Distilling Blackbox to Interpretable models for Efficient Transfer Learning

Shantanu Ghosh[1], Ke Yu[2], Kayhan Batmanghelich[1]

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

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

MICCAI 2023 Vancouver CANADA

BOSTON UNIVERSITY | BATMAN LAB

TLDR: Extracting a mixture of interpretable models from a BlackBox to provide concept-based explanations for efficient transfer learning.

Motivation

- Neural Networks fail to generalize due to scanner types, disease subtypes, patient subpopulation.
- Fine-tuning a Blackbox to a new domain can solve.
- This is data and computationally expensive.
- Whole process is not interpretable.

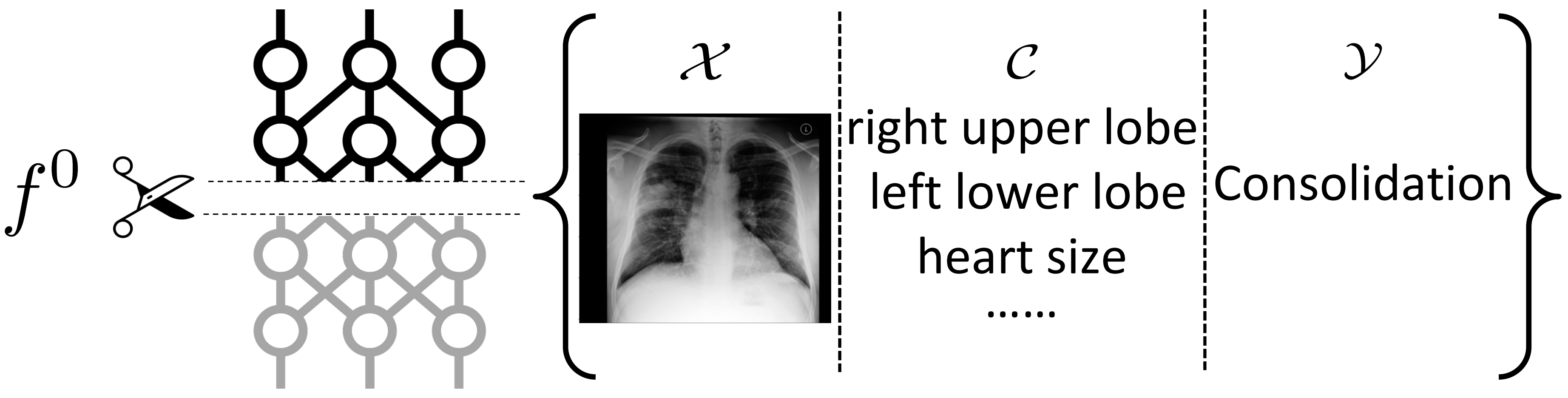
Approach by radiologist

- Search for patterns for anatomical changes to read abnormality.
- Apply generalizable logical rules for disease diagnosis.
- Whole process is interpretable.

Design choices

- Carve a mixture of interpretable models from Blackbox.
- Built on domain-invariant anatomical concepts.
- Transfer the interpretable models to an unseen domain without any concept annotation.

