Bridging the Gap: From Post Hoc Explanations to Inherently Interpretable Models for Medical Imaging

Shantanu Ghosh[1], Ke Yu[2], Forough Arabshahi[3], Kayhan Batmanghelich[1]

[1]Dept. Of Electrical and Computer Engineering, Boston University

[2]Intelligent Systems Program (ISP), University of Pittsburgh, [3]Meta AI

TLDR: Extracting a mixture of interpretable models from a BlackBox to provide instance specific concept-based explanations using First-order logic (FOL).

Post hoc explanation

Pros

- Does not alter the Black box.

Cons

- Inconsistent explanations.
- No recourse.

How to blur this gap?

Interpretable by design

Pros

- Support concept intervention.

Cons

- Harder to train.
- Sub par performance.

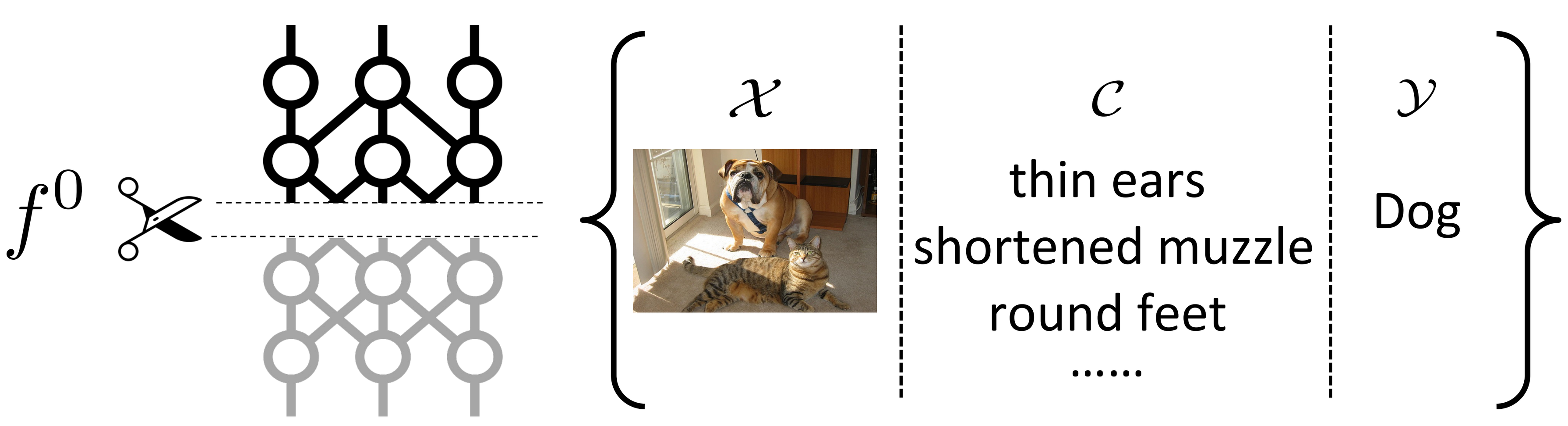
Desirable properties

- Does compromise the performance.
- Can be intervened to fix the misclassification

