An adversarial approach for the robust classification of pneumonia from chest radiographs

Joseph D. Janizek, Gabriel Erion, Alex J. DeGrave, Su-In Lee
July 24, 2020
University of Washington

Motivation

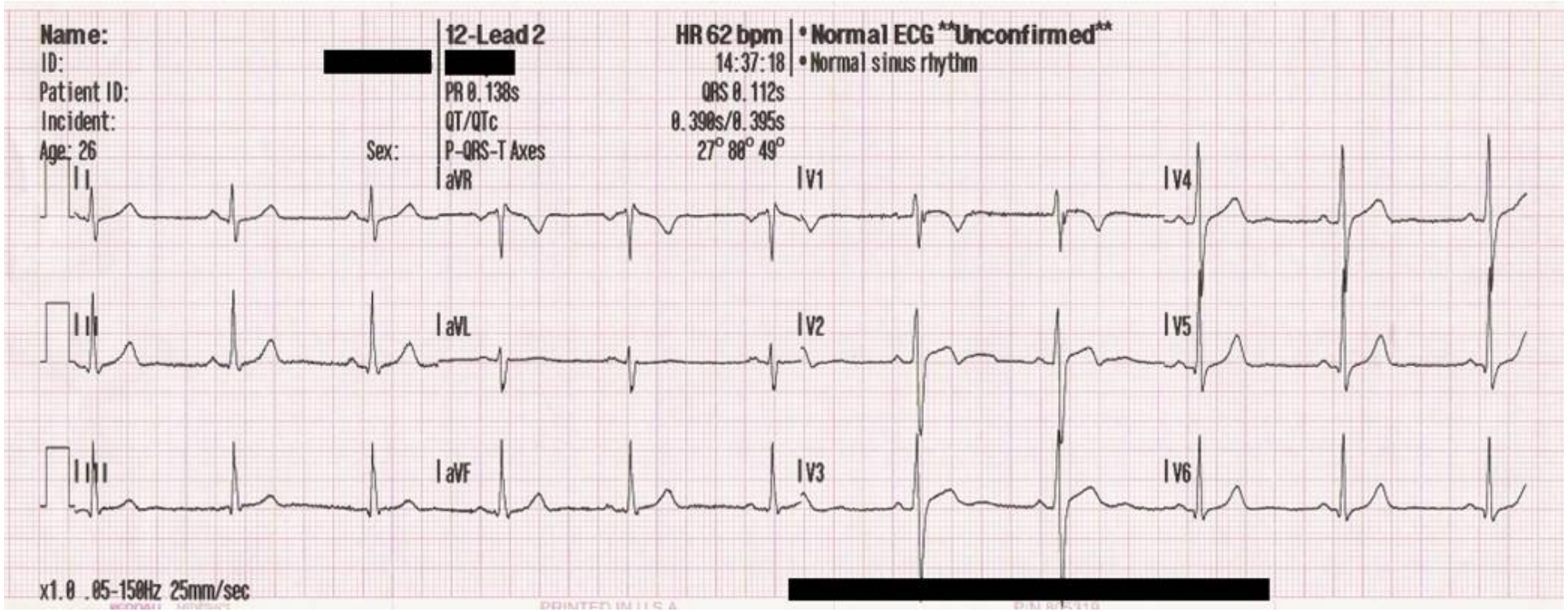
High-performance models exist for a variety of medical problems

Prediction of diabetic retinopathy from retinal fundus images

By Mikael Häggström, used with permission

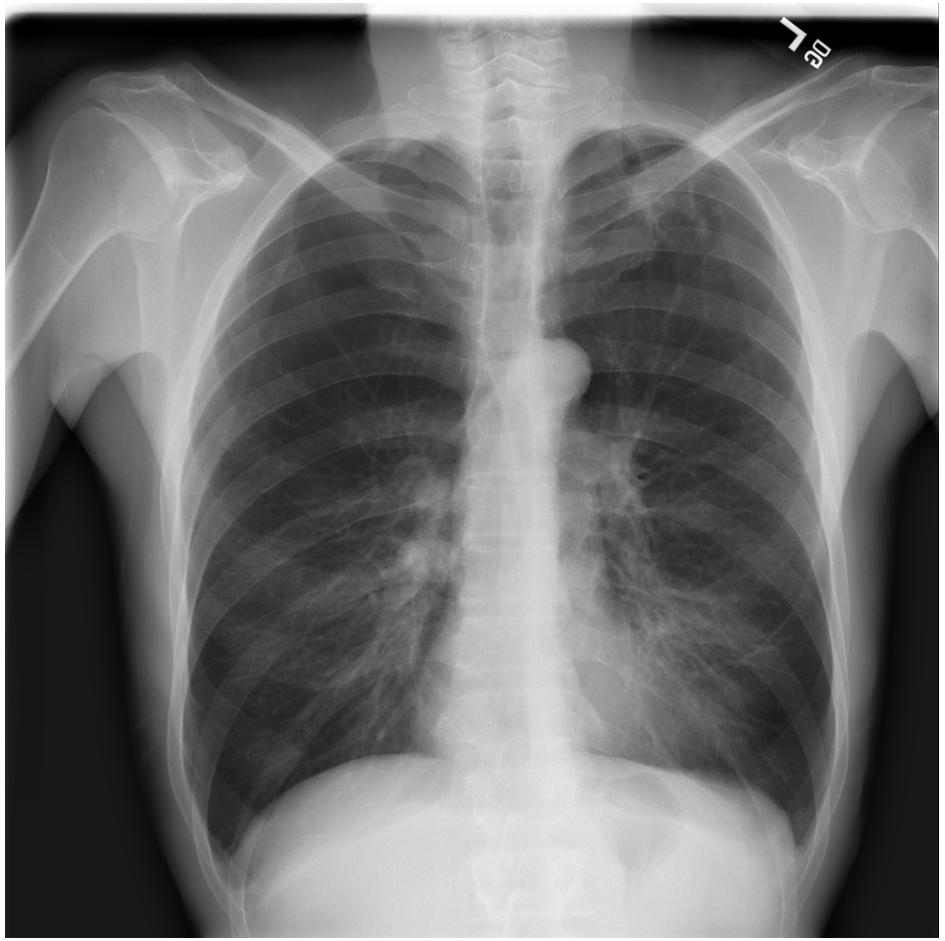
High-performance models exist for a variety of medical problems