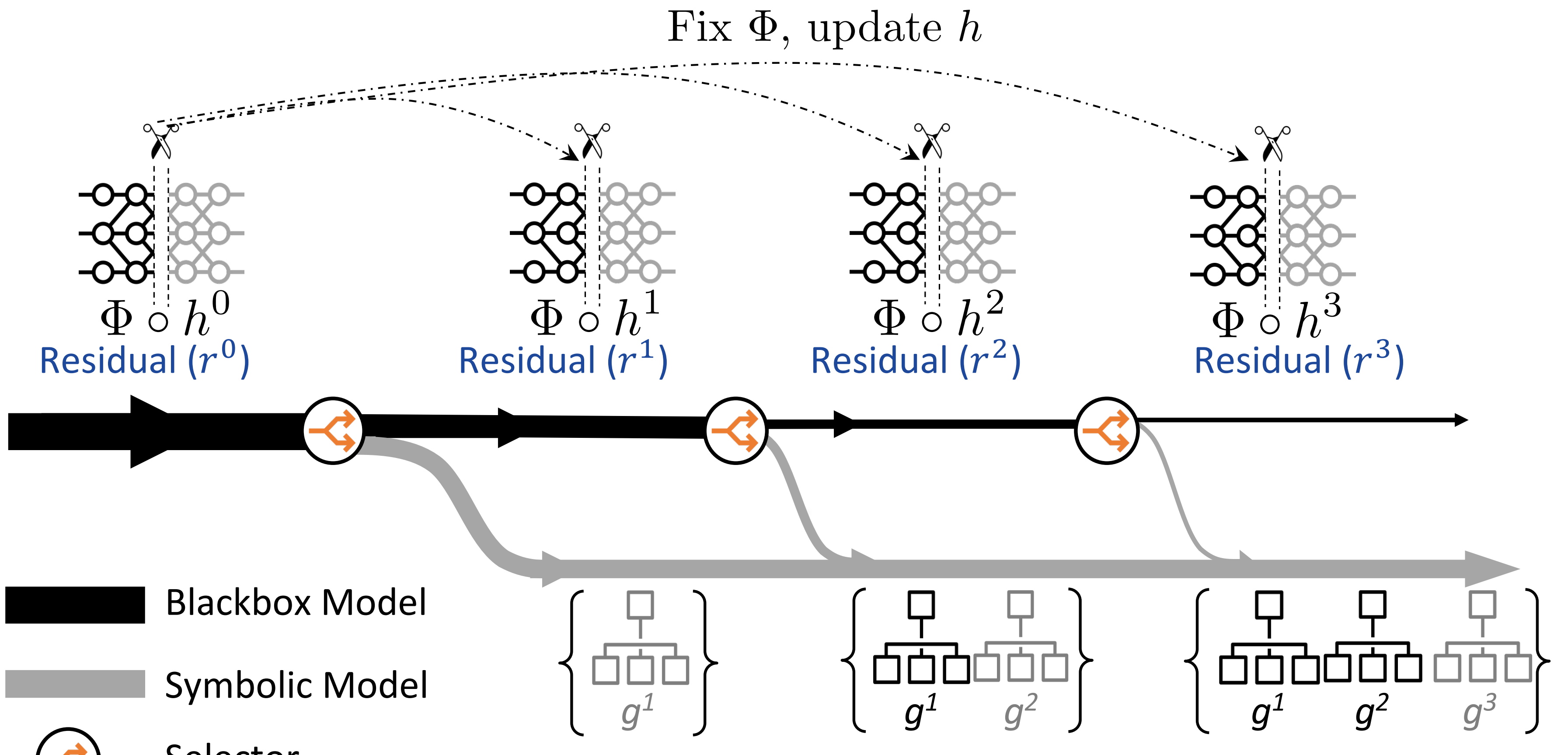
Assumption



right upper lobe
left lower lobe
heart size
……

Consolidation

Route Interpret Repeat



Residual (r^0) Residual (r^1) Residual (r^2) Residual (r^3)

