Interpretable Medical AI with Vision-Language Alignment

Shantanu Ghosh, PhD Prospectus

Jan 28, 2026

BOSTON UNIVERSITY | BATMAN LAB

The goal: **Detect** the **Systematic Mistakes**

Why?

To design better mitigation strategies(data collection, reweighting, architecture/training changes etc) to enhance **robustness**.

Dissecting Systematic Mistakes?

Class: Waterbirds
Background: **Water**

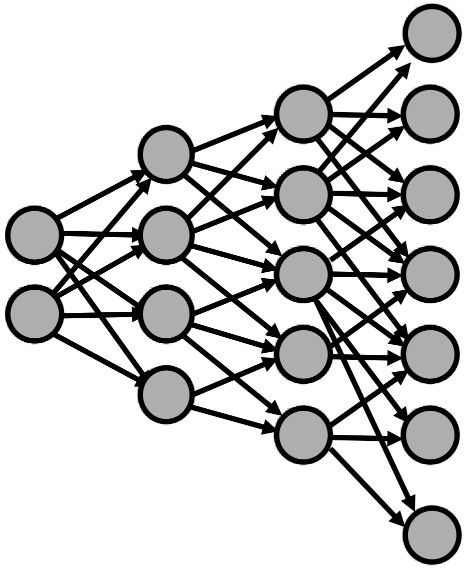
Class: Waterbirds
Background: **Land**

Class: Landbirds
Background: **Water**

Class: Landbirds
Background: **Land**

Distributionally Robust Neural Networks for Group Shifts:
On the Importance of Regularization for Worst-Case Generalization.
Sagawa et al. ICLR 2020

Dissecting Systematic Mistakes?



ResNet50

Mean Accuracy:

88.6%

Class: **Waterbirds**

Background: **Water**

Accuracy: **94.2%**

Class: **Waterbirds**

Background: **Land**

Accuracy: **68.8%**

Class: **Landbirds**

Background: **Water**

Accuracy: **80.2%**

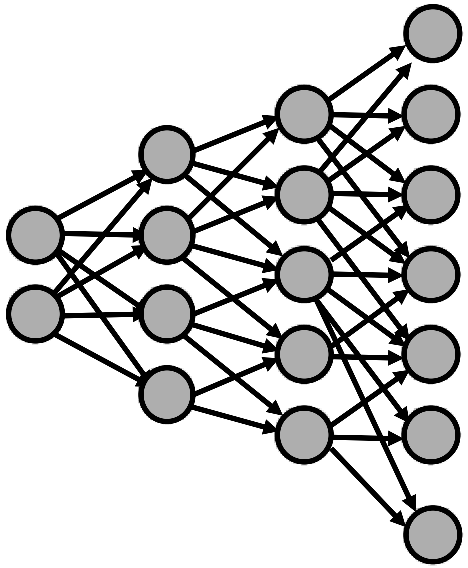
Class: **Landbirds**

Background: **Land**

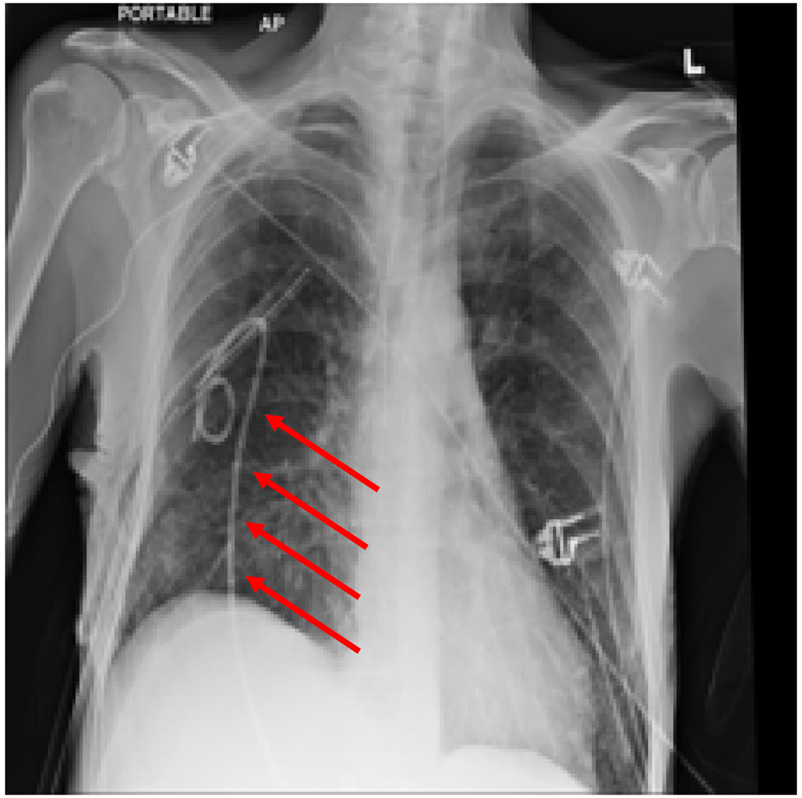
Accuracy: **99.6%**

Distributionally Robust Neural Networks for Group Shifts:
On the Importance of Regularization for Worst-Case Generalization.
Sagawa et al. ICLR 2020

Dissecting Systematic Mistakes?



ResNet50



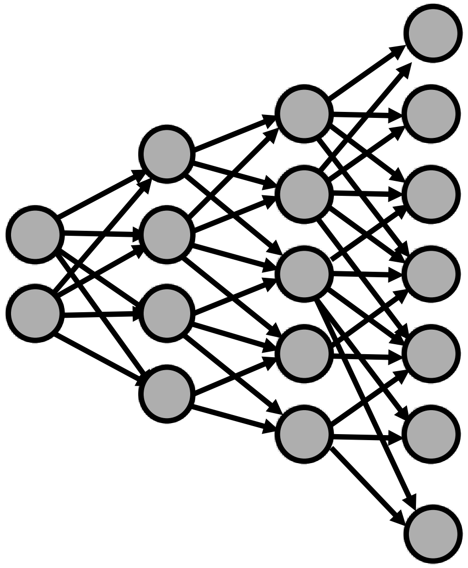
Class: **Pneumothorax**
Correlation: **Chest tube**

Class: **Pneumothorax**
Correlation: **Chest tube**

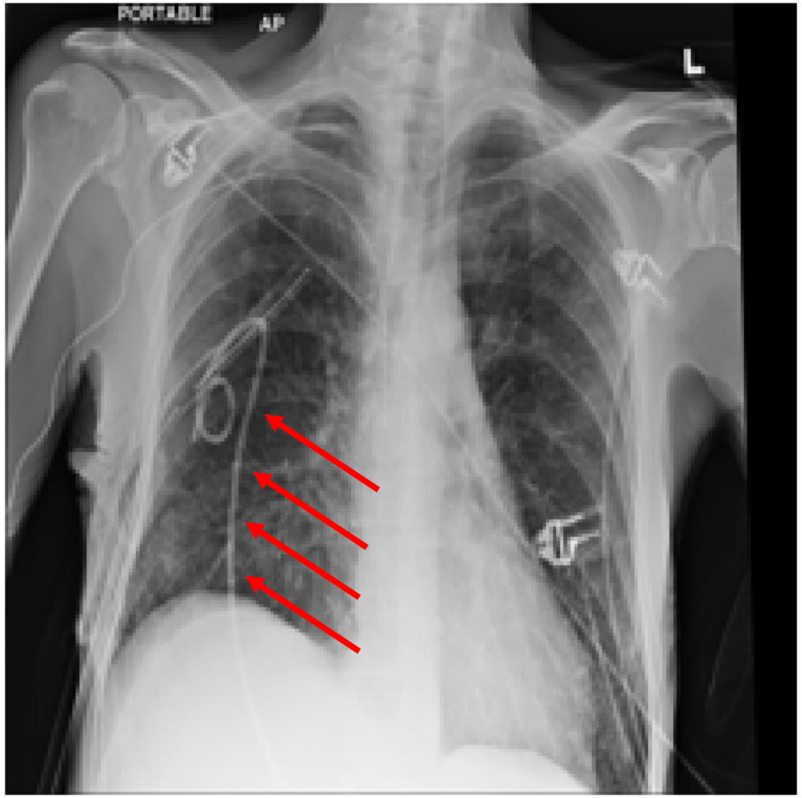
Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. Rayner et al. ML4H at NeurIPS 2019

Dissecting Systematic Mistakes?

Mean Accuracy (Pneumothorax): ~70%



ResNet50

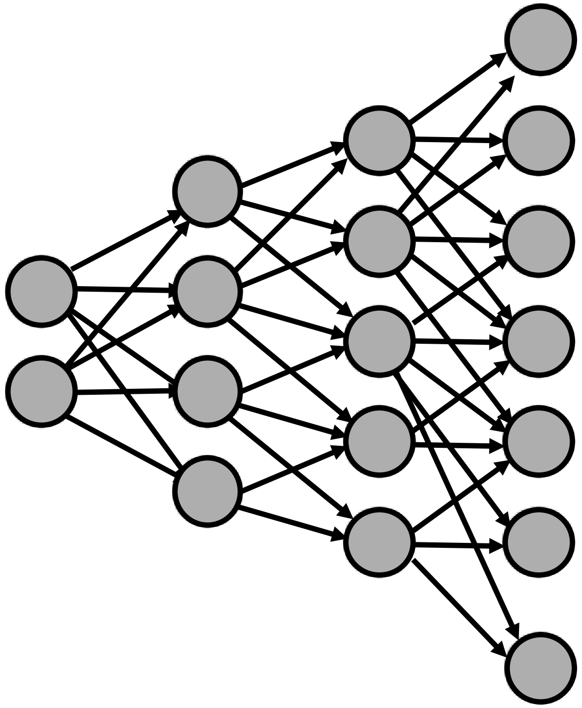


Class: **Pneumothorax**
Correlation: **Chest tube**
Accuracy: **90.4%**

Class: **Pneumothorax**
Correlation : **Chest tube**
Accuracy: **60.2%**

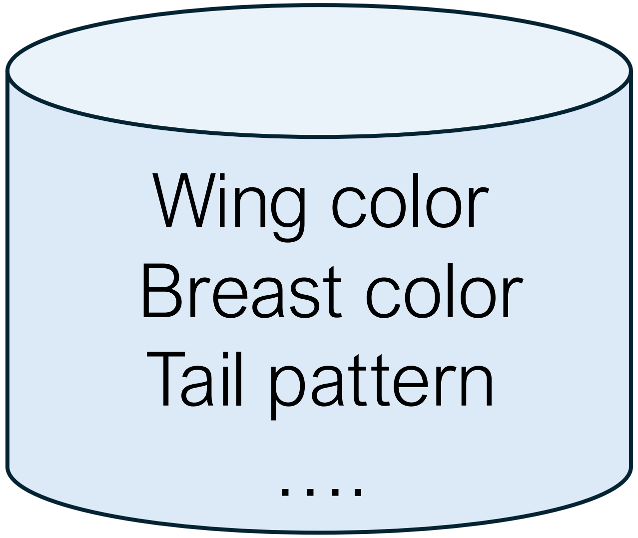
Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. Rayner et al. ML4H at NeurIPS 2019

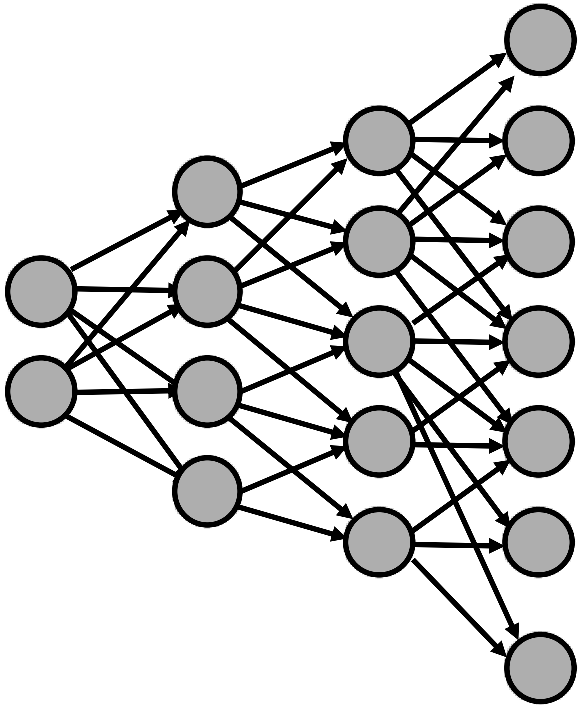
How to detect? (Aim 1)



How to detect? (Aim 1)

Concept bank

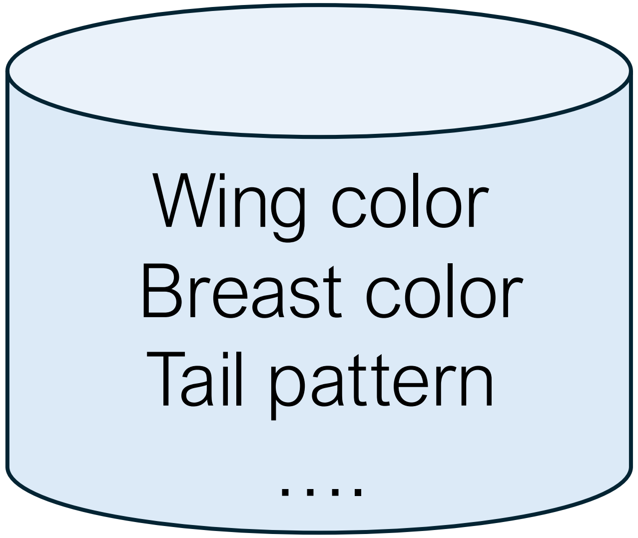


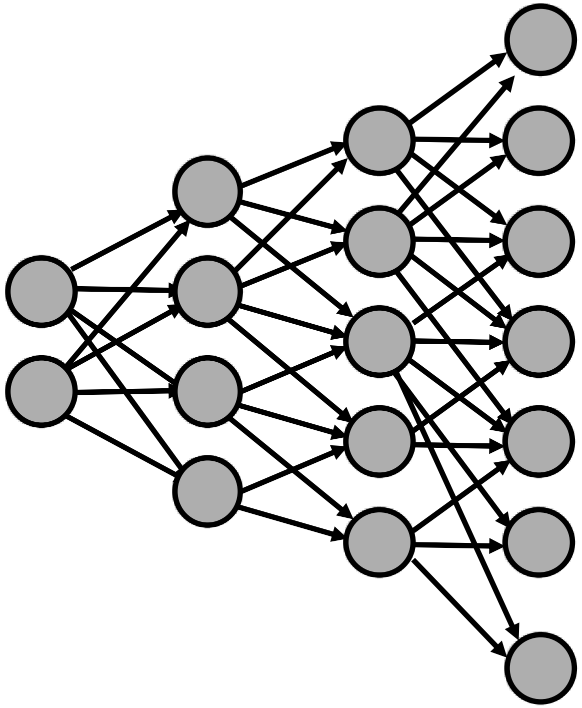


How to detect? (Aim 1)

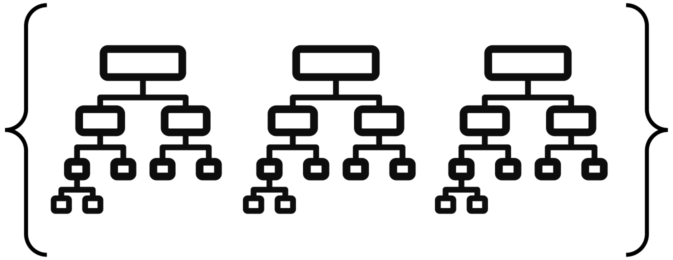
Concept bank

Wing color
Breast color
Tail pattern
....



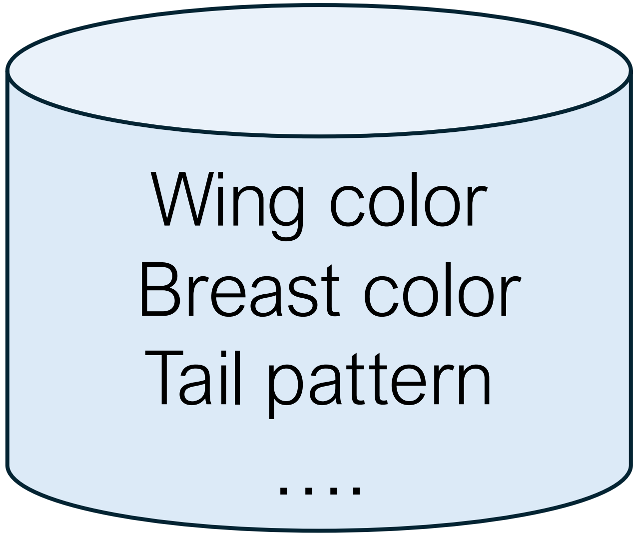


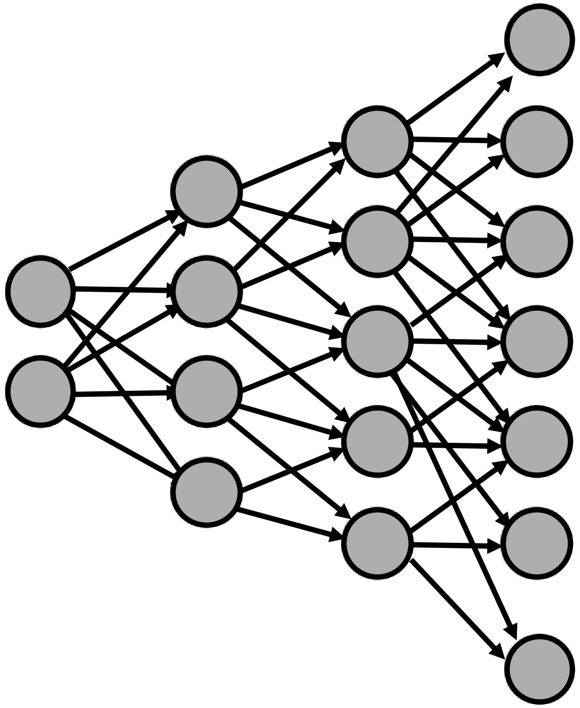
$\textbf{Olive sided Flycatcher} \leftrightarrow \text{breast\_color\_grey} \wedge \text{tail\_pattern\_solid}$



The goal: **Detect** the **Systematic Mistakes**

Concept bank

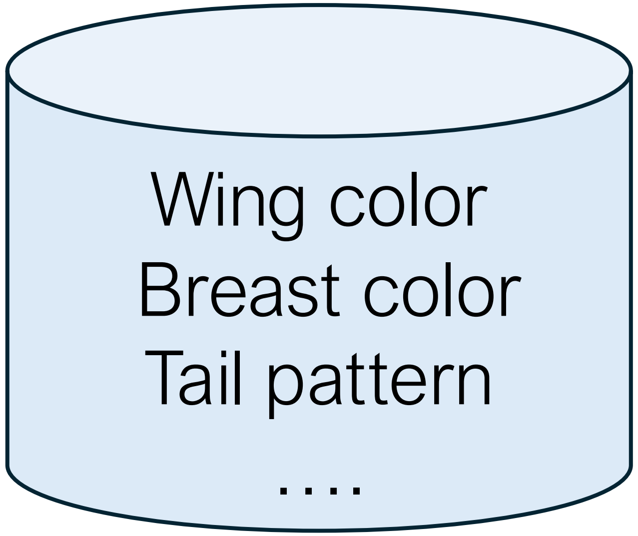




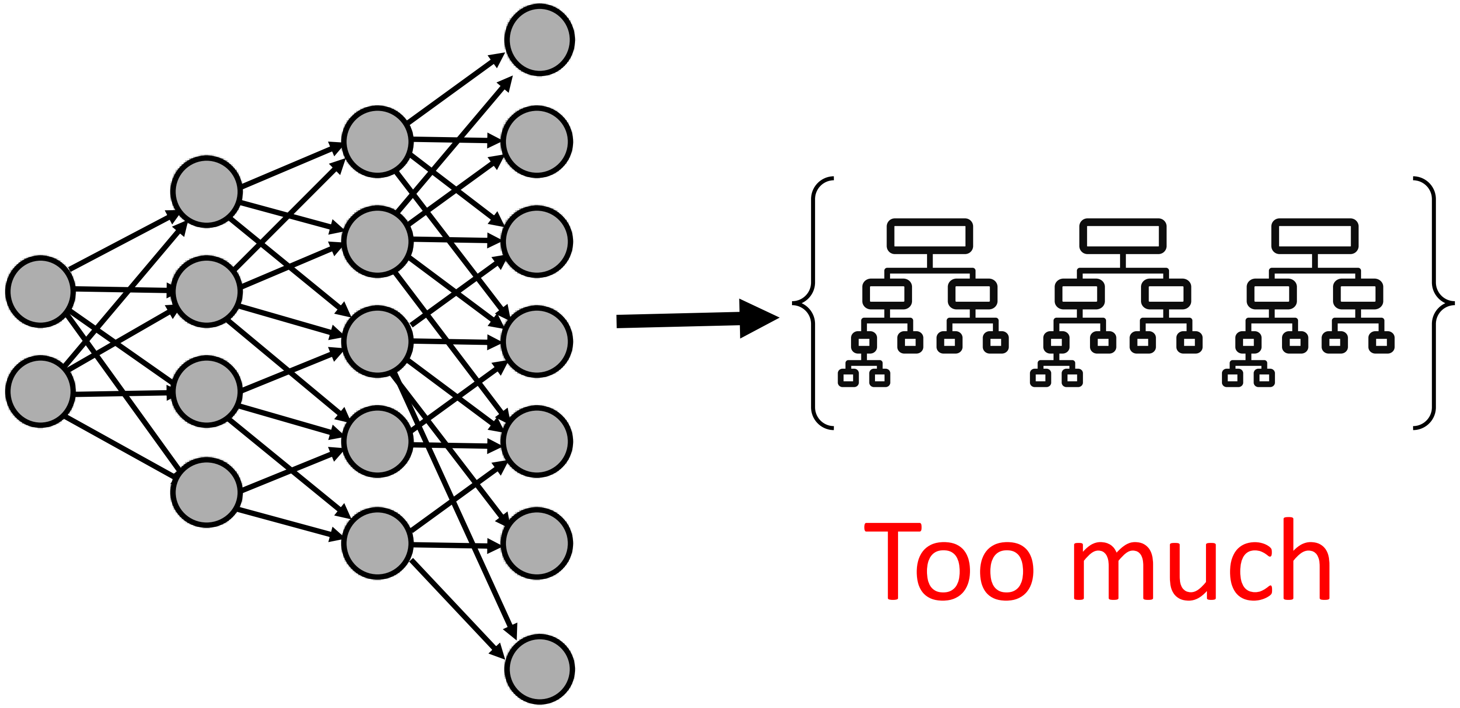
Expensive

The goal: **Detect** the **Systematic Mistakes**

Concept bank



Expensive



Too much

Captions

1. A large seagull stands on a dock against a backdrop of a harbor with boats and a blue sky
2. A digitally altered image features a large bird, possibly an albatross, superimposed over a backdrop of industrial buildings by a body of water
3. A seagull stands on rocks by the water at sunset, with a lighthouse visible in the background

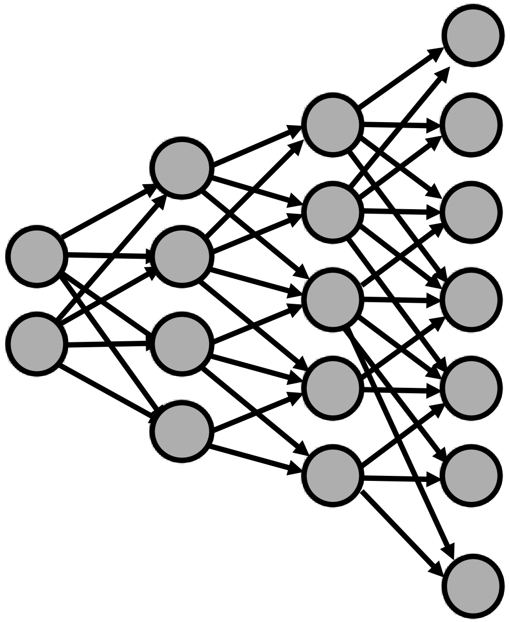
Report

1. perhaps mild increase in hydropneumothorax but with chest tube
2. other less likely possibility include expansion of known loculated hydropneumothorax (chest tube does not appear to be draining this region)
3. one of two right - sided pleural tubes has been removed in the interval
4. 3 chest tubes remain in place and there is again an area of hydro pneumothorax

How to detect? (Aim 2 & 3)

Free-text

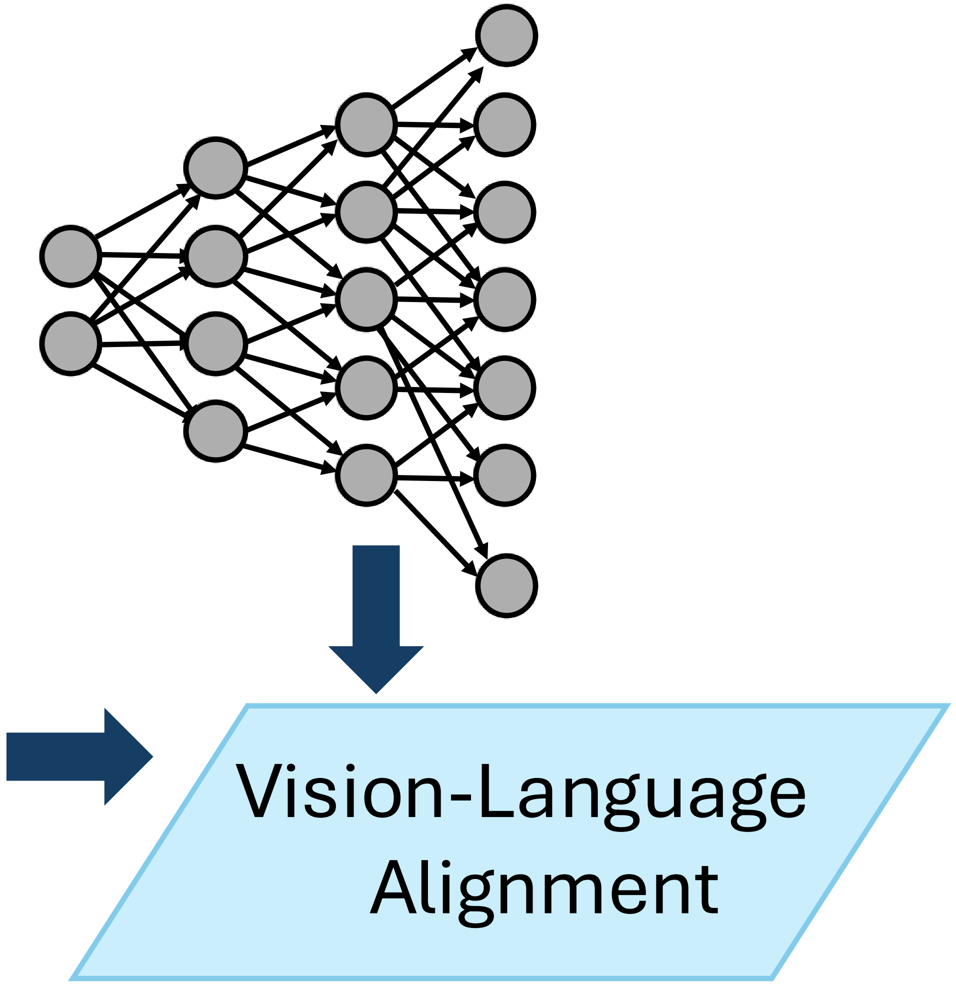
1. A large seagull stands on a dock against a backdrop of a harbor with boats and a blue sky
2. A digitally altered image features a large bird, possibly an albatross, superimposed over a backdrop of industrial buildings by a body of water
3. A seagull stands on rocks by the water at sunset, with a lighthouse visible in the background



How to detect? (Aim 2 & 3)