Prediction of arrhythmia from ECG data



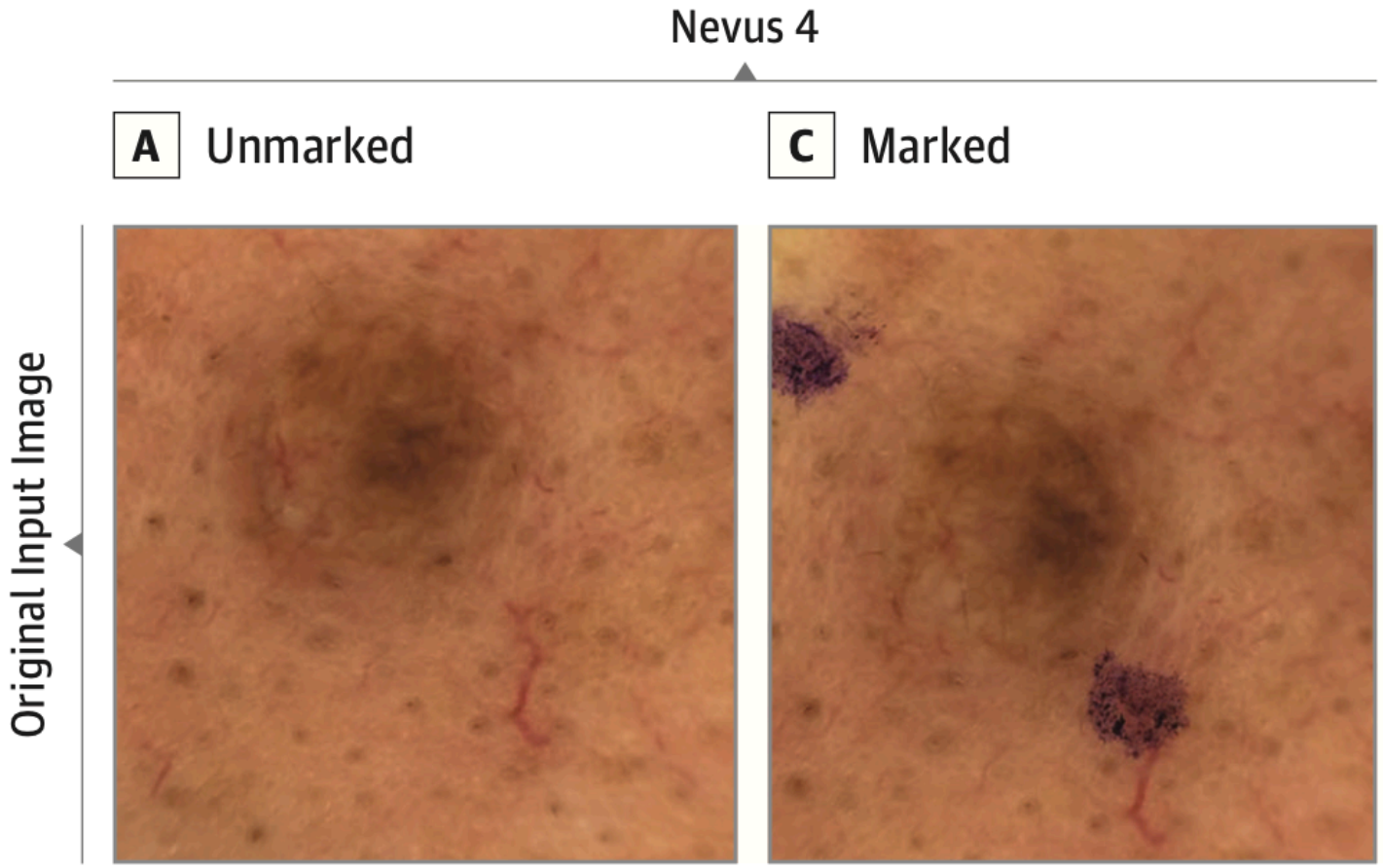
High-performance models exist for a variety of medical problems

