How Does Pruning Impact Long-Tailed Multi-Label Medical Image Classifiers?

G. Holste[1], Z. Jiang[2], A. Jaiswal[1], M. Hanna[3], S. Minkowitz[3], A.C. Legasto[3], J.G. Escalon[3], S. Steinberger[3], M. Bittman[3], T. C. Shen[4], Y. Ding[1], R.M. Summers[4], G. Shih[3], Y. Peng[3], Z. Wang[1]

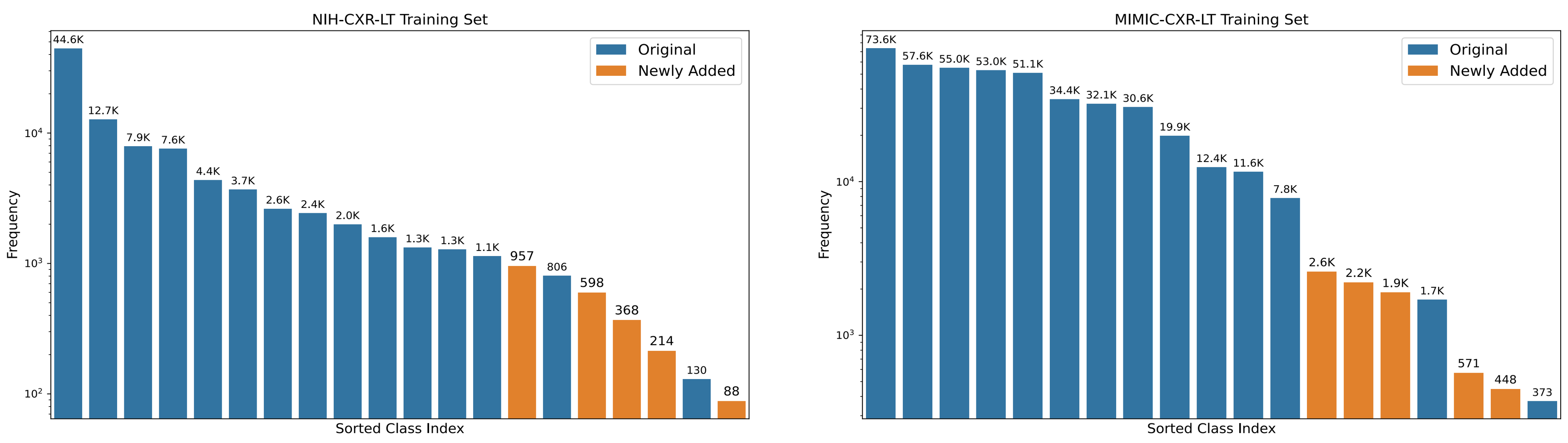
[1]The University of Texas at Austin, [2]Texas A&M University, [3]Weill Cornell Medicine, [4]NIH

MOTIVATION

- **Pruning** can reduce memory + latency with little change in overall performance
- However, unknown how pruning impacts model behavior in *long-tailed*, *multi-label* classification
  - ➢ Very common in clinical settings
  - ➢ Knowledge gap could have dangerous implications!