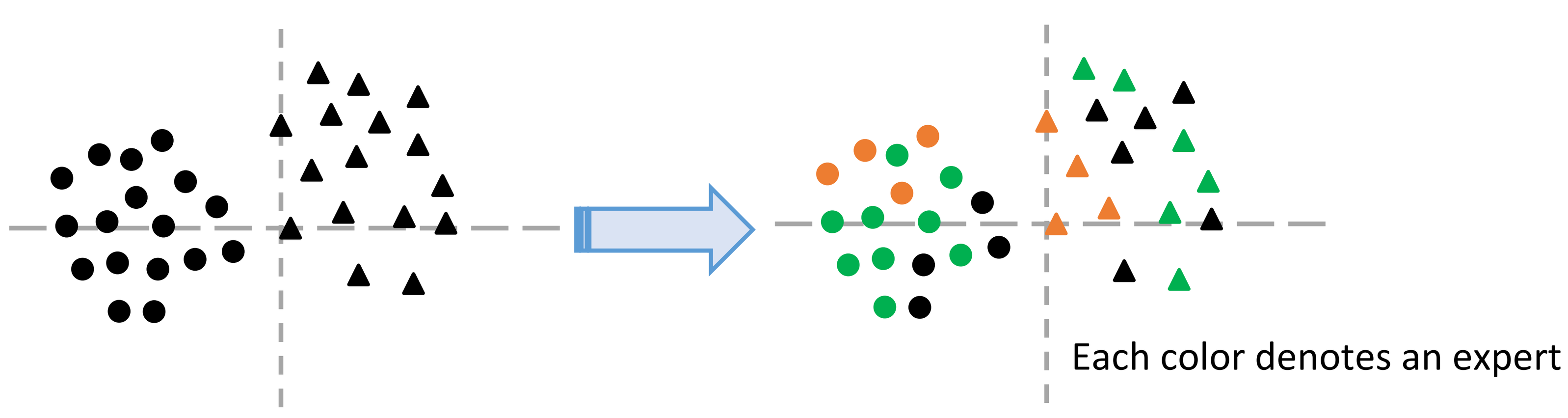
Design choices

- Carve interpretable models from Blackbox.
- Concept based
- First order logic for concept interaction