Prediction of pneumonia from chest radiographs



High-performance models may use confounders

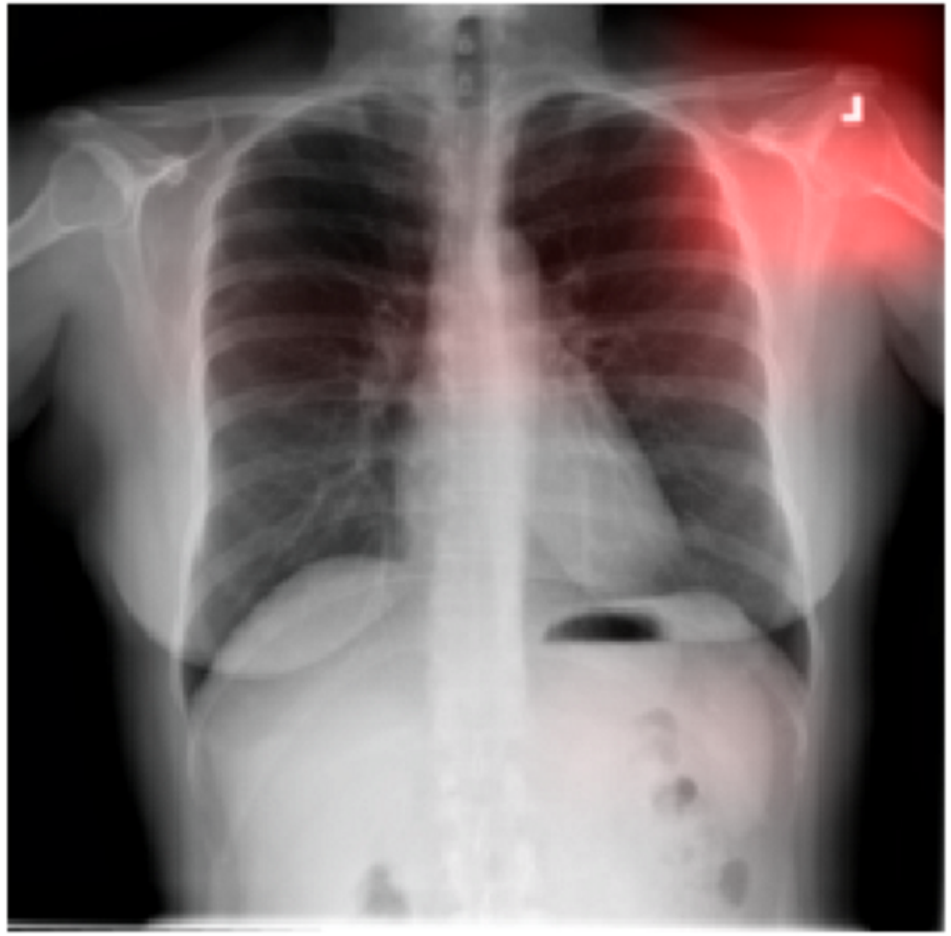
Deep learning dermatology algorithms use surgical markings to identify cancer



Winkler, Julia K., et al. "Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition." *JAMA dermatology* 155.10 (2019): 1135-1141.

High-performance models may use confounders

Radiology algorithms use hospital-specific markings to identify pneumonia and fail to generalize to new hospitals



Zech, John R., et al. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study." *PLoS medicine* 15.11 (2018).

Medical deep learning algorithms are brittle and may use confounders instead of making decisions the way we want

Goal: learn a pneumonia classifier that works in a new hospital that wasn't used for training, and that does not rely on confounding factors

Baseline Approach

Baseline classifier

DenseNet-121 (CNN)