Free-text

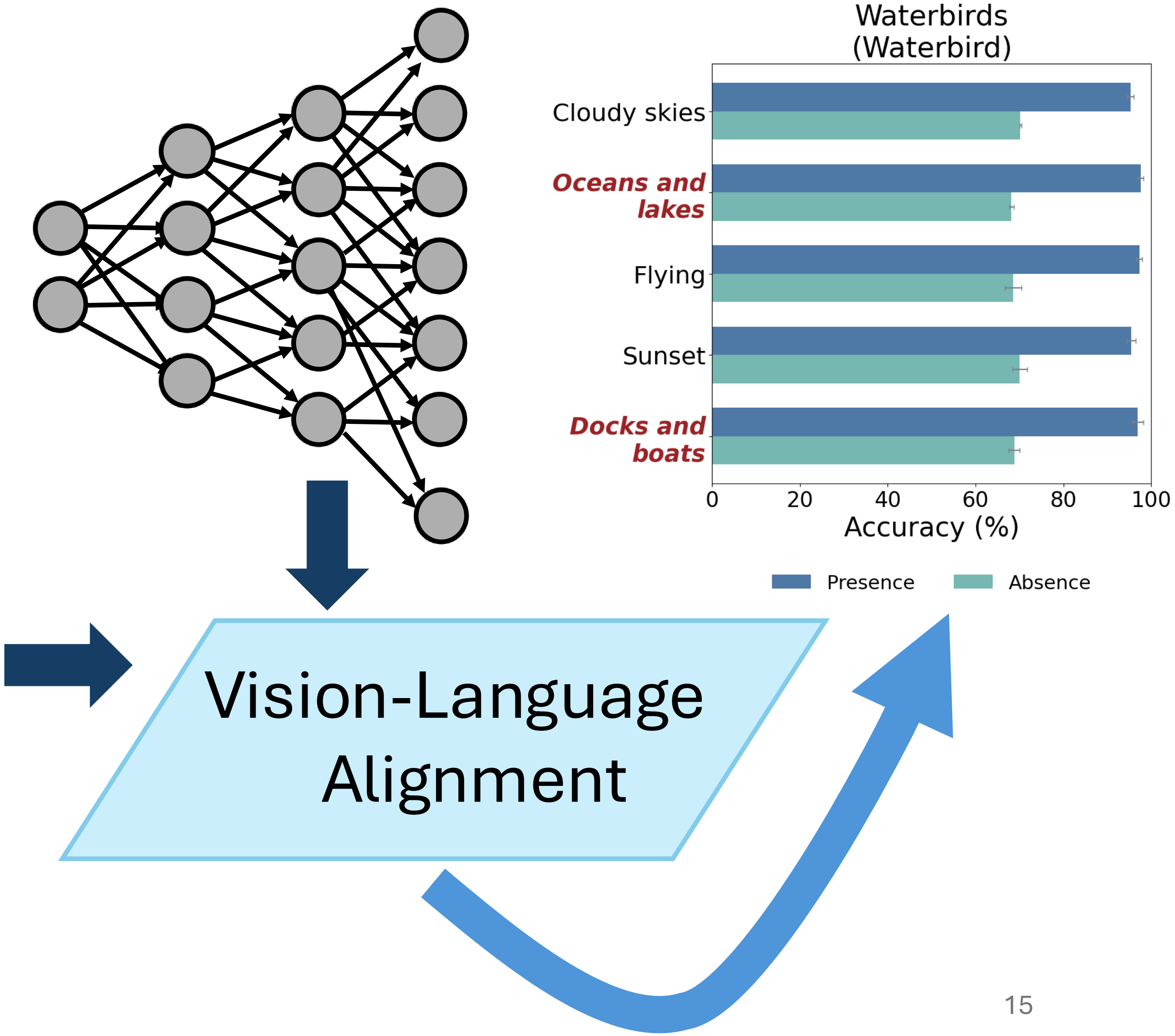
1. A large seagull stands on a dock against a backdrop of a harbor with boats and a blue sky
2. A digitally altered image features a large bird, possibly an albatross, superimposed over a backdrop of industrial buildings by a body of water
3. A seagull stands on rocks by the water at sunset, with a lighthouse visible in the background



How to detect? (Aim 2 & 3)

Free-text

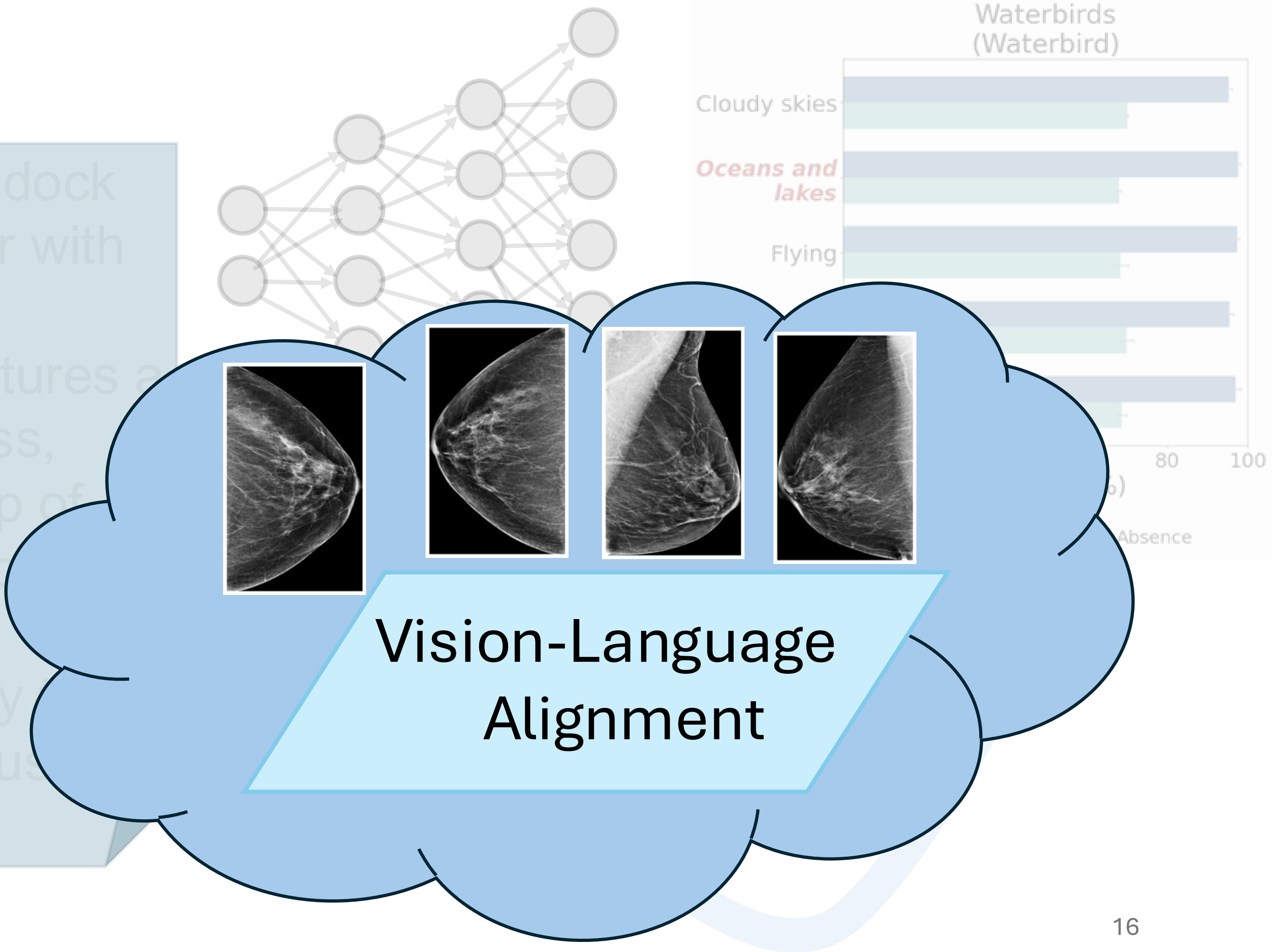
1. A large seagull stands on a dock against a backdrop of a harbor with boats and a blue sky
2. A digitally altered image features a large bird, possibly an albatross, superimposed over a backdrop of industrial buildings by a body of water
3. A seagull stands on rocks by the water at sunset, with a lighthouse visible in the background



How to detect? (Aim 2)

Free-text

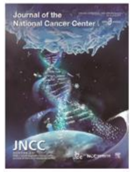
1. A large seagull stands on a dock against a backdrop of a harbor with boats and a blue sky
2. A digitally altered image features a large bird, possibly an albatross, superimposed over a backdrop of industrial buildings by a body of water
3. A seagull stands on rocks by water at sunset, with a lighthous visible in the background



Contents lists available at ScienceDirect

Journal of the National Cancer Center

journal homepage: www.elsevier.com/locate/jncc



Full Length Article

Global burden of female breast cancer: new estimates in 2022, temporal trend and future projections up to 2050 based on the latest release from GLOBOCAN

Yunmeng Zhang[1,†], Yuting Ji[1,†], Siwen Liu[1], Jingjing Li[1], Jie Wu[1], Qianyun Jin[1], Xiaomin Liu[1], Hongyuan Duan[1], Zhuowei Feng[1], Ya Liu[1], Yacong Zhang[2], Zhangyan Lyu[1], Fangfang Song[1], Fengju Song[1], Lei Yang[3], Hong Liu[4,*], Yubei Huang[1,*]

Results: In 2022, an estimated 2.3 million new BC cases and 666,000 BC-related deaths occurred globally, accounting for 23.8 % and 15.4 % of all cancer cases and deaths in women, respectively. Regionally, Eastern Asia

Effect of mammographic screening from age 40 years on breast cancer mortality (UK Age trial): final results of a randomised, controlled trial

Stephen W Duffy, Daniel Vulkan*, Howard Cuckle, Dharmishta Parmar, Shama Sheikh, Robert A Smith, Andrew Evans, Oleg Blyuss, Louise Johns, Ian O Ellis, Jonathan Myles, Peter D Sasieni*, Sue M Moss**

Summary

Background The appropriate age range for breast cancer screening remains a matter of debate. We aimed to estimate

Lancet Oncol 2020; 21: 1165–72

Challenges of early screening

Annals of Internal Medicine
ORIGINAL RESEARCH

Estimation of Breast Cancer Overdiagnosis in a U.S. Breast Screening Cohort

Marc D. Ryser, PhD; Jane Lange, PhD; Lurdes Y.T. Inoue, PhD; Ellen S. O'Meara, PhD; Charlotte Gard, PhD; Diana L. Miglioretti, PhD; Jean-Luc Bulliard, PhD; Andrew F. Brouwer, PhD; E. Shelley Hwang, MD, MPH; and Ruth B. Etzioni, PhD

ORIGINAL ARTICLE Open Access

Check for updates

Workload of diagnostic radiologists in the foreseeable future based on recent scientific advances: growth expectations and role of artificial intelligence

Thomas C. Kwee[1*] and Robert M. Kwee[2]

A Systematic Review of Fatigue in Radiology: Is It a Problem?

Nadia Stec[1]
Danielle Arje[1]
Alan R. Moody[1]
Elizabeth A. Krupinski[2]
Pascal N. Tyrrell[1,3]

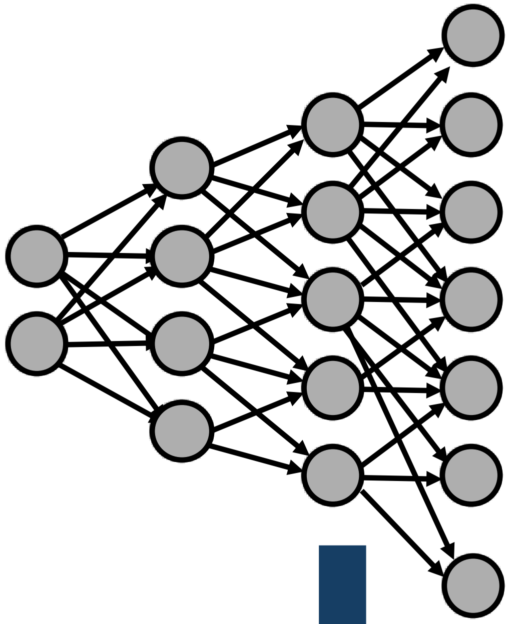
OBJECTIVE. The purpose of this study was to review current literature regarding radiologist fatigue.

MATERIALS AND METHODS. A literature search was performed using PubMed. Key words and Medical Subject Heading terms were used to generate refined queries with inclusion and exclusion criteria, focusing on fatigue and error. Results were selected according to these criteria: examined radiologist fatigue and radiologic error stemming from fatigue; experimental results measured as accuracy, error, or performance; and peer-reviewed publication. The risk of bias was addressed by including both quantitative and qualitative studies.

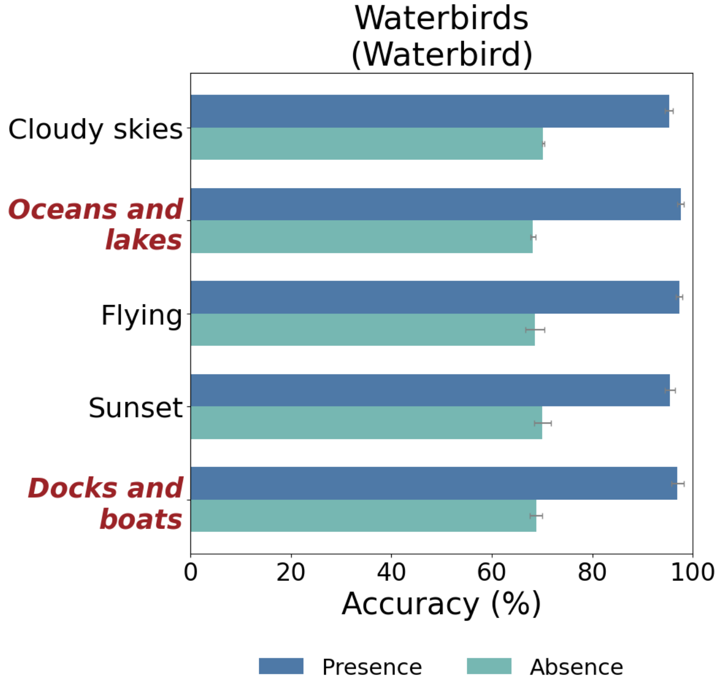
How to detect? (Aim 3)

Free-text

1. A large seagull stands on a dock against a backdrop of a harbor with boats and a blue sky
2. A digitally altered image features a large bird, possibly an albatross, superimposed over a backdrop of industrial buildings by a body of water
3. A seagull stands on rocks by the water at sunset, with a lighthouse visible in the background



Vision-Language Alignment



Aim 1

The goal: Extract mixture of **Interpretable** models from the Blackbox **Post-hoc** using **FOL**

Why Post-hoc

☑ Does not alter the Black box

✕ No intervention

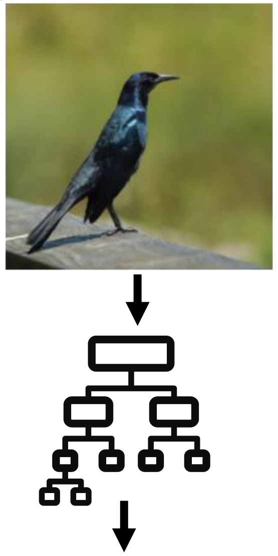
Why Post-hoc

✓ Does not alter the Black box

✗ No intervention

Why Interpretable

✓ Supports interventions



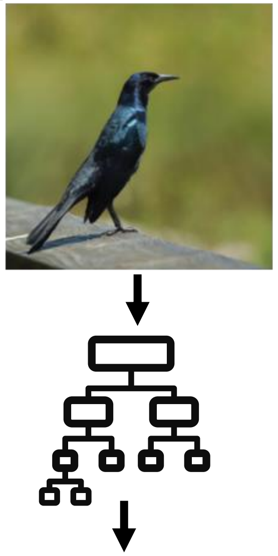
Prediction: Brewer Blackbird ✗

Why Post-hoc

- ✓ Does not alter the Black box
- ✗ No intervention

Why Interpretable

- ✓ Supports interventions



Prediction: Brewer Blackbird ✗

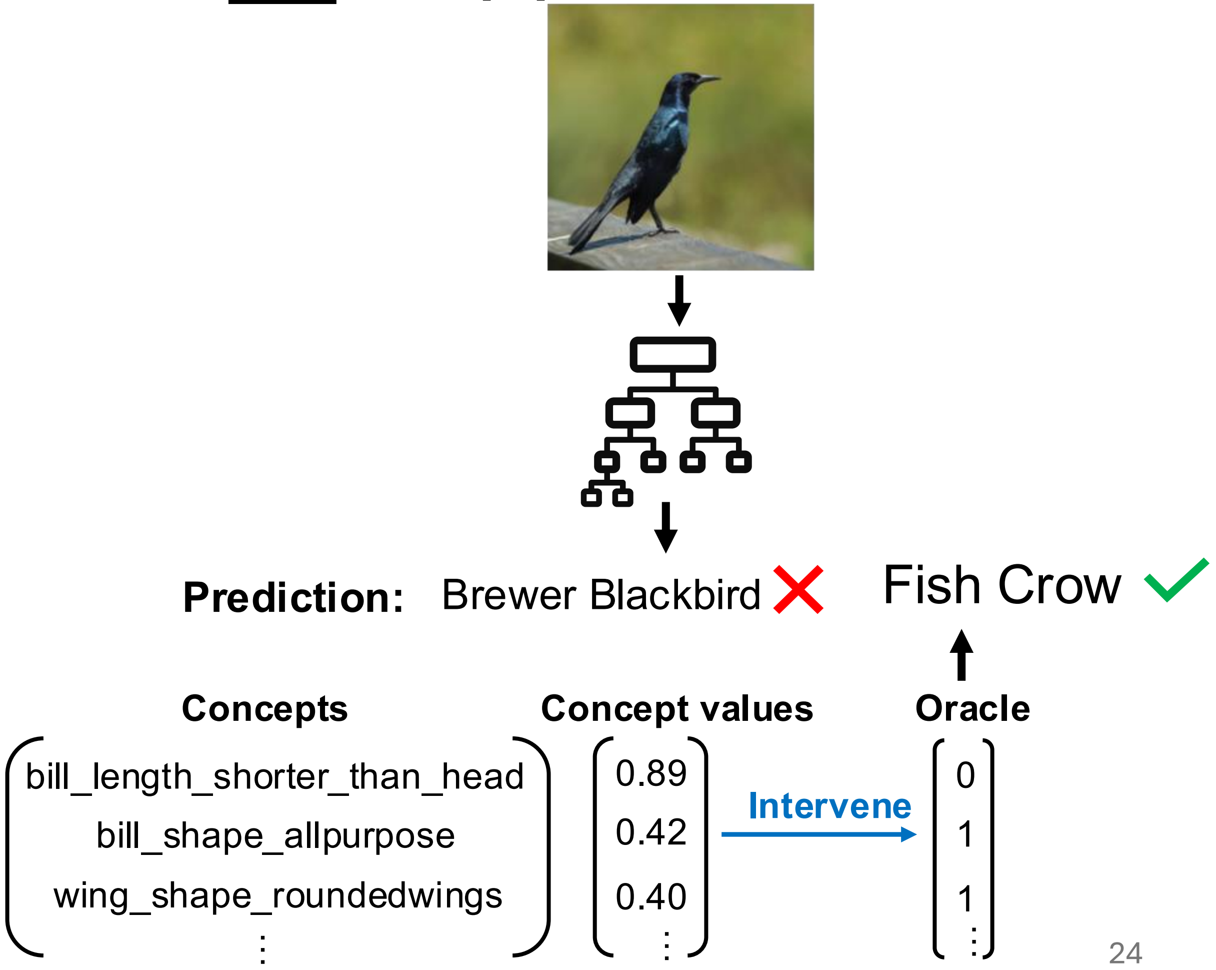
Concepts	**Concept values**
bill_length_shorter_than_head	0.89
bill_shape_allpurpose	0.42
wing_shape_roundedwings	0.40
⋮	⋮

Why Post-hoc

- ✔ Does not alter the Black box
- ✘ No intervention

Why Interpretable

- ✔ Supports interventions



Why Post-hoc

☑ Does not alter the Black box

✕ No intervention

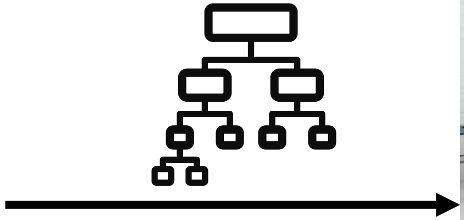
Why Interpretable

☑ Supports interventions

Why FOL?

$$\textbf{Cardiomegaly} \leftrightarrow \text{heart\_size} \land \text{enlarge}$$

Site A



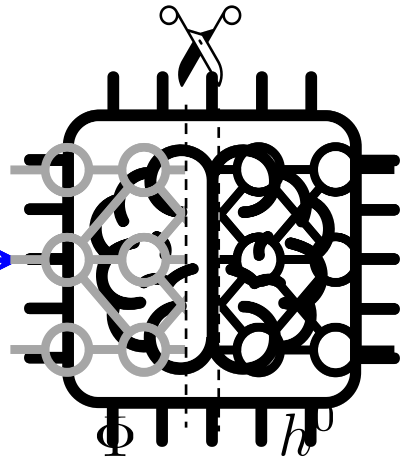
Site B

Problem Set Up

$\mathcal{X}$

$\mathcal{C}$

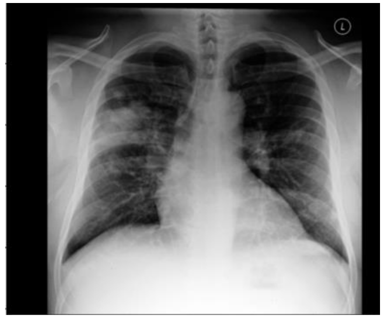
Wing color grey
Breast color white
Tail pattern
....



$\mathcal{Y}$

Olive
Sided
Flycatcher

Problem Set Up

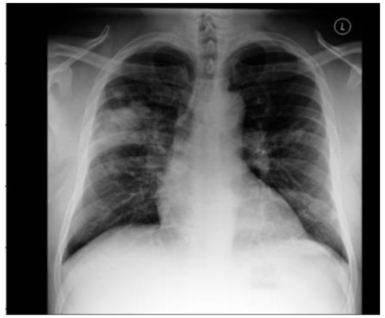


+

Report:

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

Problem Set Up



+

Report:

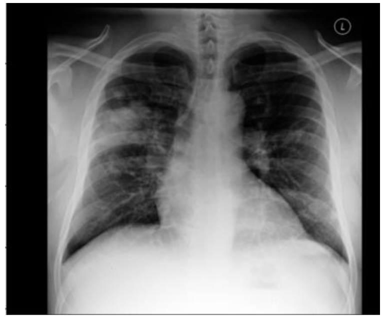
Right upper lobe consolidation with adjacent. While this may be infectious in nature, a CT scan is recommended for further clarification.

parse the reports to get the concepts

$$\left\{ \quad \mathcal{C} \quad \right\}$$

right upper lobe
left lower lobe
heart size
....

Problem Set Up



+

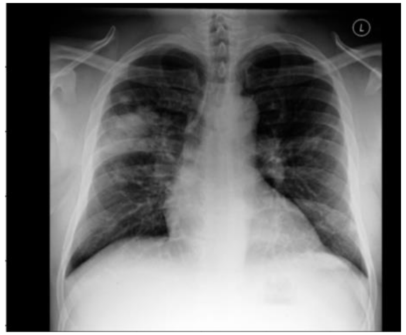
Report:

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

parse the reports to get the concepts

$\{$

$\mathcal{X}$



$\mathcal{C}$

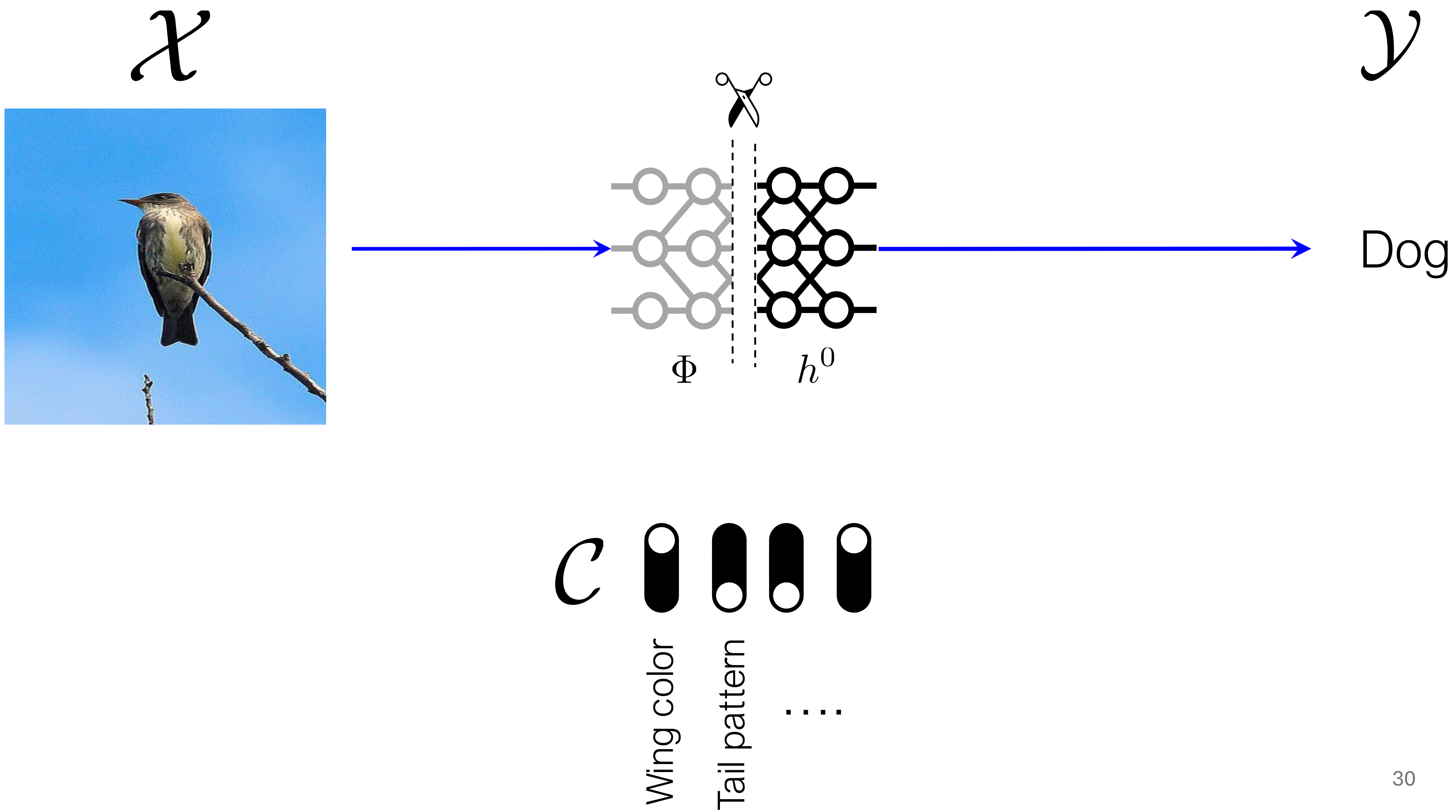
right upper lobe
left lower lobe
heart size
....

