Deep Learning for Grading Cardiomegaly Severity in Chest X-rays: An Investigation

Sema Candemir, Sivaramakrishnan Rajaraman, George Thoma, Sameer Antani

Lister Hill National Center for Biomedical Communications U.S. National Library of Medicine, NIH, Bethesda, MD, USA

sema.candemir@nih.gov, sivaramakrishnan.rajaraman@nih.gov, george.thoma@nih.gov, sameer.antani@nih.gov

Abstract—This study investigates using deep convolutional neural networks (CNN) for automatic detection of cardiomegaly in digital chest X-rays (CXRs). First, we employ and fine-tune several deep CNN architectures to detect presence of cardiomegaly in CXRs. Next, we introduce a CXR-based pre-trained model where we first fully train an architecture with a very large CXR dataset and then fine-tune the system with cardiomegaly CXRs. Finally, we investigate the correlation between softmax probability of an architecture and the severity of the disease. We use two publicly available datasets, NLM-Indiana Collection and NIH-CXR datasets. Based on our preliminary results (i) data-driven approach produces better results than prior rule-based approaches developed for cardiomegaly detection, (ii) our preliminary experiment with alternative pre-trained model is promising, and (iii) the system is more confident if severity increases.

I. Introduction

Chest radiography is commonly used in early diagnosis to detect lung and heart pathologies such as atelectasis, consolidation, pneumothorax, pleural effusion, cardiac hyperinflation [1] and is a primary tool for mass screening [2][3][4]. It is accessible, inexpensive and dose-effective compared to other imaging tools. The literature has many studies for automated analysis of chest X-rays (CXR) in radiological analysis or for population screening in endemic locations [2][3][5][4][6]. One of these is computation of radiographic indexes as indicator for cardiac diseases such as cardiomegaly.

Cardiomegaly is a medical condition in which the heart is enlarged. Figure 1 shows example CXRs from National Library of Medicine (NLM) Indiana Collection (c.f. Section IV-A). Left image is of a healthy subject; right image shows a heart that is moderate to severely enlarged. Although, 2D-echocardiography is considered as gold standard for the diagnosis of cardiomegaly [7], access to these services may not be widely available in under-resourced regions. Automated approaches could lead to use in a CXR-based triage tool.

With advances in GPU technology, computer vision systems designed with deep neural networks trained on massive amount of data have been shown to produce more accurate results than conventional approaches. Especially, convolutional neural networks (CNN) have received considerable attention for image analysis, since they preserve the spatial relationship between image pixels. In this study, we use CNN models for automatic detection of cardiomegaly in CXRs. We first apply pre-trained CNN and fine tune the models with cardiomegaly CXRs. In the second part, we introduce a CXR-based pre-trained
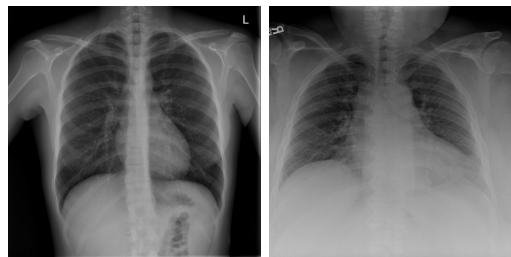

Fig. 1. Two example CXRs from NLM - Indiana Collection. Left image belongs to a healthy patient. The clinical findings for the right image is that the heart size is moderate to severely enlarged.

model. We fully train a deep CNN architecture (e.g. VGG-16) with a very large CXR dataset and fine-tune the model with cardiomegaly CXRs. The difference from the earlier fine-tuning approach is training the architecture only with CXRs instead of the ImageNet dataset. We expect better classification performance, since earlier layers of the system is trained with CXR images. In the last part, we examine correlation between the softmax probability and severity of the disease. For the experiments, we used two large publicly available datasets: NLM-Indiana Collection and National Institutes of Health (NIH)-CXR dataset [8]. The rest of the paper is organized as follows. Section II presents a brief overview of cardiomegaly detection methods developed for CXRs. In Section III, our contribution and methodology is presented, followed by results in Section IV-D. Finally, Section V concludes this study.