Blackbox Model

Symbolic Model

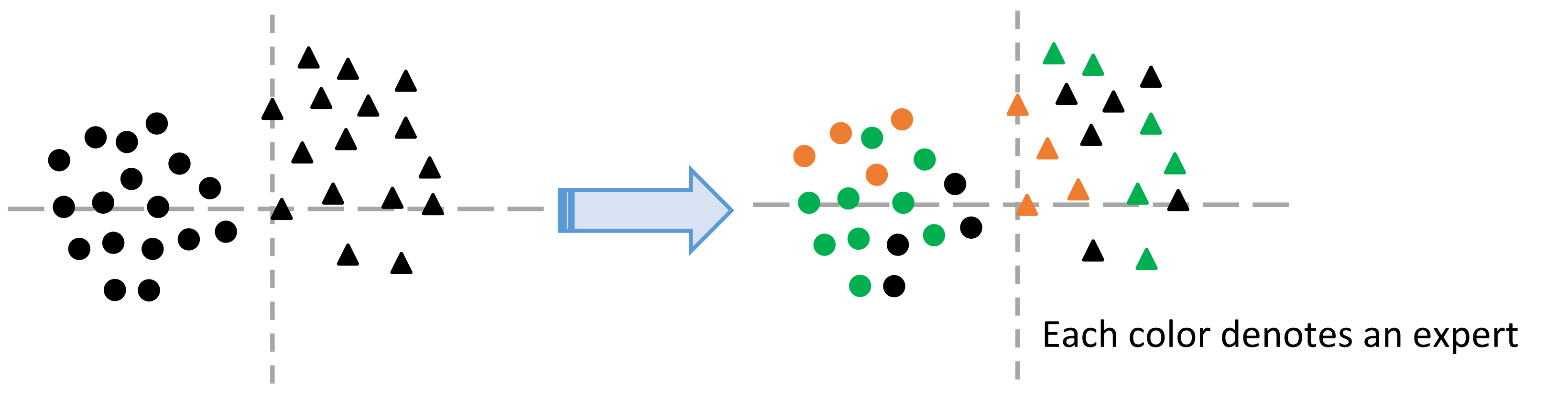
Selector

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

*** SelectiveNet [Geifman et al.] optimization**

*** Continue till at least 90% samples covered**

Carve out interpretable models from Black box



Each color denotes an expert

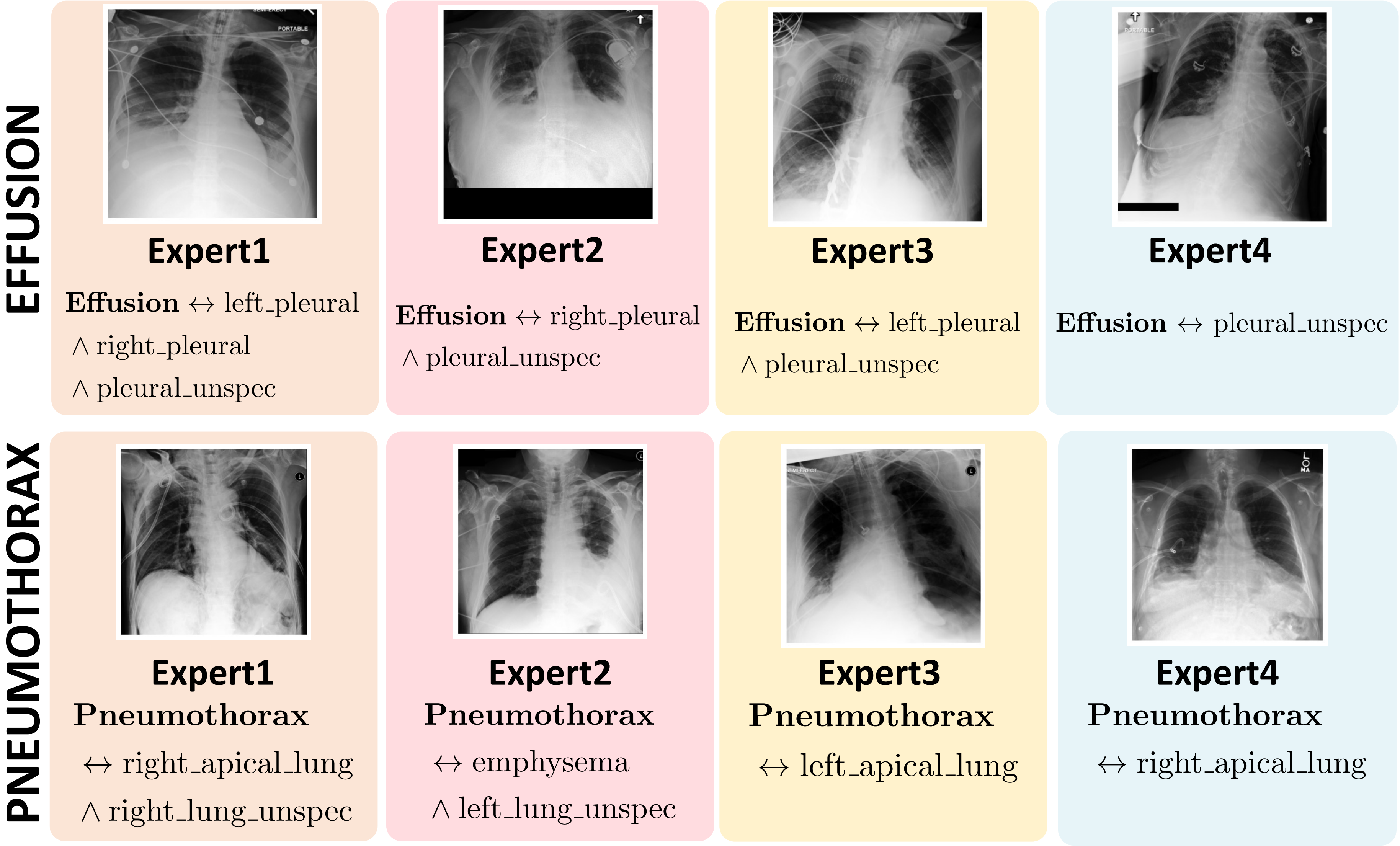
Extract concepts from MIMIC-CXR using Radgraph NLP pipeline

