RadBot-CXR: Classification of Four Clinical Finding Categories in Chest X-Ray Using Deep Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Medical imaging faces a global demand vs supply crisis. 2B people will join the middle class in the next decade. There is inequity in care across geographies (e.g. in Nigeria, much less than 2 radiologists per 1M population [3]). Imaging data has exploded 4x in the last decade. Patients suffer from physician mistakes. Among the imaging data, a few dozens of percents are X-ray, e.g. 54% in England [4].

Our approach demonstrates another step in the challenging path to include automation in the medical care process. Using a convolutional neural network (CNN) we achieved a performance which is slightly higher than radiologists in detecting four clinical finding categories in chest X-ray (CXR). The agreement of RadBot-CXR vs. radiologists is slightly higher (1-7%) than the agreement among a team of radiologists. Each image in the validation set was tagged by three radiologists.

The four finding categories are:

- General Lung Opacity: [Consolidation/ Atelectasis/ Pleural effusion/ Pulmonary mass]
- Cardiomegaly
- Hilar prominence
- Pulmonary edema

We demonstrate an automatic RadBot-CXR for detecting four clinical finding categories in CXR. The system may improve the medical care people get by:

- Shortening the time to care beginning, by detecting patients that have top priority findings.
- Improving TP in geographies that have a major deficiency in radiologist manpower.
- Reducing the FN
- Ensuring a constant level of performance, while reducing the the dependency in quality and level of physicians performance.
- Ensuring reproducible and systematic way for clinical finding detection.

Keywords CNN, Deep Learning, Medical Imaging, Algorithms, Computer Aided Diagnosis, CAD, Chest X-ray, CXR

1 Introduction

Chest X-rays (CXR's) are the most commonly performed radiology examination world-wide, with over 150 million obtained annually in the United States alone. CXR's are a cornerstone of acute triage as well as longitudinal surveillance. Despite the ubiquity of the exam and its apparent technical

Submitted to MIDL 2018. Do not distribute.

simplicity, the chest x ray is widely regarded among radiologists as among the most difficult to master[15].

Due to a shortage in supply of radiologists, radiographic technicians are increasingly called upon to provide preliminary interpretations, particularly in Europe and Africa. In the US, non-radiology physicians often provide preliminary or definitive readings of CXRs, decreasing the waiting interval at the nontrivial expense of diagnostic accuracy.

Even among expert radiologists, clinically substantial errors are made in 3-6% of studies[15, 1], with minor errors seen in 30% [2]. Accurate diagnosis of some entities is particularly challenging: early lung cancer for example is missed in 19-54% of cases, with similar sensitivity figures described for pneumothorax and rib fracture detection. The likelihood for major diagnostic errors is directly correlated with both shift length and volume of examinations being read[6], a reminder that diagnostic accuracy varies substantially even at different times of the day for a given radiologist.

Hence there exists an immense unmet need and opportunity to provide immediate, consistent and expert-level insight into every CXR. In the present work we describe a novel methodology employed in this endeavor and we present the results achieved using a robust method of clinical validation.

In order to get a broad mapping of the clinical status based on a CXR we detect four finding categories. The lungs status is tested by detecting two types of abnormality categories, both focal and diffuse. The focal category, general lung opacity, is comprised of consolidation, atelectasis, pleural effusion and pulmonary mass. The diffuse category relates to pulmonary edema. In addition to the two lung categories, the third abnormality category is cardiomegaly which maps the heart status. At last, the fourth abnormality category is hilar prominence which reflects the status of the mediastinum.

1.1 RadBot-CXR

Our approach, RadBot-CXR (Figure 1) utilizes a CNN to find the four clinical meta-findings mentioned above. Comparing the performance of RadBot-CXR to the a team of radiologists shows that the agreement of RadBot-CXR vs. radiologists is slightly better than the agreement among the radiologists team.

2 Previous Work

Automatic image analysis has been explored in the last few decades both using computer vision and machine learning approaches. On 1998 Lecun et al. [10] demonstrated the use of LeNet 5 for character recognition with a seven layer network. Later, on 2012 Krizhevsky et al. [9] introduced AlexNet, a convolutional neural network (CNN), and showed that it outperforms previous models for the task of classifying natural images. A more deeper networks were introduced later, such as Inception-V3 by Szegedy et al. [19]. Using all of the features from previous layers was introduced by Huang et al. in their DenseNet model [8].