$\mathcal{Y}$

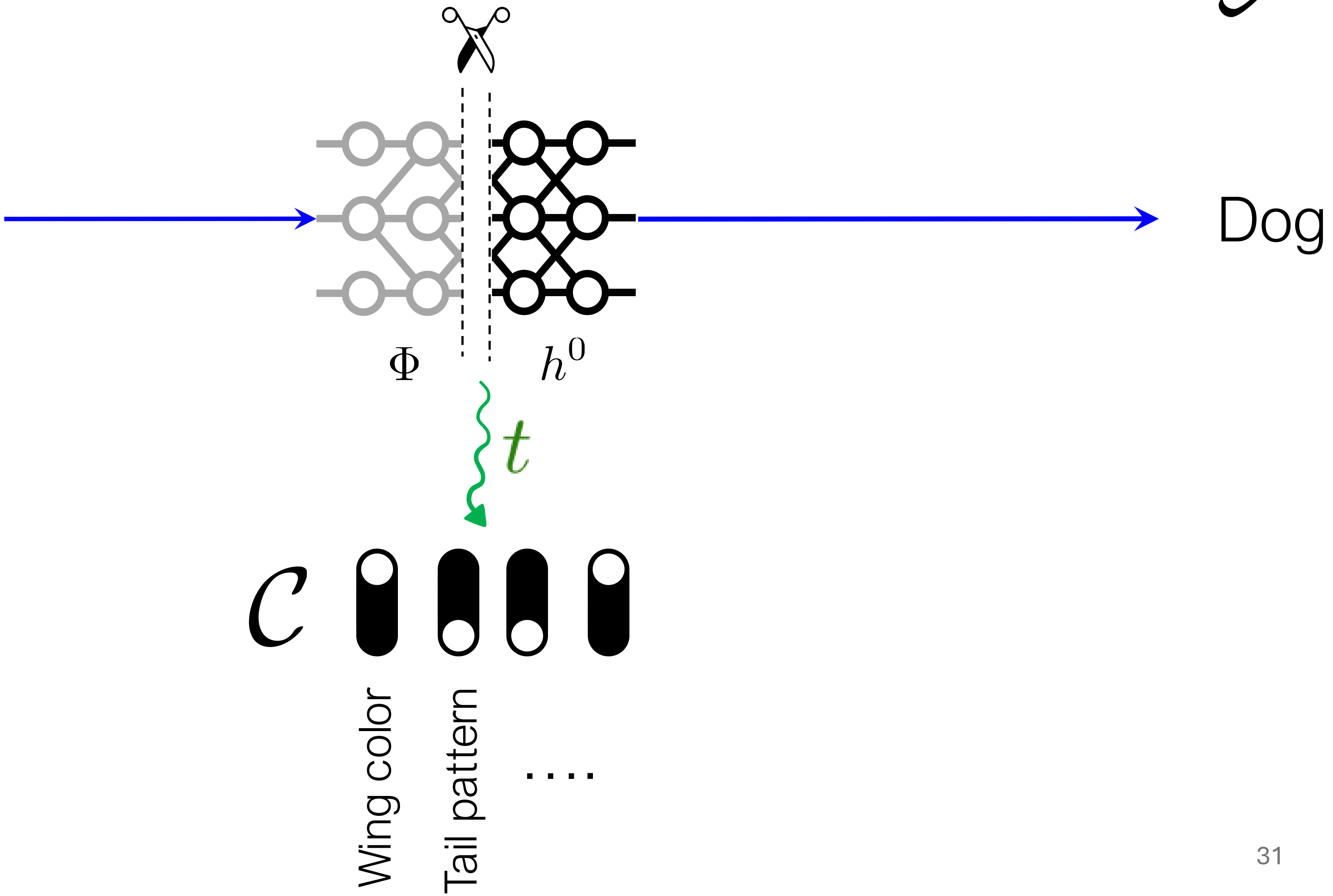
Consolidation

$\}$

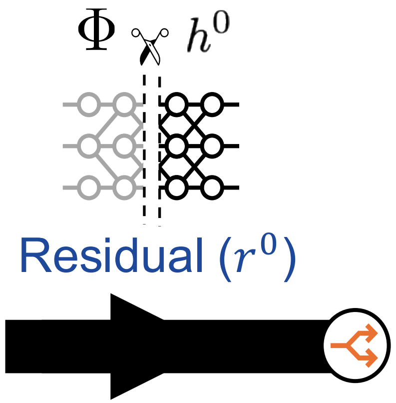
Discovering Hidden Concepts



Discovering Hidden Concepts



Carving out Interpretable Models

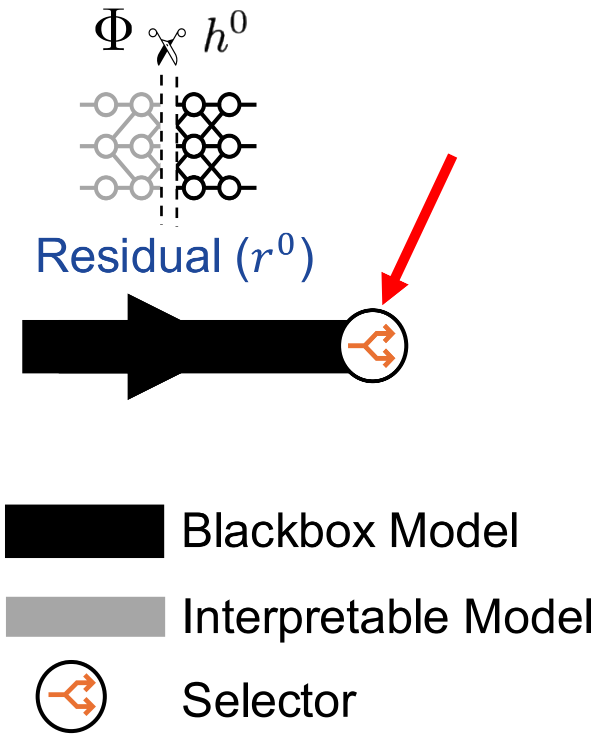


Blackbox Model

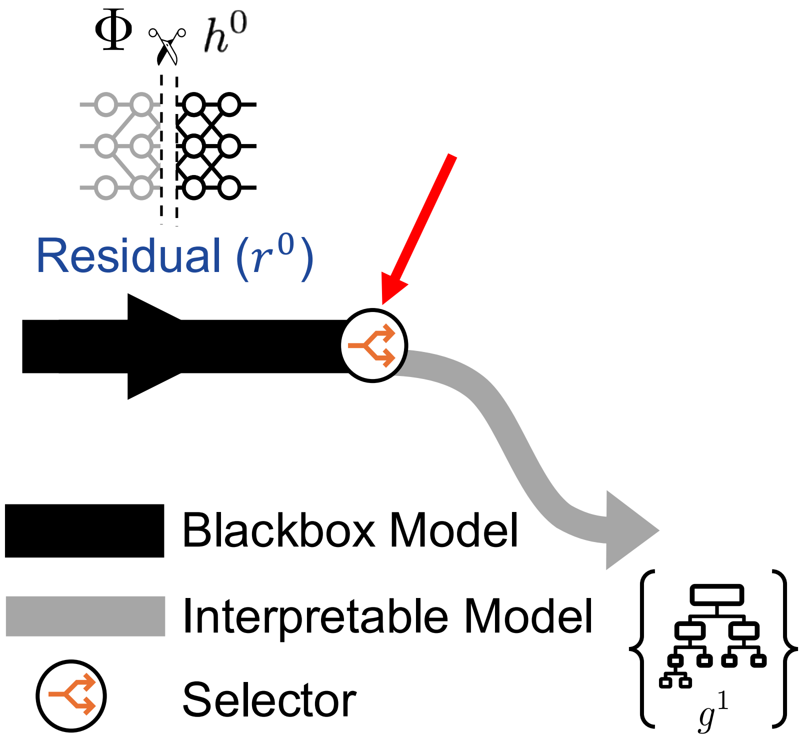
Interpretable Model

Selector

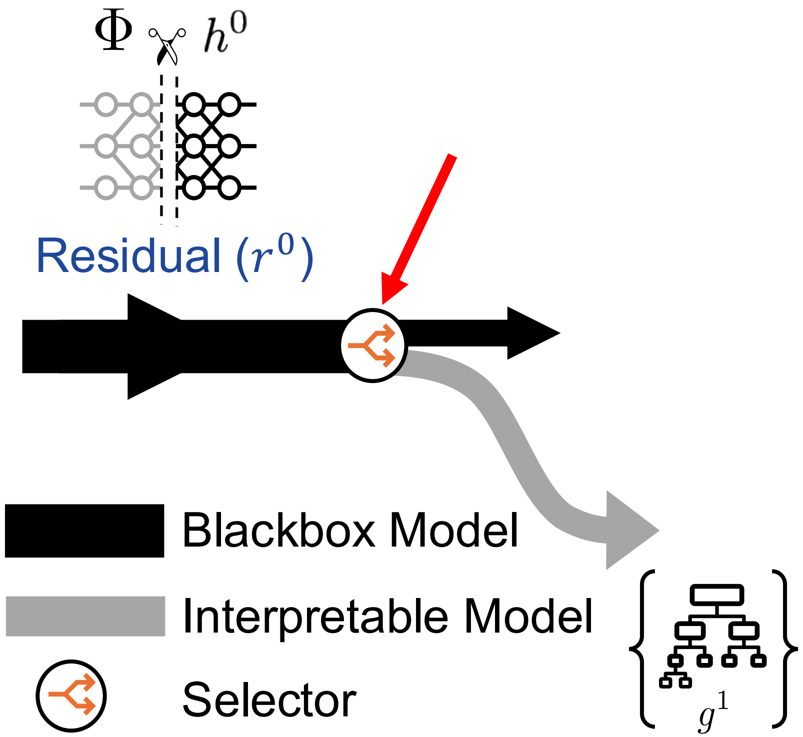
Carving out Interpretable Models



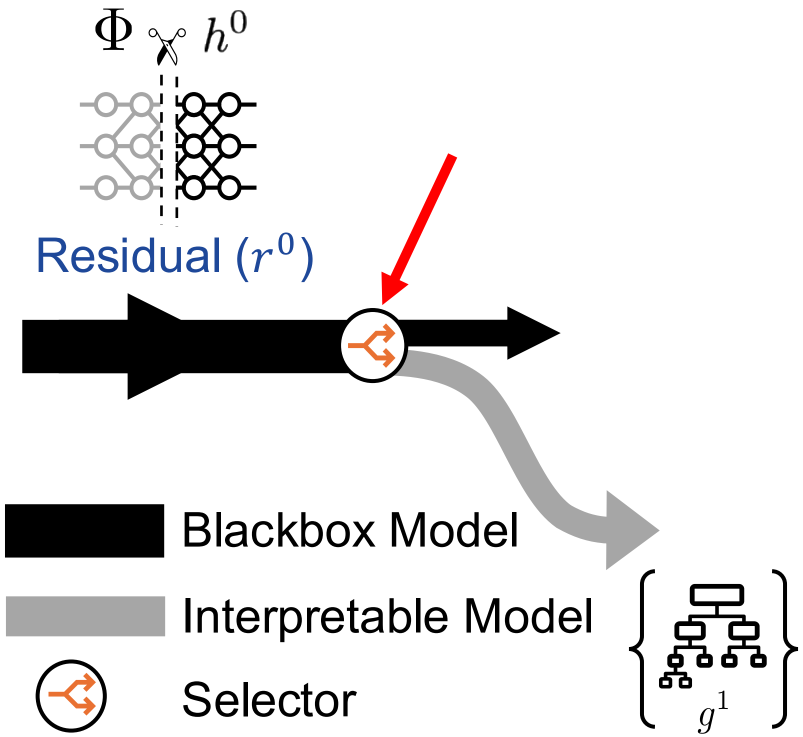
Carving out Interpretable Models



Carving out Interpretable Models



Carving out Interpretable Models

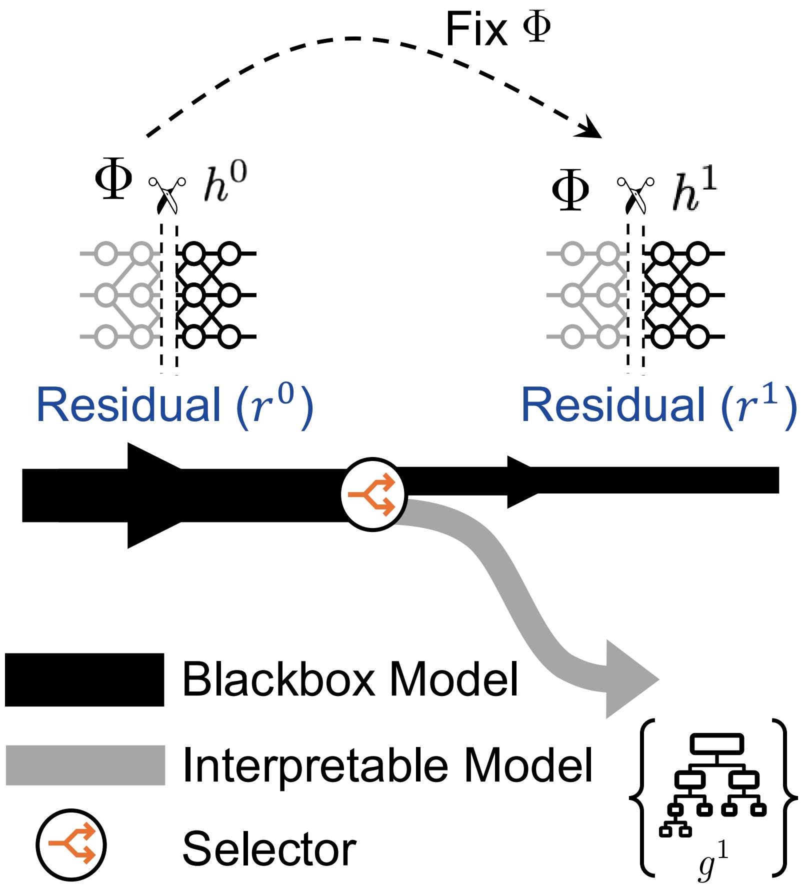


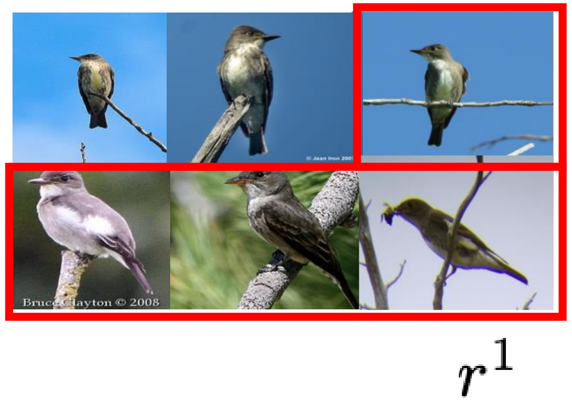
g^1

Olive sided Flycatcher $\leftrightarrow$ breast_color_grey $\wedge$ tail_pattern_solid

Each g is E-LENs (Barberio et al. AAAI 2022) to produce sample specific FOLs.

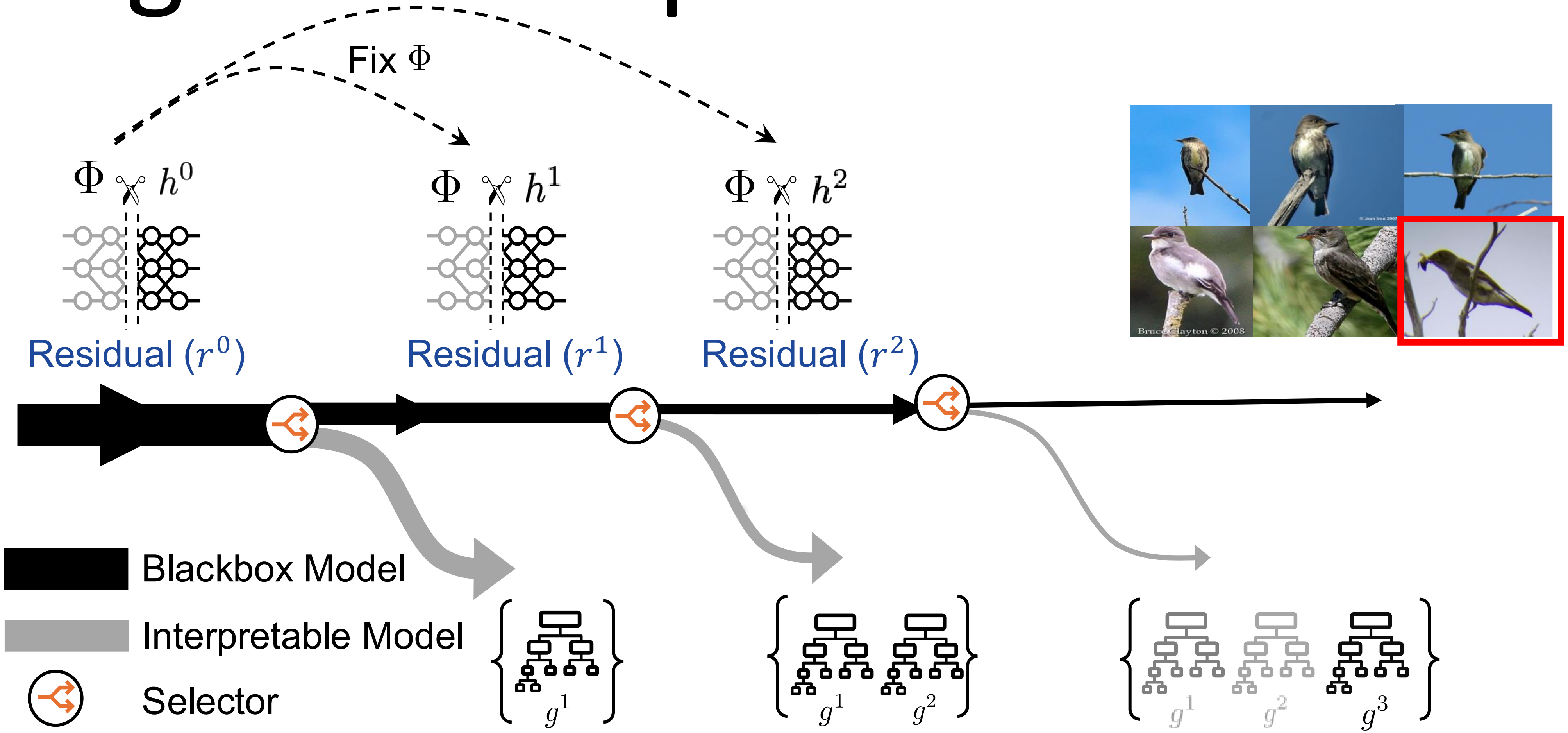
Carving out Interpretable Models





Each g is E-LENs (Barberio et al. AAAI 2022) to produce sample specific FOLs.

Carving out Interpretable Models

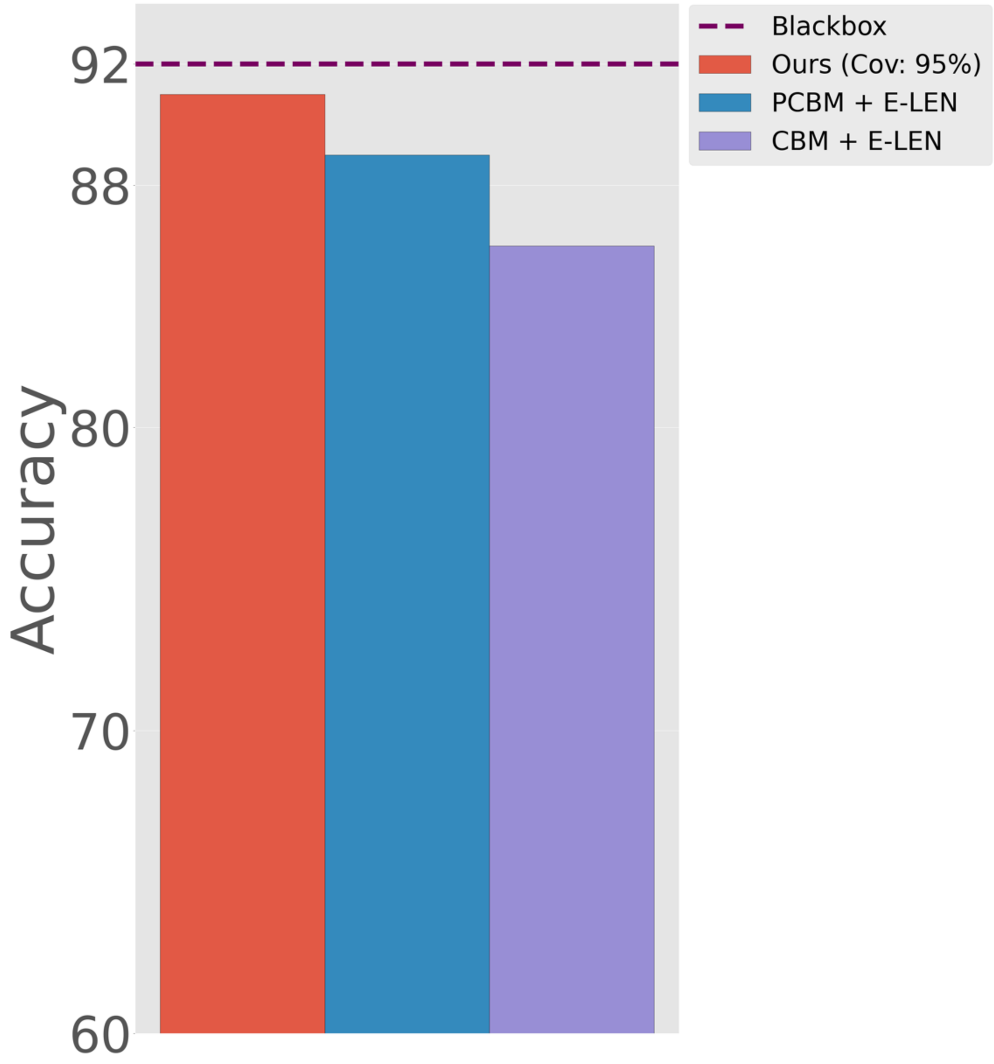


Each g is E-LENs (Barberio et al. AAAI 2022) to produce sample specific FOLs.

Dividing and Conquering a BlackBox to a Mixture of Interpretable Models: Route, Interpret, Repeat. Ghosh et al. ICML 2023.

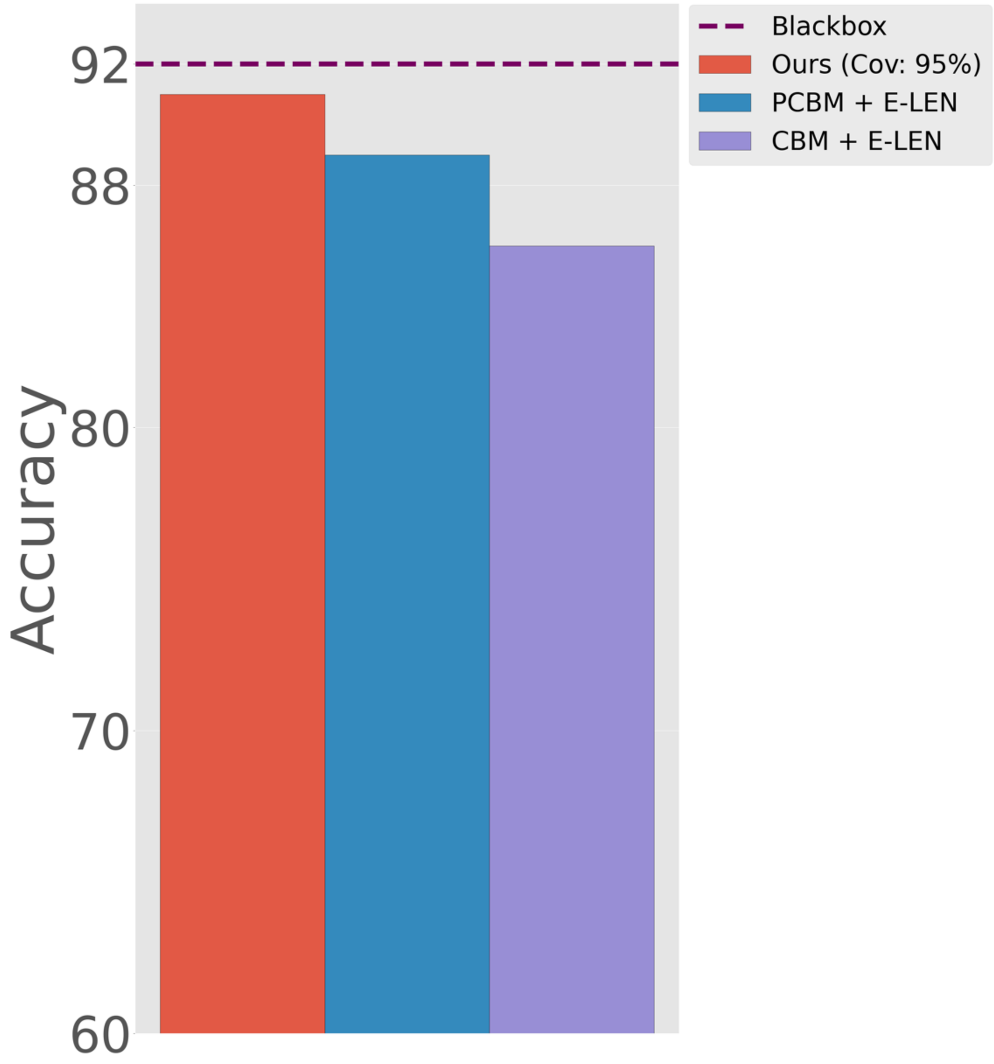
Comparing Performance

CUB-200 with ViT

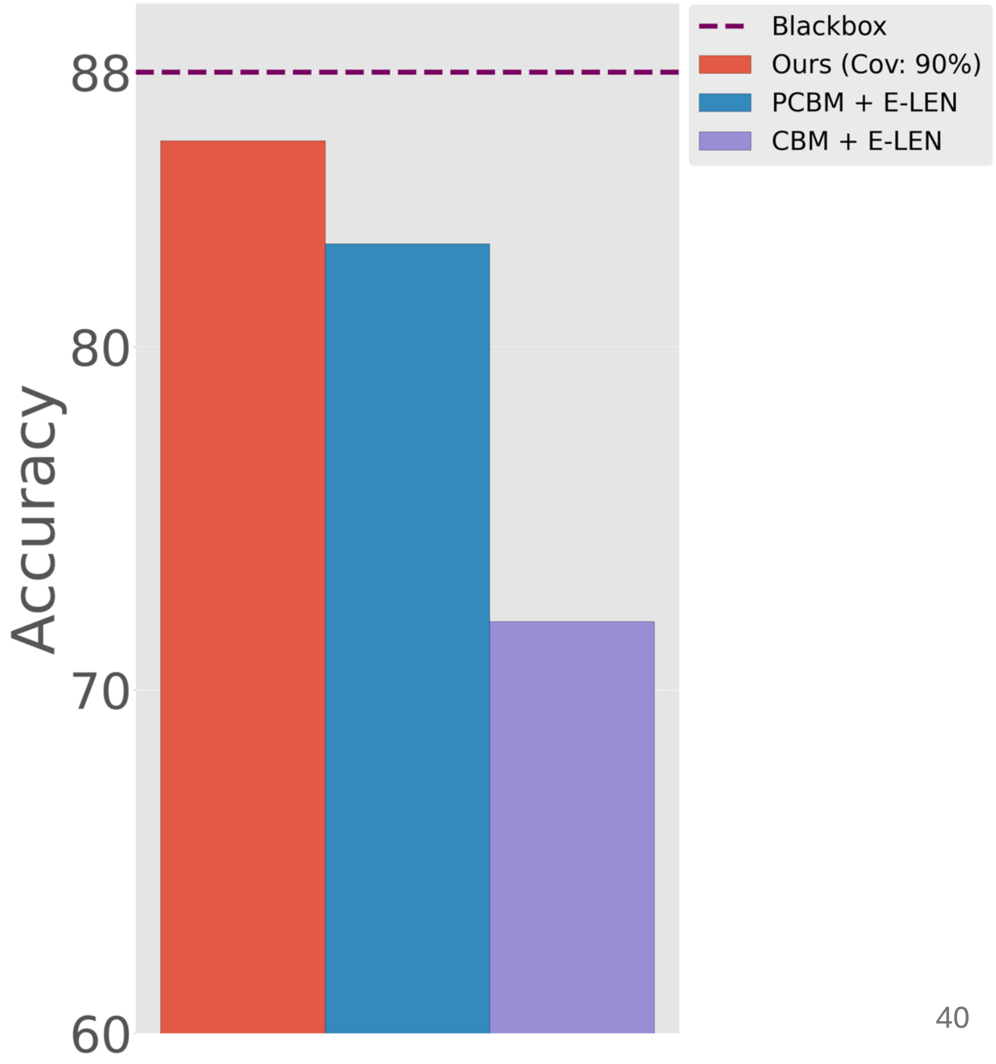


Comparing Performance

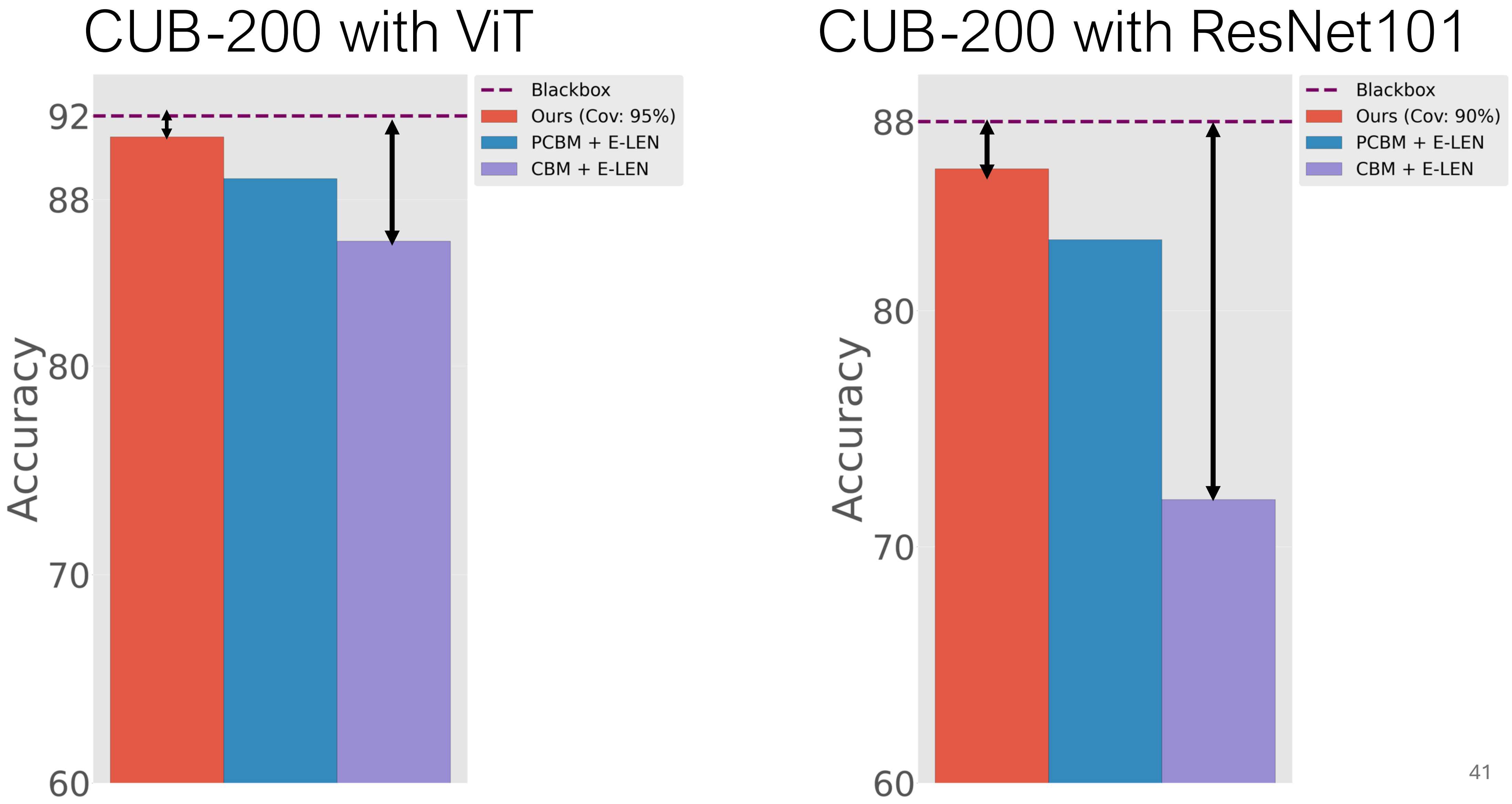
CUB-200 with ViT

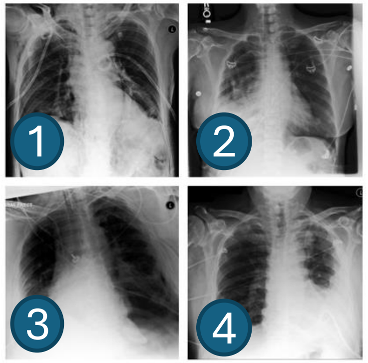


CUB-200 with ResNet101



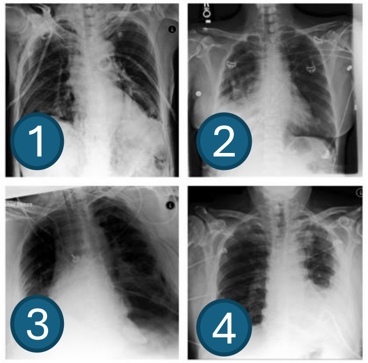
Comparing Performance



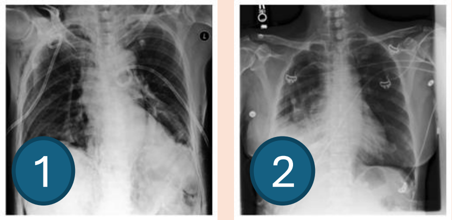


Examples on Chest X-ray

Dividing and Conquering a BlackBox to a Mixture of Interpretable Models: Route, Interpret, Repeat. **Ghosh et al.** ICML 2023.