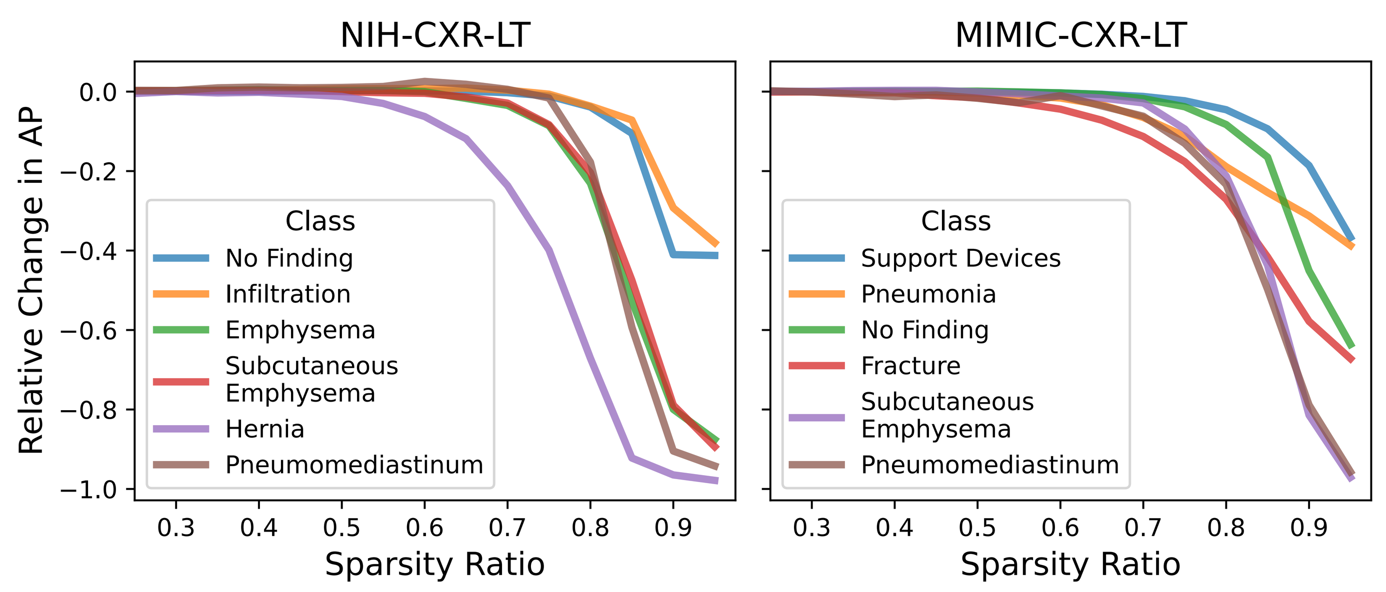
1) How does pruning impact overall performance?
2) Which classes are most impacted by pruning?
3) How does disease co-occurrence factor in?
4) Which images are most vulnerable to pruning?

DATA & METHODS



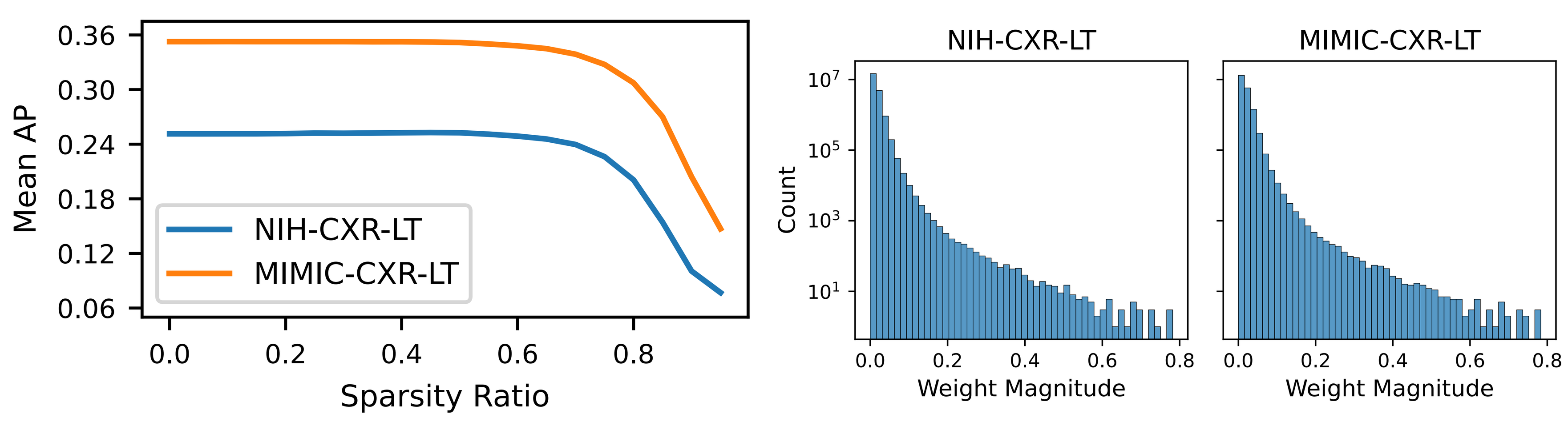
- Curate 2 long-tailed, multi-label chest X-ray datasets
  - ➢ **NIH-CXR-LT**: 112,120 images | 20 classes
  - ➢ **MIMIC-CXR-LT**: 257,018 images | 19 classes