CNNs have been also used for object detection and localization by Girshick et al.[5] and Ren et al. [14]. In addition, CNNs were later used also for segmentation by Long et al. [12].

Medical image analysis also benefited from CNNs for classification, detection, segmentation and registration. A review on the field was done by Shen et al. [16]. Classification of a CT slice for predicting diseases was demonstrated by Shin et at. [17, 18]. Latent Dirichlet allocation (LDA) was used to generate disease labels to key slices based on textual reports. The disease labels are later used to train CNNs, both AlexNet and VGG 16 & 19, to predict a disease based on an input CT slice.

Classification and detection on chest X-ray images was demonstrated by Wang et al. [20]. Disease labels for eight clinical findings were generated by a NLP approach. A CNN is then trained to classify and localize the clinical findings. Four different architectures of a CNN were tested. A recent work by Rajpurkar et at. [13] demonstrate the use of DenseNet for the classification of 14 clinical findings. For one of the finding, pneumonia, the performance of the model is also compared to radiologists. Another recent work by Li et al. [11] demonstrates the use also of localization ground truth labels to predict both existence of a clinical finding and also its location. The localization labels are supplied only for part of the data. Nevertheless, the work demonstrates an improve also in the classification performance.

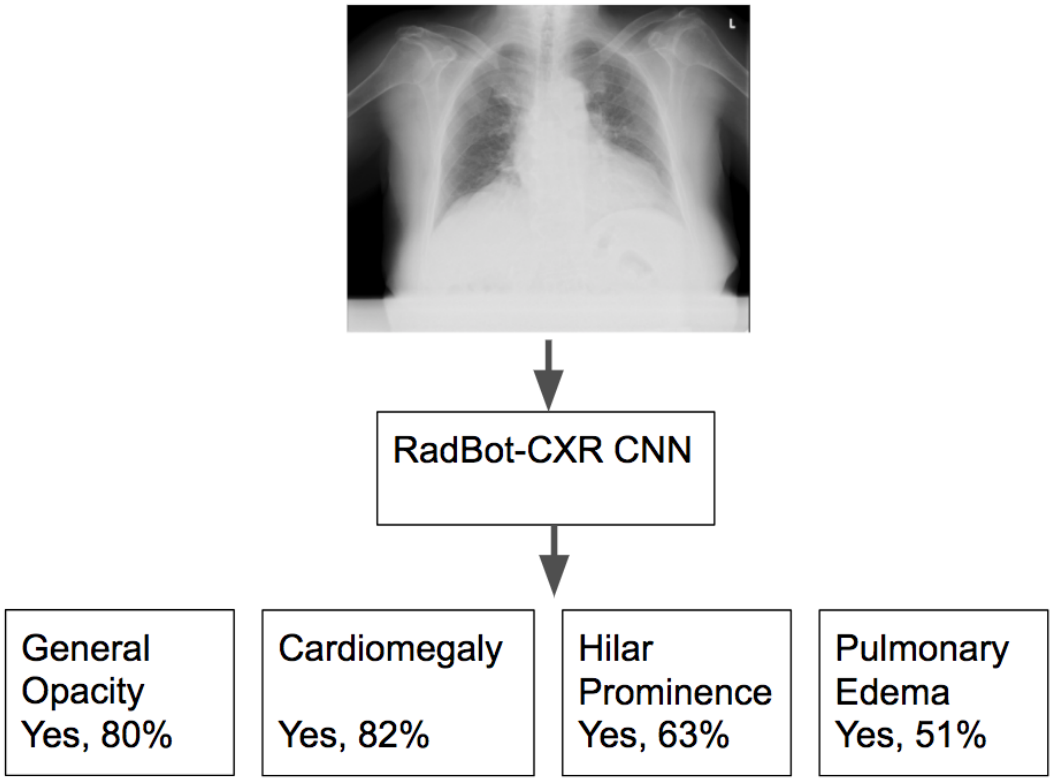


Figure 1: Input chest X-ray image is fed to RadBot-CXR CNN. The output reports the existence/non-existence for each of the meta-findings: general opacity, cardiomegaly, hilar prominence and pulmonary edema.