Examples on Chest X-ray



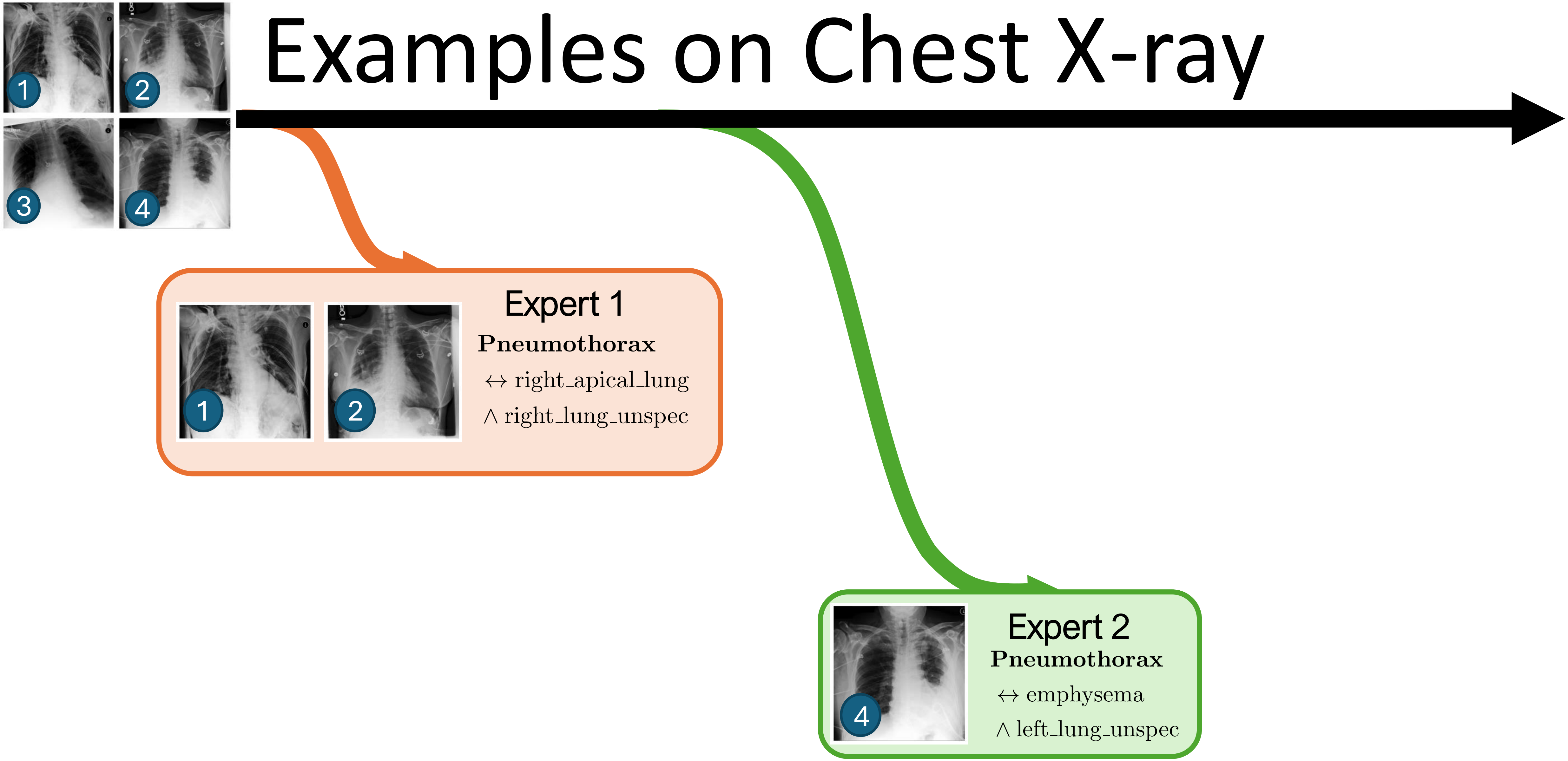
Expert 1

$$\textbf{Pneumothorax} \leftrightarrow \text{right\_apical\_lung} \wedge \text{right\_lung\_unspec}$$

* Right lung unspec refers a malignant neoplasm or cancer in an unspecified part of the right bronchus or lung.

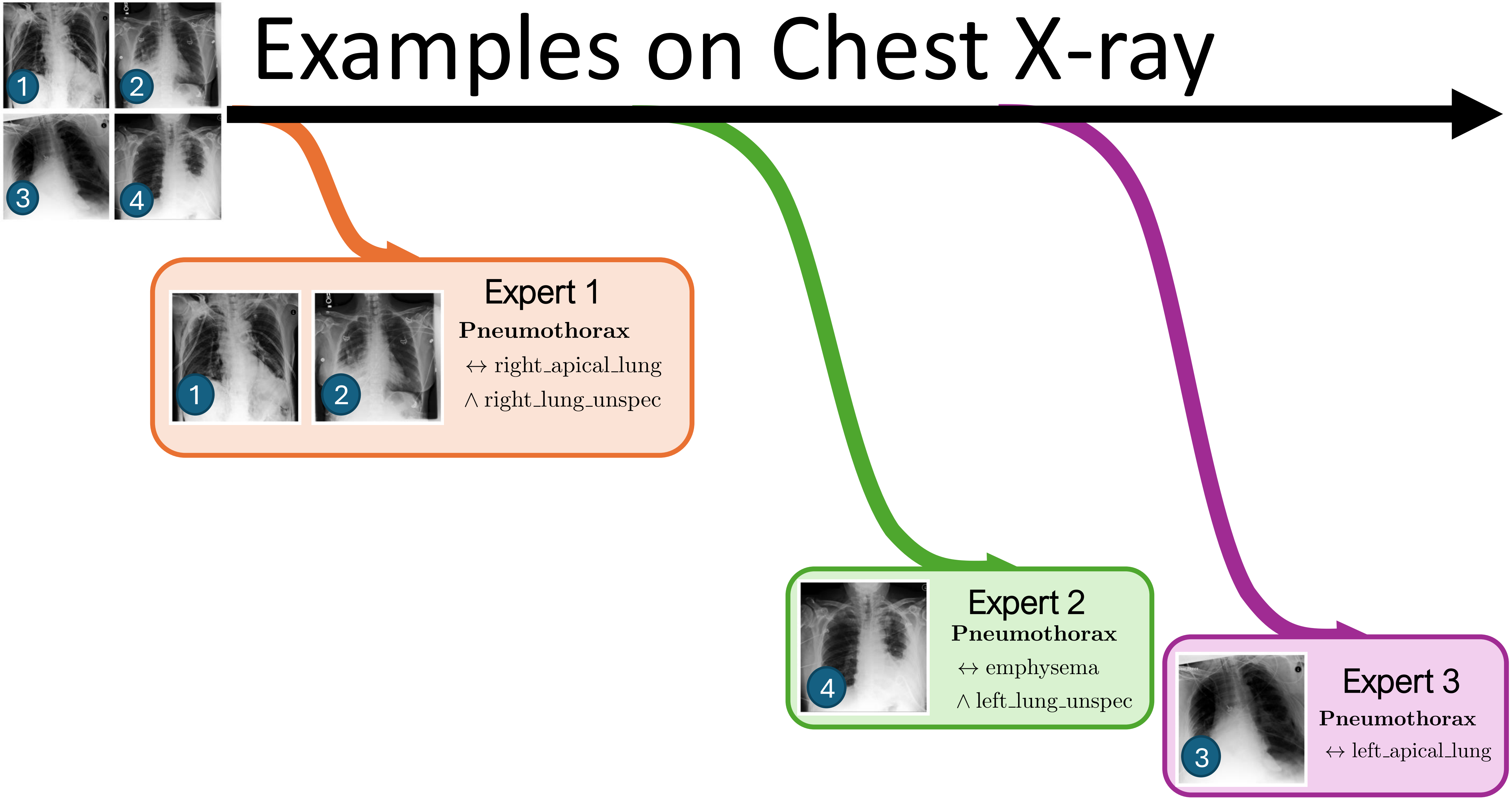
Dividing and Conquering a BlackBox to a Mixture of Interpretable Models: Route, Interpret, Repeat. **Ghosh et al.** ICML 2023.

Examples on Chest X-ray



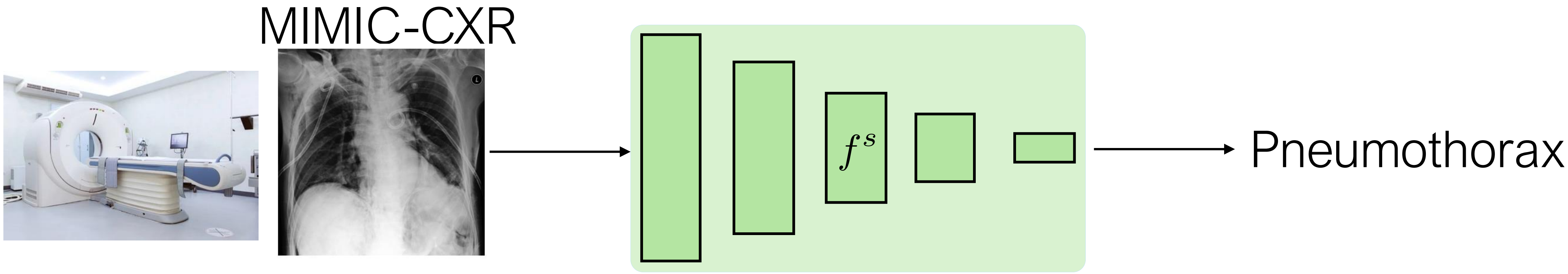
Dividing and Conquering a BlackBox to a Mixture of Interpretable Models: Route, Interpret, Repeat. **Ghosh et al.** ICML 2023.

Examples on Chest X-ray



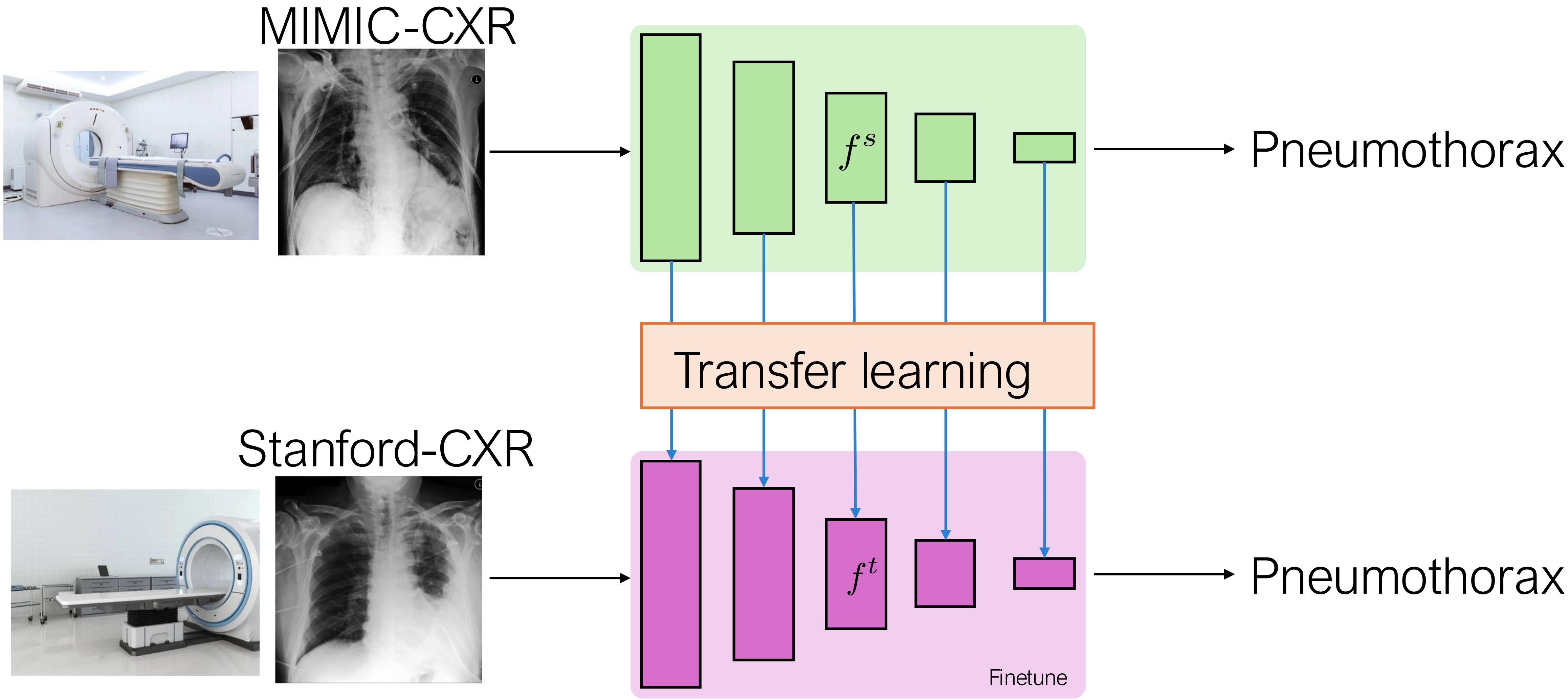
Dividing and Conquering a BlackBox to a Mixture of Interpretable Models: Route, Interpret, Repeat. **Ghosh et al.** ICML 2023.

Application: Data-Efficient Fine-tuning



Best practices for fine-tuning visual classifiers to new domains. In: Computer Vision. Chu et al. 2016. ECCV-W

Application: Data-Efficient Fine-tuning



Best practices for fine-tuning visual classifiers to new domains. In: Computer Vision. Chu et al. 2016. ECCV-W

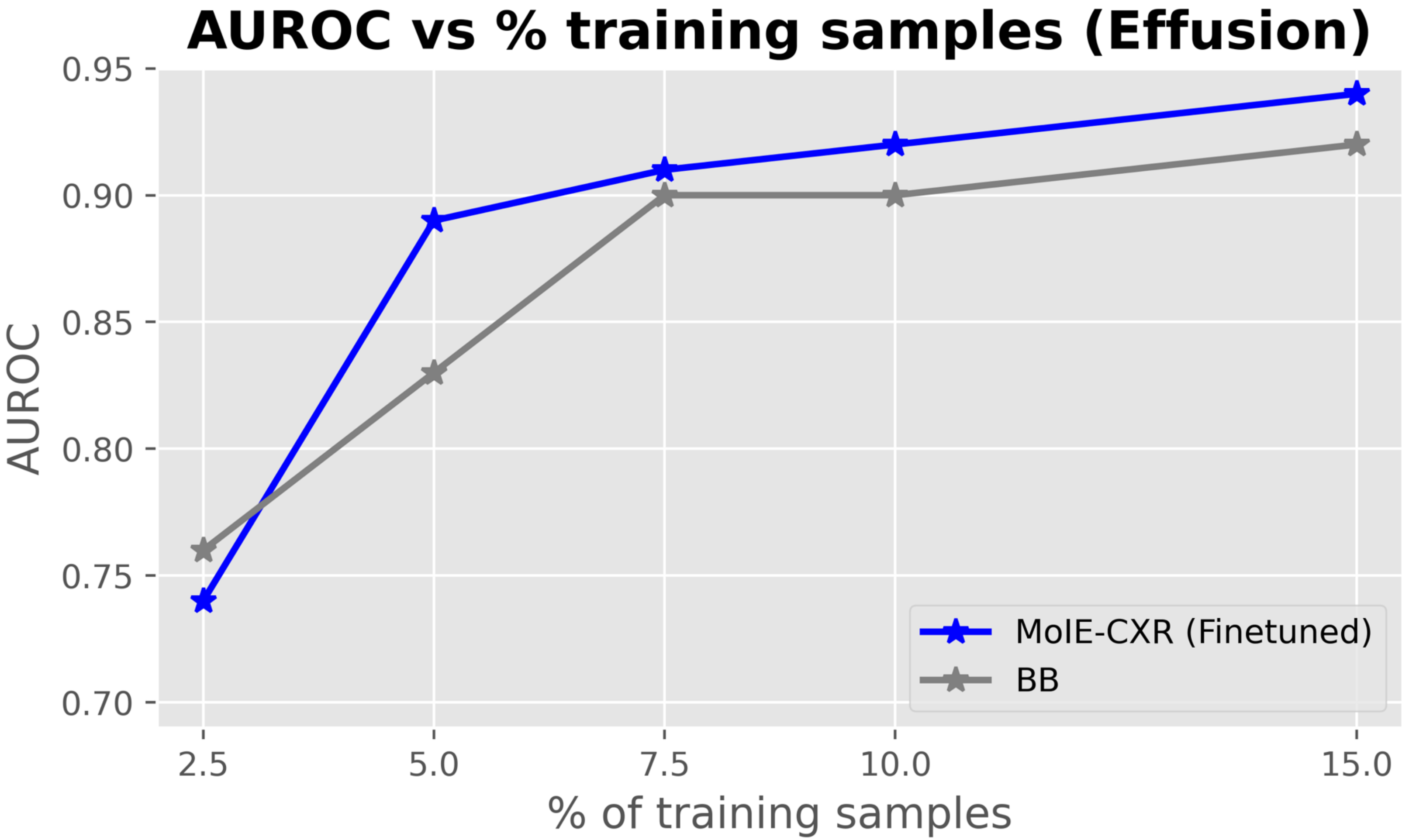
Application: Data-Efficient Fine-tuning

Data and Computationally inefficient

The clinical rules are “invariant”

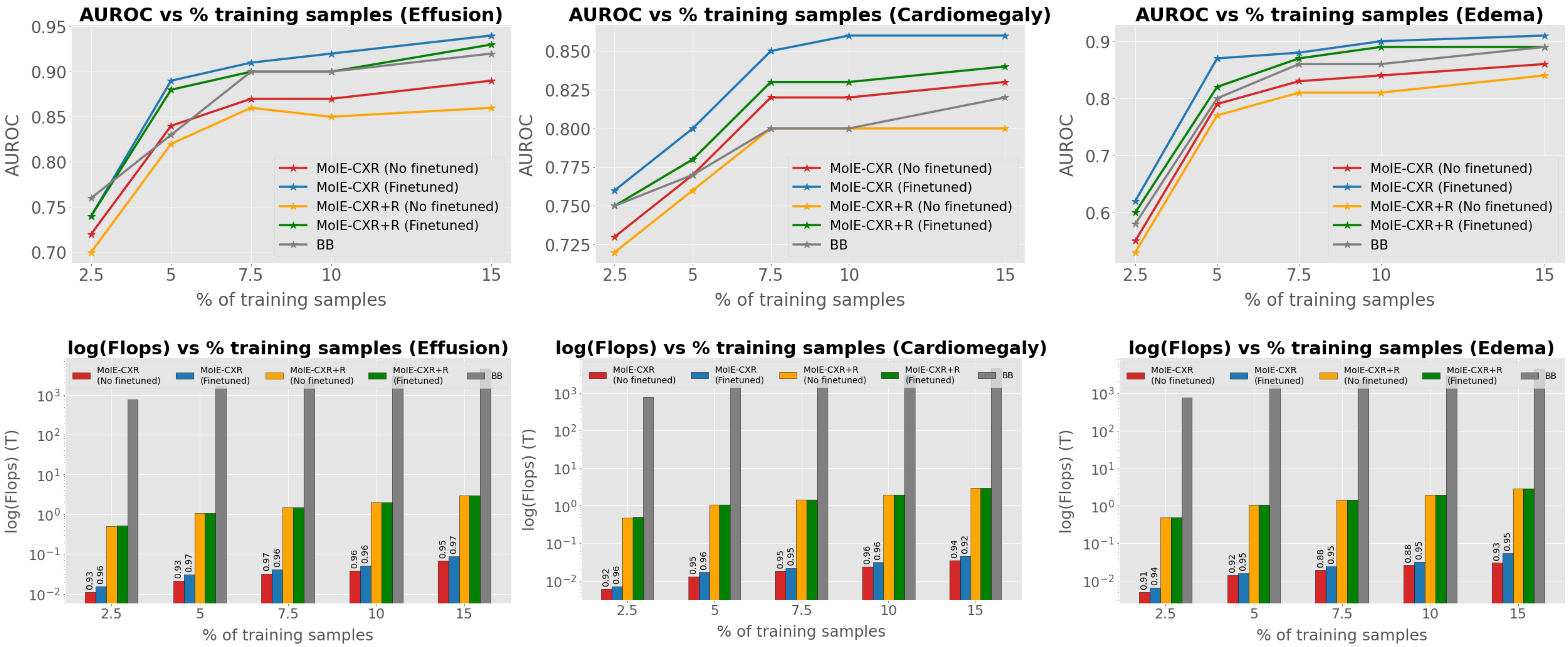
How deeply to fine-tune a convolutional neural network: a case study using a histopathology dataset. Kandel et al. 2020. Applied Sciences
Growing a brain: Fine-tuning by increasing model capacity. Wang et al. 2017 CVPR

Transferring to Stanford-CXR



Distilling BlackBox to Interpretable models for Efficient Transfer Learning. **Ghosh et al.** MICCAI 2023. Top 14%

Transferring to Stanford-CXR



Conclusion from Aim 1

1. Domain invariant rules learned: A mixture of interpretable models are ***carved out*** of a Blackbox model offering best of both worlds. **[ICML 2023]**

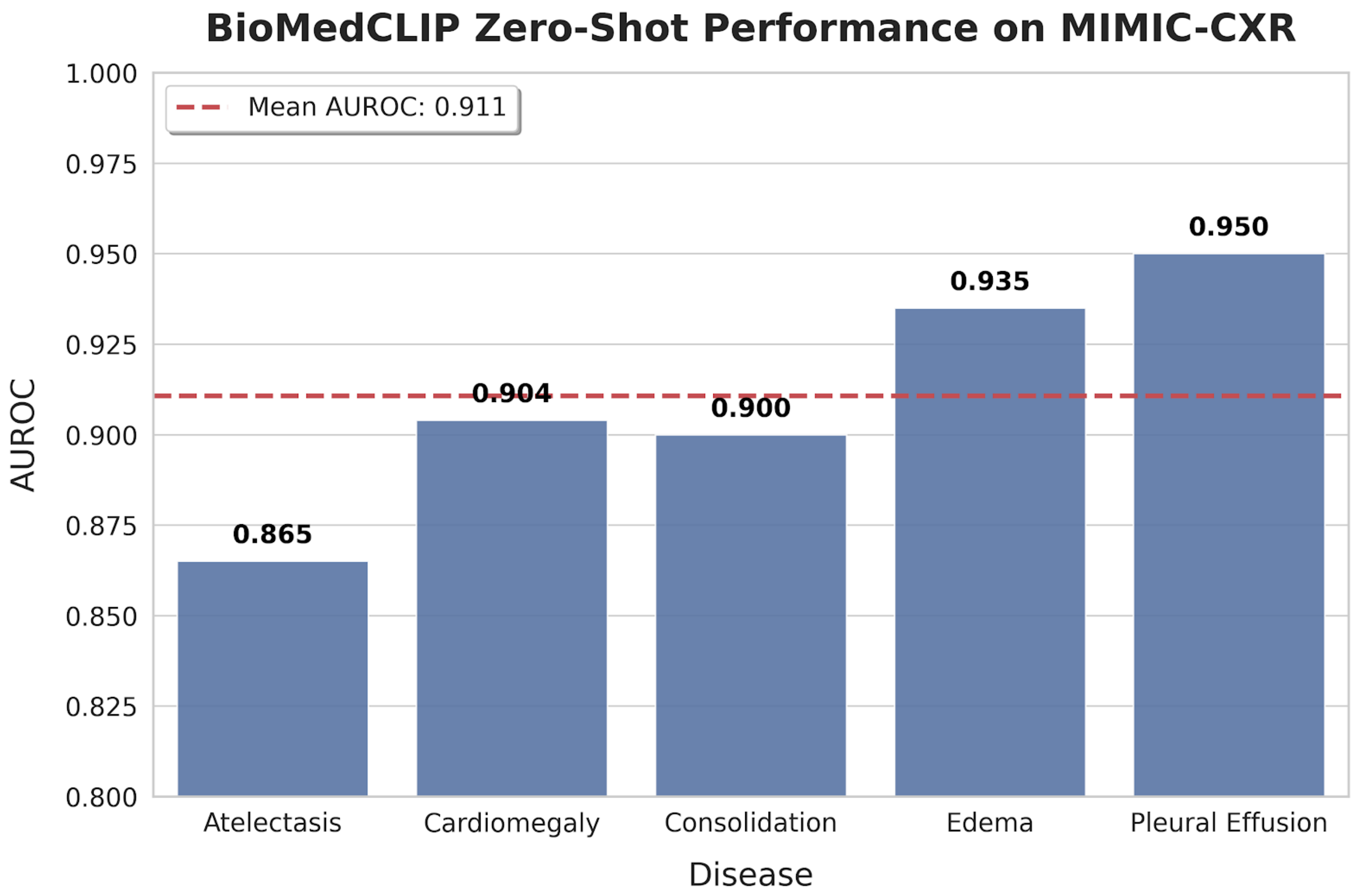
They effectively learn domain invariant rules.