Baseline classifier

CLF

Input: Frontal Chest Radiograph

Baseline classifier

CLF

0.8

Input: Frontal Chest Radiograph

Output: Probability of Pneumonia

Baseline classifier

CLF

0.8

Input: Frontal Chest Radiograph

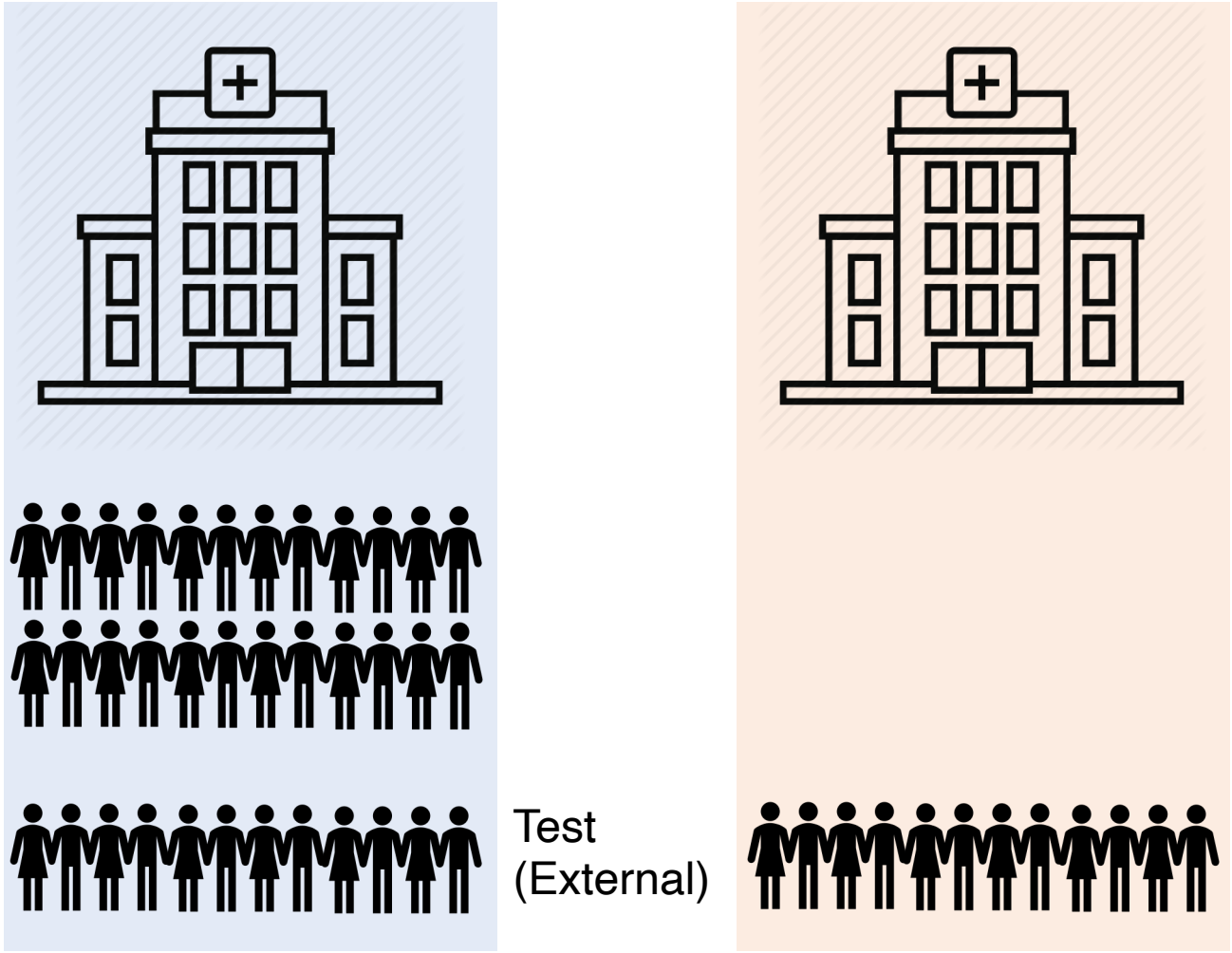
Output: Probability of Pneumonia

CLF Loss: Cross entropy between probability of pneumonia and true pneumonia label

Evaluation

CheXpert

MIMIC



Train

Test (Internal)

Test (External)