- <u>Experimental design</u>:
  - ➢ Train 30 models, evaluate by average precision (AP)
  - ➢ For each dataset and model, perform L1 pruning at sparsity ratios $k \in \{0, 0.05, 0.1, \ldots, 0.9, 0.95\}$



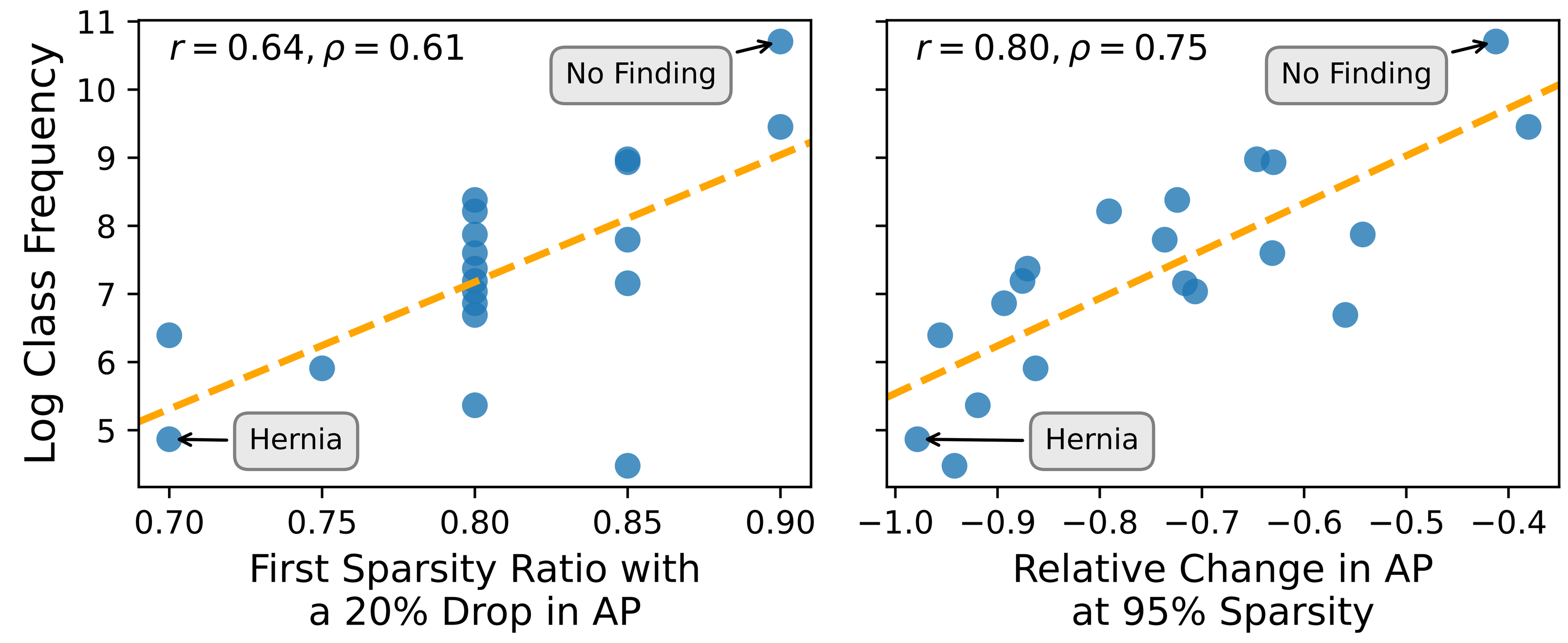
- **Forgettability curve**: For a given class, plot relative change in AP from uncompressed to k-sparse model
  - ➢ Characterizes "forgettability" of a class upon pruning
  - ➢ How do these curves relate to class frequency (*long-tailed*) and co-occurrence behavior (*multi-label*)?