2. Efficient transfer learning: Transfer Learning is more efficient ***using limited training data*** with the new interpretable model. **[MICCAI 2023. Top 14%]**

Aim 2

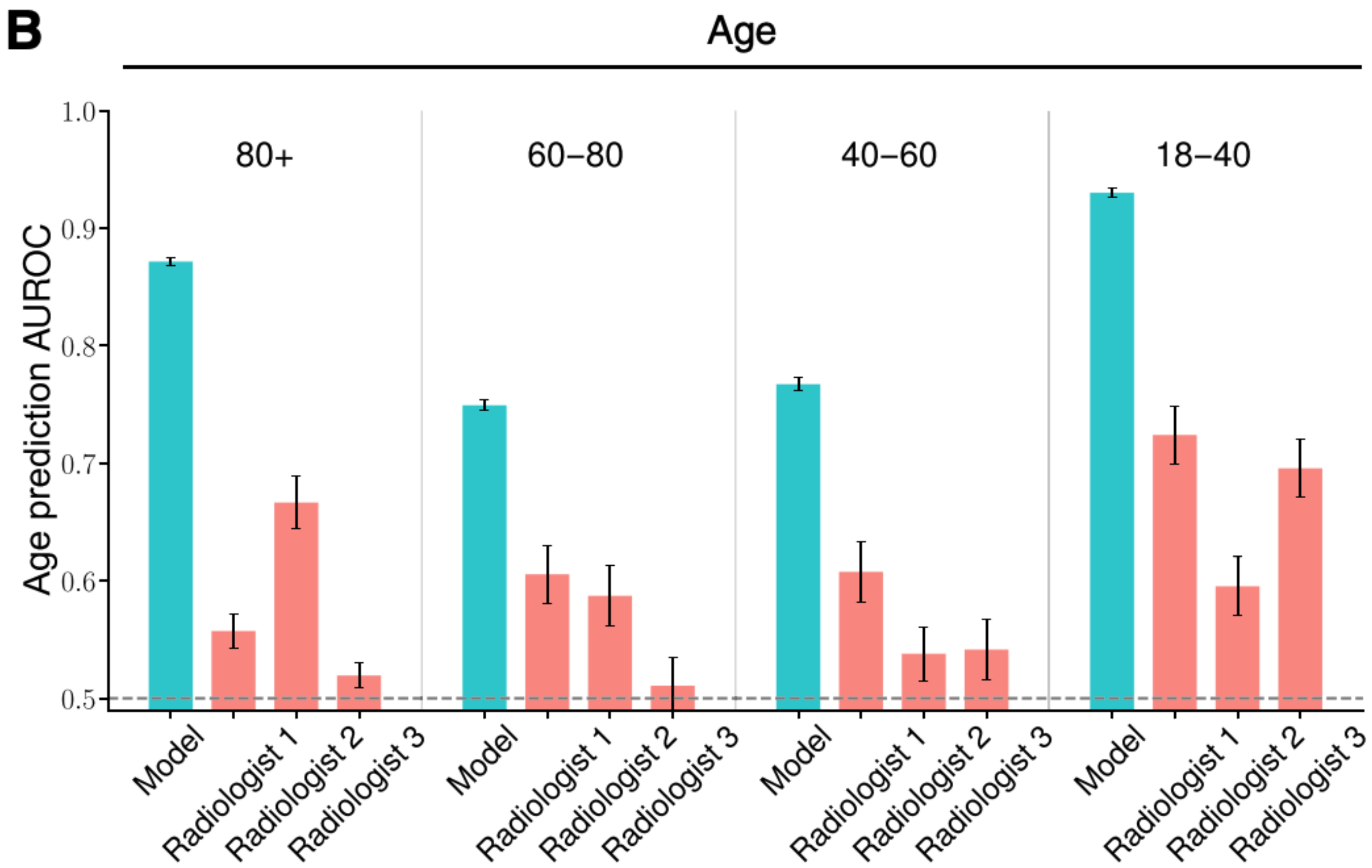
The goal: Develop a large **VLM** for **Mammography**

Why SSL based VLMs



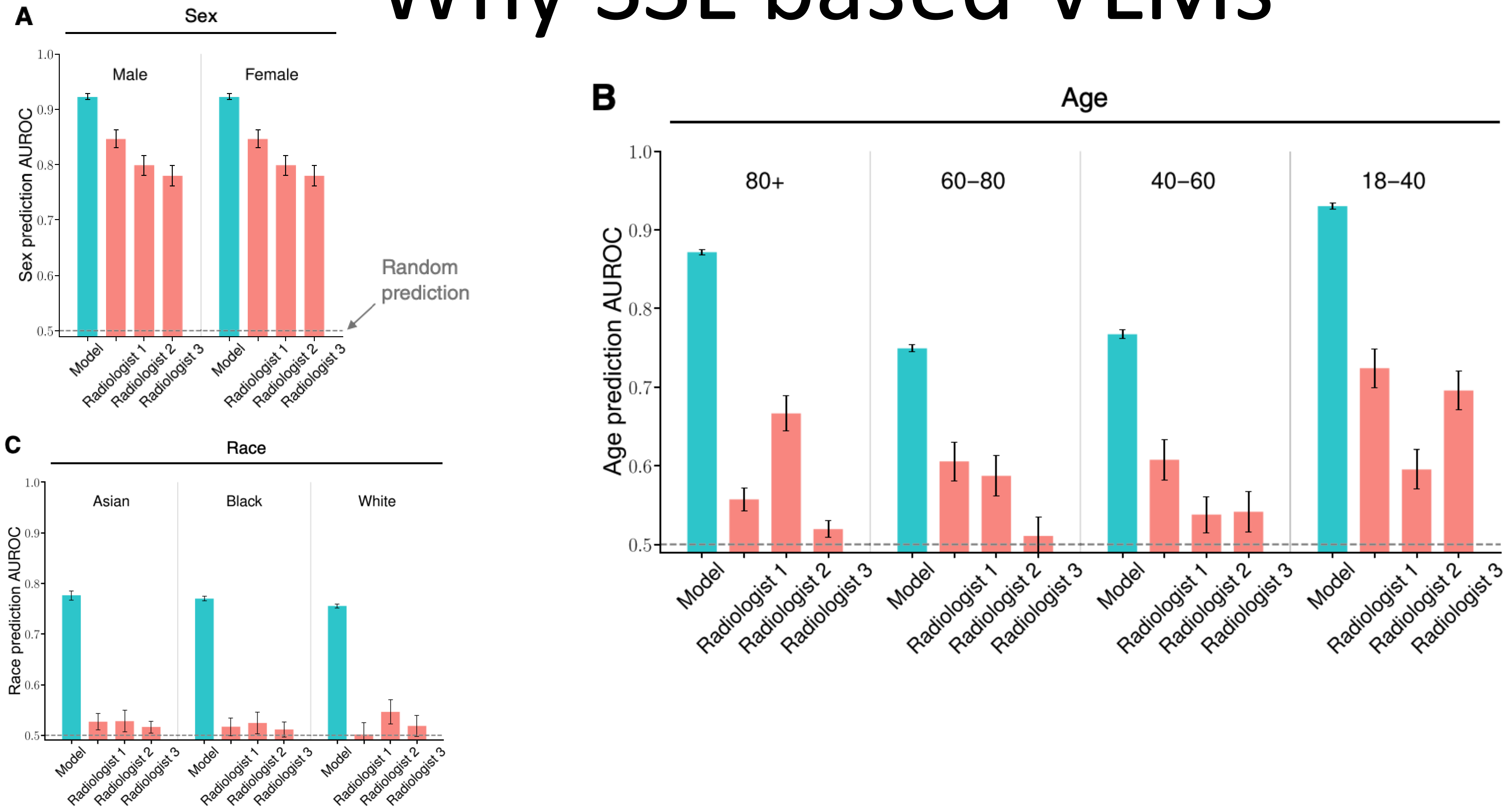
BioMedCLIP. Microsoft.

Why SSL based VLMs



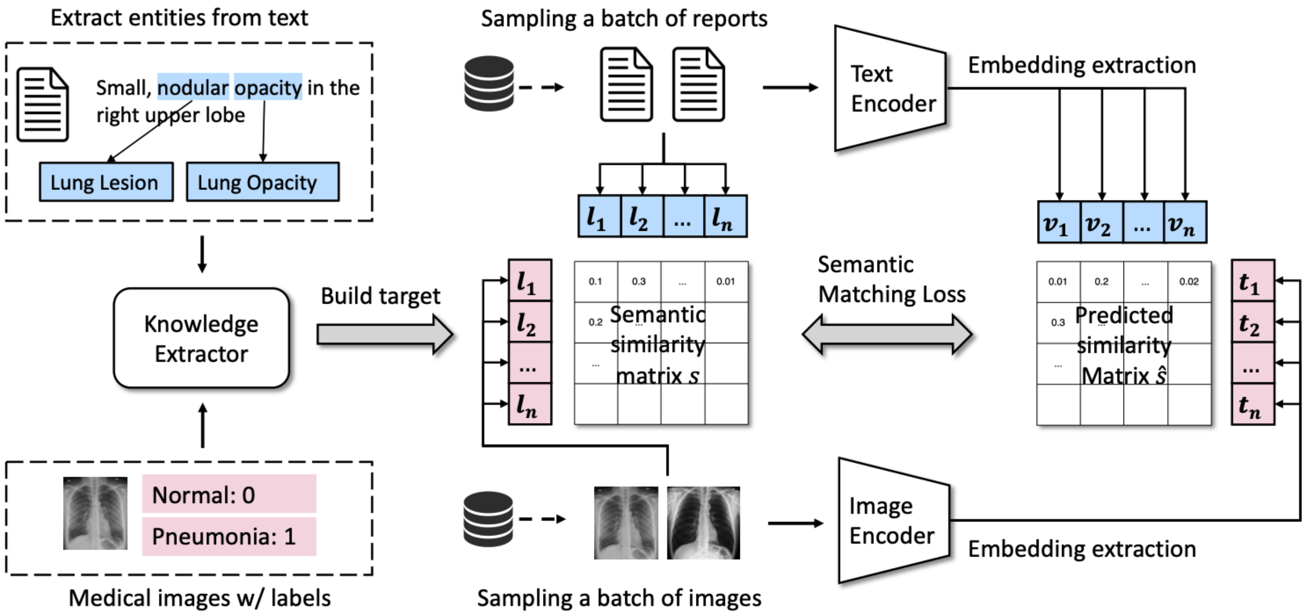
Demographic bias of expert-level vision-language foundation models in medical imaging. Yang et al. Science Advances.

Why SSL based VLMs

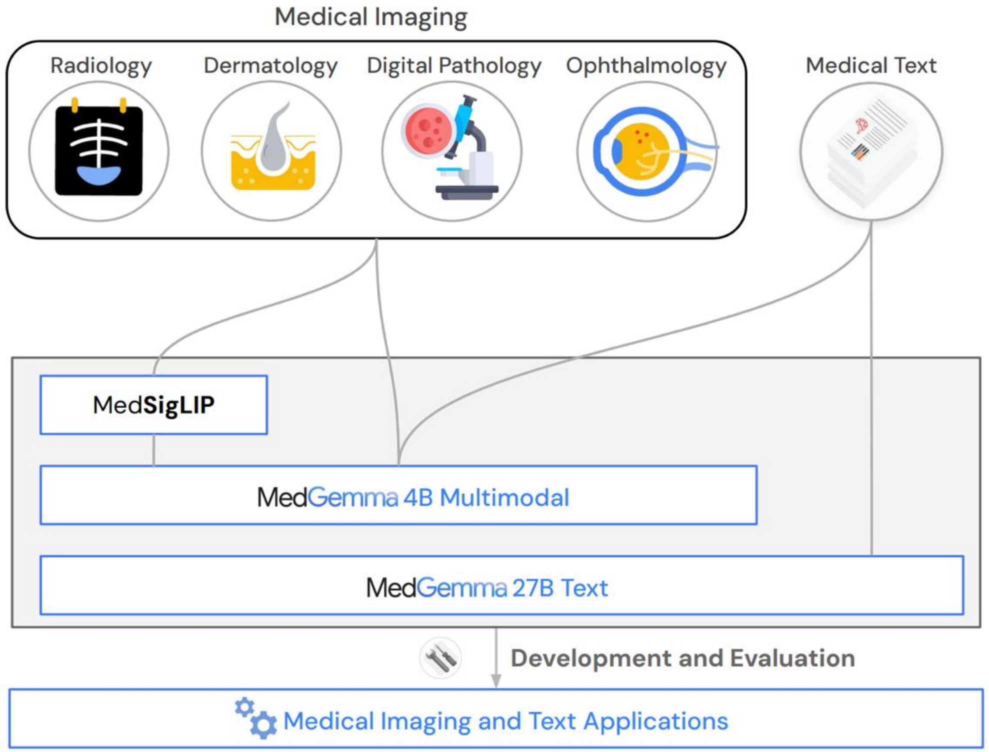


Demographic bias of expert-level vision-language foundation models in medical imaging. Yang et al. Science Advances.

VLM in Mammography: Strategies



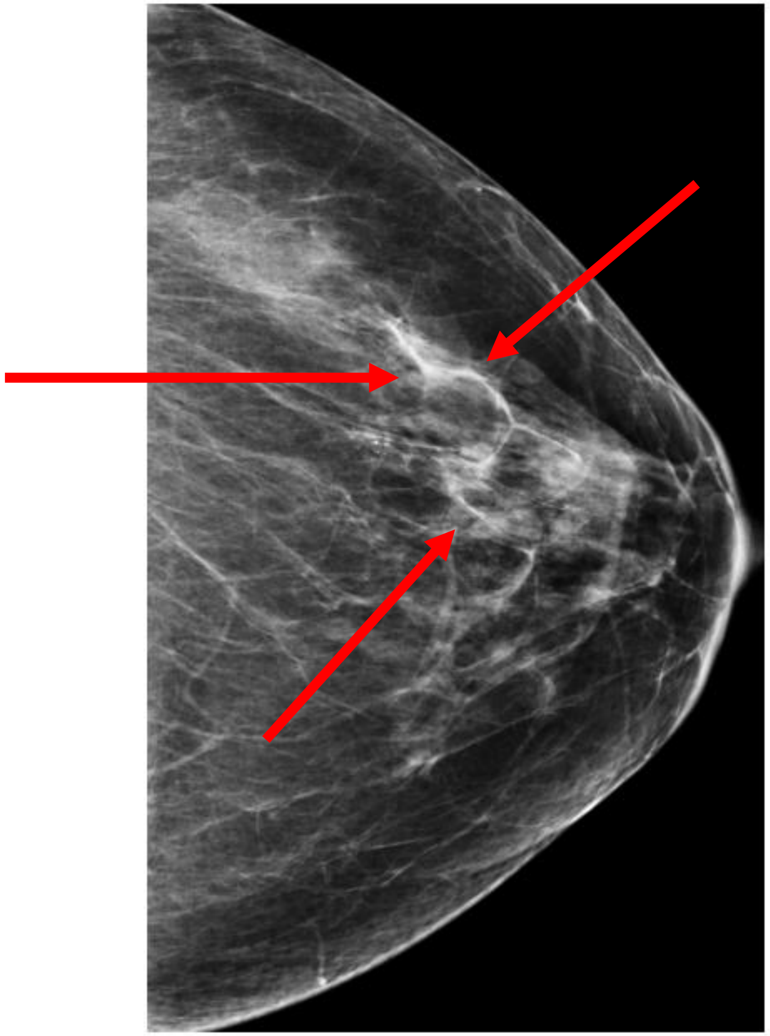
MedCLIP. EMNLP 2022



MedGemma (Google). Technical Report 2025

We need domain-specific VLMs

Mammograms are large and have subtle cues



VLMs can help

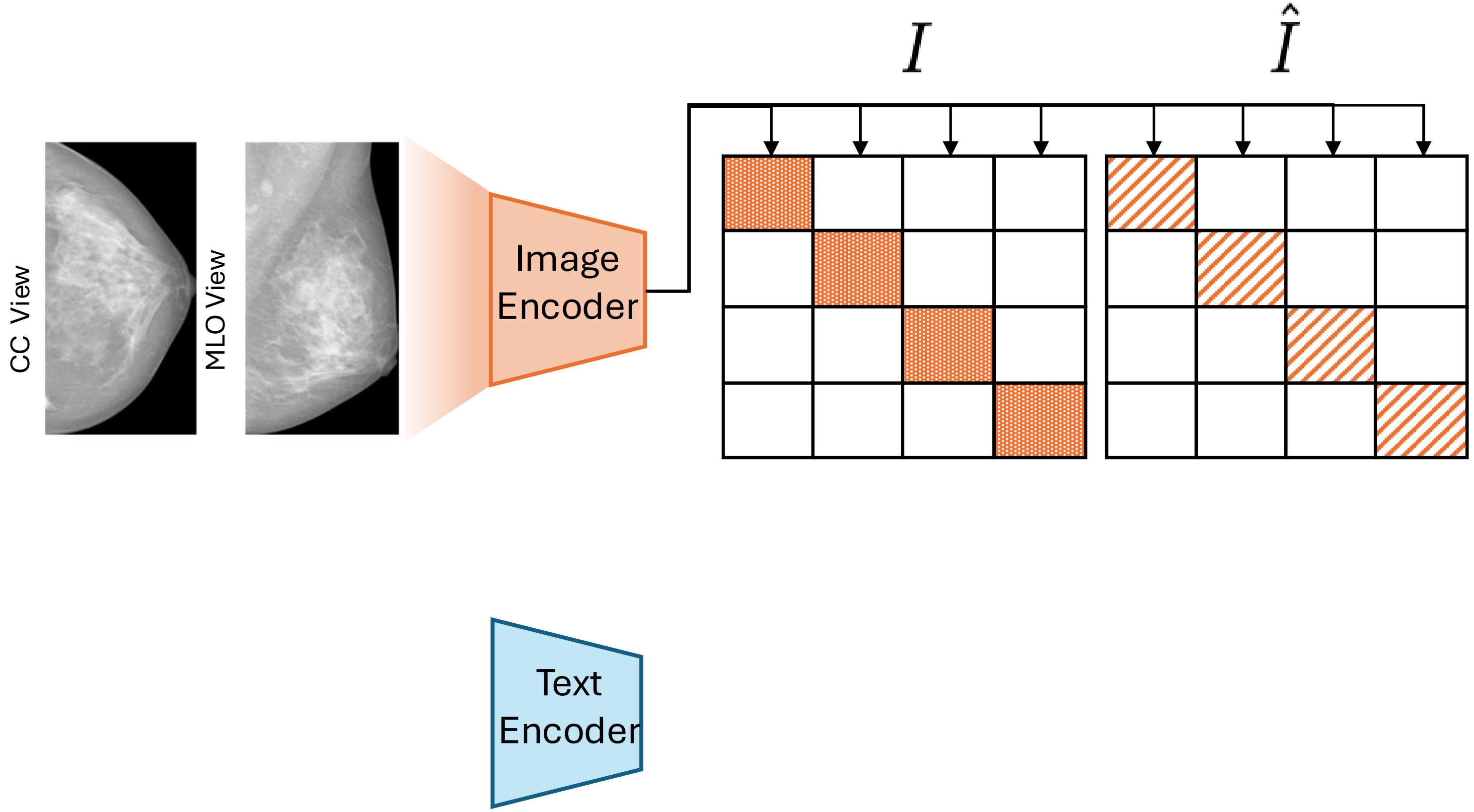
Sample screening mammogram report:

Ground truth: *Findings:* The breast tissue is heterogeneously dense, which could obscure detection of small masses. There are calcifications in the left upper, slightly outer breast. There is also an asymmetry in the left breast. Otherwise, no suspicious masses, clustered microcalcifications or areas of architectural distortion are seen. *Impression*: BI-RADS: 0. Calcifications and asymmetry in the left breast, which needs additional imaging.

Mammo-FM: pretraining

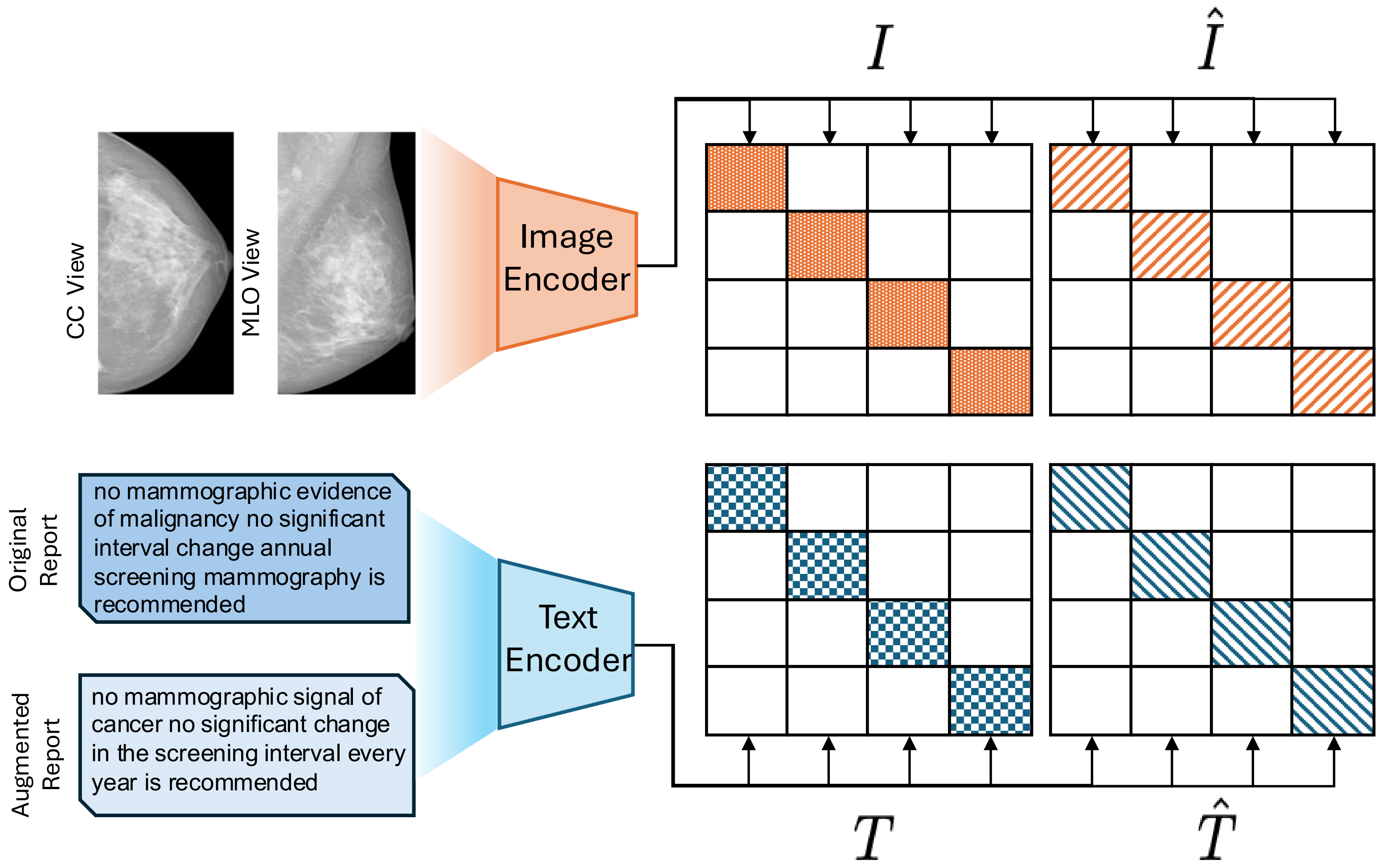
Mammo-CLIP. Ghosh et al. MICCAI 2024.

Mammo-FM: pretraining



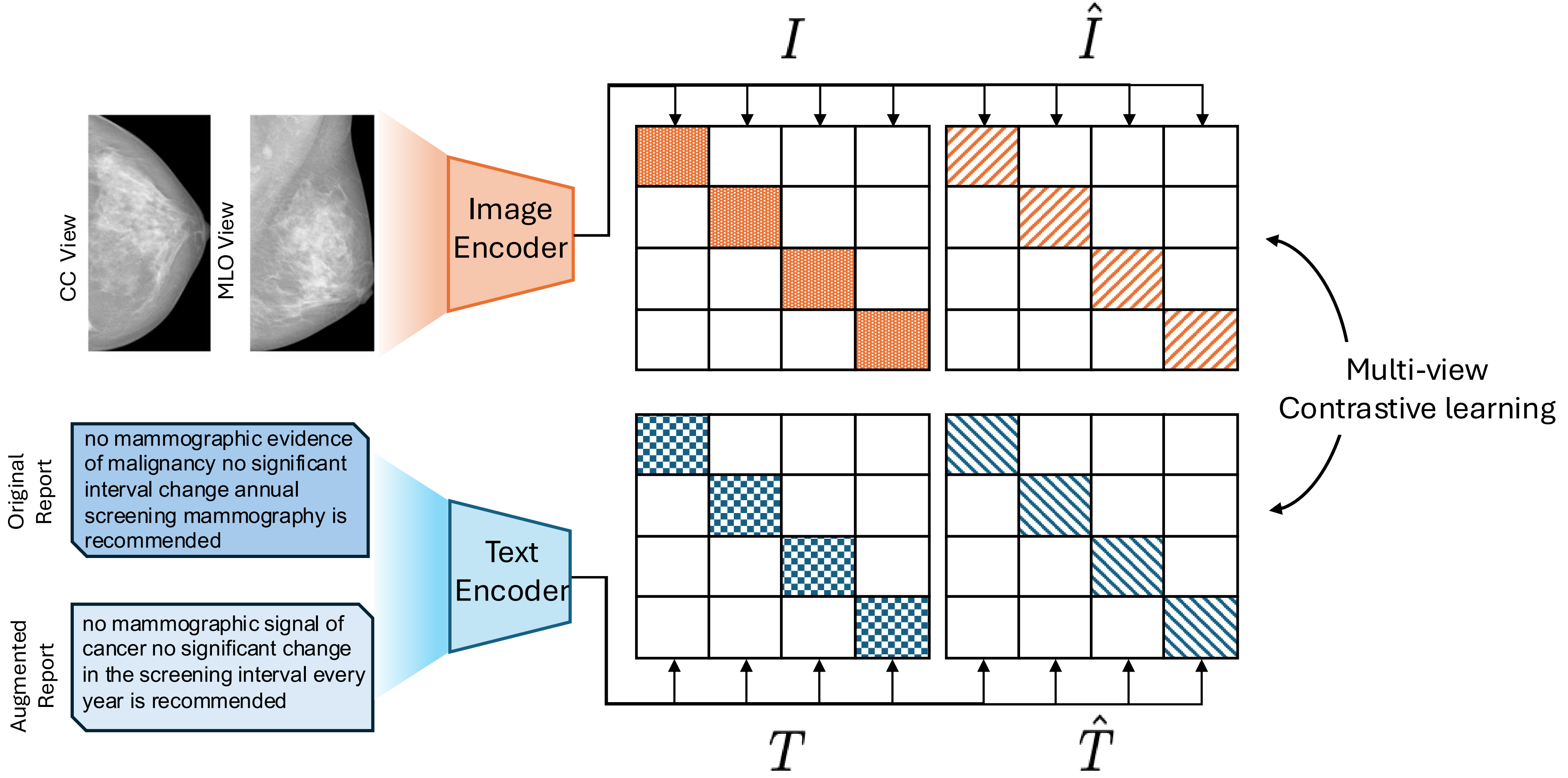
Mammo-CLIP. Ghosh et al. MICCAI 2024.

Mammo-FM: pretraining



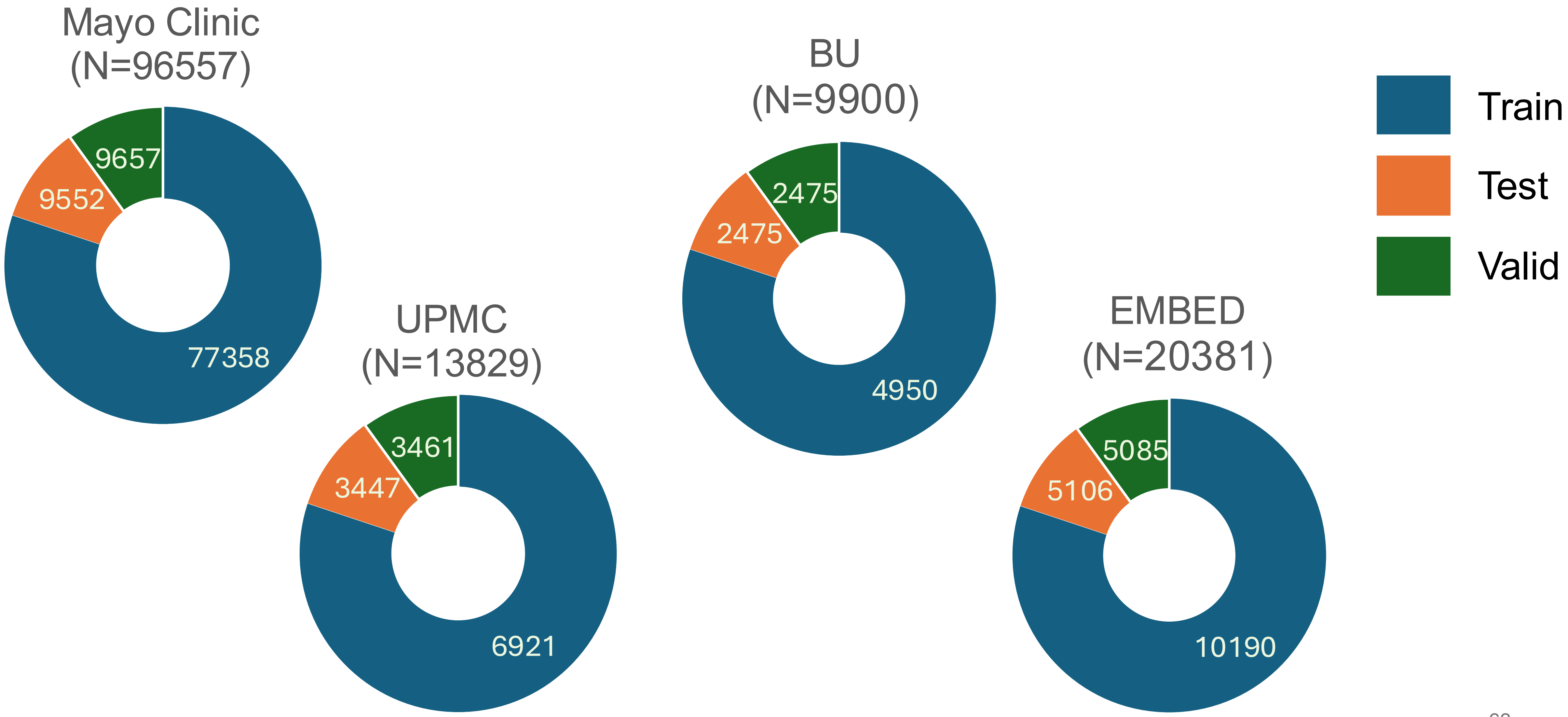
Mammo-CLIP. Ghosh et al. MICCAI 2024.