3 RadBot-CXR

Our model is an inception-v3 variant CNN of 27 layers.

The input is a frontal (posterior-anterior, PA) CXR image, 8-12 bits, of size at most 8.6 [MP] (3000*3000). Their original resolution is in the range (0.1,0.2) [mm/pix]. We first resized all images to 1024x1024 [pixels] and saved them in a local cache. We apply the following preprocessing. First, the raw pixel values are updated according to the *RescaleSlope, RescaleIntercept, PhotometricInterpretation* dicom tags. Next, the pixel values are stretched to the range (0,1). An algorithm to remove black frame from the image is then applied.

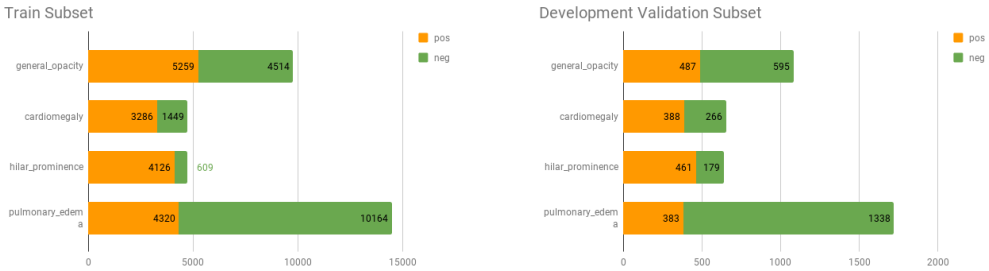
The input size of the network is 559*469 [pixels]. This is chosen to optimize the aspect ratio according to aspect ratio distribution in the DB. An interpolation is used to convert from size 1024*1024 [pixels] to 559*469 [pixels].

The output of the network are multiple sigmoids. For each of the findings there exist one or three sigmoids. For unilateral finding such as cardiomegaly there is a single sigmoid. For bilateral findings, e.g. pulmonary edema, there exist three sigmoids. One for the existence of the finding in the whole image, and additional two sigmoids for the existence on of the finding on the left or the right sides.

4 Dataset

4.1 CXR Images

We have a dataset that contains 1.5M frontal (PA) CXR studies. Each study has also a report that was written by a radiologist in the medical center. For the development a dataset is built as follows. Our goal is to have a few thousands of examples for each of the clinical findings, both positives and negatives. We use textual search on the reports to pick a few thousands of studies that are positive.



(a) *Train* subset (b) *Development validation* subset

Figure 2: *Train* & *development validation* subset samples.

This is done for each of the findings. Next, a random sample from the population is picked to add a few thousand of studies that most of them are negatives. The negatives are picked with gender and age prevalence according to the prevalence in the positives.

We use in the dataset only adult images from age 18 and above. Our original database has a tendency to the elder age, with mean 56.9 [y].

The dataset is then split to three subsets, train (80%), validation (10%) & test (10%). The anonymized patient id is used to associate each study with one of the subsets, to ensure that all studies of any specific patient reside in the same subset.

4.2 Image Tagging

A team of six radiologists analyzed the dataset images and tagged them. We use a web-based tagging application where each image is shown together with a set of questions.

The first question refers to the relevancy of the image. An image is defined as relevant if it is an upright PA chest x-ray.

The other questions refer to the various findings in the image. Each question refers to a single finding and asks whether the image lacks the finding or if it exists in the image. If it exists, in some of the findings a few options regarding the severeness and state of the finding, are given. Moreover, in bilateral findings, e.g. general opacity, the side, left and/or right is also tagged.

Each of the images in the train subset is tagged by a single radiologist, whereas, in the validation subset each image is tagged by three radiologists. This is done to increase the confidence in the tagging by having multiple votes for each image. The radiologists do not see the tagging done by their peers.