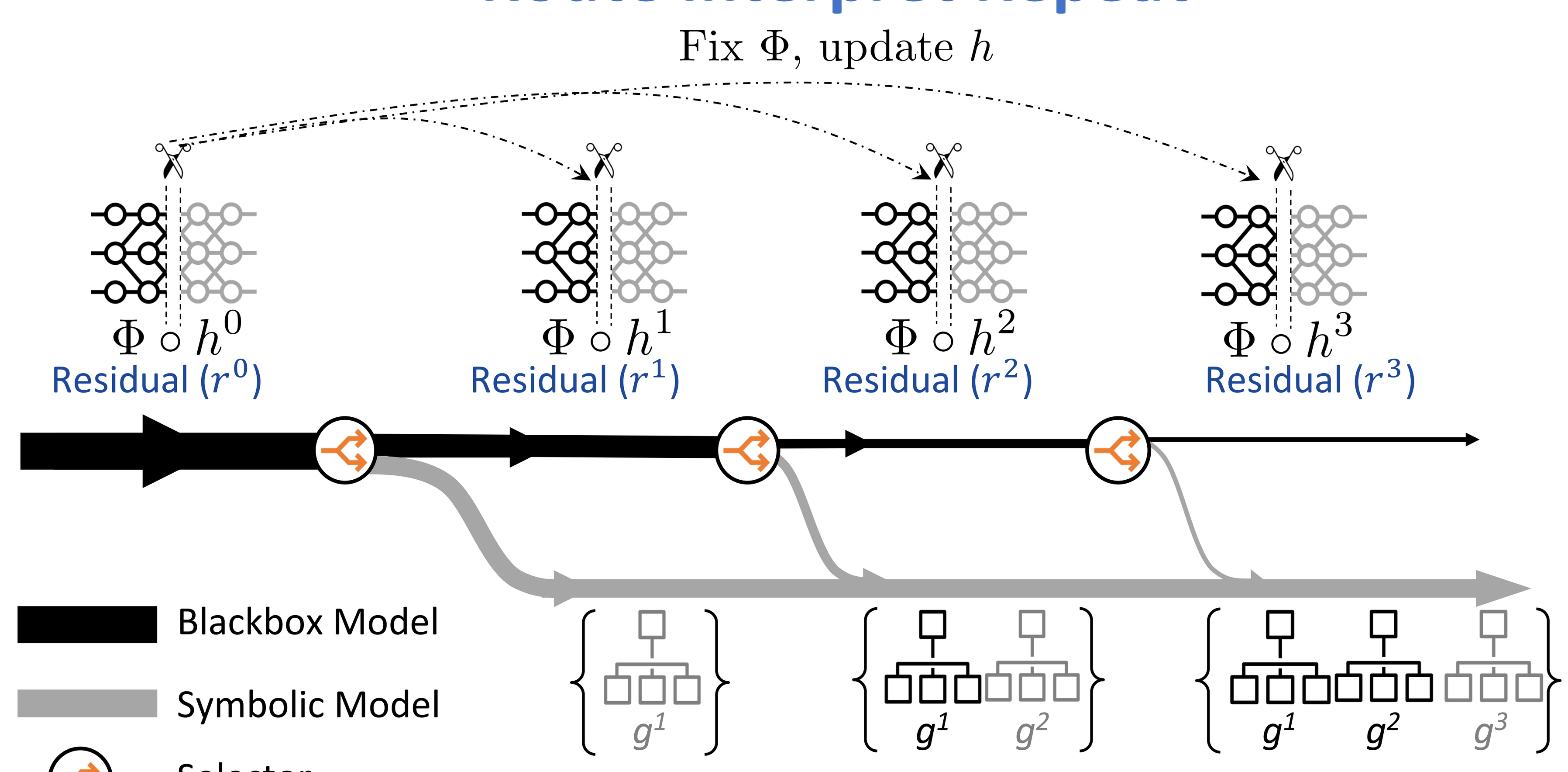
Assumption



Carve out interpretable models from Black box



Route Interpret Repeat



Each g is E-LEN [Barberio et al.], constructing FOL

- *** SelectiveNet [Geifman et al.] optimization**
- *** Continue till at least 90% samples covered**

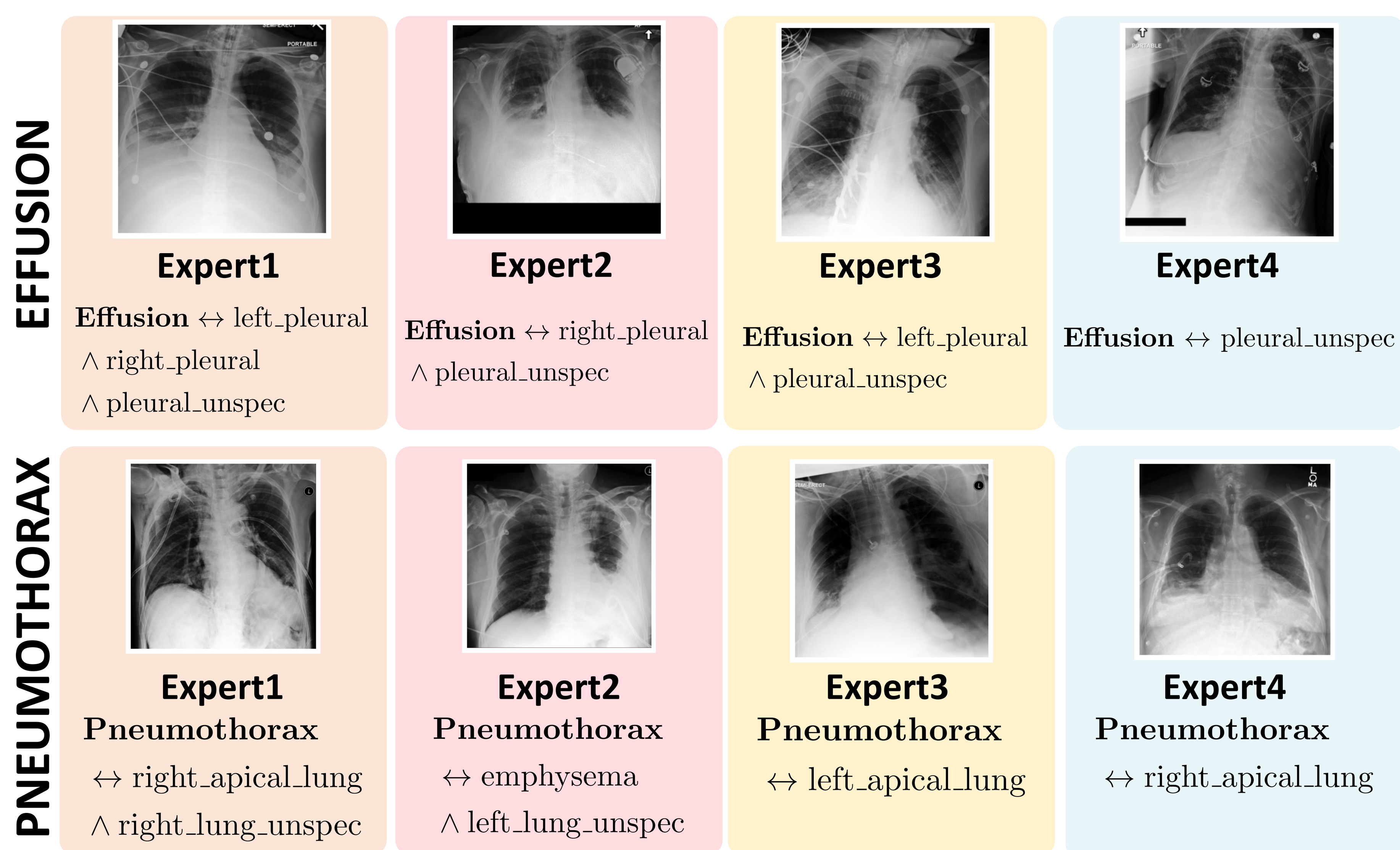
Extract concepts from MIMIC-CXR using Radgraph NLP pipeline

Ke Yu et al.