Mammo-FM: pretraining



Mammo-CLIP. Ghosh et al. MICCAI 2024.

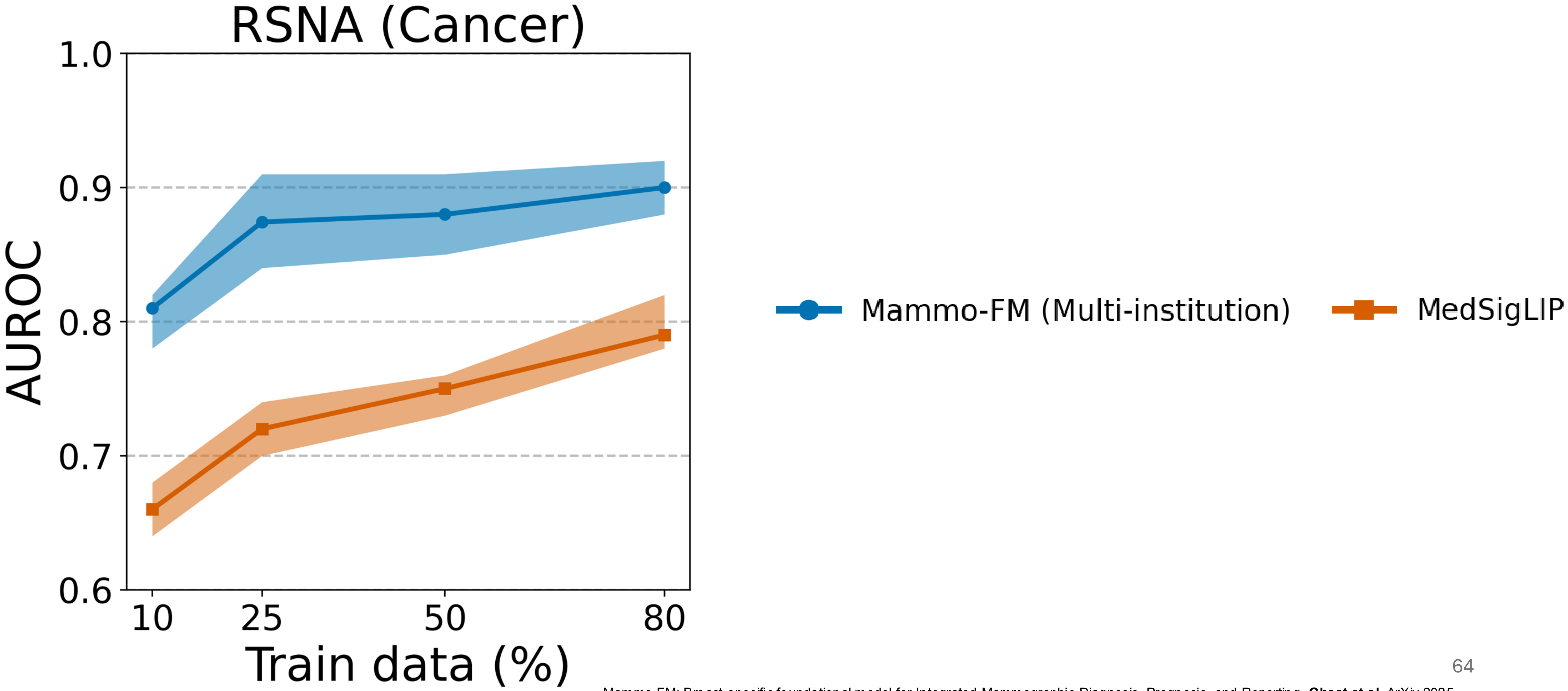
Mammo-FM: pretraining



Mammo-FM: Breast-specific foundational model for Integrated Mammographic Diagnosis, Prognosis, and Reporting. **Ghost et al.** ArXiv 2025

Mammo-FM: data efficiency

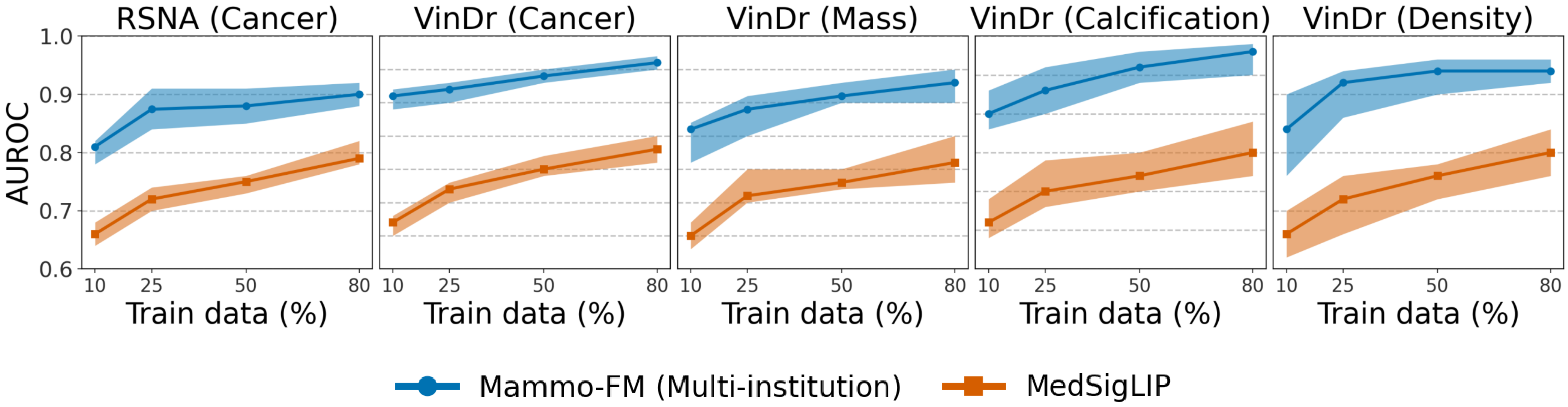
Out of distribution evaluation



Mammo-FM: Breast-specific foundational model for Integrated Mammographic Diagnosis, Prognosis, and Reporting. **Ghost et al.** ArXiv 2025

Mammo-FM: data efficiency

Out of distribution evaluation



Mammo-FM: Breast-specific foundational model for Integrated Mammographic Diagnosis, Prognosis, and Reporting. **Ghost et al.** ArXiv 2025

Conclusion from Aim 2

1. Robust mammography features learned: We learn generalized features for mammography by pre-training on the ***largest*** and ***most diverse*** mammography datasets. **[MICCAI 2024. Top 11%]**

It's diagnostic performance is better than the SOTA ***generalist industrial models***.

2. Two Applications: It helps to ***interpret*** the risks from any SOTA ***risk predictors***.

We develop the ***1st*** ***report generator*** for mammography using Mammo-FM. **[ArXiv 2025]**

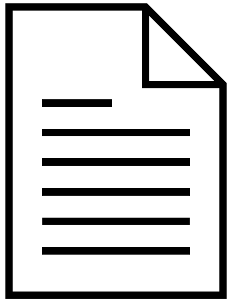
Aim 3

The goal: **Detect** the **Systematic Mistakes** using **Language**

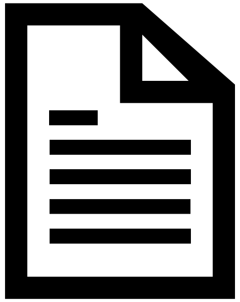
Aim 3

The goal: **Detect** the **Systematic Mistakes** using **Language**

Captions



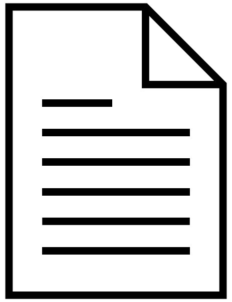
Reports



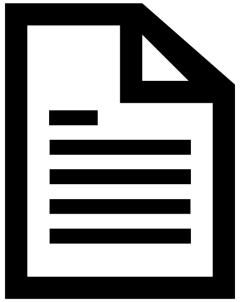
Aim 3

The goal: **Detect** the **Systematic Mistakes** using **Language**

Captions



Reports



Vision-Language Alignment

Natural image

CLIP *(ICML 2020)*

CXRs

1. ***GLORIA*** *(ICCV2021)*

2. ***MedCLIP*** *(EMNLP 2022)*

3. ***CXR-CLIP*** *(MICCAI 2023)*

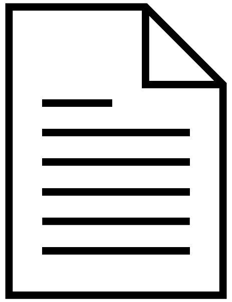
Mammo

?

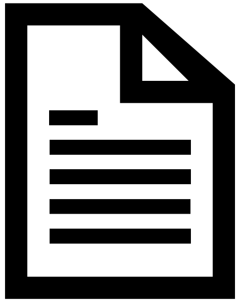
Aim 3

The goal: **Detect** the **Systematic Mistakes** using **Language**

Captions



Reports



Vision-Language Alignment

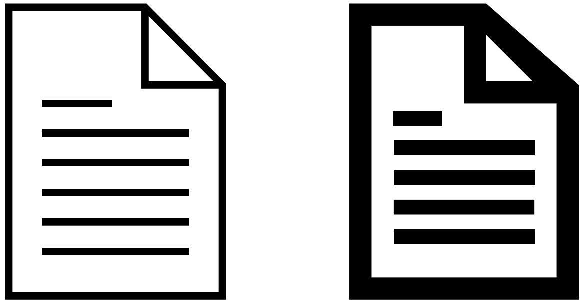
Natural image	CXRs	Mammo
CLIP *(ICML 2020)*	*1.* ***GLORIA*** *(ICCV2021)* *2.* ***MedCLIP*** *(EMNLP 2022)* *3.* ***CXR-CLIP*** *(MICCAI 2023)*	***Mammo-FM*** *(MICCAI 2024, ArXiv 2025)*

Aim 3

(Visual + non-Visual)

The goal: **Detect** the **Systematic Mistakes** using **Language**

Captions Reports



Vision-Language Alignment

Natural image

CLIP *(ICML 2020)*

CXRs

1. ***GLORIA*** *(ICCV2021)*

2. ***MedCLIP*** *(EMNLP 2022)*

3. ***CXR-CLIP*** *(MICCAI 2023)*

Mammo

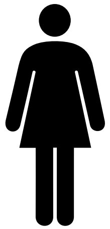
Mammo-FM *(MICCAI 2024, ArXiv 2025)*

Tracing non-visual mistakes

Population

Age: [32-88]
Race: 80% Non-Hispanic White, 20% Asian
….

Individual

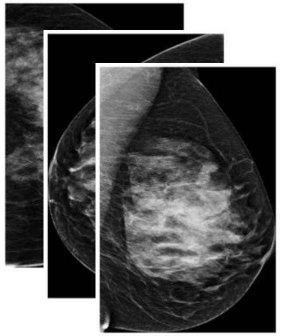


Reason for Visit: [….]
Blood Pressure: […]
Lab Test: [….]
…

Site

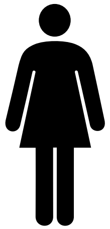
Manufacturer: [….]
X-ray Dosage: […]
Aperture Setting: [….]
…

Preprocessing

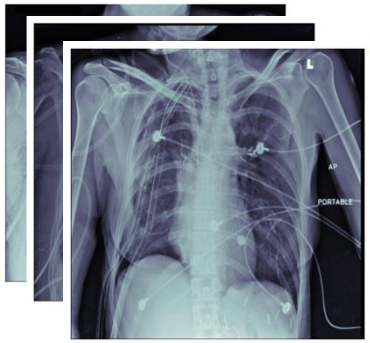


Photometric Interpretation: [Monochrome 1 vs Monochrome2]
Crop ratio: […]
…

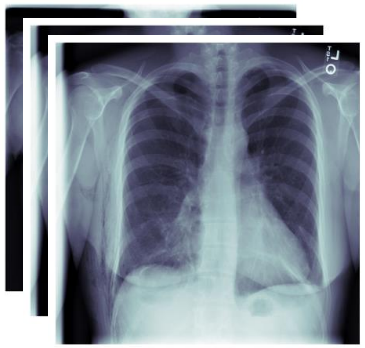
How a human would do?



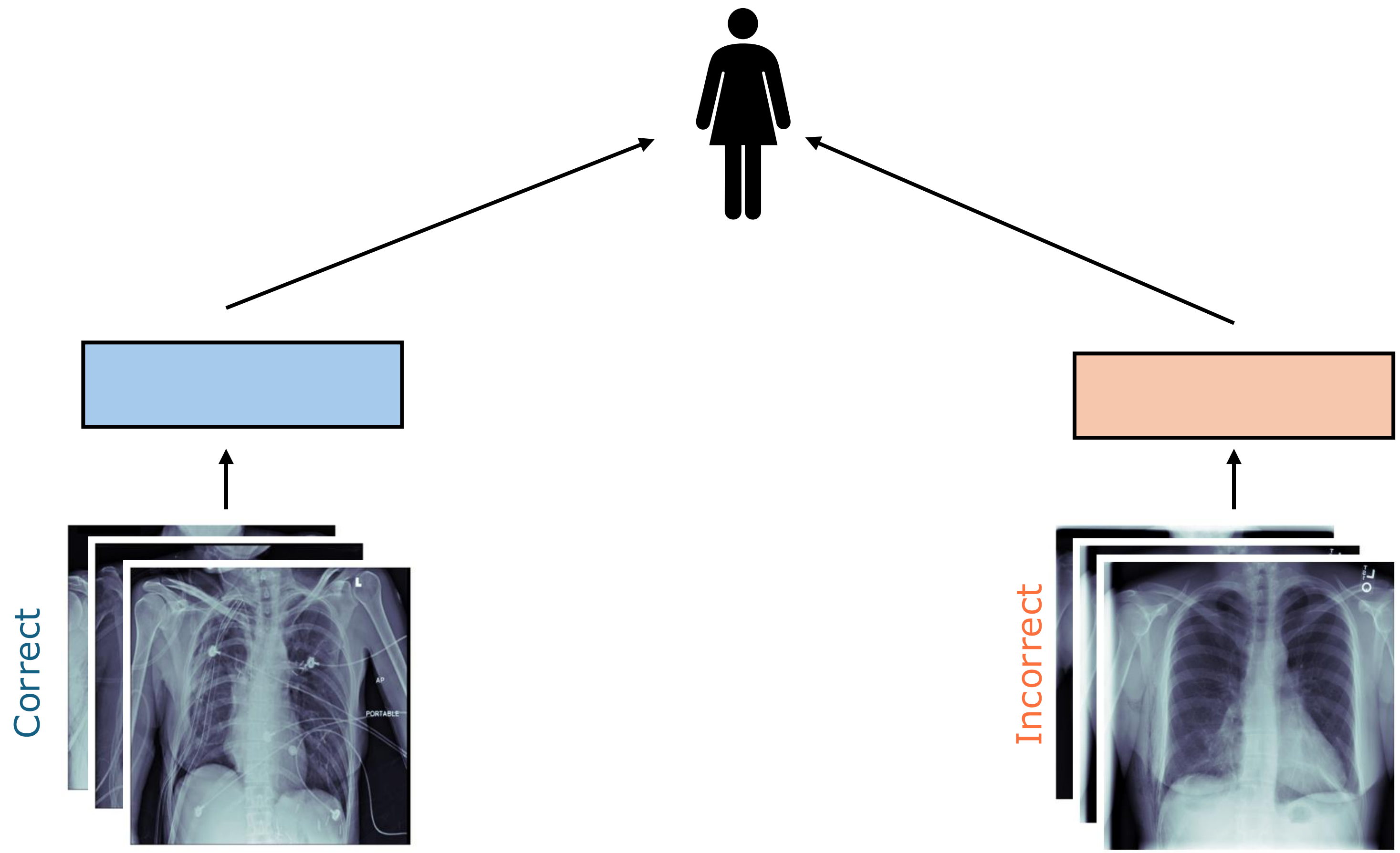
Correct



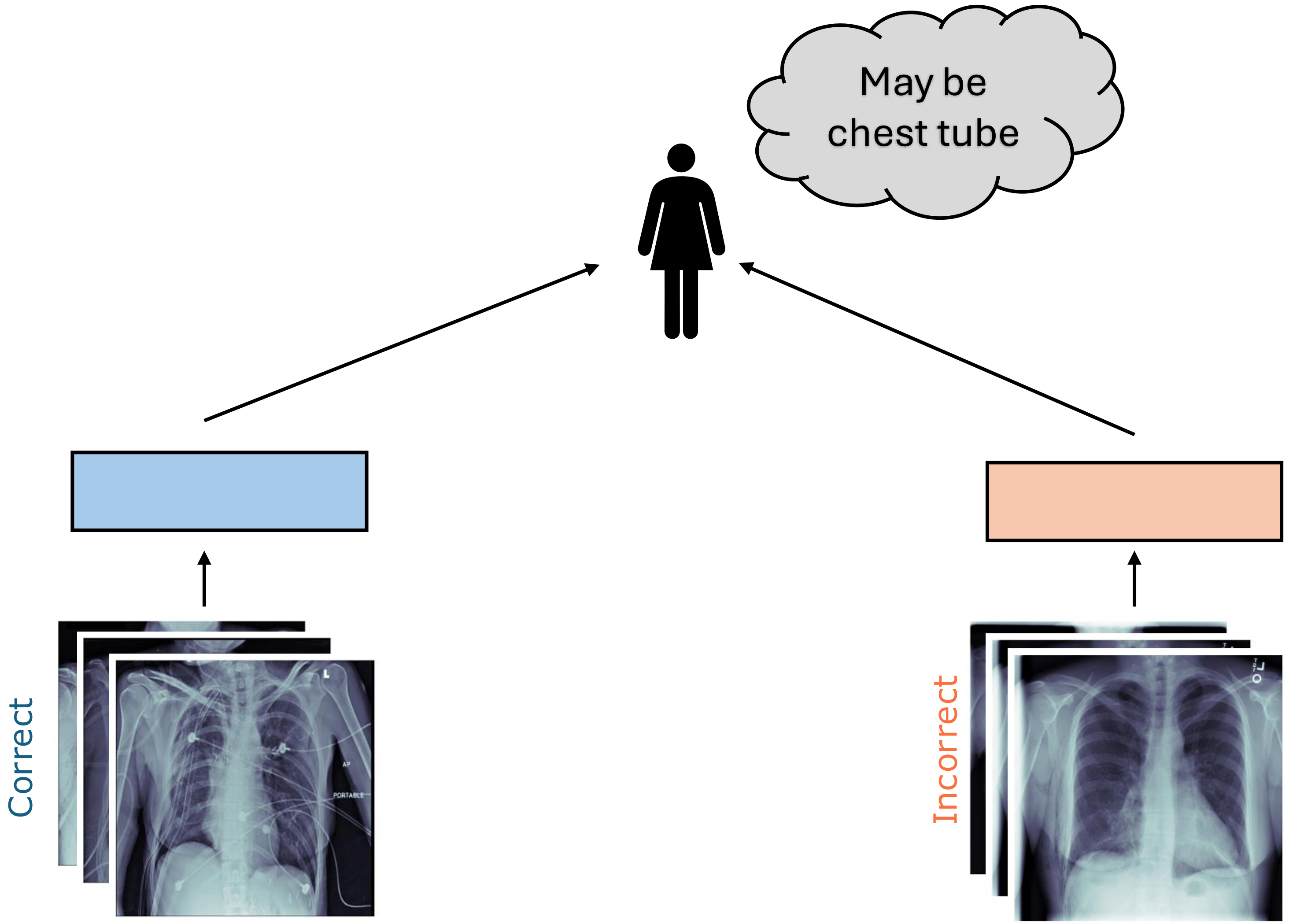
Incorrect



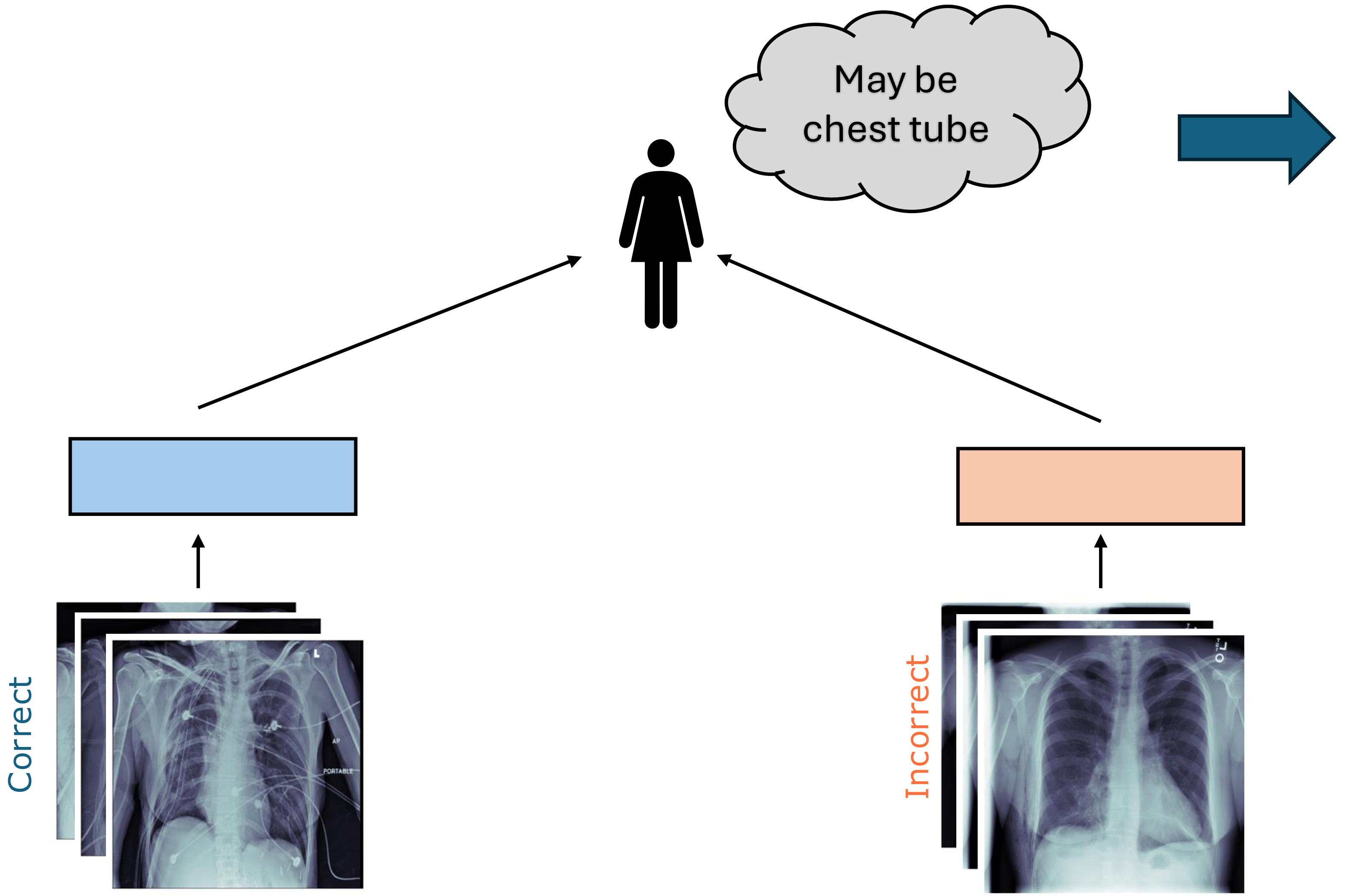
How a human would do?



How a human would do?

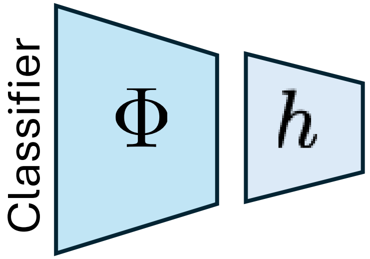


How a human would do?



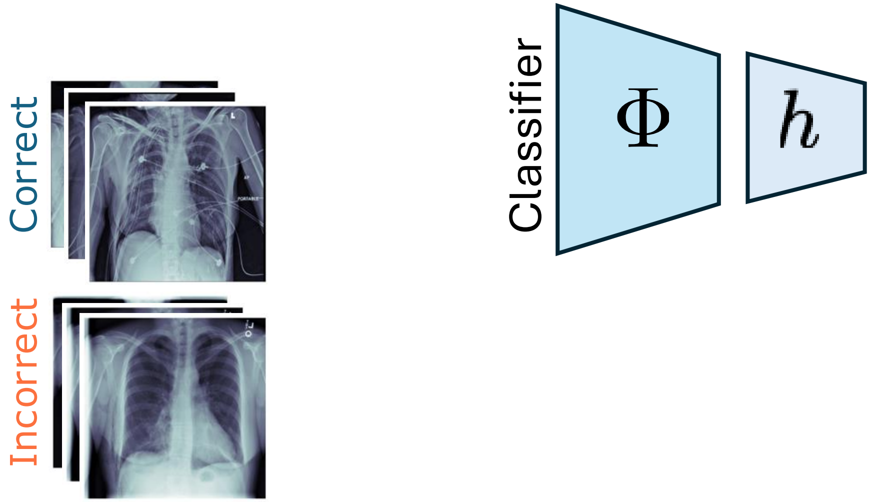
Validate:
Check the
performance gap

Ladder



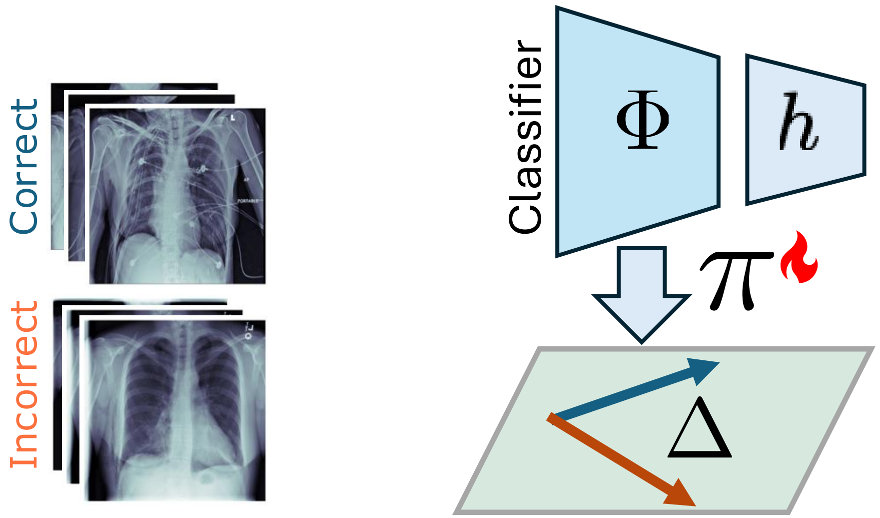
LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers. **Ghosh et al.** ACL 2025.

