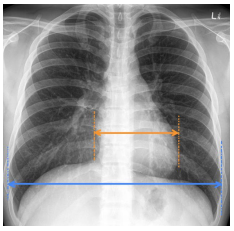

Figure 1: The orange line indicates the maximal horizontal cardiac diameter and the blue line indicates the maximal horizontal thoracic diameter

Our intuition is that, by learning the information that is used by radiologists to identify diseases we can provide an explanation to the model predictions. Complementing an end-to-end cardiomegaly prediction model such as the one trained on (1) with a model trained to calculate the CTR can help to flag these uncertain cases in the same way radiologists do it.

2 Related work

Cardiothoracic ratio prediction with deep learning has been explored using different neural network architectures and datasets. In (3), authors use a private dataset consisting of 5000 x-ray images to train a U-Net architecture (4) to predict 3 different classes of mask: left lung, right lung and heart, and later apply a post-processing step to measure the CTR on the masks. Furthermore on (5), authors follow the same approach and try different U-Net variations such as Attention U-net (6), SE-Resnext

CS230: Deep Learning, Spring 2021, Code: https://github.com/ignacioberdinas/cs-230-project

and EfficientNet U-Net (7). Both (3) and (5) use private datasets that are not publicly available for training.

In this work, we draw from publicly available datasets. We use the Shenzhen and Montgomery datasets (8) to train the lung segmentation. To train the heart bounding-box prediction, we use the VinBigData dataset. Finally, we use the trained models to calculate the CTR on the CheXpert dataset (1).

It is important to consider the generalization across different chest x-ray datasets. In CheXpedition(9), Rajpurkar et al. find that generalization is not guaranteed and that further work is needed to validate that the model still performs as expected.

Lastly, it is important to consider the approach used to generate the dataset. CheXpert(1) labels are the output of an automated rule-based labeler that extracts them from the text radiology reports. Newer methods (10) (11) achieve higher quality labels from the same radiology reports(12), but the labels resulting from this newer methods are not yet public.

3 Datasets

In our approach, we train models for two tasks: lung segmentation and heart detection. The output of these models is then used for the CTR calculation and cardiomegaly classification.

For the **lung segmentation**, we used the Shenzhen and Montgomery datasets(8). Both datasets were initially labeled for the purpose of Tuberculosis detection, but they also have masks for lung segmentation. The Shenzhen dataset consists of 662 x-rays and was collected at the Shenzhen No.3 Hospital in China. The Montgomery dataset consists of 138 x-rays and was collected by the Department of Health and Human Services in Montgomery County, USA.

For the **heart detection**, we used the dataset from the Kaggle competition "VinBigData Chest X-ray Abnormalities Detection". It contains bounding boxes of illneses that can be identified via chest x-ray for 15,000 images. From these, we took the 2,300 samples tagged as cardiomegaly, that contain heart bounding-boxes.

For the **final cardiomegaly classification task**, we used the CheXpert dataset (1), which contains 224,316 chest radiographs of 65,240 patients. The chest radiographs are gray-scale images, rescaled from the original radiographs to a height of 320 pixels while preserving the aspect ratio. For this project we only used the images tagged as 'No finding', 'Cardiomegaly' and 'Uncertain Cardiomegaly' and we end up with 9,547 x-rays.

We also sampled 500 images (200 'Cardiomegaly', 200 'No finding', 100 'Cardiomegaly uncertain') from the 9,547 subset we extracted from CheXpert and labeled them with both the heart bounding box and the lungs mask. This dataset was used for validation purposes, as it allowed us to validate each step of the pipeline with the same set of images; and it was also used for fine-tuning the models on the CheXpert dataset to tackle the limitations of cross dataset model generalization.

4 Methodology

4.1 Models

The proposed methodology consists of two different models, as shown on figure 2: For the lung segmentation, we train a U-Net (4) to predict lung masks, while we use a Faster R-CNN network (13) to predict bounding boxes for the heart.

