

NIH Chest X-ray Dataset of 14 Common Thorax Disease Categories:

(1, Atelectasis; 2, Cardiomegaly; 3, Effusion; 4, Infiltration; 5, Mass; 6, Nodule; 7, Pneumonia; 8, Pneumothorax; 9, Consolidation; 10, Edema; 11, Emphysema; 12, Fibrosis; 13, Pleural_Thickening; 14 Hernia)

Background & Motivation: Chest X-ray exam is one of the most frequent and cost-effective medical imaging examination. However clinical diagnosis of chest X-ray can be challenging, and sometimes believed to be harder than diagnosis via chest CT imaging. Even some promising work have been reported in the past, and especially in recent deep learning work on Tuberculosis (TB) classification. To achieve clinically relevant computer-aided detection and diagnosis (CAD) in real world medical sites on all data settings of chest X-rays is still very difficult, if not impossible when only several thousands of images are employed for study. This is evident from [2] where the performance deep neural networks for thorax disease recognition is severely limited by the availability of only 4143 frontal view images [3] (Openi is the previous largest publicly available chest X-ray dataset to date).

In this database, we provide an enhanced version (with 6 more disease categories and more images as well) of the dataset used in the recent work [1] which is approximately 27 times of the number of frontal chest x-ray images in [3]. Our dataset is extracted from the clinical PACS database at National Institutes of Health Clinical Center and consists of ~60% of all frontal chest x-rays in the hospital. Therefore we expect this dataset is significantly more representative to the real patient population distributions and realistic clinical diagnosis challenges, than any previous chest x-ray datasets. Of course, the size of our dataset, in terms of the total numbers of images and thorax disease frequencies, would better facilitate deep neural network training [2]. Refer to [1] on the details of how the dataset is extracted and image labels are mined through natural language processing (NLP).

Details: ChestX-ray dataset comprises 112,120 frontal-view X-ray images of 30,805 unique patients with the text-mined fourteen disease image labels (where each image can have multi-labels), mined from the associated radiological reports using natural language processing. Fourteen common thoracic pathologies include Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural_thickening, Cardiomegaly, Nodule, Mass and Hernia, which is an extension of the 8 common disease patterns listed in our CVPR 2017 paper. *Note that original radiology reports (associated with these chest x-ray studies) are not meant to be publicly shared for many reasons. The text-mined disease labels are expected to have accuracy >90%. Please find more details and benchmark performance of trained models based on 14 disease labels in our arxiv paper: [1705.02315](https://arxiv.org/abs/1705.02315)*

Contents:

1. 112,120 frontal-view chest X-ray PNG images in 1024*1024 resolution (under images folder)
2. Meta data for all images (Data_Entry_2017.csv): Image Index, Finding Labels, Follow-up #, Patient ID, Patient Age, Patient Gender, View Position, Original Image Size and Original Image Pixel Spacing.

3. Bounding boxes for ~1000 images (BBox_List_2017.csv): Image Index, Finding Label, Bbox[x, y, w, h]. [x y] are coordinates of each box's topleft corner. [w h] represent the width and height of each box.
4. Two data split files (train_val_list.txt and test_list.txt) are provided. Images in the ChestX-ray dataset are divided into these two sets on the patient level. All studies from the same patient will only appear in either training/validation or testing set.

If you find the dataset useful for your research projects, please cite our CVPR 2017 paper:

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald M. Summers. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly- Supervised Classification and Localization of Common Thorax Diseases, IEEE CVPR, pp. 3462-3471,2017

```
@InProceedings{wang2017chestxray,
  author    = {Wang, Xiaosong and Peng, Yifan and Lu, Le and Lu, Zhiyong and Bagheri, Mohammadhadi
and Summers, Ronald},
  title     = {ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised
Classification and Localization of Common Thorax Diseases},
  booktitle = {2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)},
  pages     = {3462--3471},
  year      = {2017}
}
```

Questions Comments: (xiaosong.wang@nih.gov; le.lu@nih.gov; rms@nih.gov)

Limitations: 1) The image labels are NLP extracted so there would be some erroneous labels but the NLP labelling accuracy is estimated to be >90%. 2) Very limited numbers of disease region bounding boxes. 3) Chest x-ray radiology reports are not anticipated to be publicly shared. Parties who use this public dataset are encouraged to share their “updated” image labels and/or new bounding boxes in their own studied later, maybe through manual annotation.