In the tagging we found that 7% of the images were tagged as non-relevant (an image is considered non-relevant if at least one of the radiologists tagged it as non-relevant).

The findings may have a partial correlation, i.e. an image may have more than one finding. Clinicaly we know that there exist partial correlation between the findings. In addition, an image that serves as a positive example for one finding may serve also as a negative example for other finding.

Figure 2 summarizes the amount of samples we have for each of the findings. The *train* subset has 4.7K-14.5K samples for each finding. Among them 30-87% are positive samples, and 13-70% are negative samples. The *development validation* subset has 0.6K-1.7K samples for each finding. Among them 22-72% are positive samples, and 28-78% are negative samples. Another subset, *internal validation*, will be described in the next section.

5 Experiments

5.1 training

The tagged samples are used to train a model using the following procedure. A random batch of 32 images is chosen. Random augmentation is applied on each image separately. Types of augmentations: brightness, contrast, crop. Horizontal and vertical flips are excluded due to the no-symmetric nature of the chest. Slight rotations were not tested.

The train procedure continues by iteratively choosing a random batch, and updating the weights by the rmsprops algorithm [7]. Every 1000 iterations the performance of the model is evaluated using the following metrics. The evaluation is done for each of the four findings separately.

One metric is agreement. We have three tagging/votes for each image in the *development validation* subset. We define pair agreement as follows. Given a set of taggings by two taggers, their pair agreement is the fraction of cases the two taggers agreed. For example:

- 100 images were tagged
- 100 were tagged as relevant by both taggers
- 40 were tagged as positive by both taggers
- 50 were tagged as negative by both taggers
- Pair Agreement = (40 + 50)/100 = 0.90

We further define *cxr7_algo__rads* agreement as the average pair agreements of cxr7_algo and a radiologist. For example:

- cxr7_algo__rad1 agreement: 0.80
- cxr7_algo__rad2 agreement: 0.90
- cxr7_algo__rad3 agreement: 1.00
- Average: 0.90

In a similar manner we also define inter_rads agreement as the average pair agreements of radiologist *i* and a radiologist *j*.

Another metric is sensitivity & specificity. We measure the point of equal sensitivity & specificity.

We conducted a few hundreds of training sessions. Various variants of architectures were tested.

Running on Xeon E5-2650 v2 machine with 8 Tesla K80 gpu's we have a training rate of 62 [images/sec]. A training session has 60k iterations, which result in 8.6 [h].

5.2 Choosing a Model & Model Calibration

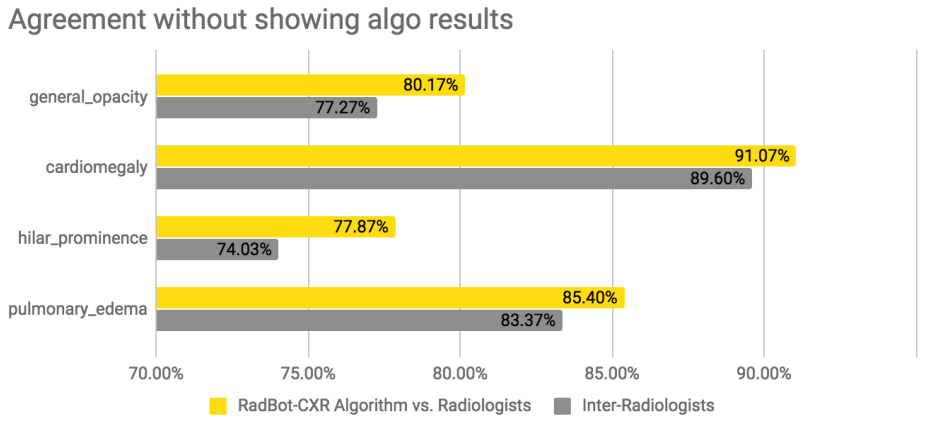
Considering a few hundreds of training sessions with a few dozens of models for each (measured every 1000 iterations), result in a few thousands of putative models. For each of the four findings we choose an optimal model according to the metric of equal point sensitivity & specificity.

We then calibrate each of the chosen models separately for each findings. An expert radiologist used both an ROC curve and a scores bar display to calibrate the threshold for each finding. The scores bar display shows the images (with their scores) sorted by their scores.