Lung Segmentation: The U-net architecture uses a resnet-50 backbone pre-trained on ImageNet as a feature extractor. It is trained using data augmentation techniques: rotation, horizontal flip and scale rotate. The loss function we optimize in this task is binary-cross-entropy dice loss. First, we do a pre-training on the Schenzen and Montgomery datasets for 5 epochs using Adam Optimizer with a learning rate $\lambda = 5 * 10^{-5}$. After that, we do a fine tuning with the 400 labeled examples from CheXpert with a learning rate $\lambda = 5 * 10^{-6}$ for 5 epochs.

Heart bounding boxes: The Faster R-CNN network also uses resnet-50 as a feature extractor. It uses a multi-task loss that optimizes the class prediction and the bounding box regression simultaneously,

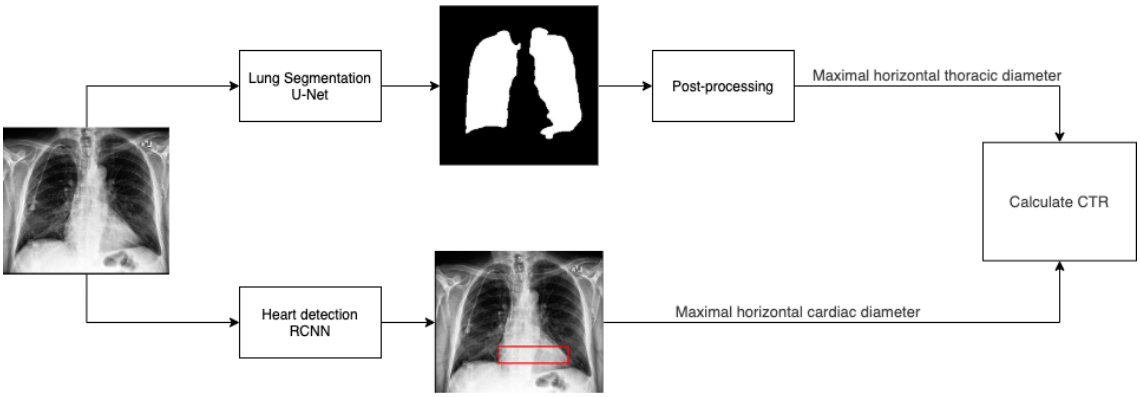

Figure 2: Process diagram

as described in the Fast R-CNN paper (14). The model is first trained on the "VinBigData Chest X-ray Abnormalities Detection" using the Adam optimizer with a learning rate $\lambda = 0.05$ for 10 epochs. After that, we do a fine tuning with the 400 labeled examples from CheXpert with a learning rate $\lambda = 0.01$ for 10 more epochs.

Fine tuning: The fine tuning is motivated by two different aspects: First, we want to address the data distribution differences between the datasets used to train each of the models and the one used to validate our Cardiomegaly classification approach. Second, for the heart detection model, all labeled examples are of abnormal hearts, while the dataset we use to fine-tune contains instances of both normal and abnormal hearts, to avoid the model being biased to predict the wider bounding boxes that are characteristic of abnormal hearts. *From now on, we will refer as 'Baseline' to the model without the fine-tuning step.*

4.2 Post-processing and CTR calculation

For the calculation of the maximal horizontal thoracic diameter we do a post-processing on the predicted lungs mask: To improve the predictions we detect all the contours present on the predicted mask and only keep the two biggest contours; this step helps clean small patches of mask that can introduce noise in the width calculation. We then calculate the maximal mask width by taking the biggest horizontal width of the mask in pixels and dividing it by the width of the image.

The maximal horizontal cardiac diameter is calculated by taking the width of the predicted bounding box and dividing it by the width of the image.

Both the thoracic and the cardiac diameters are calculated as a percentage of the image width. Finally, we calculate the CTR by dividing the cardiac diameter by the thoracic diameter: CTR = Cardiac diameter / Thoracic diameter .

5 Evaluation

5.1 Lung segmentation

To study the results of the lung segmentation we use the 100 labeled samples and analyze the prediction in two different aspects. First, we calculate the Dice coefficient to measure the similarity between the expected and predicted masks. Second, we evaluate how the model performs at measuring the width of the lungs, after applying the same post-processing step that we would apply for the CTR calculation to get the predicted width and compare it to the expected width.