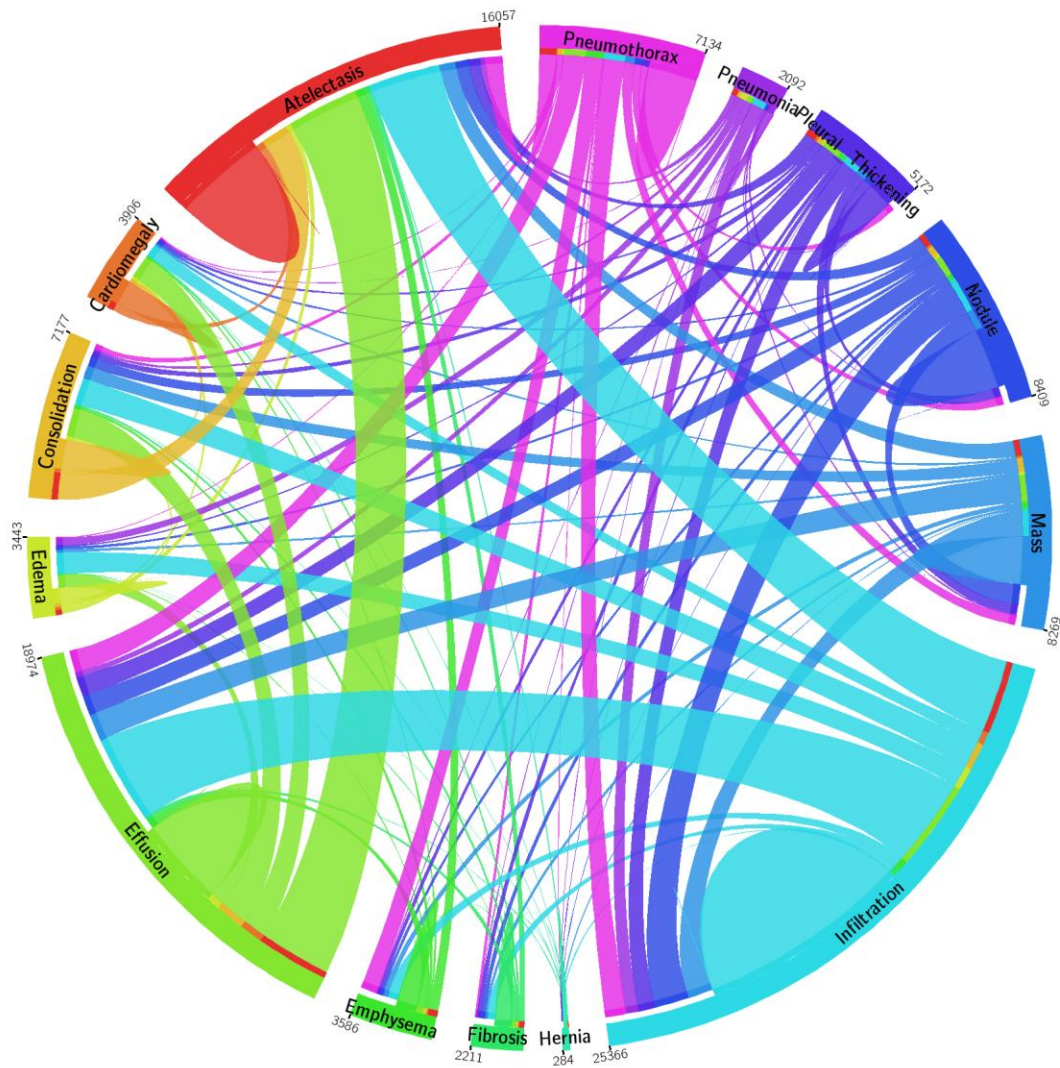
Acknowledgement: This work was supported by the Intramural Research Program of the NIH Clinical Center (clinicalcenter.nih.gov) and National Library of Medicine (www.nlm.nih.gov). We thank NVIDIA Corporation for the GPU donations.

Reference:

- [1] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald Summers, ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, IEEE CVPR, pp. 3462-3471, 2017
- [2] Hoo-chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, Ronald M. Summers, Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation, IEEE CVPR, pp. 2497-2506, 2016

[3] Open-i: An open access biomedical search engine. <https://openi.nlm.nih.gov>

A, Distributions of 14 disease categories with co-occurrence statistics:

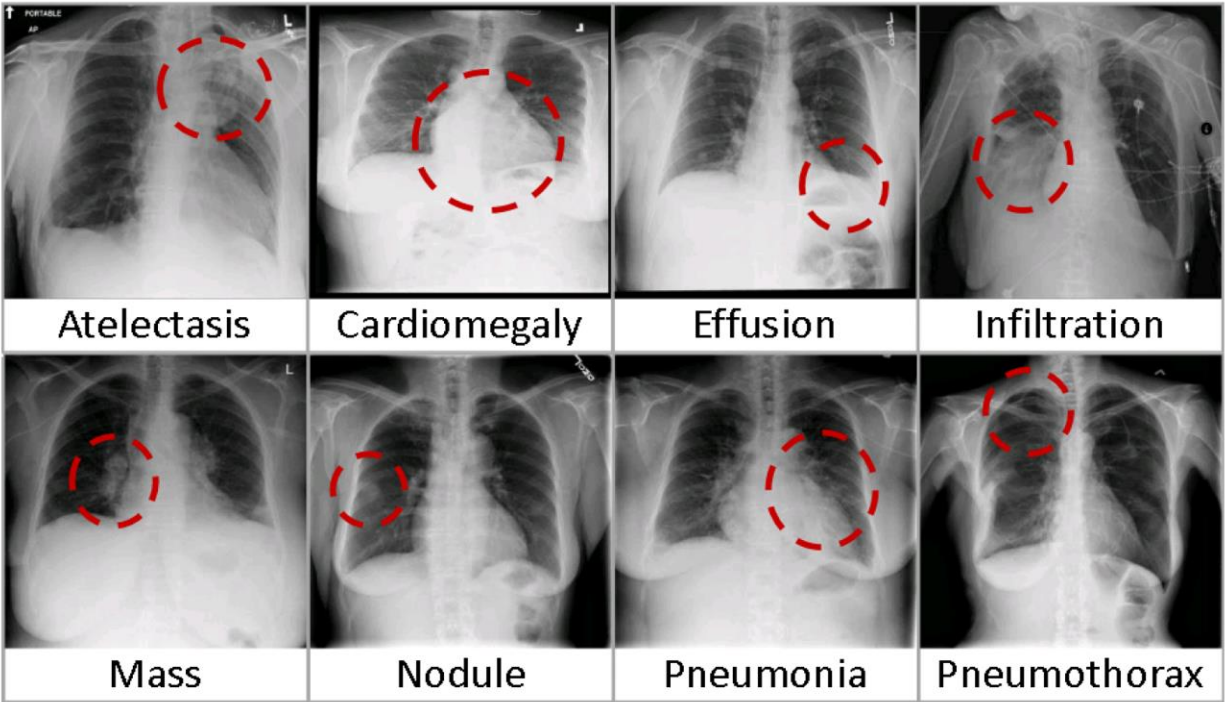


A.2 co-occurrence matrix of the fourteen thorax diseases in this chest X-ray dataset

4212	369	3269	3259	727	585	243	772	1222	221	423	220	495	40
369	1094	1060	583	99	108	36	48	169	127	44	51	111	7
3269	1060	3959	3990	1244	909	253	995	1287	592	359	188	848	21
3259	583	3990	9552	1151	1544	571	943	1220	979	447	345	749	33
727	99	1244	1151	2138	894	62	424	602	128	212	115	448	25

585	108	909	1544	894	2706	63	340	428	131	115	166	410	10
243	36	253	571	62	63	307	34	114	330	21	11	45	2
772	48	995	943	424	340	34	2199	222	33	746	80	289	9
1222	169	1287	1220	602	428	114	222	1314	162	103	79	251	4
221	127	592	979	128	131	330	33	162	634	30	9	64	3
423	44	359	447	212	115	21	746	103	30	895	36	151	4
220	51	188	345	115	166	11	80	79	9	36	727	176	8
495	111	848	749	448	410	45	289	251	64	151	176	1127	8
40	7	21	33	25	10	2	9	4	3	4	8	8	110
11535	2772	13307	19871	5746	6323	1353	5298	4667	2303	2516	1686	3385	227

B. Eight visual examples of common thorax diseases



C. Two Samples of disease localization using weakly supervised deep neural networks

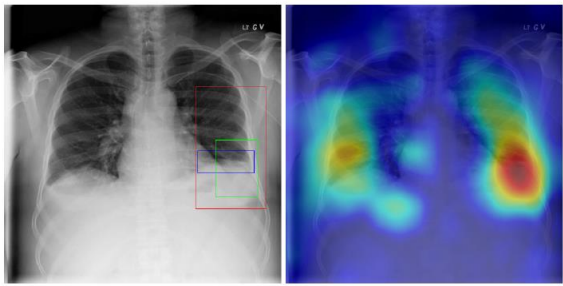
Radiology report	Keyword	Localization Result
findings include: 1. left basilar atelectasis/consolidation. 2. prominent hilum (mediastinal adenopathy). 3. left pic catheter (tip in atriocaval junction). 4. stable, normal appearing cardiomeastinal silhouette. impression: small right pleural effusion otherwise stable abnormal study including left basilar infiltrate/atelectasis, prominent hilum, and position of left pic catheter (tip atriocaval junction).	Effusion; Infiltration; Atelectasis	

Table 4. A sample of chest x-ray radiology report, mined disease keywords and localization result from the “Atelectasis” Class. Correct bounding box (in green), false positives (in red) and the ground truth (in blue) are plotted over the original image.

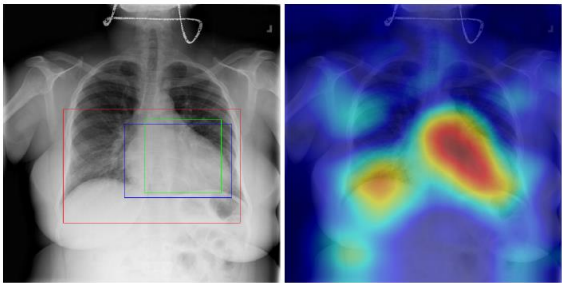
Radiology report	Keyword	Localization Result
findings include: 1. cardiomegaly (ct ratio of 17/30). 2. otherwise normal lungs and mediastinal contours. 3. no evidence of focal bone lesion. dictating	Cardiomegaly	

Table 5. A sample of chest x-ray radiology report, mined disease keywords and localization result from the “Cardiomegaly” Class. Correct bounding box (in green), false positives (in red) and the ground truth (in blue) are plotted over the original image.