CheXseen: Unseen Disease Detection for Deep Learning Interpretation of Chest X-rays

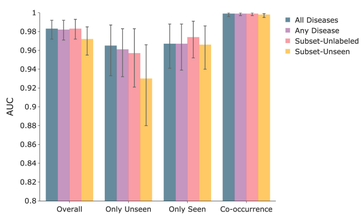
Siyu Shi*[1], Ishaan Malhi*[1], Kevin Tran[1], Andrew Y. Ng[1], Pranav Rajpurkar[2]
1. Department of Computer Science, Stanford University; 2. Department of Biomedical Informatics, Harvard Medical School

Motivation

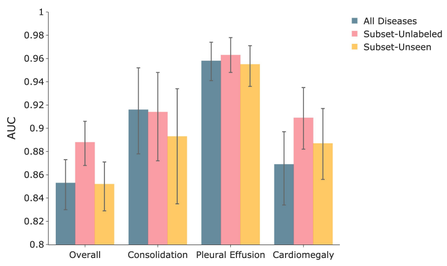
1. Most medical imaging deep learning models are trained and evaluated on a few common diseases.
2. However, a successful deployment requires medical AI to maintain performance in the presence of unseen diseases and also flag unseen diseases for human interpretation.
3. We evaluated the performance of deep learning algorithms in the presence of unseen diseases that the model is not trained to detect, and proposed several strategies to detect unseen diseases in chest X-rays.

Can a model detect "no disease" and seen diseases in the presence of unseen diseases?

- Models tend to falsely classify diseases outside of the subset (unseen diseases) as "no disease".

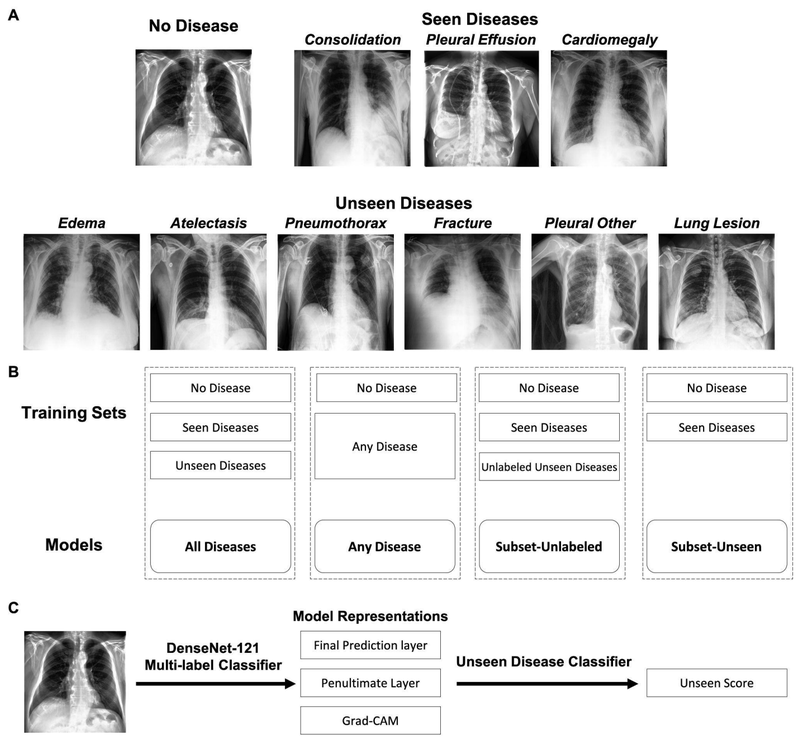


- The performance of the Subset-Unlabeled model, which is trained with unseen disease examples, but not unseen disease labels is comparable to both controls.



- Models are still able to detect seen diseases even when co-occurring with unseen diseases at a level comparable to the All Diseases model.

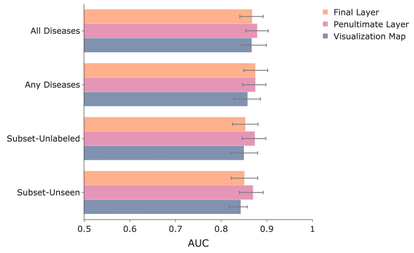
Overview of the experimental setup



A. Chest X-ray image labels included "No Disease", three seen diseases, and six unseen diseases.
B. Training data setup for the four multi-label models. The Subset-Unlabeled model was trained with all images in the train set, with the labels of "No Disease" and three seen diseases, while excluding the labels of six unseen diseases. The Subset-Unseen model was trained with only the images that do not have any of the six unseen diseases. The All Diseases model was trained with all images and all ten labels ("seen" diseases, "unseen" diseases and "no disease"), and served as a control. The Any Disease model was trained with all images for either having "any disease" or "no disease", and served as another control.
C. Three outputs from the models: final prediction layer, penultimate (intermediate) layer, and visualization map were used to train unseen disease classifiers, to predict the "unseen score" (whether an unseen disease was present during testing).

Can feature representations learned by models be used to detect the presence of unseen diseases given a small labeled set of unseen diseases?

- Classifiers trained with the penultimate layer output performed the best, followed by the final layer and the visualization map.



- The unseen scores from the Subset-Unlabeled model has higher performance than those from the Subset-Unseen model, likely because the former learns representations of the unlabeled diseases during training.

Discussion

- In this study, we evaluate the performance of deep learning models in the presence of diseases not labeled for or present during training.
- Our results can inform the safe clinical deployment of deep learning models trained on a non-exhaustive set of disease classes.

References

Jeremy Irvin, Pranav Rajpurkar, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. 2019
Abhijit Bendale and Terrance E Boult. Towards open set deep networks. Towards open set deep networks. 2016
Tianshi Cao, Chinwei Huang,et al. A benchmark of medical out of distribution detection. 2020