RESULTS

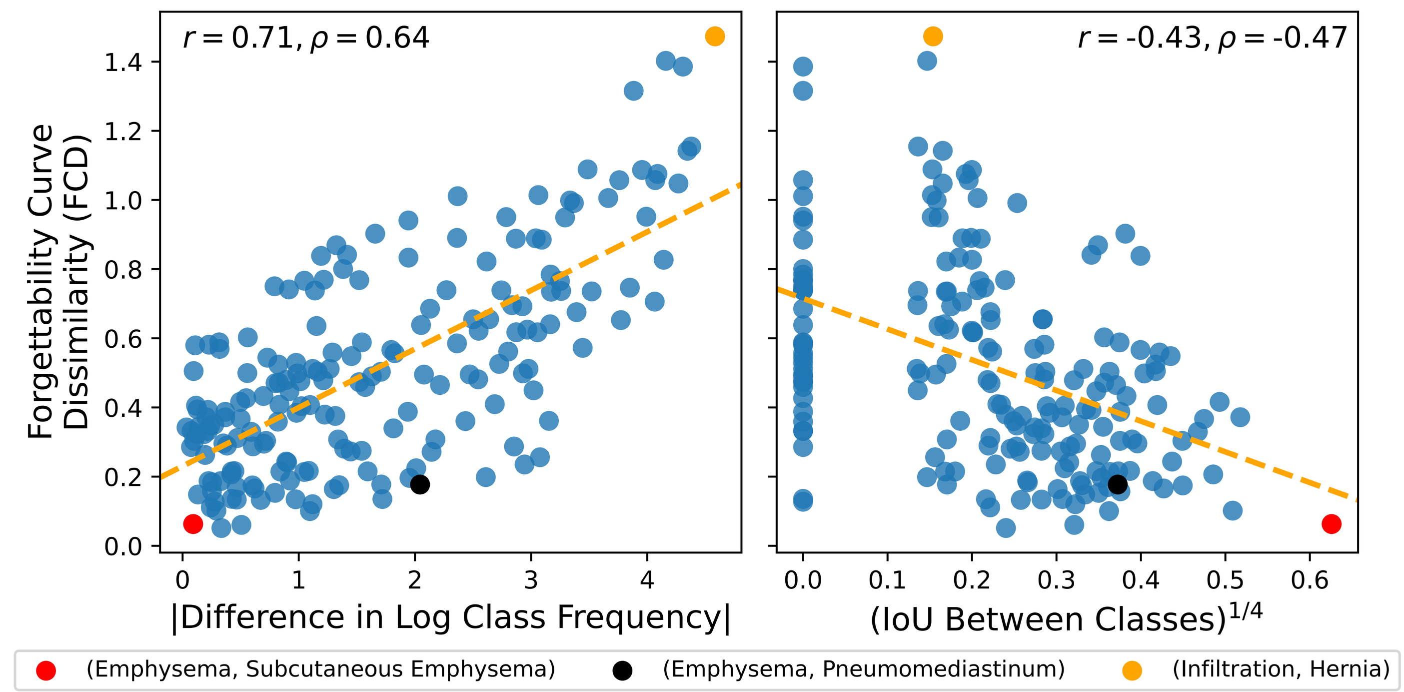


1) Up to 65% of weights can be pruned with no significant impact on overall performance
- ➢ ResNet 50 is overparameterized for this task
- ➢ Learned weights are naturally sparse, indicating only a small subset of neurons are needed for modeling

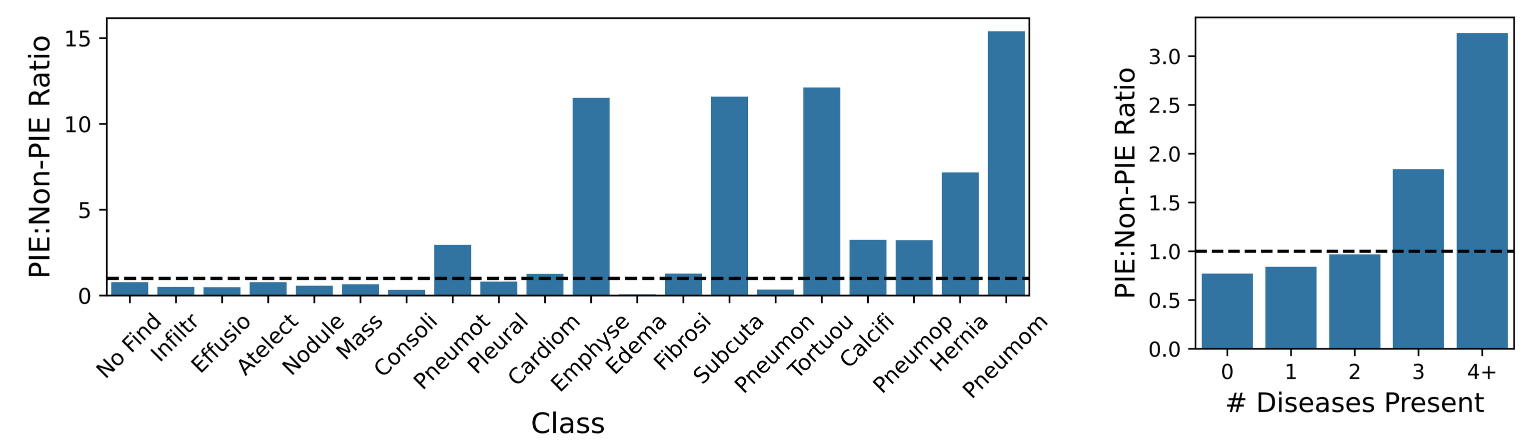
RESULTS (CONT'D)