With the fine-tuning, the Dice score rises from **0.903** to **0.942**. What is more important is that we observe a significant reduction of all absolute error statistics for widths calculation, as can be seen in table 1.

	Lung width		Heart width		CTR	
	Baseline	Finetuned	Baseline	Finetuned	Baseline	Finetuned
RMSE	0.0312	0.0145	0.0621	0.0261	0.1057	0.0344
Min	0.0003	0.0000	0.0006	0.0001	0.0021	0.0000
Max	0.1847	0.0562	0.1493	0.0835	0.5913	0.0999
Mean	0.0195	0.0109	0.0517	0.0185	0.0789	0.0251
Median	0.0136	0.0078	0.0478	0.0130	0.0704	0.0185
STD	0.0244	0.0095	0.0344	0.0184	0.0703	0.0235

Table 1: Statistics of the errors for the lung width, heart width and CTR validated against the 100 labeled samples from CheXpert. Lung and Heart width are divided by the width of the image so that errors can be compared across different image sizes.

5.2 Heart detection

To study the results of the heart detection, we do the same as for the Lung segmentation. First, we calculate the Precision at different IoU thresholds to measure the accuracy of the model. After this, we evaluate how the model performs at measuring the width of the heart.

In table 1, we see that the heart detection model benefits more from the fine-tuning compared to the lung segmentation model. The benefits in the heart detection model can be seen in the increase of precision at higher IoU thresholds, as shown in table 2. We observe that the baseline model fails to detect the heart at higher IoU thresholds; this aspect improves after the fine-tuning. In table 1, we can see how the model performs well, correctly predicting the width, even if the precision at IoU 0.9 is still relatively low.

IoU thresholds	0.5	0.75	0.9
Baseline precision	0.94	0.20	0.00
Fine-tuned precision	0.99	0.87	0.39

Table 2: Heart bounding box prediction precision values at different IoU Thresholds for both the baseline model and the fine-tuned model

5.3 Cardiothoracic ratio calculation

We calculate the CTR using the outputs of the lung segmentation and heart detection steps. These results are validated in two different ways: First, we evaluate the error of the CTR calculation against our validation dataset (the 100 validation images that we labeled from CheXpert). Second, we evaluate the classification accuracy of this approach against the CheXpert labels (all of the CheXpert images labeled as 'No finding', 'Cardiomegaly' or 'Uncertain Cardiomegaly'). It is important to consider that CheXpert labels have this 'Uncertain cardiomegaly' category that for the accuracy and AUROC score we will treat as 'Cardiomegaly'; this is known as **U-Ones** strategy (1).

In table 1 we see the impact of how the errors add up when including more than one machine learning model on a pipeline. While the errors of each model are low, an error of 0.07 (the median error) is large when considering the mean 'Cardiomegaly' CTR and 'No finding' CTR: the difference between the means for 'Cardiomegaly' and 'No finding' CTRs is only 0.1443.

	AUROC score
Our approach (Baseline)	0.8418
CheXpert (1)	0.854
Ye et al. (15)	0.870
Pham et al.(16)	0.910

	AUROC score
Our approach (Baseline)	0.8419
Our approach (Fine-tuned)	0.909

Table 3: AUC scores for cardiomegaly classification. For the table on the left we use all the samples to calculate the AUROC score. For the table on the right we remove the samples used for the fine-tuning. From the total 9,547 x-rays used for validation, we only use 400 for fine-tuning.

The whole pipeline achieves an accuracy of **78%** on Cardiomegaly classification without the fine-tuning of the models. After fine-tuning, the model accuracy improves to **85%**.