Evaluation

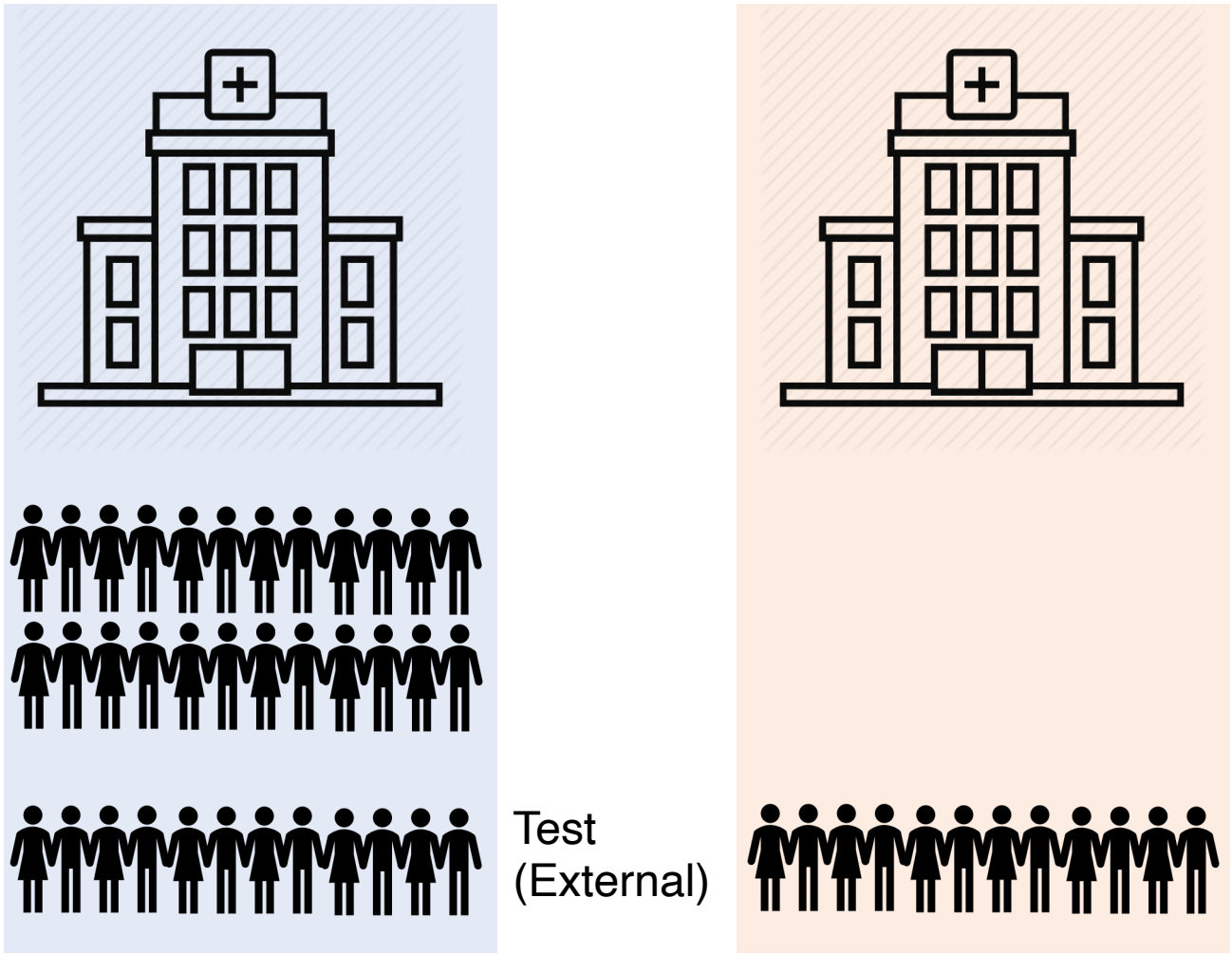
CheXpert

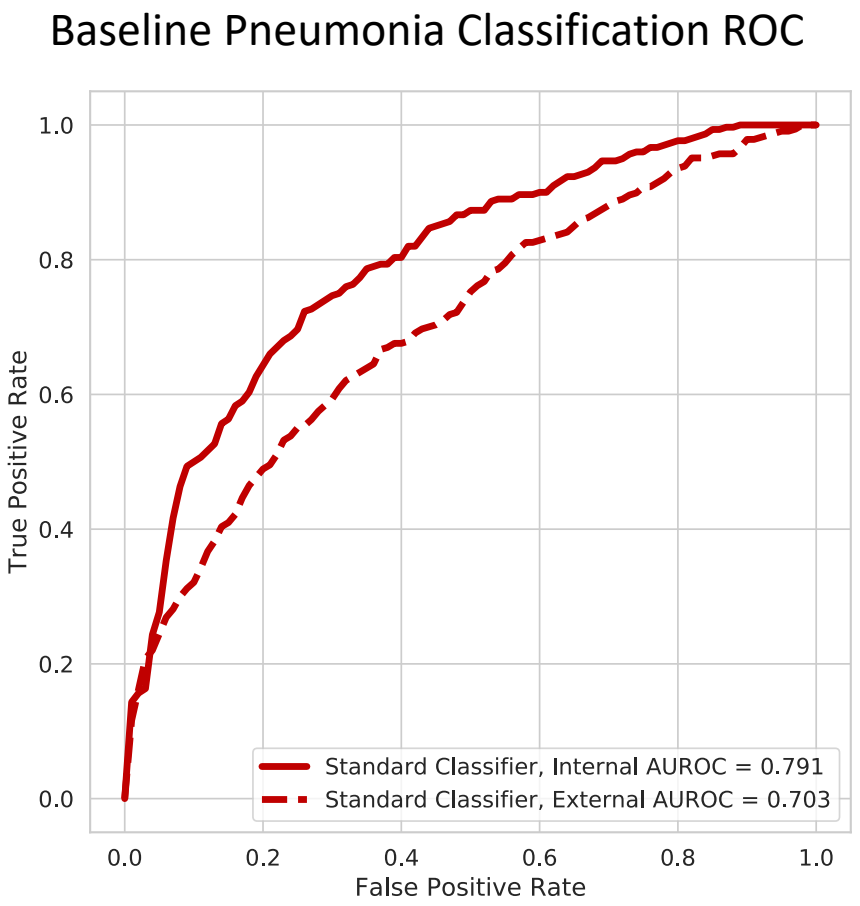
MIMIC

Train

Test (Internal)

Test (External)



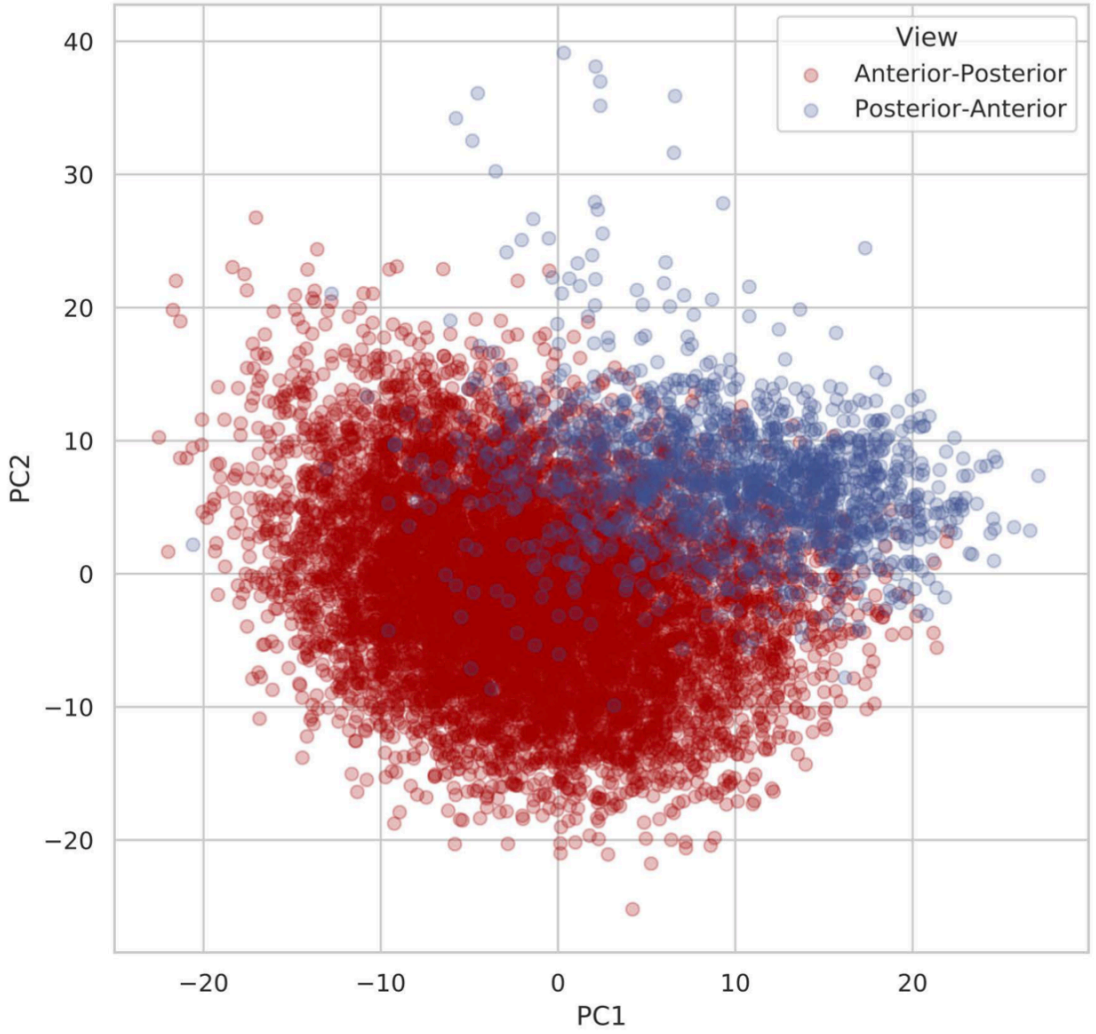


A standard model performs significantly worse when tested on images taken from a different hospital than its training data

Why?