II. Background

Rule-based Approaches: A well-known radiographic index is the cardiothoracic ratio (CTR) which is defined as the ratio between the maximum transverse cardiac diameter and the maximum thoracic diameter measured between the inner margins of ribs [9]. There is no consensus for optimal CTR for cardiomegaly [10][11]. Generally, CTR higher than 0.5 is considered as a sign of heart enlargement [12][13]. One of the earlier methods for automated cardiomegaly detection is presented in [14]. The authors measured the maximum diameter of heart shadow and the maximum diameter of the rib-cage shadow from CXRs with vertical intensity histogram analysis. In [15], researchers compute CTR by fitting a Fourier shape to hearth boundary profiles. In [16] [17], the CTR computation is used as a clinical application of lung boundary detection algorithm. Although, traditional CTR is widely accepted as

109

a standard index for heart analysis in CXRs, literature has alternative radiographic indexes. In [18], 2D-CTR is proposed as the ratio between the pixel counts of the cardiac outline and the whole thorax. In [19], CTAR is defined as the ratio of the area of heart region to the area of lung region. The comparison of radiographic indexes are investigated in [20].

Machine Learning Approaches: To our knowledge, there is only one study that investigated the performance of a traditional machine learning approach by training a support vector machine with radiographic indexes [20]. However, this approach uses limited image feature (only radiographic indexes) without considering other image characteristic in CXRs using a general image descriptor [21][22]. With the recent advancement in artificial intelligence, researchers developed deep learning based algorithms for medical image analysis [23][24]. However, the studies which focus on cardiomegaly detection are limited. In [25], several pre-trained models are tested on NLM-Indiana dataset, however, the system is trained only with approximately 250 CXRs. Although researchers compare their system with earlier rule-based approaches, they did not comment on reasons why deep learning architectures tend to perform better than rule-based systems. The other reported scores for cardiomegaly detection are in [8] where researchers trained a multi-labeled deep CNN architecture and reported detection scores for 8 pathologies in NIH-CXR dataset including cardiomegaly. In [26], researchers trained 121-layer CNN with NIH-CXR dataset for pneumonia detection, and tested dense-net for 14 pathologies including cardiomegaly. However, pathology labels in NIH-CXR dataset are text-mined from radiologist reports and contain errors. Therefore, system scores reported on NIH dataset are not considered reliable.

Our contributions are i) investigating pre-trained CNN model's performance by training them with the largest cardiomegaly set (combination of NIH-CXR and NLM-Indiana collection) and reporting on radiologist annotated NLM-Indiana collection; ii) discussing why data-driven approach performs better than earlier rule-based system, iii) introducing CXR-based pre-training model; and iv) investigating the correlation between softmax probability of CNN and disease severity. We anticipate that (iii) and (iv) can be applied for other diseases which can be detected on CXRs.

III. CNNs FOR CARDIOMEGALY DETECTION IN CXRs

CNN is a feed-forward artificial neural network which is composed of a set of convolutional, pooling and fully connected layers. The architecture convolves image with filters and creates stack of filtered images. These convolutional layers model image patterns at multiple levels; earlier layers learn low level features such as lines, corners, and deeper layers learn global features.

A. Fine-tuning pre-trained models

Training a deep neural network architecture requires large amount of annotated data which is generally limited in biomedical imaging domain due to need for expert annotation.

One solution for working with limited data is using pre-trained models. The pre-trained models are trained with 1.2 million general images with 1000 categories from ImageNet [27]. For our problem, low level features are used from the pre-trained model; higher level features are learned by fine-tuning the system with the limited dataset. In this section, we report the performance of pre-trained models for cardiomegaly detection. We adopt the following pre-trained models that have different depths and architecture: (i) AlexNet [28] as a relatively shallow network, (ii) VGG-16 and VGG-19 [29] as middle size networks, and (iii) InceptionV3 [30] as a very deep network. We use the layers of pretrained networks (with their learned weights) till first fully connected layer of the architecture. Then, we insert two fully connected layers, two drop-out layers among these fully connected layers in order to reduce the over-fitting, and a final output layer which produces the probability of cardiomegaly. We then fine-tune the added layers with the cardiomegaly dataset. Experimental results are in Section IV-D.