In figure 3 we can observe the distribution of the CTR predictions. It is interesting to analyze the distribution means for each of the classes. We observe that the mean of the 'Uncertain Cardiomegaly'

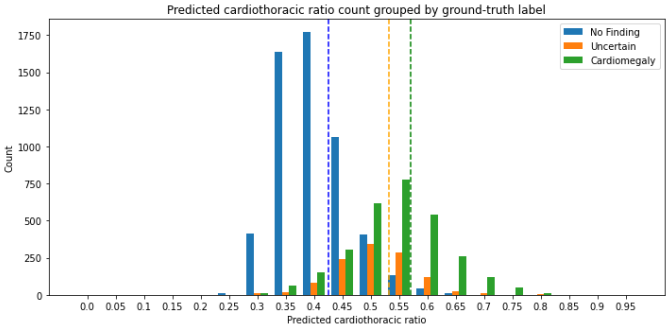

Figure 3: Distribution of the predictions of the CTR for the fine-tuned pipeline by label. The dotted lines indicate the mean for each distribution.
No finding: 0.423
Uncertain: 0.530
Cardiomegaly: 0.566

samples (**0.530**) is closer to the mean of the 'Cardiomegaly' samples (**0.566**), while the mean for the 'No finding' and healthy hearts (**0.423**) is well below the classification threshold of 0.5, which explains why the radiology reports find this group of cases as uncertain.

For table 3, we treat the CTR ratio as the probability that the person suffers from Cardiomegaly, interpreting the ratio as a probability lets us do a comparison with other end-to-end approaches that report the AUROC score of their methods. While our baseline experiment performs worse than other end-to-end approaches, after fine-tuning we achieve an AUROC of **0.909**. We can't directly compare this to the other end-to-end approaches, as the validation dataset differs due to the 400 samples (of 9,547) that we removed to use for fine-tuning. However, considering that the baseline AUROC score only changed in the fourth significant digit when removing those samples, we can estimate that our fine-tuned model AUROC score is close to the state-of-the-art approach from Pham et al.(16), which is the second best solution in the CheXpert leaderboard.

6 Discussion

We achieve a similar AUROC score to state-of-the-art end-to-end approaches (16) and surpass the CheXpert baseline(1) (taking into account the considerations about the fine-tuning, detailed in the previous section). The proposed approach has the added benefit of being able to make a prediction in a way that is easily verifiable by the radiologist. This avoids the issue of some end-to-end computer vision models that take spurious "shortcuts"(17) over signal when classifying. We find this approach useful as a complementary method to the end-to-end model to mitigate these issues.

We show that by fine-tuning the models on a small amount of samples, we can surpass the limitations on generalization discussed on CheXpedition(9).

Furthermore, during error analysis, we observed that some of the false-positives exhibited signs of cardiomegaly. We think that these samples might be errors of the automatic labeler. We believe this method can help in the evaluation of the results of automatic labelers by highlighting instances where the CTR (a measurement used by radiologists) doesn't agree with the label obtained by the automatic labeler.

Finally, we show that our approach has a high accuracy at detecting the heart width and lungs shape. We believe that the intermediate results of this approach: lungs mask, heart bounding box and cardio-thoracic ratio may be useful as input features for other tasks.

7 Future work

As explored in (12) the labels we take in this work as ground-truth are the product of an automatic radiology report labeler. Future work includes validating this approach against labels directly assigned by radiologists and against newer sets of labels created by novel automatic radiology report labelers such as (10)(11).

Furthermore, this approach can be tested with other illnesses such as Atelectasis, which is the complete or partial collapse of a lung or a section of a lung, and might be classified by measuring the area of each lung.

8 Contributions

All work is my own. Thanks to Huizi Mao for the advice on the cross dataset generalization limitations that led to trying out fine-tuning. Thanks to Sofia Distefano, 4th year medical school student, for providing the domain knowledge and helping with the labelling.

References

[1] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. L. Ball, K. S. 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," *CoRR*, vol. abs/1901.07031, 2019.

[2] Y. B. Mensah, K. Mensah, S. Asiamah, H. Gbadamosi, E. A. Idun, W. Brakohiapa, and A. Oddoye, "Establishing the Cardiothoracic Ratio Using Chest Radiographs in an Indigenous Ghanaian Population: A Simple Tool for Cardiomegaly Screening," *Ghana Medical Journal*, vol. 49, pp. 159–164, Sept. 2015.

[3] Z. Li, Z. Hou, C. Chen, Z. Hao, Y. An, S. Liang, and B. Lu, "Automatic cardiothoracic ratio calculation with deep learning," *IEEE Access*, vol. 7, pp. 37749–37756, 2019.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[5] T. Gupte, M. Niljikar, M. Gawali, V. Kulkarni, A. Kharat, and A. Pant, "Deep learning models for calculation of cardiothoracic ratio from chest radiographs for assisted diagnosis of cardiomegaly," 2021.