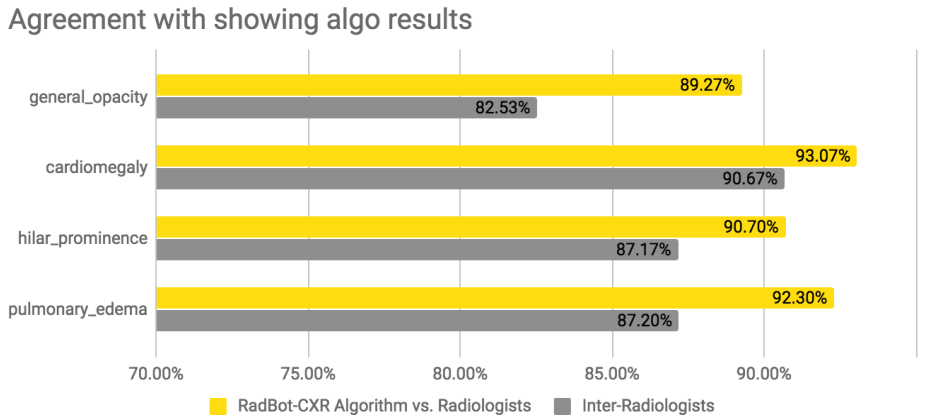
5.3 RadBot-CXR Performance

The optimal four models are chosen using the agreement metric while measuring on the *development validation* subset. In order to measure the performance of the chosen models and verify their generalization we conducted the following experiment.

We choose a random set of 20k images from the general population in our DB that do not belong to the train subset. We then run the algorithm on this set. Inference running time on Xeon E5-2695 v4 machine is 330 [msec/image] when running on 4 GTX 1080Ti gpu's or 1 [sec/image] when running on a cpu. We then build an *internal validation* subset in the following manner.



(a) Algorithm-radiologists agreement compared to the inter-radiologists agreement when the algorithm results are not shown. The agreement of the algorithm-radiologists is higher than the inter-radiologists by 1-4%.



(b) Algorithm-radiologists agreement compared to the inter-radiologists agreement when the algorithm results are shown. The agreement of the algorithm-radiologists is higher than the inter-radiologists by 2-7%.

Figure 3: Agreement measurements.

Internal validation, Part I

- Choose randomly 126 images wherein the algorithm reports no findings
- For each finding, choose randomly 31 images wherein the algorithm reported the finding
- Total: 126 + 4*31 = 250 images

Part II is build in a similar manner, while excluding the images already chosen in part I.

The tagging of the *internal validation* was done by three *expert* radiologists. Part I is used to measure the agreement when the radiologists do not know the algorithm result. In addition, part II is used to measure the agreement when the radiologists do know the algorithm result.

Comparison of the algorithm-radiologists agreement compared to the inter-radiologists agreement is shown in Figure 3a for part I. The inter-radiologists agreement is 74-90%, while the algorithm-radiologists agreement is 78-91%. The agreement of the algorithm-radiologists is higher than the inter-radiologists by 1-4%.

We further test the algorithm performance on part II while showing the radiologists the algorithm results. Comparison of the algorithm-radiologists agreement compared to the inter-radiologists