B. CXR-based pre-trained models for CXRs

As we mentioned in Section III-A, when training deep neural networks with limited data we use a pre-trained model and apply fine-tuning to the final layers with the available training data. However, all pre-trained models are trained with ImageNet which contains general images, thus have different visual features than chest X-rays. We propose a new pre-trained model for CXRs and test the model for cardiomegaly detection. Recently, NIH released a large CXR dataset [8] which contains 112,120 frontal view X-ray images of 30,805 patients. The dataset contains several lung abnormalities (c.f. Section IV-A). We apply end-to-end training to a CNN architecture (e.g. VGG-16) only with NIH-CXR dataset and then fine-tune the final layers of this model with limited cardiomegaly CXRs. Thus, earlier layers learn low level features from general abnormal CXR images, and final layers learn more specific features to cardiomegaly from cardiomegaly CXR images. The proposed idea is illustrated in Figure 2. We present the idea for cardiomegaly classification problem, however, it can be generalized to other pulmonary diseases in CXRs.
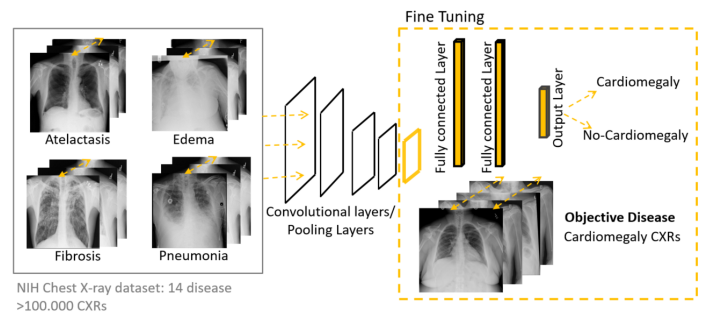

Fig. 2. The system is trained with NIH-CXR dataset which has several lung abnormalities. This pre-trained system is fine-tuned with cardiomegaly CXRs.

110

C. Automated grading of disease level

Cardiomegaly diagnosis has several levels of severity such as borderline, moderate and severe. One way to detect the disease level is training a multi-class classification system with images at different levels. However, there is not enough labeled data to train such a system. As an alternative, we investigate the correlation between softmax probability and disease level. Using radiologist report, we classify cardiomegaly images in NLM-Indiana collection into severity classes, then, we measure the softmax probability of CXR in each class. We expect higher softmax probability for severe cardiomegaly cases than mild cases.

IV. Experiment

A. Datasets

We use two largest and publicly available chest X-ray datasets: NLM-Indiana Collection and NIH-CXR dataset.

NLM-Indiana Collection is a set of images de-identified at source that are collected from various hospitals affiliated with the Indiana University School of Medicine (NIH IRB#5357). The set contains approximately 4000 frontal and lateral CXRs with several lung abnormalities and corresponding radiologist reports. Among these, 283 CXRs have cardiomegaly at various levels such as borderline, moderate, and severe, shown in Table I. The set is publicly available through Open-i® [31], which is a multimodal biomedical literature search engine developed by NLM. For experiments, we use 283 CXRs with cardiomegaly and 283 CXRs with normal cases. Using available radiology readings, we classify CXRs according to cardiomegaly severity. Each CXR is placed into one of the following categories: *borderline*, *mild*, *moderate* and *severe*. Some reports do not have severity designation, therefore, we could not classify their severity.

Severity	Borderline	Mild	Moderate	Severe	Non-classified
# of CXRs	47	107	18	4	107

TABLE I
CXR IMAGES WITH CARDIOMEGALY IN NLM INDIANA COLLECTION

NIH CXRs dataset [8] is currently the largest publicly available CXR dataset which contains 112,120 frontal-view X-ray images of 30,805 patients. The dataset contains several lung abnormalities including atelectasis, cardiomegaly, effusion, pulmonary infiltration and pneumothorax. The disease labels are text mined from the radiology reports by the data providers using natural language processing claiming 90% accuracy [8]. However they do not assign a severity grade to the disease label. Further, due to admitted text-mining errors the NIH dataset may be unsuitable for testing. Therefore, we use this set only to train CNN architectures. For our study, we extract CXRs labeled as "cardiomegaly" and obtained 2762 CXRs.

B. Evaluation

We use accuracy, sensitivity, specificity, and area under curve metrics. Accuracy is the ratio of number of correctly classified CXRs to the number of CXRs in the dataset. Sensitivity is the ratio of the correctly classified cardiomegaly patients. Specificity is the ratio of the correctly classified normal patients. Area under receiver operating characteristics curve (AUC) is the total area under the receiver operating characteristics curve which is the plot of sensitivity against 1-specificity.