[6] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention u-net: Learning where to look for the pancreas," 2018.

[7] B. Baheti, S. Innani, S. Gajre, and S. Talbar, "Eff-unet: A novel architecture for semantic segmentation in unstructured environment," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1473–1481, 2020.

[8] S. Jaeger, S. Candemir, S. Antani, Y. X. Wáng, P. X. Lu, and G. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quant Imaging Med Surg*, vol. 4, pp. 475–477, Dec 2014.

[9] P. Rajpurkar, A. Joshi, A. Pareek, P. Chen, A. Kiani, J. Irvin, A. Y. Ng, and M. P. Lungren, "Chexpedition: Investigating generalization challenges for translation of chest x-ray algorithms to the clinical setting," 2020.

[10] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren, "Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert," 2020.

[11] S. Jain, A. Smit, S. Q. Truong, C. D. Nguyen, M.-T. Huynh, M. Jain, V. A. Young, A. Y. Ng, M. P. Lungren, and P. Rajpurkar, "Visualchexbert: Addressing the discrepancy between radiology report labels and image labels," *arXiv preprint arXiv:2102.11467*, 2021.

[12] S. Jain, A. Smit, A. Y. Ng, and P. Rajpurkar, "Effect of radiology report labeler quality on deep learning models for chest x-ray interpretation," 2021.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015.

[14] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440–1448, 2015.

[15] W. Ye, J. Yao, H. Xue, and Y. Li, "Weakly supervised lesion localization with probabilistic-cam pooling," 2020.

[16] H. H. Pham, T. T. Le, D. Q. Tran, D. T. Ngo, and H. Q. Nguyen, "Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels," 2020.

[17] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," *Nature Machine Intelligence*, May 2021.

A Improvement of prediction distribution after finetuning

To evaluate the improvement after fine-tuning, we plotted the distribution of the predictions split by ground-truth class. We observe that previously to fine-tuning the model was biased to predict higher CTR: this is because the training dataset for the heart detection only contained bounding boxes for unhealthy/abnormally big heart. After fine-tuning the standard deviation of each distribution is lower and greatly improves the bias towards higher CTR.

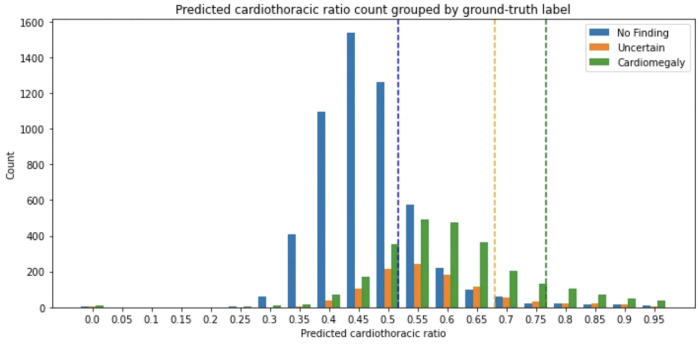

Figure 4: Distribution of predictions before fine-tuning. The dotted lines indicate the mean of each distribution

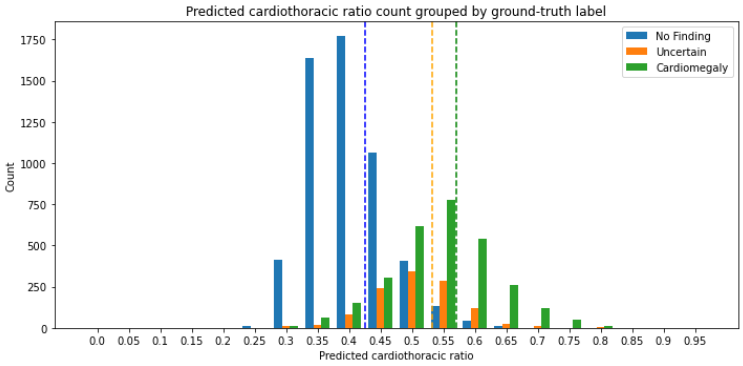

Figure 5: Distribution of predictions after fine-tuning. The dotted lines indicate the mean of each distribution

B Analysis of model robustness

To study the model robustness we plot the ROC curve for different slices of data. In figure 6 we compare the fine-tuned model to the baseline model.