Ke Yu et al., MICCAI, 2022

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



EFFUSION

Expert1
Effusion ↔ left_pleural ∧ right_pleural ∧ pleural_unspec

Expert2
Effusion ↔ right_pleural ∧ pleural_unspec

Expert3
Effusion ↔ left_pleural ∧ pleural_unspec

Expert4
Effusion ↔ pleural_unspec

PNEUMOTHORAX

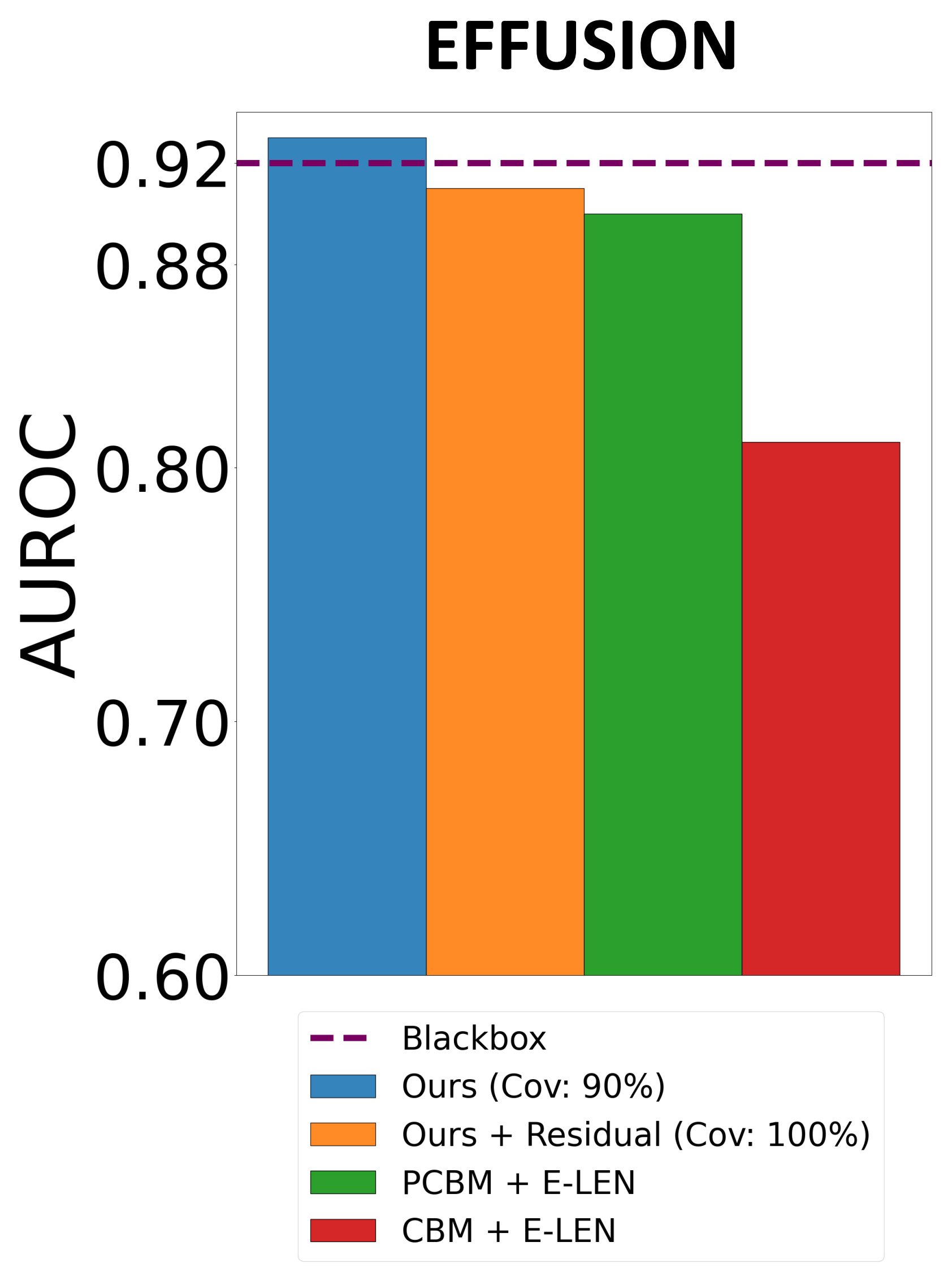
Expert1
Pneumothorax ↔ right_apical_lung ∧ right_lung_unspec