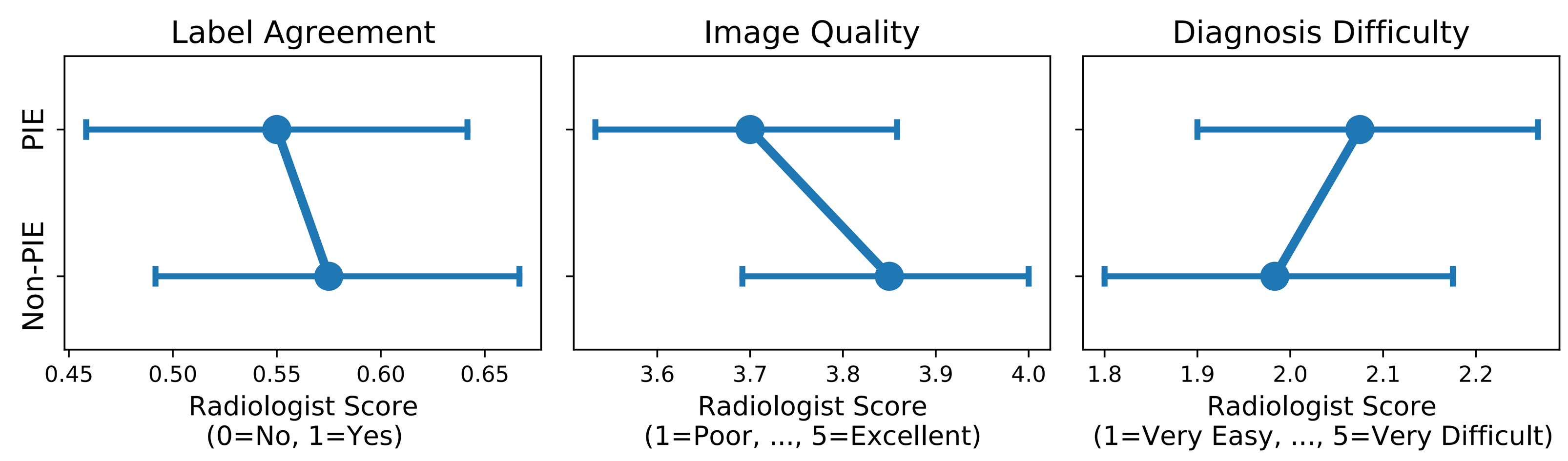


2) Rare classes are (i) forgotten *earlier* and (ii) more *severely forgotten* at high sparsity



3) A disease's forgettability can be explained by prevalence and co-occurrence behavior
- ➢ **FCD** = MSE between two forgettability curves
- ➢ Diseases w/ larger differences in prevalence exhibit more distinct "forgetting trajectories" (lower FCD)
- ➢ The more two diseases co-occur, the more similar their forgettability curves (higher FCD)





4) Pruning can identify images with complex disease presentation, label noise, and low image quality
- ➢ **PIE** = image where original and pruned model disagree
  - Bottom 5th percentile of correlation between predictions
- ➢ Rare classes are 3-15x overrepresented in PIEs
- ➢ Images with 3+ diseases are ~2x overrepresented in PIEs
- ➢ In human reader study, radiologists found PIEs to have:
  - more label noise, lower quality, + higher diagnosis difficulty

FUTURE WORK

- Do these findings hold for other architectures, datasets, imaging modalities, + compression methods?
- Are PIEs (a) valuable "hard examples" that deserve upweighting or (b) noisy examples that could be removed?

PAPER