Ladder



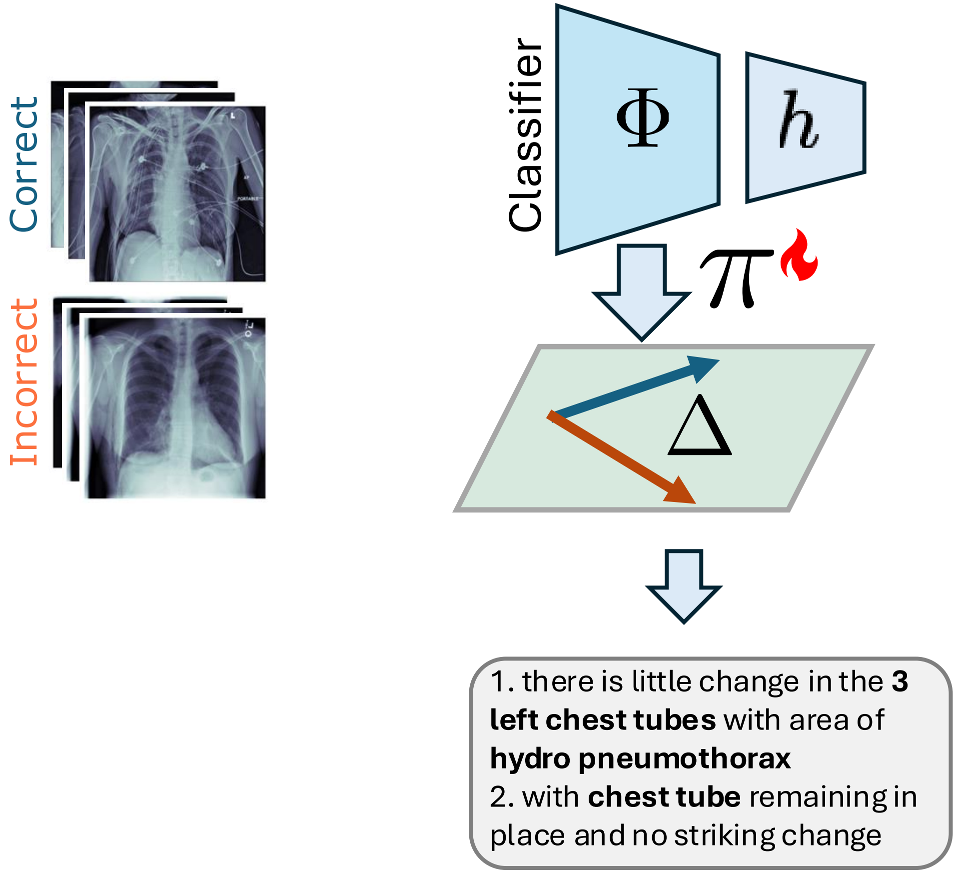
LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers. **Ghosh et al.** ACL 2025.

Ladder



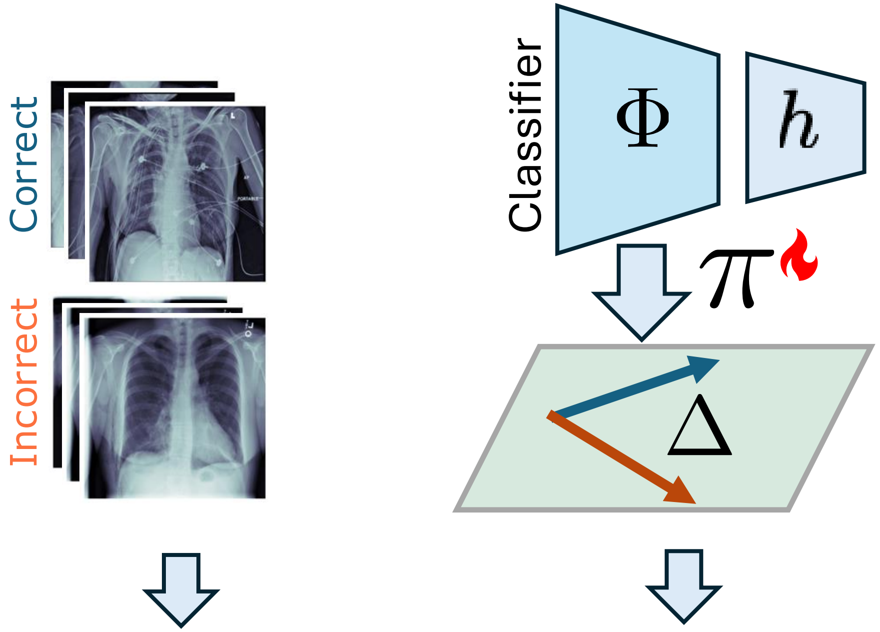
LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers. **Ghosh et al.** ACL 2025.

Ladder



LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers. **Ghosh et al.** ACL 2025.

Ladder



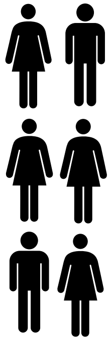
Metadata (e.g, DICOMS)

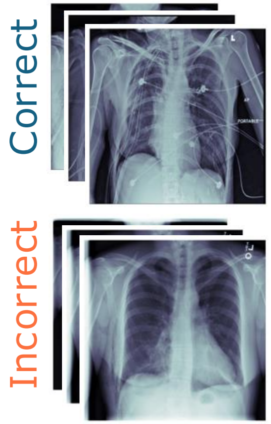
Manufacturer: [….]
X-ray Dosage: […]
Aperture Setting: [….]
…

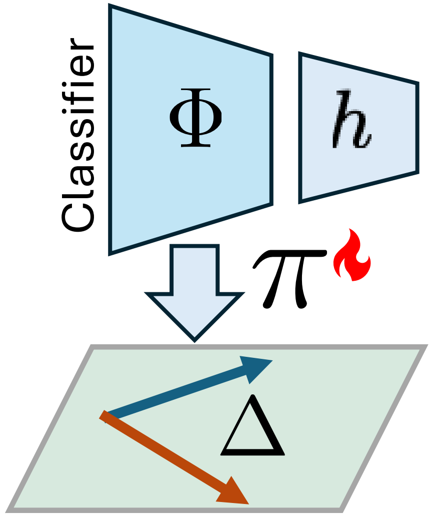
1. there is little change in the **3 left chest tubes** with area of **hydro pneumothorax**
2. with **chest tube** remaining in place and no striking change

LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers. **Ghosh et al.** ACL 2025.

Ladder







Patient Data (EHR)

Age: [....]
Blood Pressure: [...]
Lab Test: [....]
...

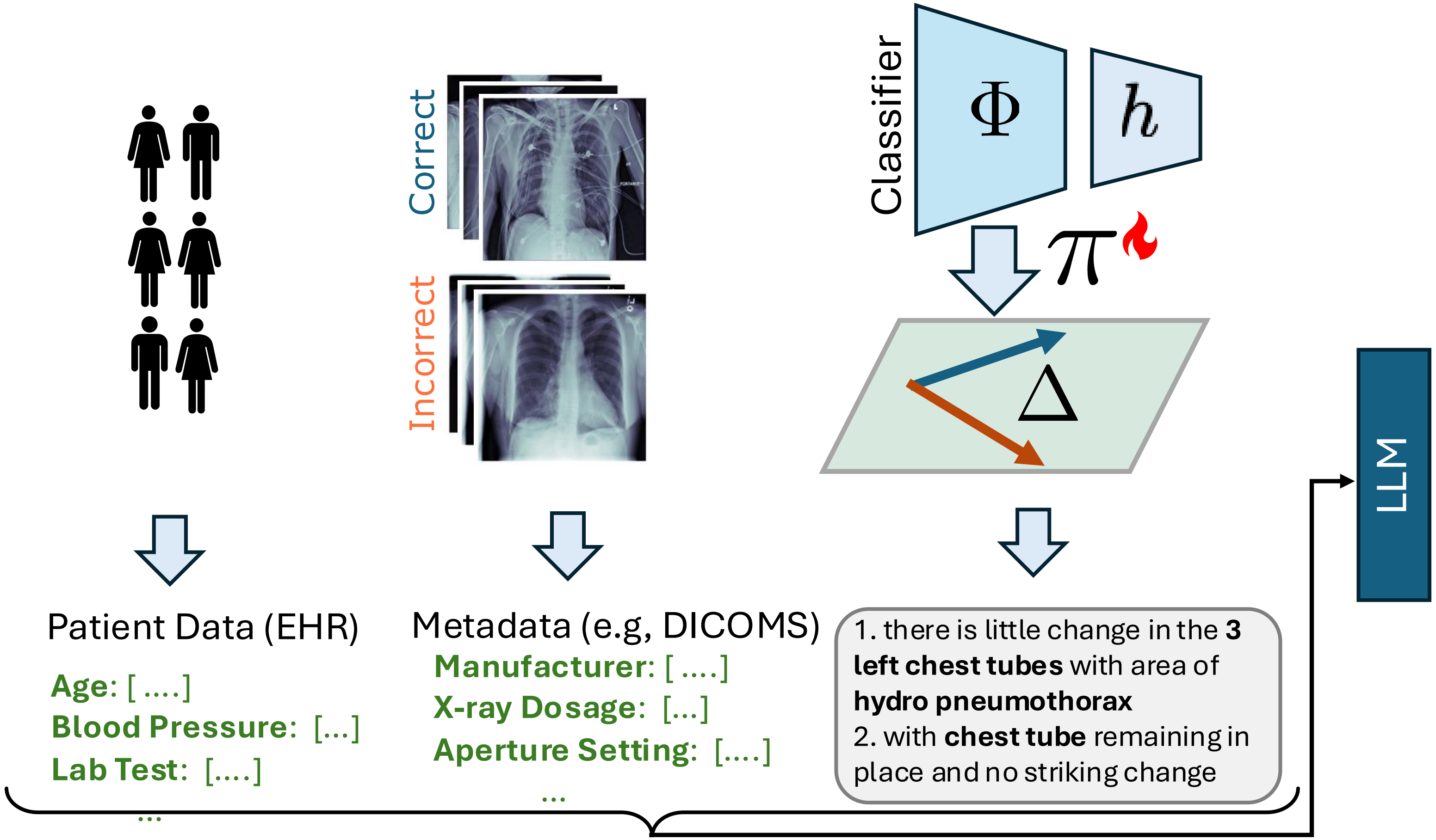
Metadata (e.g, DICOMS)

Manufacturer: [....]
X-ray Dosage: [...]
Aperture Setting: [....]
...

1. there is little change in the **3 left chest tubes** with area of **hydro pneumothorax**
2. with **chest tube** remaining in place and no striking change

LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers. **Ghosh et al.** ACL 2025.

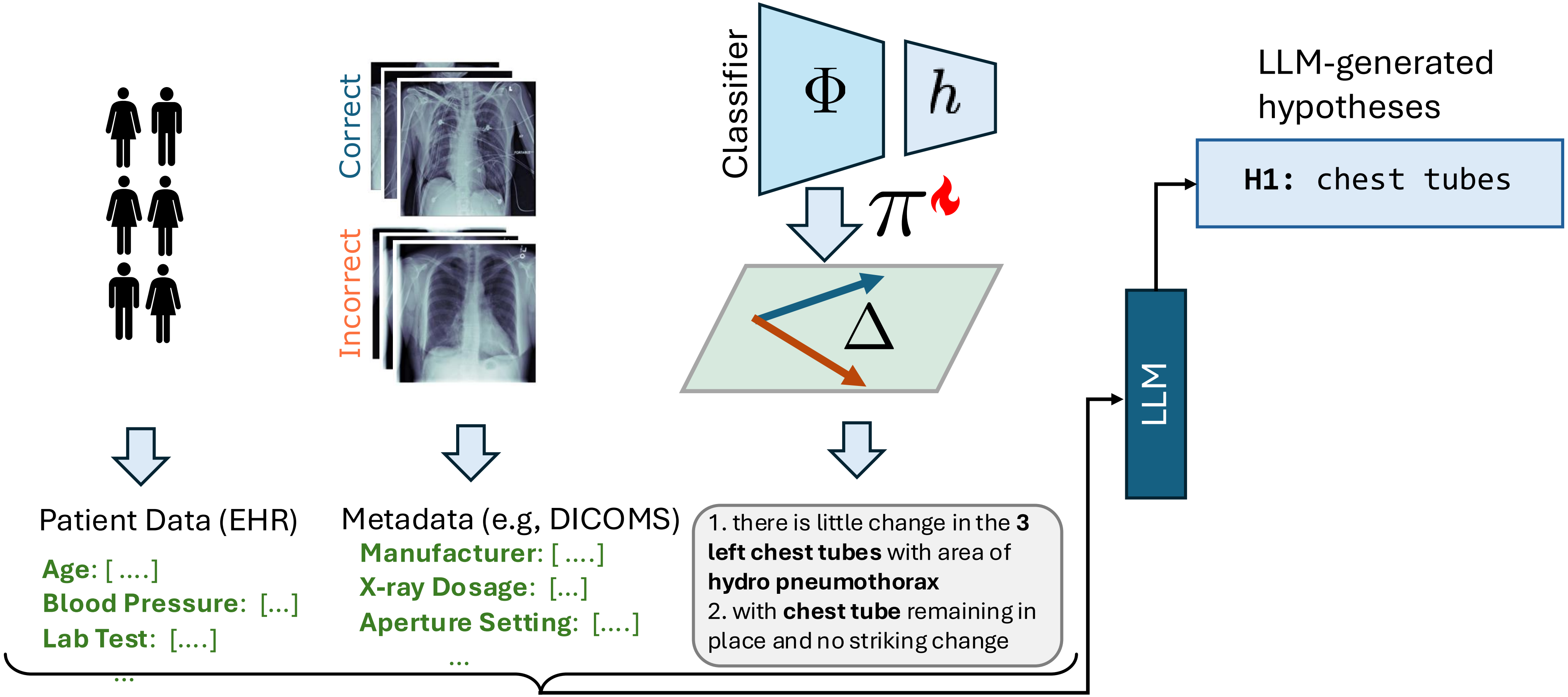
Ladder



The DICOMs and EHR is an illustrative example here. NIH-CXR does not have that. We use RSNA (Cancer) dataset to perform this experiment

LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers. **Ghosh et al.** ACL 2025.

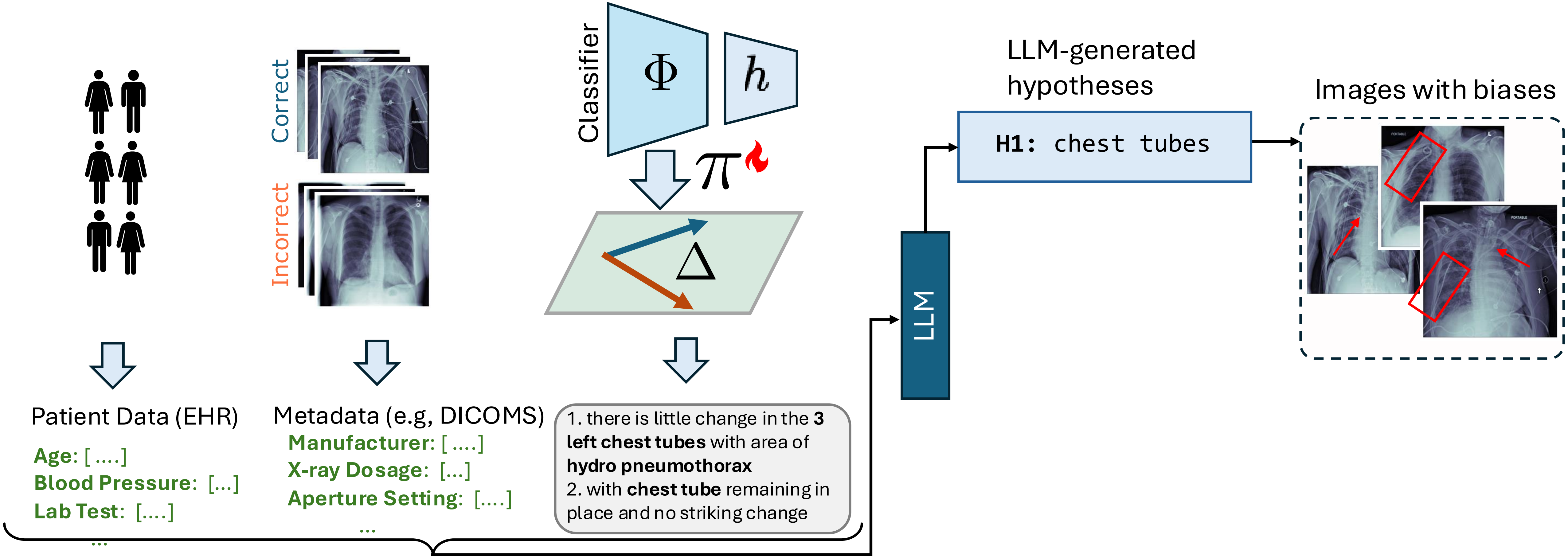
Ladder



The DICOMs and EHR is an illustrative example here. NIH-CXR does not have that. We use RSNA (Cancer) dataset to perform this experiment

LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers. **Ghosh et al.** ACL 2025.

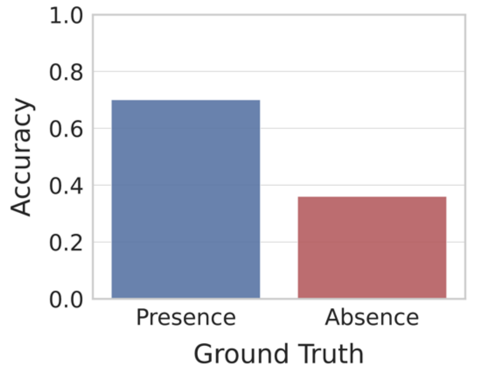
Ladder

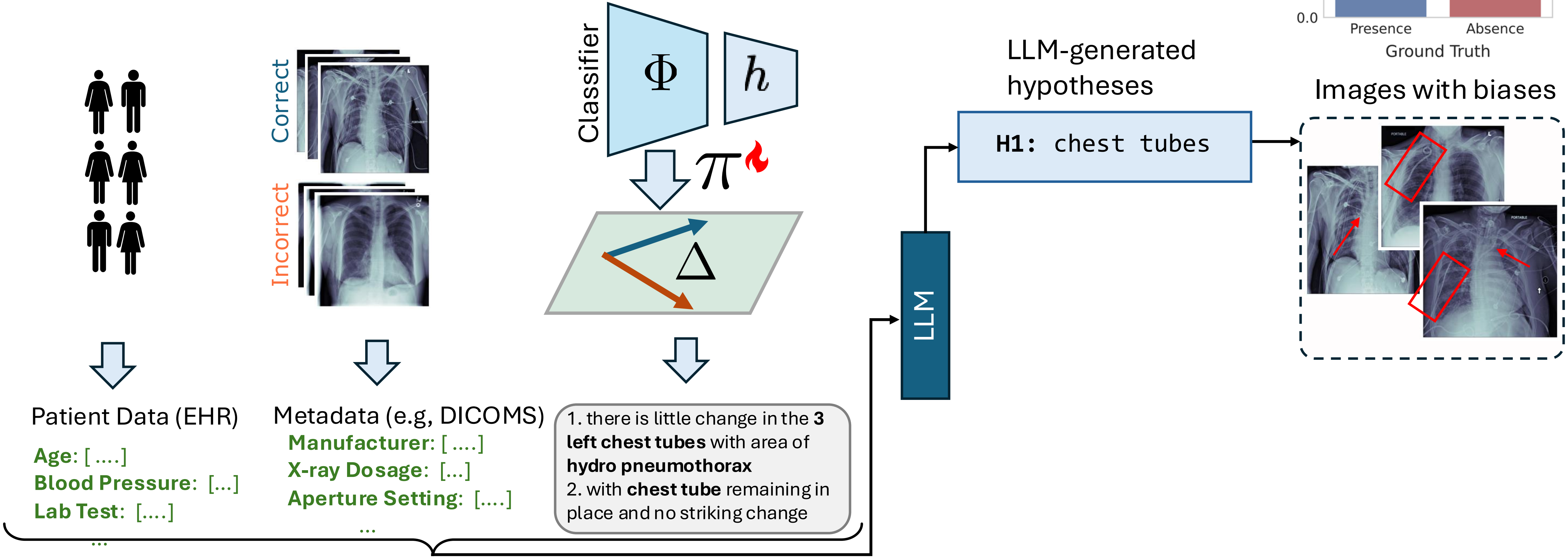


The DICOMs and EHR is an illustrative example here. NIH-CXR does not have that. We use RSNA (Cancer) dataset to perform this experiment

LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers. **Ghosh et al.** ACL 2025.

Ladder

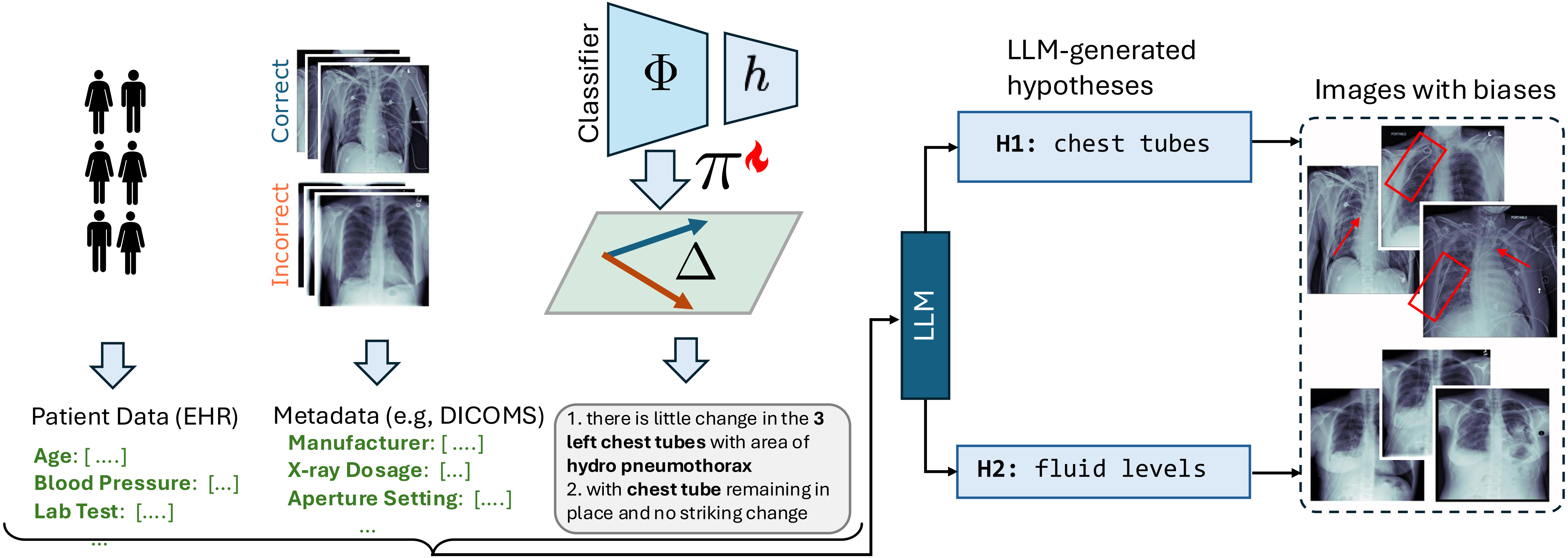




The DICOMs and EHR is an illustrative example here. NIH-CXR does not have that. We use RSNA (Cancer) dataset to perform this experiment

LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers. **Ghosh et al.** ACL 2025.

Ladder



The DICOMs and EHR is an illustrative example here. NIH-CXR does not have that. We use RSNA (Cancer) dataset to perform this experiment

LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers. **Ghosh et al.** ACL 2025.

Qualitative Results

Dataset: Waterbirds **Target: Waterbird**

Bias: Specific background elements like docks and boats

LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers. **Ghosh et al.** ACL 2025.

Qualitative Results

Dataset: Waterbirds **Target: Waterbird**

Bias: Specific background elements like docks and boats



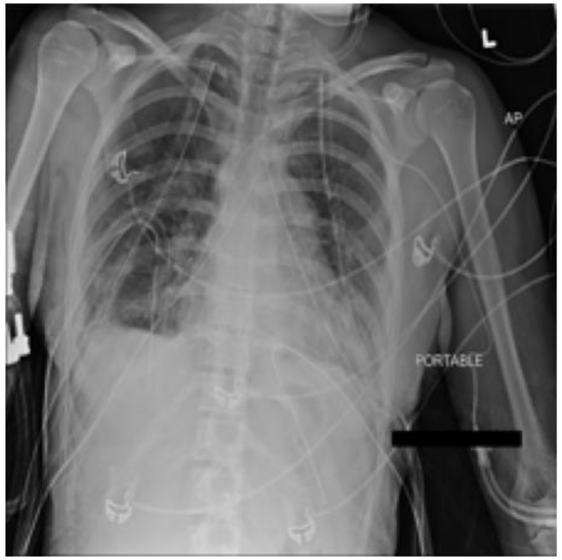

Presence: 97.2 % **Classifier Performance** **Absence: 68.8%**

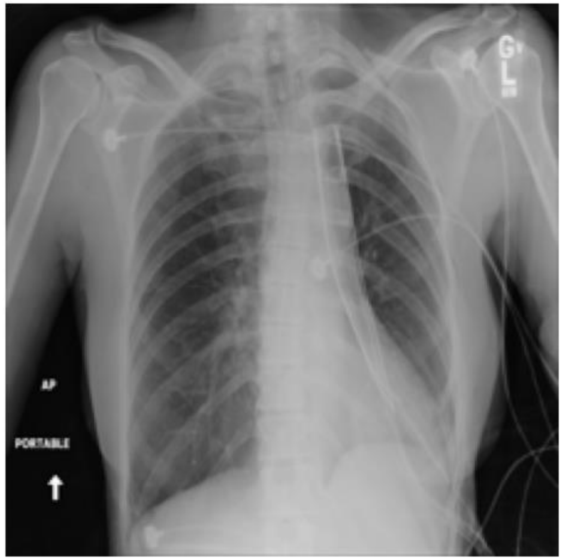
LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers. **Ghosh et al.** ACL 2025.

Qualitative Results

Dataset: NIH **Target: Pneumothorax**

Bias: Chest tubes



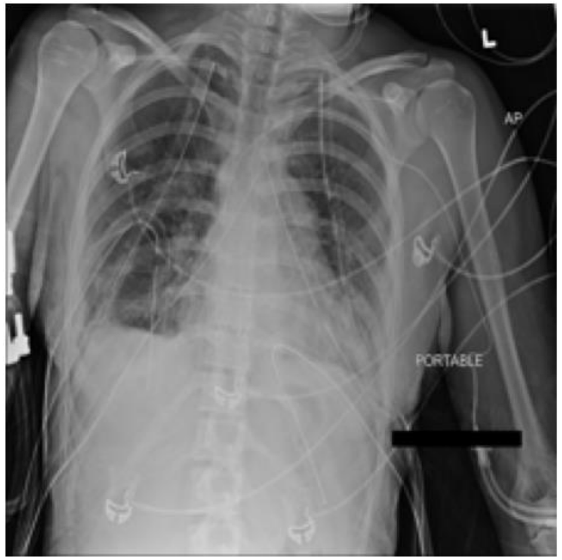


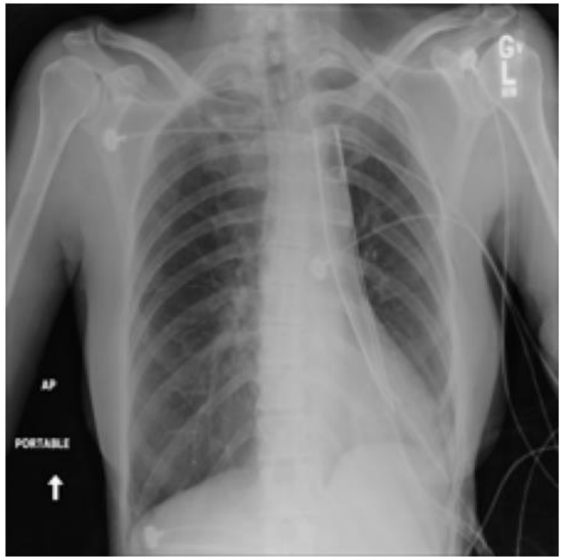
LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers. **Ghosh et al.** ACL 2025.