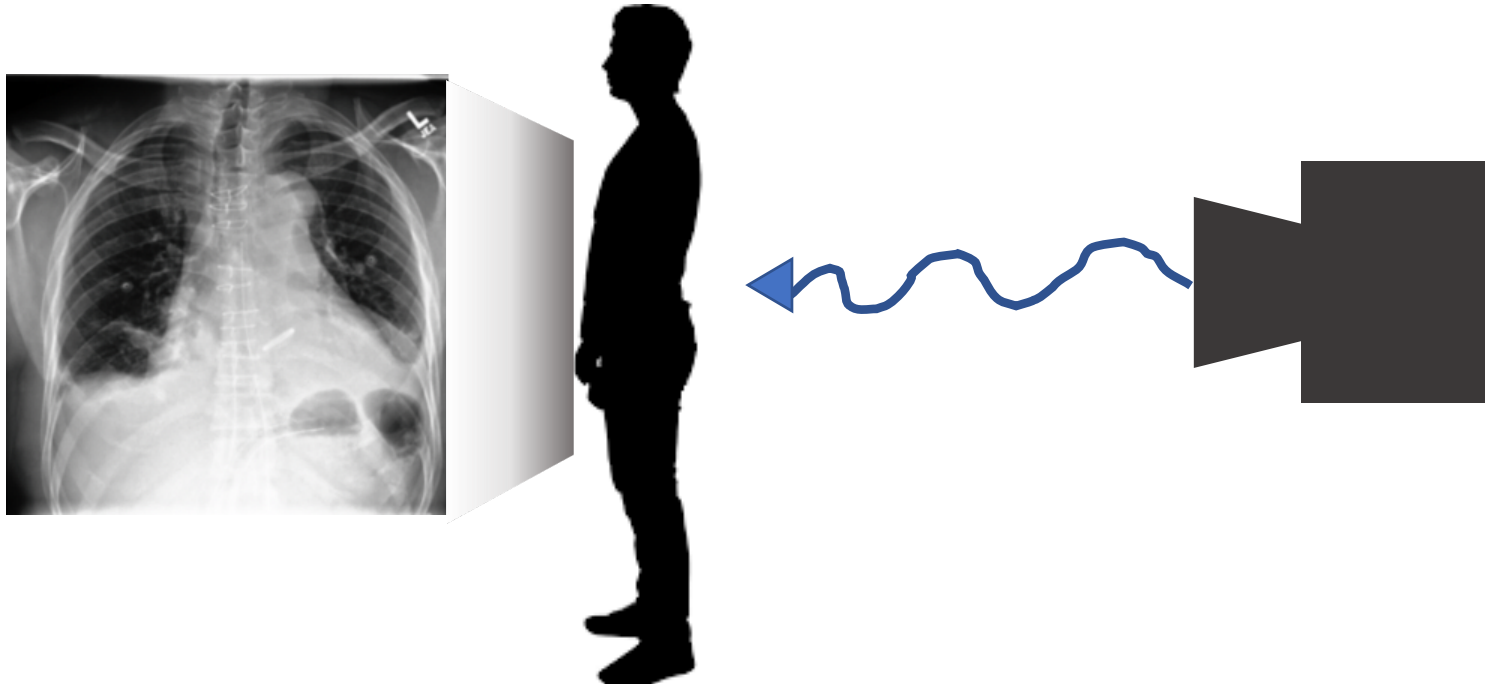
Deep convolutional networks easily separate view in embedding space

- Take embedding from last hidden layer
- Run PCA to reduce dimension from 1024 hidden nodes
  → first 2 PCs



Two types of frontal radiograph

- PA View
- Back to front
- Fixed Scanner
- Must be standing
- Tends to be associated with *healthier* patients



Two types of frontal radiograph

- AP View
- Front to back
- Portable Scanner
- Can be brought to patients in their bed
- Tends to be associated with *sicker* patients

Differences between AP and PA are not limited to one region

- Image may contain label like “PORTABLE”
- Heart shadow magnified
- Scapular edges visible over the lung fields
- Relative underinflation of lungs

How to tell a classifier to be invariant to these factors?

How to tell a classifier to be invariant to these factors?

Optimize it to fool an adversary that tries to tell whether prediction was made for AP or PA view

De-confounding classifier

Step 1: Pretrain classifier

CLF

0.8

Input: Frontal Chest Radiograph

Output: Probability of Pneumonia

CLF Loss: Cross entropy between probability of pneumonia and true pneumonia label