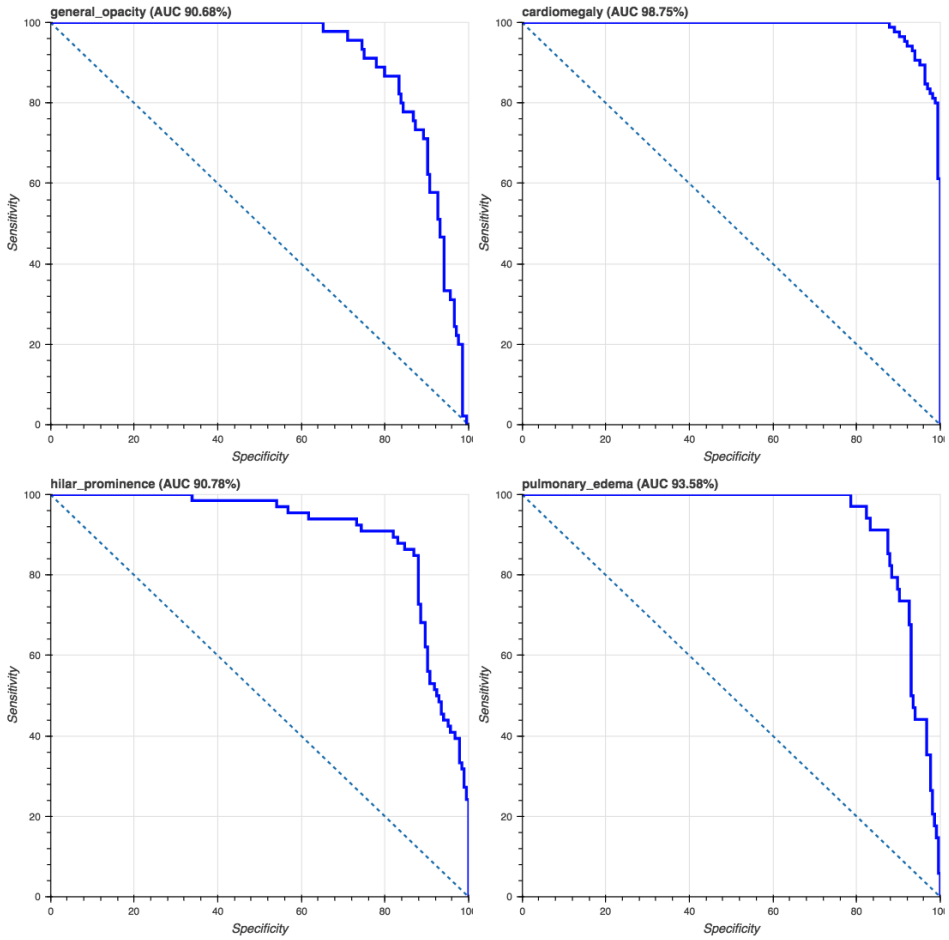


Figure 4: ROC measured on the *internal validation* for each of the meta-findings: general opacity, cardiomegaly, hilar prominence and pulmonary edema.

agreement is shown in Figure 3b for part II. The inter-radiologists agreement is 83-91%, while the algorithm-radiologists agreement is 89-93%. The agreement of the algorithm-radiologists is higher than the inter-radiologists by 2-7%.

We compare the performance measured on the two sets, part I and part II. When showing the algorithm results both the algorithm-radiologists & the inter-radiologists agreements increase. The algorithm-radiologists agreement increases from 78-91% to 89-93%. Similarly, the inter-radiologists agreement increases from 74-90% to 83-91%. The inter-radiologists agreement increases by 1-13%, where hilar prominence agreement has the maximal increase by 13%. The algorithm-radiologists agreement increases by 2-13%, where again the hilar prominence agreement has the maximal increase by 13%. General opacity and pulmonary edema agreements increase by 9% and 7% respectively.

The agreement is a metric that is more appropriate than ROC to predict the value of such an algorithm in the field. Nevertheless, to compare ourselves to previous works that report AUC we demonstrate the performance also using this metric. Figure 4 shows the ROC graphs for the four finding categories, with AUC's 90.7-98.8%. As aground truth label we use the majority among the three radiologists tagging.

We compare our AUC's to Li et al. [11] in Table 1. It is noted that our AUC's are measured on a different dataset since a preliminary assessment on a single finding showed a low agreement between our radiologists team and the NIH labels. Regarding cardiomegaly and pulmonary edema our AUC's (98.8% and 93.6%) are higher by 11% and 5% respectability. Li et al. do not detect hilar prominence, while our AUC is 90.8%. Finally, our general opacity met-findings AUC is 90.7%. Li et al. detect

Table 1: AUC's measurements with rough comparison to Li et al. [11], measured on different datasets. Regarding cardiomegaly and pulmonary edema our AUC's are higher by 11% and 5% respectability

finding	**RadBot-CXR (ours) AUC [%]**	**Li et at. AUC [%]**
Cardiomegaly	98.8	87.4
Pulmonary Edema	93.6	88.2
Hilar Prominence	90.8	—
General Opacity	90.7	ATL 79.6, CNS 79.5, EFF 86.7, MSS 83.1

it as multiple findings with AUC's 79.5-86.7%. It is reasonable to assume their AUC for detecting general opacity meta-finding would be higher than 86.7%.