C. Parameters

First, we apply histogram equalization to the images. Next, we downsize the images to the size of the input size of pre-trained model. We also apply augmentation to increase the number of training images. The organ ratios in CXRs are important features for cardiomegaly. Therefore, we carefully augment images by using only translation and rotation. We follow 3-fold cross validation approach. We use the entire NIH-CXR dataset for training. We randomly divide the NLM-Indiana collection into 3 subsets. We add one subset into training set, and the rest of the images (2 subsets) are used for testing. We repeat the process for each subset and compute the average of obtained scores. During training, we tune layers through validation set within the training set (10%). We use stochastic gradient descent as loss function. All experiments were performed using Keras framework with Tensorflow library for numerical computations in Python 3.5.

D. Experiment Results

1) Fine-tuning pre-trained models: Table II lists evaluation scores of pre-trained models to detect the presence of cardiomegaly. Our earlier study [20] investigated the radiographic index performance for cardiomegaly classification on NLM-Indiana dataset. We also listed these results for a comparison between rule-based and data-driven approaches. One reason of lower accuracy of rule-based systems is that there is no consensus for the optimal value of radiographic indexes. However, data-based approaches could learn the best discriminatory features with their optimal value in training process. For cardiomegaly case, these features would be the optimal measurements to detect the heart enlargement. The other reason is that radiographic index computation needs heart and lung boundaries which introduces additional error during boundary detection.

2) CXR-based pre-trained models: We investigate the performance of CXR-based pre-trained model. We train a VGG-16 architecture from scratch with NIH-CXR dataset. This dataset contains several lung pathologies including cardiomegaly. We then fine-tune the final convolutional layer and fully connected layers with only cardiomegaly CXRs. The cardiomegaly detection performance of standard pre-trained and CXR-based pre-trained models are tested on NLM-Indiana dataset and evaluation results are listed in Table III. We obtained higher overall accuracy and higher specificity with 500 training iterations. The gain in specificity is compensated

111

Rule-based approach Radiographic Index	Accuracy	Sensitivity	Specificity		
CTR [20]	0.736	0.872	0.60		
2D-CTR [20]	0.708	0.704	0.712		
CTAR [20]	0.756	0.856	0.656		
Conventional Machine Learning	Accuracy	Sensitivity	Specificity		
SVM Classifier [20]	0.765	0.771	0.764		
Data-Driven Approach Pre-trained Models	Accuracy	Sensitivity	Specificity	F1	AUC
Fine-tuned AlexNet	0.8764	0.8911	0.8675	0.8773	0.9436
Fine-tuned VGG-16	**0.8824**	0.9258	0.8392	**0.8873**	**0.9487**
Fine-tuned VGG-19	0.8189	**0.9487**	0.6892	0.8399	0.9178
InceptionV3	0.5627	0.1943	**0.9311**	0.3063	0.6151

TABLE II

Evaluating pre-trained models: **training set:** NIH set + 30% of Indiana Collection; **test set:** 70% of Indiana Collection.

by a small drop in sensitivity. Even at this performance, the solution is valuable for global-health applications, particularly for under-resourced regions, where specificity (finding normals as normal) is as important as sensitivity (finding abnormal as abnormal). This classifier would act like a triage tool minimizing the burden on the under-resourced radiology analysis system. However, we conjecture that with a higher number of training iterations we will see an overall improvement.

Pre-trained VGG-16	Accuracy	Sensitivity	Specificity
Trained with ImageNet	0.8824	**0.9258**	0.8392
CXR-based pre-trained model	**0.8986**	0.8881	**0.9091**

TABLE III

COMPARISON BETWEEN ARCHITECTURE PRE-TRAINED WITH IMAGE-NET AND SAME ARCHITECTURE PRE-TRAINED WITH NIH-CXR DATASET

3) Automated grading of disease level: We investigate the correlation between the severity and the softmax probability of the CNN architecture; a fine-tuned VGG-16. We measure the softmax proability of VGG-16 for each CXRs in each severity classes. The distribution of softmax values of each severity class are plotted in Figure 3. Each box represents one severity class. As expected, severe cases spread in the high end of the softmax probability with smaller variability and higher median average. Moderate and mild cases spread with larger variability towards lower end of the probability. Borderline cases have the largest distribution variability with several outliers towards lower end of the probability. The average probability of each severity class is listed in Table IV. Based on the average scores, the system confidence increases with the severity of the disease.