De-confounding classifier

Step 2: Pretrain adversary

CLF

0.8

ADV

De-confounding classifier

Step 2: Pretrain adversary

CLF → **0.8** → ADV

Input: Probability of pneumonia

De-confounding classifier

Step 2: Pretrain adversary

CLF → **0.8** → ADV → **0.2**

Input: Probability of pneumonia

Output: Probability of PA view

De-confounding classifier

Step 2: Pretrain adversary

CLF → **0.8** → ADV → **0.2**

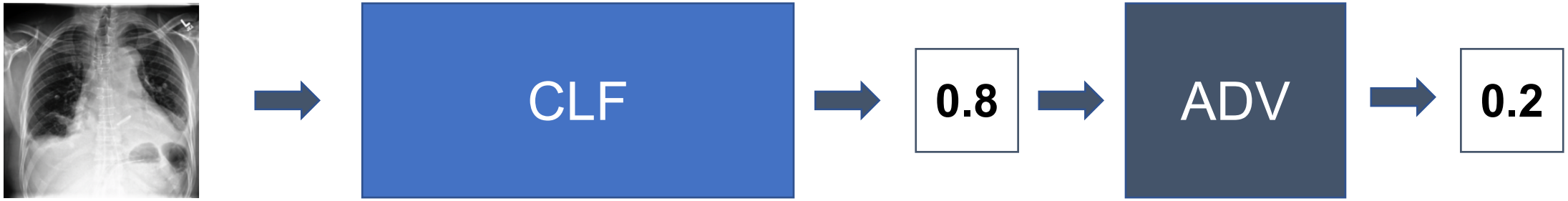
Input: Probability of pneumonia

Output: Probability of PA view

ADV Loss: Cross entropy between probability of PA view and true view label

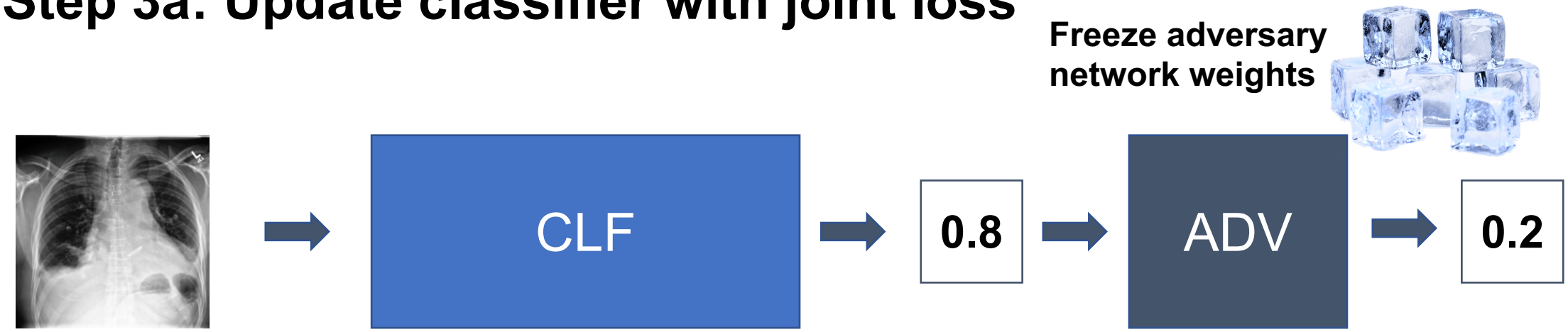
De-confounding classifier

Step 3: Joint optimization



De-confounding classifier

Step 3a: Update classifier with joint loss



De-confounding classifier

Step 3a: Update classifier with joint loss

Freeze adversary network weights

CLF → **0.8** → ADV → **0.2**

CLF Loss: Cross entropy between probability of pneumonia and true pneumonia label

ADV Loss: Cross entropy between probability of PA view and true view label

Joint Loss = CLF Loss – $\lambda \times$ ADV Loss

De-confounding classifier

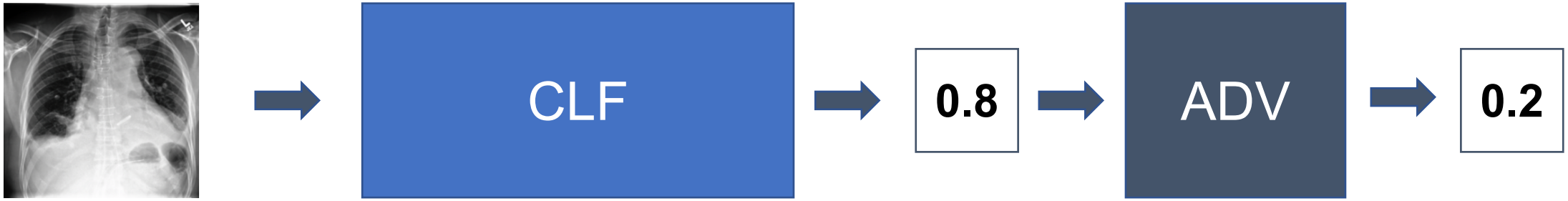
Step 3b: Train adversary until convergence