6 Future Work

Our research could be continued in a few interesting directions. One approach is to predict also the location of the finding in addition to the finding classification task. This approach seems more interesting for general opacity, while for cardiomegaly and hilar prominence it's less valuable since their location is known to the radiologists.

Another approach is to incorporate segmentation of specific organs to improve the classification task, either by accelerating the training and/or by improving the performance. This is expected on behalf of the upgraded information the net gets.

An additional interesting approach is to cover additional finding categories. There are a few dozens of findings that can inferred from a single frontal X-ray.

7 Conclusion

There is a big gap between the needs for medical care and the current availability. 2B people will join the middle class in the next decade. There is inequity in care across geographies (e.g. in Nigeria, much less than 2 radiologists per 1M population [3]. Medical imaging data has exploded 4x in the last decade. Patients suffer from physician mistakes. Among the imaging data, a few dozens of percents are X-ray, e.g. 54% in England [4].

RadBot-CXR was trained with thousands of images tagged by radiologists. It is trained to detect four (clinical) finding categories:

- General Lung Opacity: [Consolidation/Atelectasis/Pleural effusion/Pulmonary mass]
- Cardiomegaly
- Hilar Prominence
- Pulmonary Edema (interstitial/alveolar)

Detecting four (clinical) findings categories gives a broad mapping of the clinical status.

The performance of the algorithm was measured by expert radiologists tagging. The algorithm-radiologists agreement is slightly higher than the inter-radiologists agreement by 1-7%. When showing the algorithm results the algorithm-radiologists agreement increases from 78-91% to 89-93%.

Our approach demonstrate another step in the challenging path to include automation in the medical care process. Using a CNN we achieved a performance which slightly higher than radiologists, in detecting four clinical finding categories in chest X-ray (CXR).

References

[1] Adrian Brady, Risteárd Ó Laoide, Peter McCarthy, and Ronan McDermott. Discrepancy and error in radiology: concepts, causes and consequences. *The Ulster medical journal*, 81(1):3–9, 1 2012.

[2] Michael A. Bruno, Eric A. Walker, and Hani H. Abujudeh. Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction. *RadioGraphics*, 35(6):1668–1676, 10 2015.

[3] E U Ekpo, N O Egbe, and B E Akpan. Radiographers' performance in chest X-ray interpretation: the Nigerian experience. *The British Journal of Radiology*, 88(1051):20150023, 7 2015.

[4] NHS England. Diagnostic imaging dataset: Standardised ccg rates 2016/17, 2017.

[5] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[6] Tarek N. Hanna, Christine Lamoureux, Elizabeth A. Krupinski, Scott Weber, and Jamlik-Omari Johnson. Effect of Shift, Schedule, and Volume on Interpretive Accuracy: A Retrospective Analysis of 2.9 Million Radiologic Examinations. *Radiology*, page 170555, 11 2017.

[7] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent, 2012.

[8] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[10] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[11] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Fei-Fei Li. Thoracic disease identification and localization with limited supervision. *arXiv preprint arXiv:1711.06373*, 2017.

[12] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[13] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 11 2017.

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[15] P J Robinson, D Wilson, A Coral, A Murphy, and P Verow. Variation between experienced observers in the interpretation of accident and emergency radiographs. *The British journal of radiology*, 72(856):323–30, 4 1999.

[16] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

[17] Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, and Ronald M. Summers. Interleaved text/image Deep Mining on a large-scale radiology database. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1090–1099. IEEE, 6 2015.

[18] Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Seff, Jianhua Yao, and Ronald M. Summers. Interleaved Text/Image Deep Mining on a Large-Scale Radiology Database for Automated Image Interpretation. *Journal of Machine Learning Research*, 17(107):1–31, 2016.

[19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[20] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471. IEEE, 2017.