Severity	Severe	Moderate	Mild	Borderline
Avg. Softmax Prob.	0.9844	0.9237	0.8588	0.7701

TABLE IV

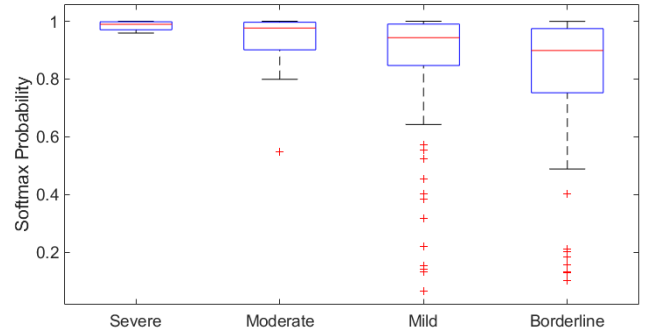AVERAGE SOFTMAX PROBABILITY FOR EACH SEVERITY CLASS

Fig. 3. Correlating softmax probability of CXRs with disease levels. Average softmax probability of each class is listed in Table IV.

V. CONCLUSIONS AND DISCUSSION

In this study, we investigated the performance of deep CNNs for automatic cardiomegaly detection. We first used pre-trained models and compared their cardiomegaly detection performance with radiographic indexes. We obtained higher accuracy with CNN models since system learns the discriminatory features and their best values from the data. In addition, radiographic index computation needs heart and lung boundaries which introduces possible segmentation errors. In data-driven approach, system learns the features from the whole image without boundary detection stage. In the second part, we introduced CXR-based pre-trained models for pulmonary classification in CXRs and tested the idea for cardiomegaly classification. We observed a significant improvement in specificity. We expect this to improve as we train further. We anticipate that CXR-based pre-trained models can be extended for other pathologies in CXRs as well as other imaging modalities (e.g. CT, MRI). As the final part of the study, we observed the correlation between softmax probability and severity of the disease. The distribution of each severity class and average probabilities show that system confidence increases with severity.

112

There are thoracic dimension and thoracoabdominal configuration differences between genders. The lung region, rib-cage dimensions and diaphragm length are smaller in females [32]. Further, thoracic dimension changes with aging. The mean transverse cardiac diameter increases gradually with age and males having slightly greater cardiac diameter than females [33]. In clinical examination of CXR, radiologist considers age and gender information in their decision making process in addition to visual clues in CXR. To our knowledge, any automated cardiomegaly detection approach in literature explicitly includes meta data to their system. However, age and gender information are implicitly contained in CXRs. The data-driven approaches could learn these differences and implicitly consider age-gender information during classification. For example, dense opacity of breast tissue and breast contour are visible differences in female CXRs. Lung size, rib-cage dimensions and diaphragm lengths are dimension-related differences between genders [32], and can be clue for an automated system to detect the gender [34]. The greater inclination of ribs in females [32] creates different textural pattern in female CXRs. In addition to lung size, lung shape differences between different age groups [35][36] and bone appearance [37] can be clue for the patient age. As a future study, we will systematically investigate these factors in addition to image-based clues.

REFERENCES

[1] J. Corne and K. Pointon, *Chest X-Ray Made Easy*, Churchill Livingstone; 3 edition, 2009.

[2] "National Library of Medicine, Chest X-ray Screening Project," http://archive.nlm.nih.gov/repos/chestImages.php/, [Accessed: 08/24/2015].

[3] S. Jaeger et al., "Automatic tuberculosis screening using chest radiographs," *IEEE Trans. on Medical Imaging*, vol. 33, pp. 233–245, 2014.

[4] S Rajaraman et al., "A novel stacked generalization of models for improved tb detection in chest radiographs," in *Proc. of IEEE Engineering in Medicine and Biology Conference*, 2018, pp. 718–721.

[5] S. Candemir et al., "Lung segmentation in chest radiographs using anatomical atlases with non-rigid registration," *IEEE Trans. on Medical Imaging*, vol. 33, no. 2, pp. 577–590, 2014.

[6] S Rajaraman et al., "Comparing deep learning models for population screening using chest radiography," in *Proc. of SPIE Medical Imaging: Computer-Aided Diagnosis*, 2018, vol. 10575.

[7] U. Sinha et al., "Comparative study of cardiac size by chest x-ray and echocardiography," *J. of the Anatomical Society of India*, vol. 62, no. 1, pp. 28 – 32, 2013.

[8] X. Wang et al., "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3462–3471.

[9] C. S. Danzer, "The cardiothoracic ratio: an index of cardiac enlargement.," *The American Journal of the Medical Sciences*, vol. 157, no. 4, pp. 513–554, 1919.