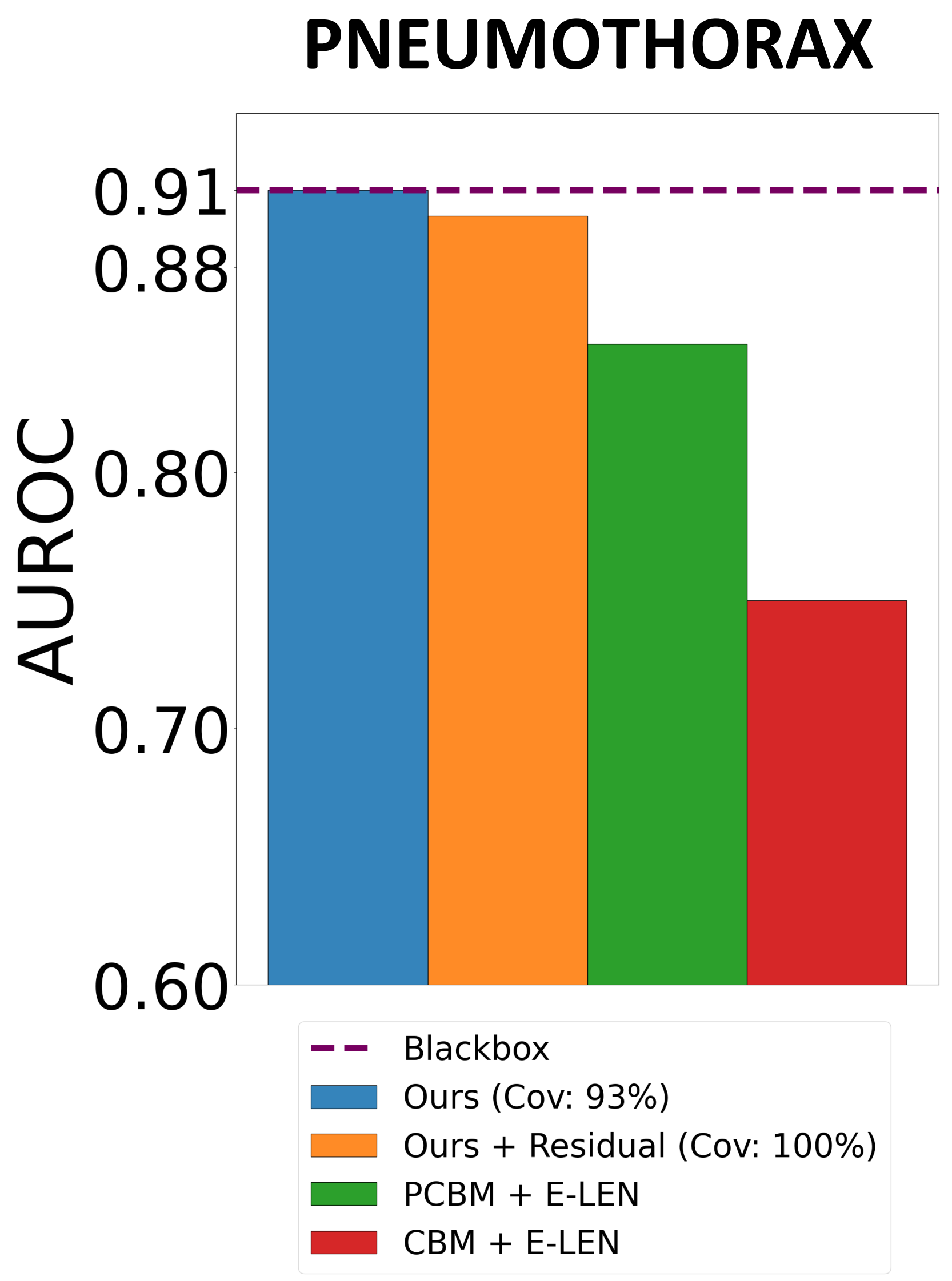
Expert2
Pneumothorax ↔ emphysema ∧ left_lung_unspec