In figure 7 we plot the ROC curve slicing the dataset by sex, we observe that the model accuracy doesn't change for the slices.

In figure 8 we slice the dataset by an extra label from the CheXpert dataset that indicates if the patient has any type of device connected (eg: pacemaker). We observe that the model is robust to the presence of support devices. There is a third category, which indicates that the automatic labeler is uncertain about the presence of support devices. In this case the model performance takes a hit; inspecting those samples we observe that most of the patients have been labeled as 'Uncertain cardiomegaly' and the model predicts CTR ratios close to 0.5.

In figure 9 we slice the dataset by the number of observations. We refer to each distinct illness or uncertainty of illness as an observation and add them up for each sample (Healthy patients also appear as one observation). For observation numbers higher than 2 we report the precision as the model didn't have any "False positive".

- Precision 1 observation: 0.879. Number of samples: 4,577
- Precision 2 observations: 0.816. Number of samples: 2,391

- Precision 3 observations: 0.762. Number of samples: 933
- Precision 4 observations: 0.817. Number of samples: 802
- Precision more than 5 observations: 0.861. Number of samples: 303

We observe the precision drops as more observations the patient has. The precision at more than 5 observations is similar to the precision at 1 observation but it is important to consider it is calculated with much lower amount of samples.

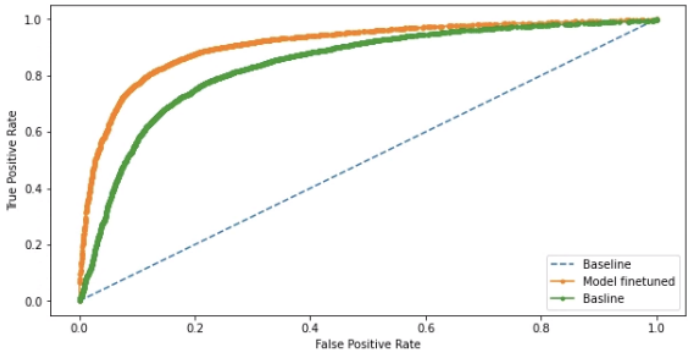

Figure 6: ROC for the baseline and fine-tuned models. Model fine-tuned AUC: 0.909. Baseline: 0.842

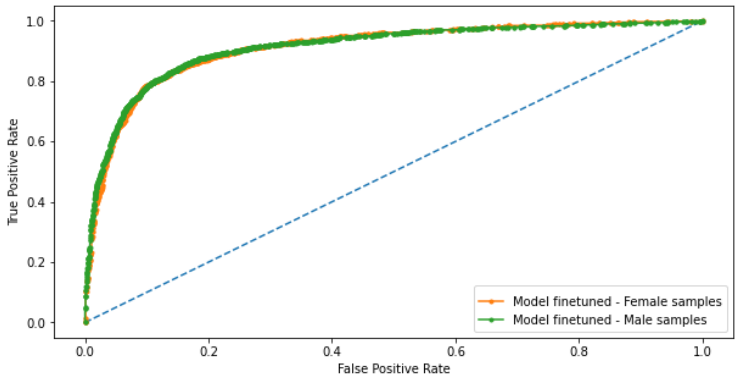

Figure 7: ROC plot slicing by sex. Model fine-tuned - Female samples: 0.909. Model fine-tuned - Male samples: 0.909