[10] DK Edwards et al., "The cardiothoracic ratio in newborn infants," *American Journal of Roentgenology*, vol. 136, no. 5, pp. 907–913, 1981.

[11] G. E. Gómez, "Importance of the relation between the anthropometric index and the transverse cardiac diameter for appraising the size of the heart," *Radiology*, vol. 57, no. 2, pp. 217–226, 1951.

[12] W. H. Frishman et al., "Cardiomegaly on chest x-ray: prognostic implications from a ten-year cohort study of elderly subjects: a report from the bronx longitudinal aging study," *American heart journal*, vol. 124, no. 4, pp. 1026–1030, 1992.

[13] M.T. Kearney et al., "A prognostic index to predict long-term mortality in patients with mild to moderate chronic heart failure stabilised on angiotensin converting enzyme inhibitors," *European Journal of Heart Failure*, vol. 5, no. 4, pp. 489–497, 2003.

[14] H.C. Becker et al., "Digital computer determination of a medical diagnostic index directly from chest x-ray images," *IEEE Trans. on Biomedical Engineering*, , no. 3, pp. 67–72, 1964.

[15] N. Nakamori et al., "Effect of heart-size parameters computed from digital chest radiographs on detection of cardiomegaly: potential usefulness for computer-aided diagnosis," *Investigative Radiology*, vol. 26, no. 6, pp. 546–550, 1991.

[16] B. Ginneken et al., "Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database," *Medical Image Analysis*, vol. 10, no. 1, pp. 19–40, 2006.

[17] D. Seghers et al., "Minimal shape and intensity cost path segmentation," *IEEE Trans. on Medical Imaging*, vol. 26, pp. 1115–1129, 2007.

[18] F.J. Browne et al., "Extraction of the two-dimensional cardiothoracic ratio from digital pa chest radiographs: correlation with cardiac function and the traditional cardiothoracic ratio," *Journal of digital imaging*, vol. 17, no. 2, pp. 120–123, 2004.

[19] M.A. Hasan et al., "Automatic evaluation of cardiac hypertrophy using cardiothoracic area ratio in chest radiograph images," *Computer methods and programs in biomedicine*, vol. 105, no. 2, pp. 95–108, 2012.

[20] S. Candemir et al., "Automatic heart localization and radiographic index computation in chest x-rays," in *Proc. of SPIE Medical Imaging, vol.9785*, 2016.

[21] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[22] S. Candemir, E. Borovikov, KC Santosh, S. Antani, and G. Thoma, "Rsilc: rotation-and scale-invariant, line-based color-aware descriptor," *Image and Vision Computing*, vol. 42, pp. 1–12, 2015.

[23] G. Litjens et al., "A survey on deep learning in medical image analysis," *arXiv preprint arXiv:1702.05747*, 2017.

[24] D. Ravì et al., "Deep learning for health informatics," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 4–21, 2017.

[25] M. T. Islam et al., "Abnormality detection and localization in chest x-rays using deep convolutional neural networks," *arXiv preprint:1705.09850*, 2017.

[26] P. Rajpurkar et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint:1711.05225*, 2017.

[27] "ImageNet," http://www.image-net.org/, [Accessed: 12/18/2017].

[28] A. Krizhevsky et al., "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint:1409.1556*, 2014.

[30] C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[31] "National Library of Medicine, Open-i Project," https://openi.nlm.nih.gov, [Accessed: 13/01/2017].

[32] F. Bellemare et al., "Sex differences in thoracic dimensions and configuration," *American journal of respiratory and critical care medicine*, vol. 168, no. 3, pp. 305–312, 2003.

[33] YB Mensah et al., "Establishing the cardiothoracic ratio using chest radiographs in an indigenous ghanaian population: a simple tool for cardiomegaly screening," *Ghana Medical Journal*, vol. 49/3, 2015.

[34] Z. Xue et al., "Using deep learning for detecting gender in adult chest radiographs," in *Proc. of SPIE - Medical Imaging*, 2018, vol. 10579.

[35] V. Kovalev et al., "Mining lung shape from x-ray images," in *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, 2009, pp. 554–568.

[36] S. Candemir et al., "Lung boundary detection in pediatric chest x-rays," in *Proc. of SPIE Medical Imaging*, 2015, vol. 9418.

[37] H. Thodberg et al., "The bonexpert method for automated determination of skeletal maturity," *IEEE Trans. on Medical Imaging*, vol. 28, no. 1, pp. 52–66, 2009.