Expert3
Pneumothorax ↔ left_apical_lung

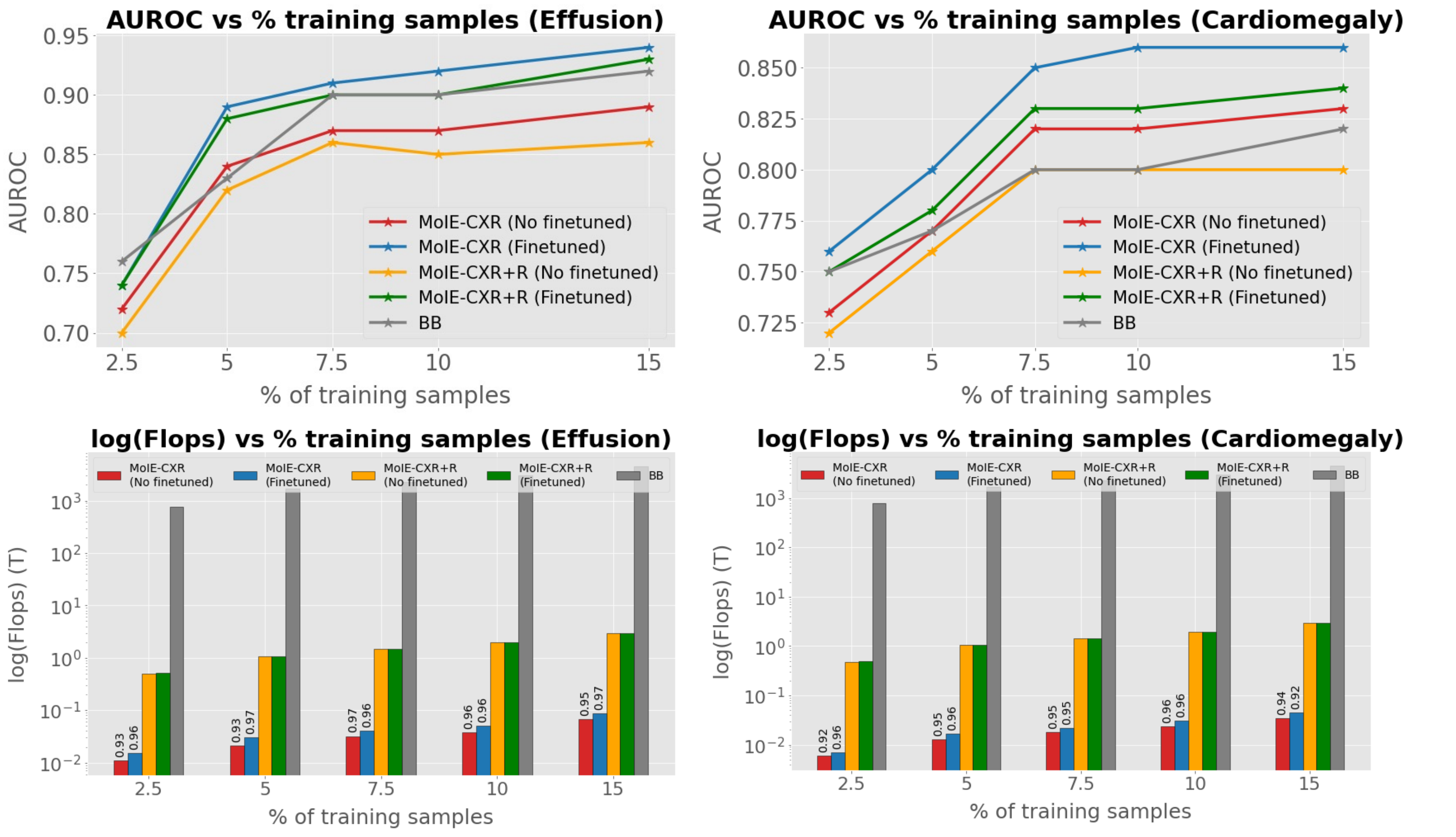
Expert4
Pneumothorax ↔ right_apical_lung

Not compromising the accuracy in MIMIC-CXR





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


AUROC vs % training samples (Effusion)
AUROC vs % training samples (Cardiomegaly)
log(Flops) vs % training samples (Effusion)
log(Flops) vs % training samples (Cardiomegaly)
AUROC
log(Flops) (T)
% of training samples
MoIE-CXR (No finetuned)
MoIE-CXR (Finetuned)
MoIE-CXR+R (No finetuned)
MoIE-CXR+R (Finetuned)
BB