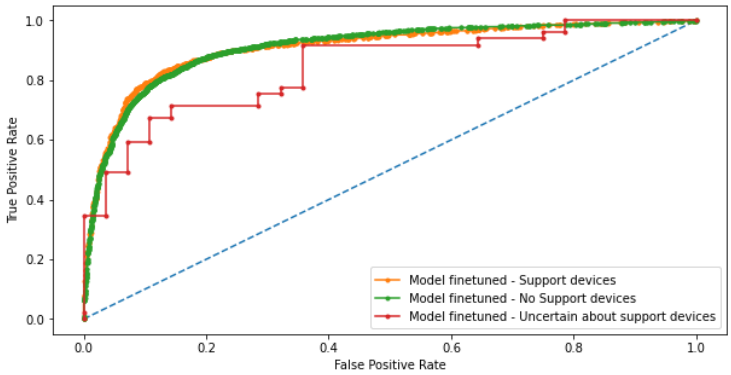

Figure 8: ROC plot slicing by support devices

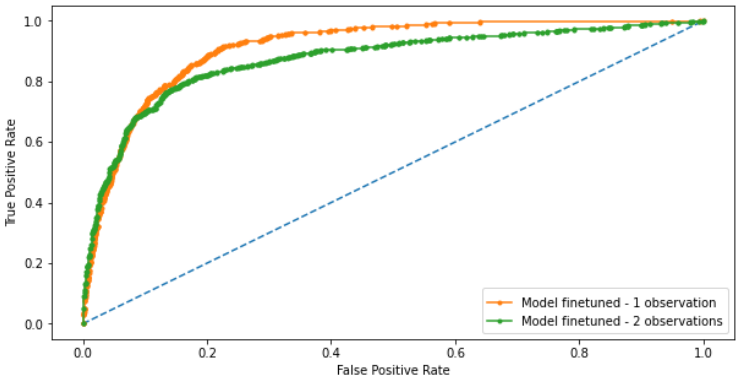

Figure 9: ROC plot slicing by amount of observations. Model fine-tuned - 1 observation AUC: 0.909. Model fine-tuned - 2 observations AUC: 0.872

C Fine-tuning motivation

On figure 10 we show a random sample of images from the Schenzen and Montgomery datasets. We can observe that all the images contain the chest at the same zoom level. There is a difference in x-ray penetration (a configuration of the x-ray image) which makes the lungs change in opacity. There are no zoomed-out images that show part of the head or arms.

On figure 11 we show a random sample of the VinBigData dataset used to train the heart detection. In this case, we see the images have more variation that the the Schenzen and Montgomery datasets.

On figure 12 we show a random sample from CheXpert. This is the dataset with the most variations on how the x-rays are taken. This increase in variations degrades the performance on CheXpert when validating our approach and motivates the finetuning of our models on the images we labeled.

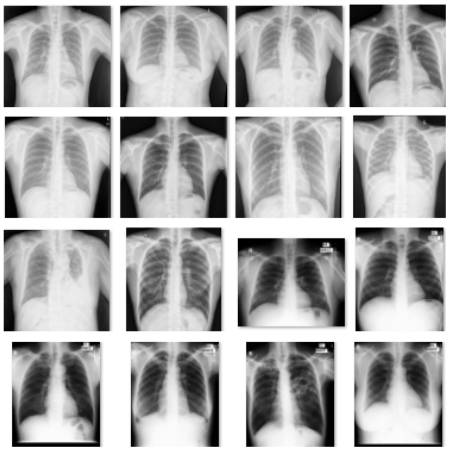

Figure 10: Randomly sampled images from Schenzen and Montgomery datasets

D Sampling and analysis of errors

We sample some of the predictions to visualize at what part of the pipeline the model is failing. On figures 13 15 14 we have 3 instances of samples that have a CTR higher than 0.5 that is classified as "No finding" in the CheXpert dataset. We believe this are errors in the CheXpert labels and not errors of the models.

On figures 16 17 we see the model predicts a bounding box of the incorrect width and fails to identify the right tip of the heart correctly.

Finally in figure 18 we observe an outlier, a sample where the model completly fails to recognize the heart.