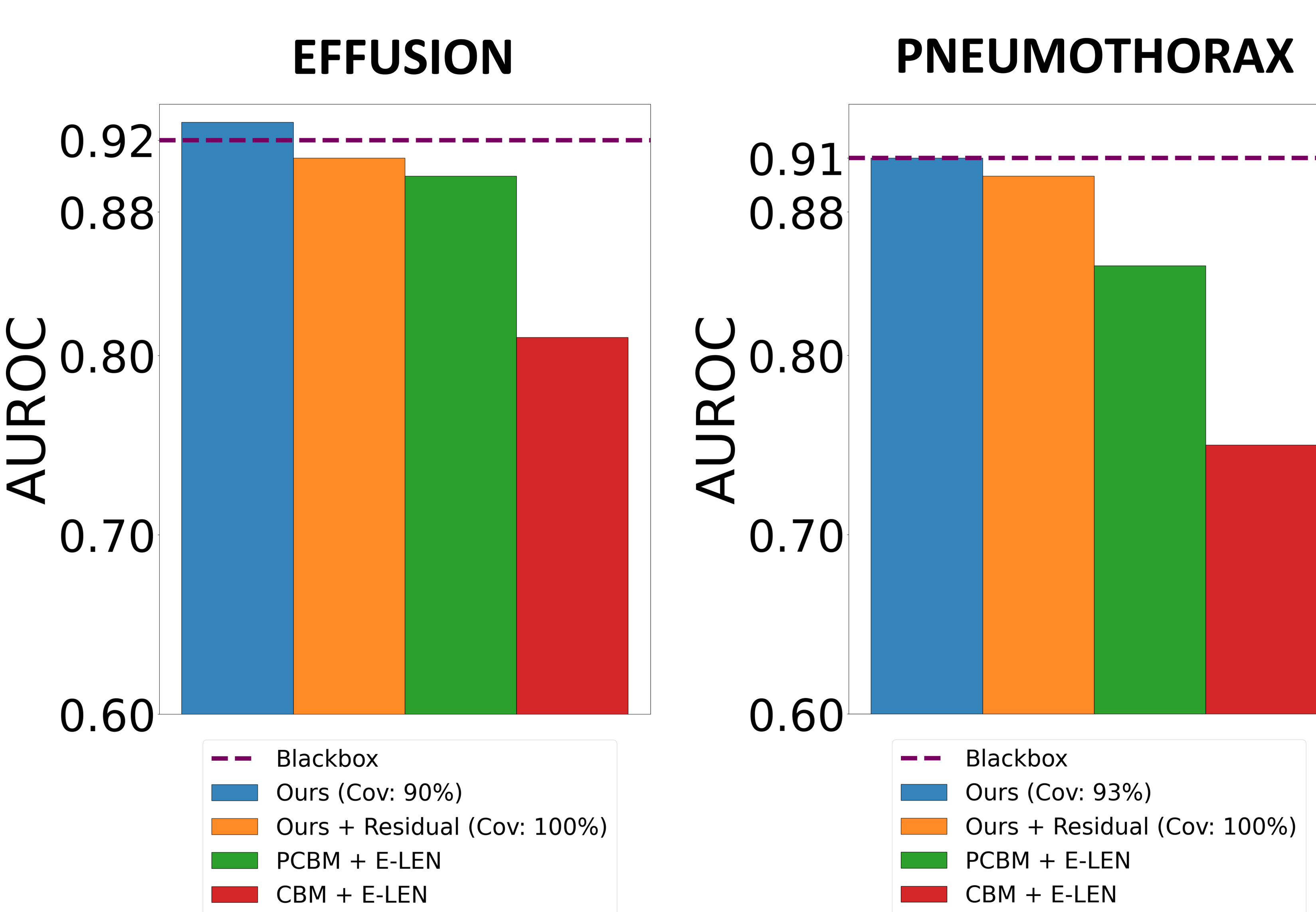
Report:

Right upper lobe consolidation with adjacent. While this may be infectious in nature, a CT scan is recommended for further clarification.