CLF → **0.8** → ADV → **0.2**

Freeze classifier network weights

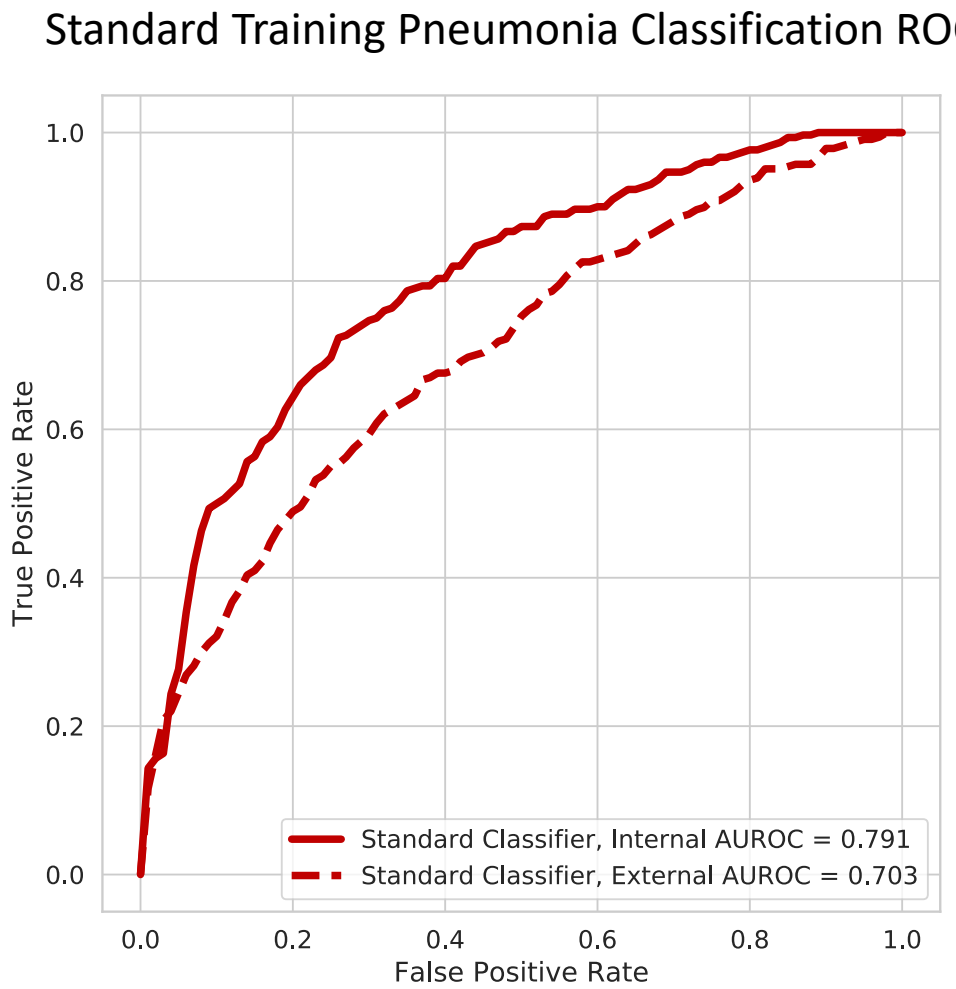
ADV Loss: Cross entropy between probability of PA view and true view label

De-confounding classifier



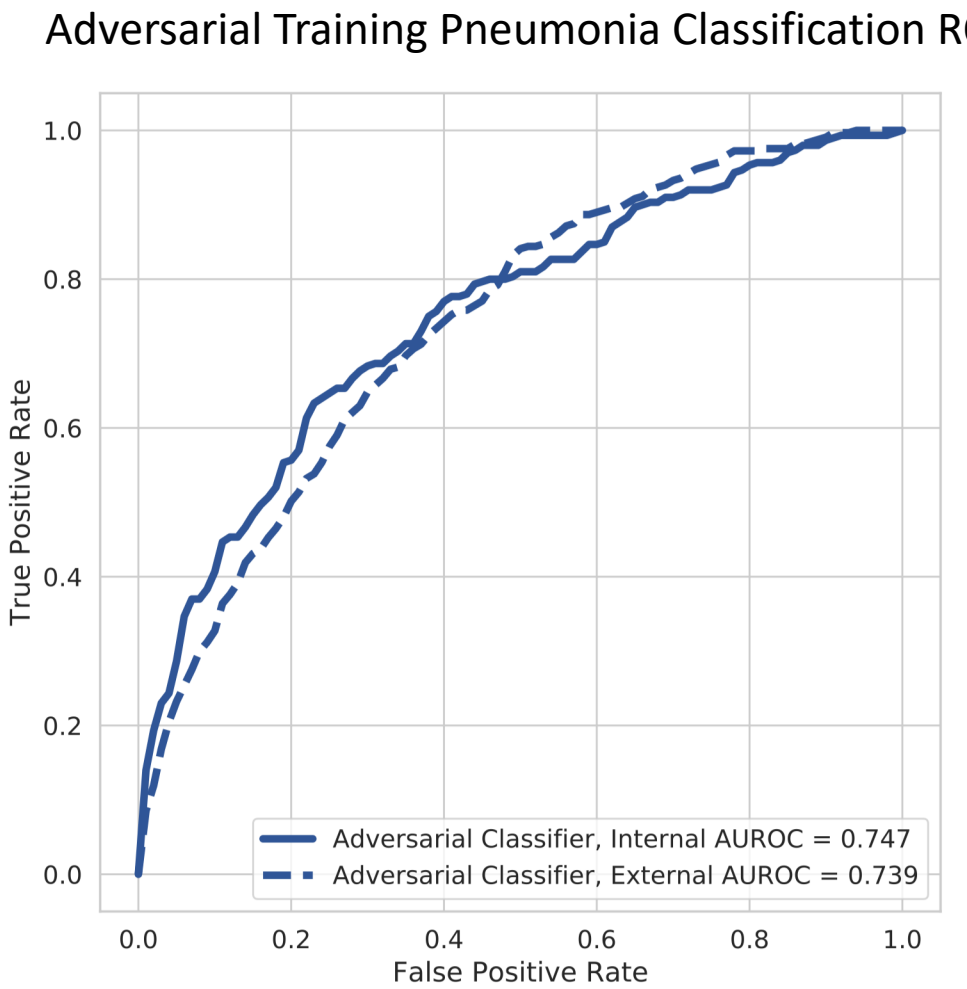
Learn a network that *classifies pneumonia accurately* while *fooling the adversary* that tries to predict view from score

Can adversarial training improve generalization performance?



Generalization gap: -0.09
External AUROC: 0.70

Adversarial training improves generalization performance



Generalization gap: -0.01
External AUROC: 0.74

A simple approach to learn a classifier that is invariant to a known confounder, even when that confounder can’t be localized to a single region of the image

Thank you!
Code: github.com/suinleelab/cxr_adv
Contact: jjanizek.github.io