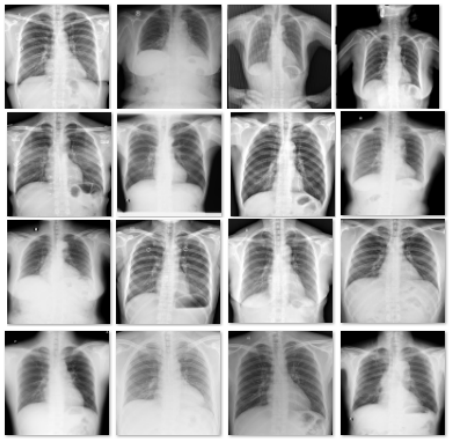

Figure 11: Randomly sampled VinBigData images

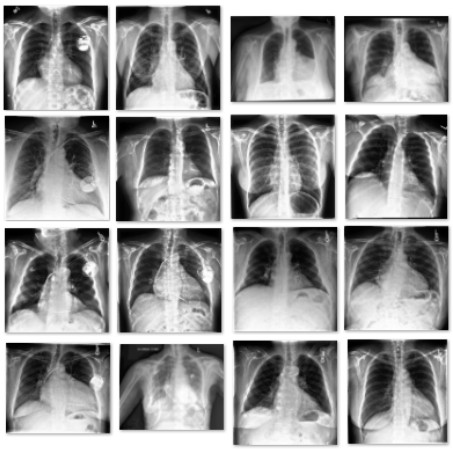

Figure 12: Randomly sampled CheXpert images

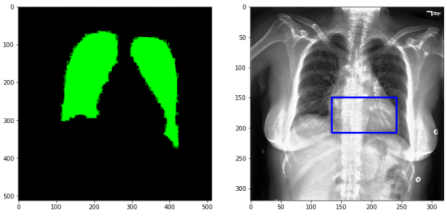

Figure 13: Predicted CTR 0.55. CheXpert label: No finding

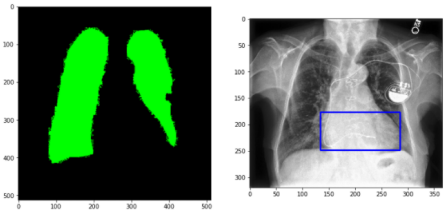

Figure 14: Predicted CTR 0.61. CheXpert label: No finding

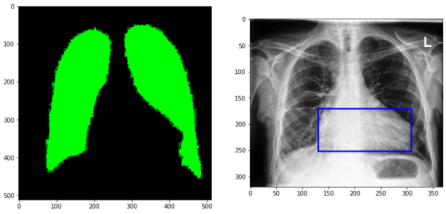

Figure 15: Predicted CTR 0.59. CheXpert label: No finding

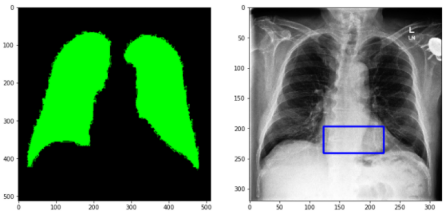

Figure 16: Predicted CTR 0.35. CheXpert label: Cardiomegaly

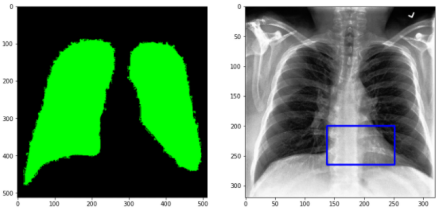

Figure 17: Predicted CTR 0.38. CheXpert label: Cardiomegaly

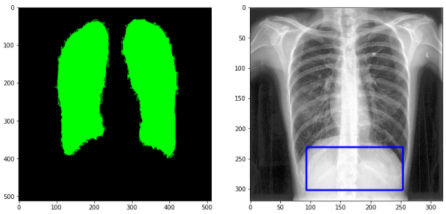

Figure 18: Model fails to detect the heart, detects the diaphragm

E Hyper parameter tuning

Hyper parameters were tuned by checking the validation metric each epoch and saving the models with the best score, recording both the score of the model and the epoch number. There were two main hyper parameters to be tuned for each model. The learning rate λ_1 for the initial pre-training and the learning rate λ_2 for the finetuning step. It was important to pick λ_2 such that the model takes advantage of the learned information in the pre-training to carry out the fine-tuning task. Picking a wrong second learning rate wastes the knowledge learned in the pre-training. This was achieved by doing multiple runs of the training procedure with different learning rates until the best combination of epochs and learning rates was found.