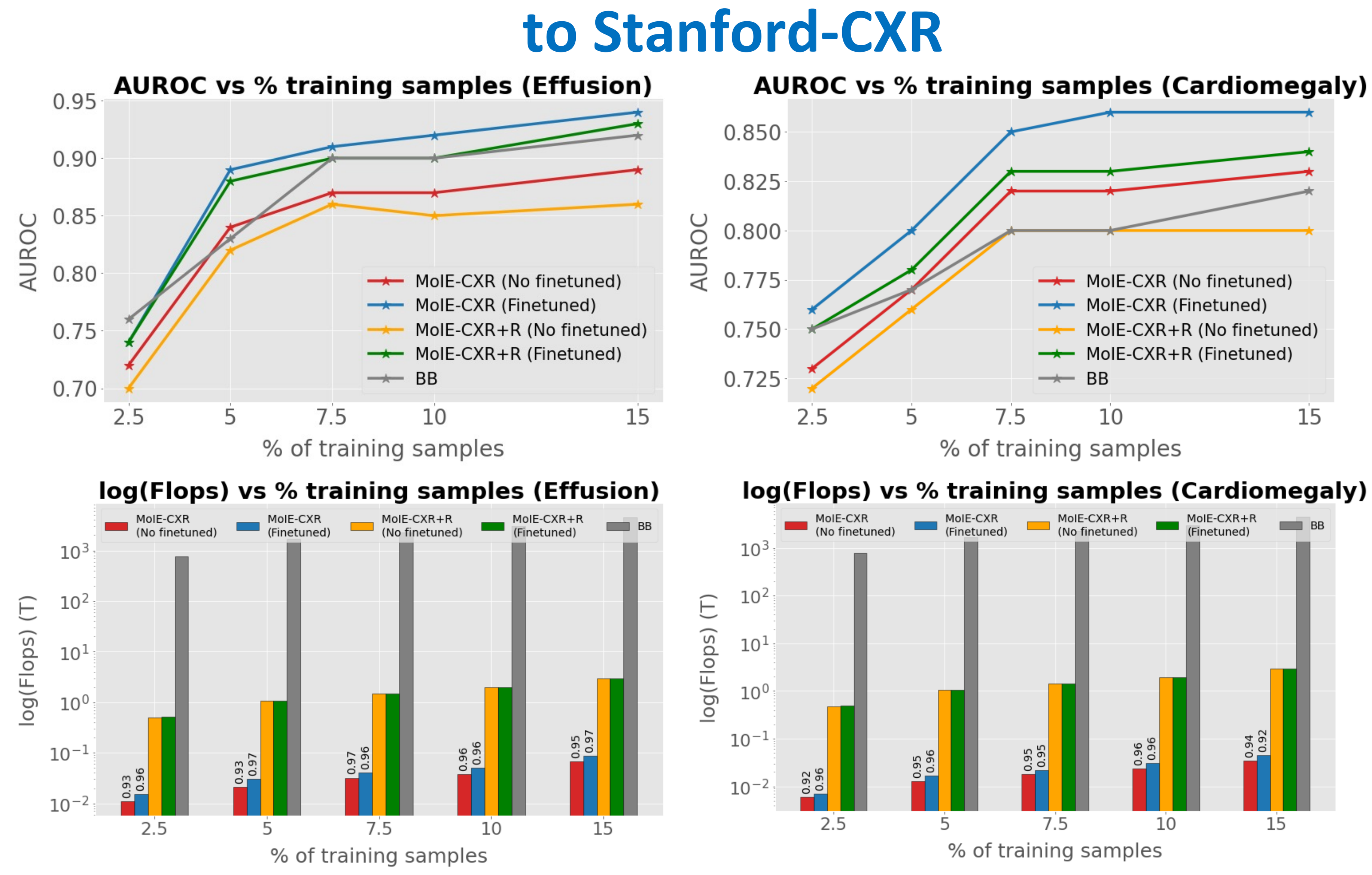
Diversity in local explanations



Not compromising the accuracy in MIMIC-CXR



Transferring the first 3 experts of MIMIC-CXR to Stanford-CXR



We will also be in MICCAI 2023 to present **Distilling BlackBox to Interpretable models for Efficient Transfer Learning** (Early accept, top ~ 14%)