Qualitative Results

Dataset: NIH **Target: Pneumothorax**

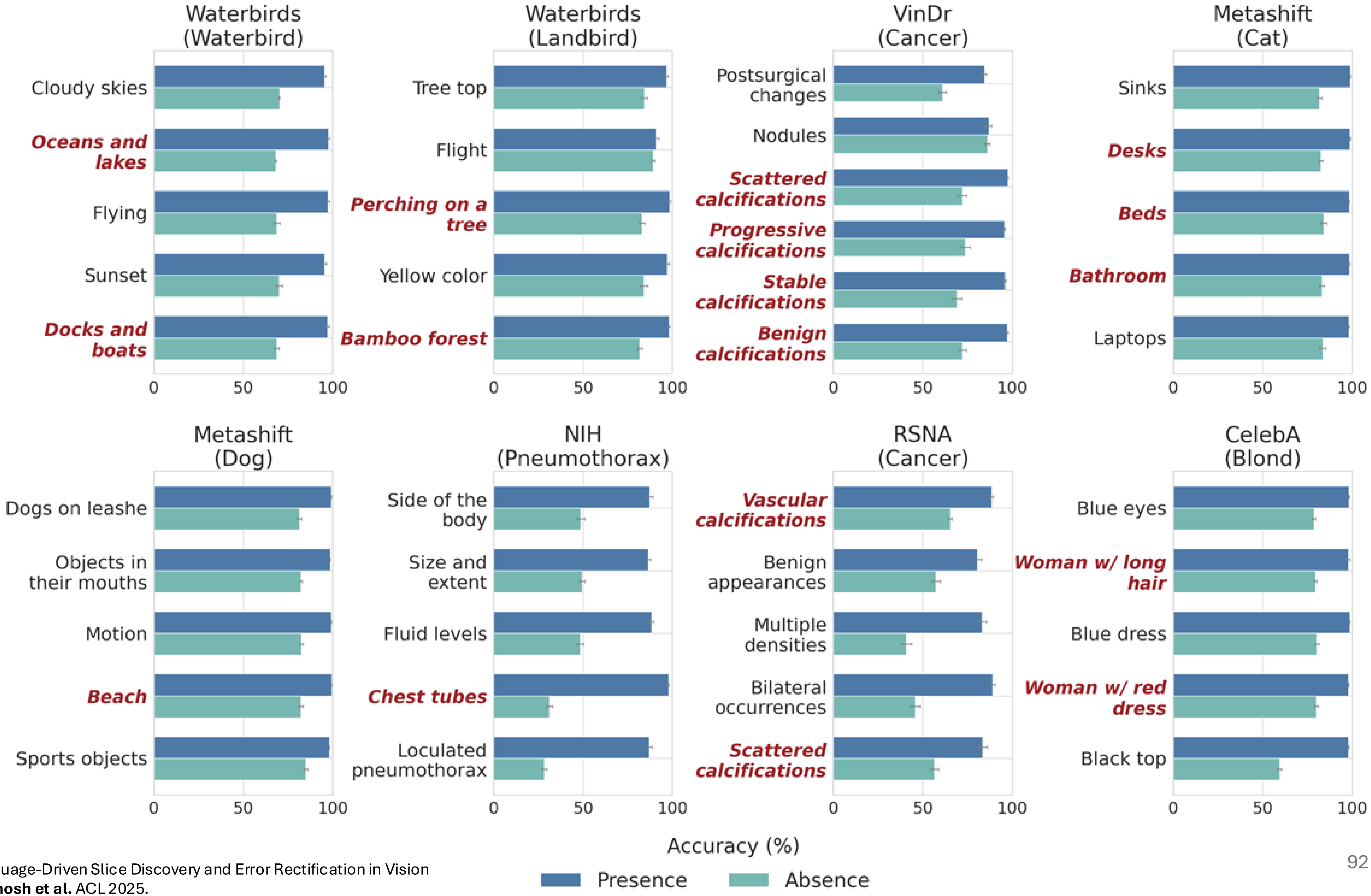
Bias: Chest tubes





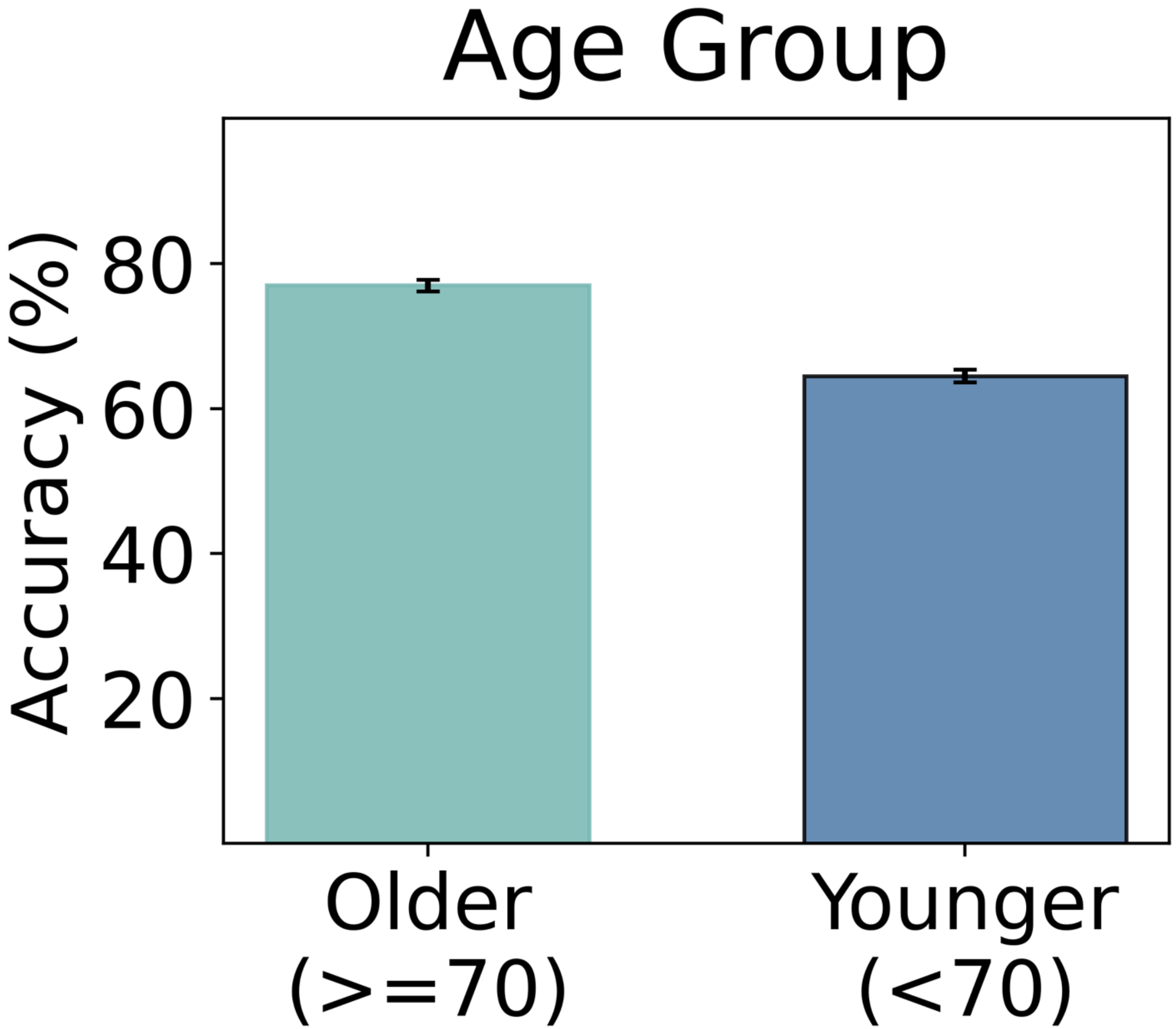
Presence: 76.2 % **Classifier Performance** **Absence: 34.8%**

LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers. **Ghosh et al.** ACL 2025.



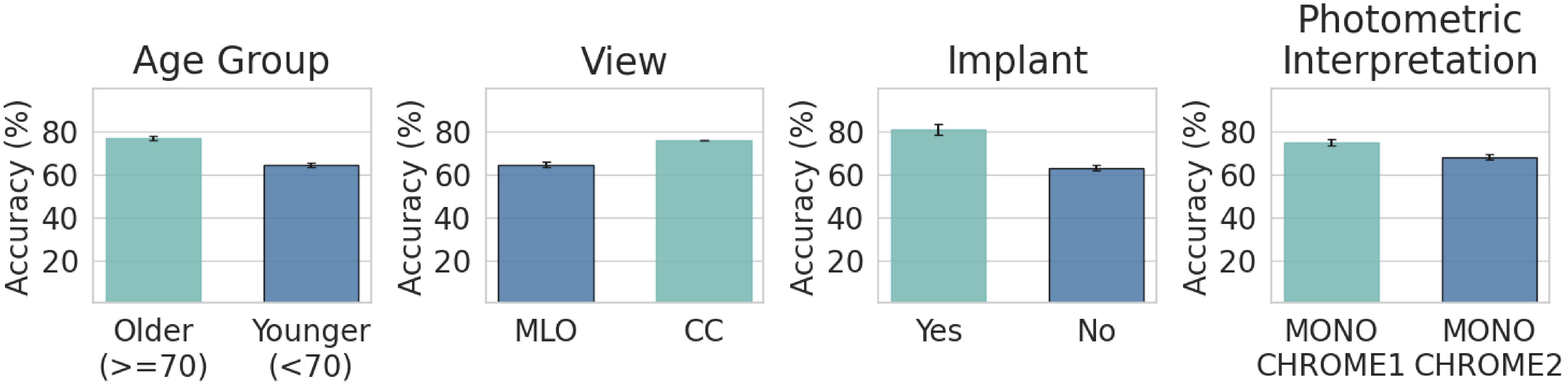
LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers. **Ghosh et al.** ACL 2025.

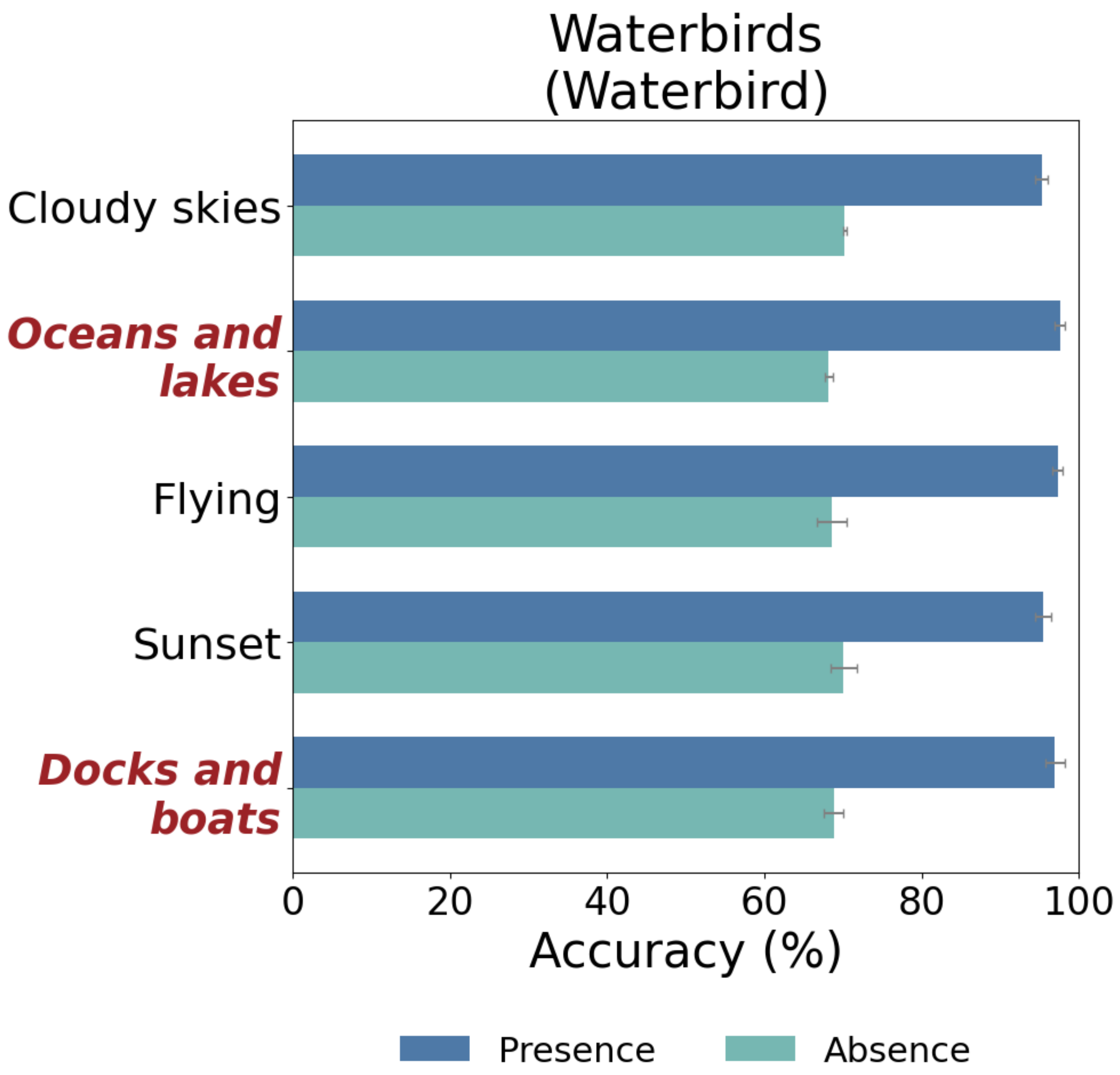
Detecting non-visual mistakes



Detecting non-visual mistakes

Dataset: RSNA **Target: Breast cancer**





$$\{\underbrace{A_1, A_2, \ldots A_k}_{\text{Explained by data distribution shift}}, \underbrace{A_{k1} \ldots A_n}_{\text{Completely unexplained}}\}$$

Explained by data distribution shift

(Reject the hypotheses)

Completely unexplained

Lets a take step back

$$G_a \;=\; \mathbb{E}[m(R,Y) \mid A=a] \;-\; \mathbb{E}[m(R,Y)]$$

Lets a take step back

$$G_a = \overbrace{\mathbb{E}[m(R, Y) \mid A = a]}^{\text{Performance of the subgroup of interest, e.g. Age or younger patient}} - \mathbb{E}[m(R, Y)]$$

Lets a take step back

$$G_a = \overbrace{\mathbb{E}[m(R,Y) \mid A = a]}^{\text{Performance of the subgroup of interest, e.g. Age or younger patient}} - \overbrace{\mathbb{E}[m(R,Y)]}^{\text{Avg Performance of the model}}$$

Lets a take step back

$$G_a = \underbrace{\mathbb{E}[m(R,Y) \mid A = a]}^{\text{Performance of the subgroup of interest, e.g. Age or younger patient}} - \overbrace{\mathbb{E}[m(R,Y)]}^{\text{Avg Performance of the model}}$$

Metric (accuracy, recall)

Prediction from model

Ground-truth label

Lets a take step back

$$G_a = \overbrace{\mathbb{E}[m(R,Y) \mid A = a]}^{\text{Performance of the subgroup of interest, e.g. Age or younger patient}} - \overbrace{\mathbb{E}[m(R,Y)]}^{\text{Avg Performance of the model}}$$

Metric (accuracy, recall)

Prediction from model

Ground-truth label

m

X: Covariates or images

Y: Label

A: Subgroups

Understanding challenges to the interpretation of disaggregated evaluations of algorithmic fairness. Pfohl et al. NeurIPS 2025.

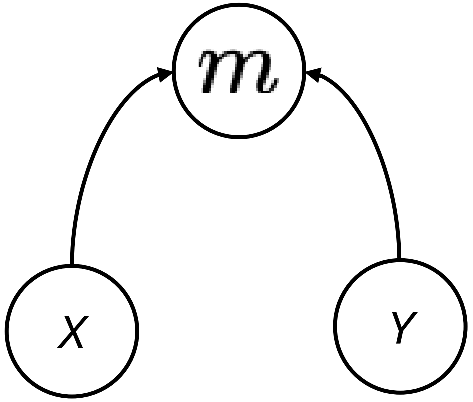
Lets a take step back

$$G_a = \overbrace{\mathbb{E}[m(R,Y) \mid A = a]}^{\text{Performance of the subgroup of interest, e.g. Age or younger patient}} - \overbrace{\mathbb{E}[m(R,Y)]}^{\text{Avg Performance of the model}}$$

Metric (accuracy, recall)

Prediction from model

Ground-truth label



X: Covariates or images

Y: Label

A: Subgroups

Understanding challenges to the interpretation of disaggregated evaluations of algorithmic fairness. Pfohl et al. NeurIPS 2025.

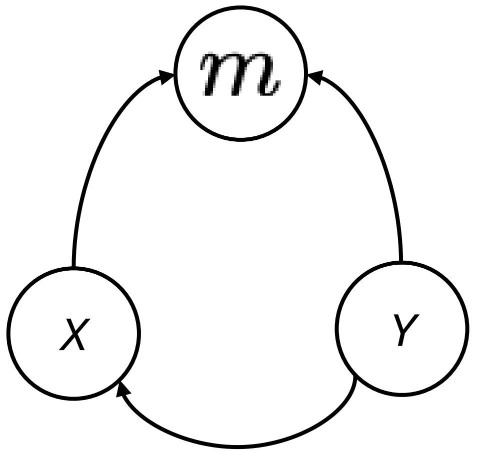
Lets a take step back

$$G_a = \overbrace{\mathbb{E}[m(R,Y) \mid A = a]}^{\text{Performance of the subgroup of interest, e.g. Age or younger patient}} - \overbrace{\mathbb{E}[m(R,Y)]}^{\text{Avg Performance of the model}}$$

Metric (accuracy, recall)

Prediction from model

Ground-truth label



X: Covariates or images

Y: Label

A: Subgroups

Understanding challenges to the interpretation of disaggregated evaluations of algorithmic fairness. Pfohl et al. NeurIPS 2025.

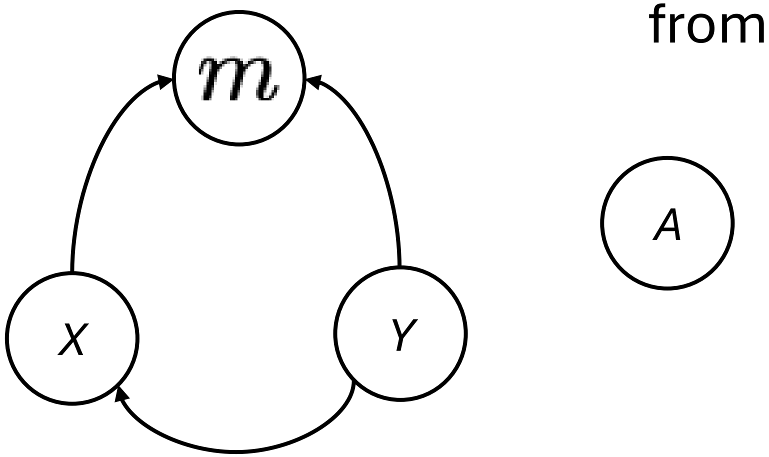
Lets a take step back

$$G_a = \overbrace{\mathbb{E}[m(R,Y) \mid A = a]}^{\text{Performance of the subgroup of interest, e.g. Age or younger patient}} - \overbrace{\mathbb{E}[m(R,Y)]}^{\text{Avg Performance of the model}}$$

Metric (accuracy, recall)

Prediction from model

Ground-truth label



X: Covariates or images
Y: Label
A: Subgroups

Lets a take step back

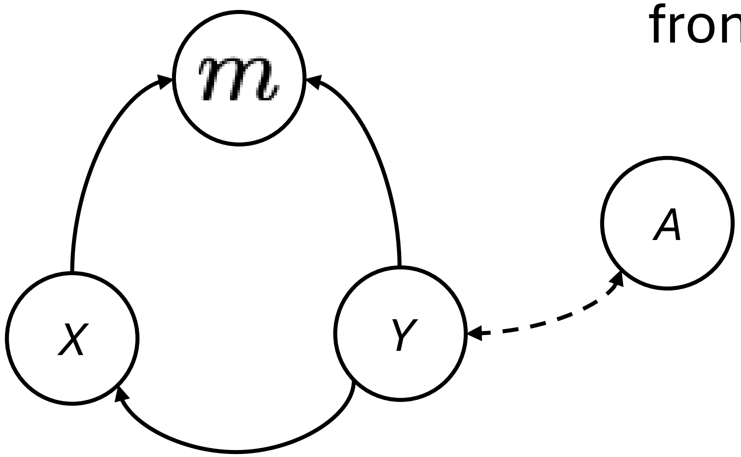
$$G_a = \overbrace{\mathbb{E}[m(R,Y) \mid A = a]}^{\text{Performance of the subgroup of interest, e.g. Age or younger patient}} - \overbrace{\mathbb{E}[m(R,Y)]}^{\text{Avg Performance of the model}}$$

Metric (accuracy, recall)

Prediction from model

Ground-truth label

Label Shift



X: Covariates or images

Y: Label

A: Subgroups

Unobserved confounding

Causal direction

Understanding challenges to the interpretation of disaggregated evaluations of algorithmic fairness. Pfohl et al. NeurIPS 2025.

Lets a take step back

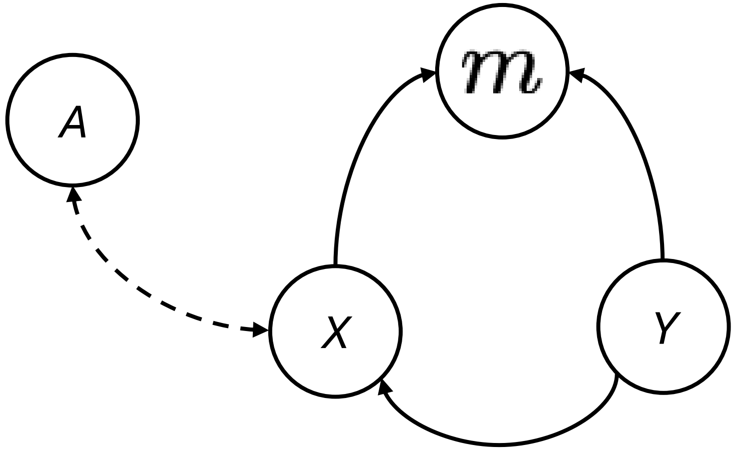
$$G_a = \overbrace{\mathbb{E}[m(R,Y) \mid A = a]}^{\text{Performance of the subgroup of interest, e.g. Age or younger patient}} - \overbrace{\mathbb{E}[m(R,Y)]}^{\text{Avg Performance of the model}}$$

Metric (accuracy, recall)

Prediction from model

Ground-truth label

Presentation Shift



X: Covariates or images

Y: Label

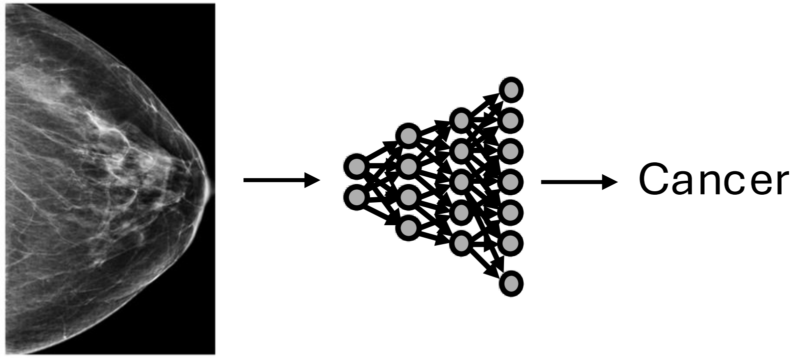
A: Subgroups

Unobserved confounding

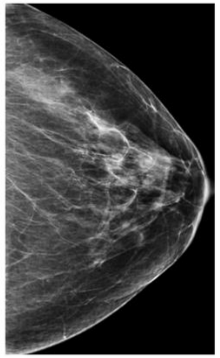
Causal direction

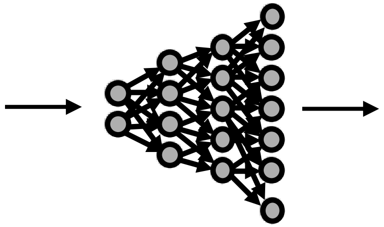
Understanding challenges to the interpretation of disaggregated evaluations of algorithmic fairness. Pfohl et al. NeurIPS 2025.

A toy example



A toy example



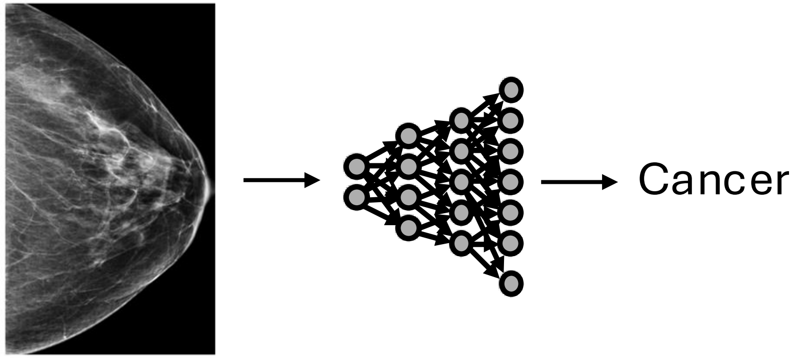


Cancer

Population (on cancer patients)

Avg Model error = 11 %

A toy example



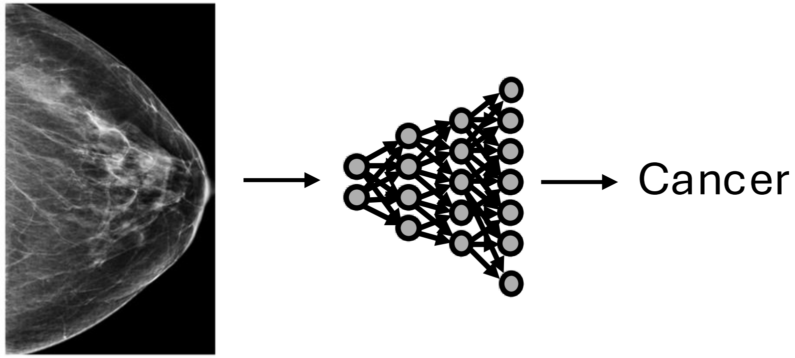
Population (on cancer patients)

Avg Model error = 11 %

Subgroup (on younger patients)

Model error = 29%

A toy example



Population (on cancer patients)

Avg Model error = 11 %

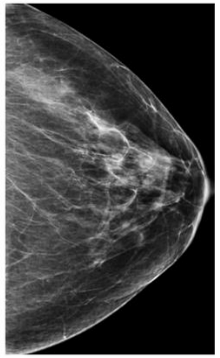
Significant Gap

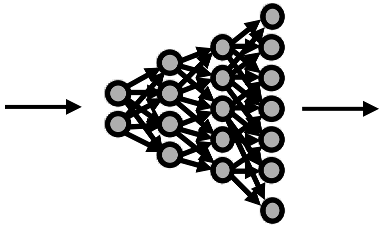
Subgroup (on younger patients)

Model error = 29%

Conclusion: The model is struggling on younger patients

A toy example





Cancer

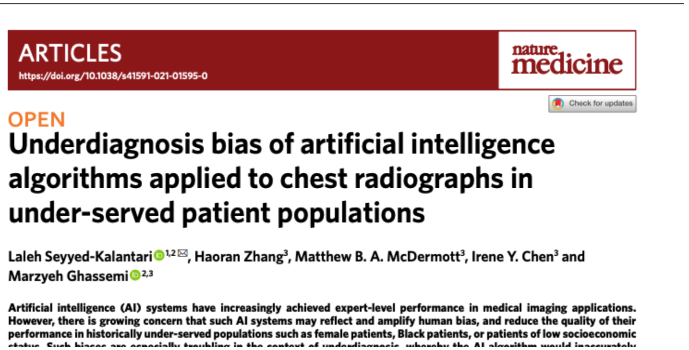
Population (on cancer patients)

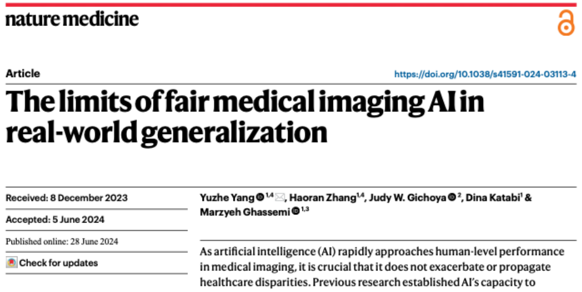
Avg Model error = **11 %**

Significant Gap

Subgroup (on younger patients)

Model error = **29%**

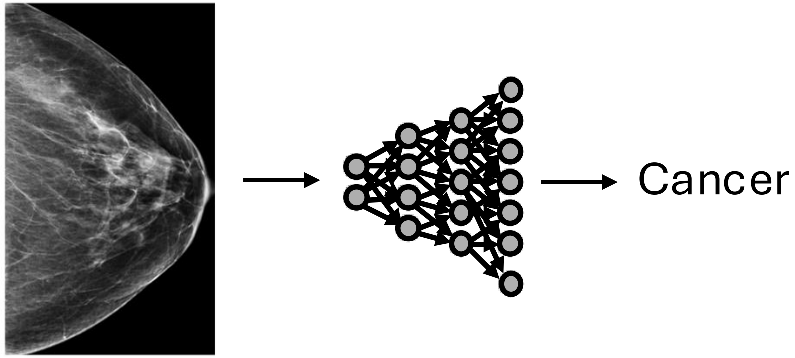

ARTICLES
https://doi.org/10.1038/s41591-021-01595-0
nature medicine
Check for updates
OPEN
Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations
Laleh Seyyed-Kalantari[1,2], Haoran Zhang[3], Matthew B. A. McDermott[3], Irene Y. Chen[3] and Marzyeh Ghassemi[2,3]
Artificial intelligence (AI) systems have increasingly achieved expert-level performance in medical imaging applications. However, there is growing concern that such AI systems may reflect and amplify human bias, and reduce the quality of their performance in historically under-served populations such as female patients, Black patients, or patients of low socioeconomic


nature medicine
Article
https://doi.org/10.1038/s41591-024-03113-4
The limits of fair medical imaging AI in real-world generalization
Received: 8 December 2023
Accepted: 5 June 2024
Published online: 28 June 2024
Check for updates
Yuzhe Yang[1,4], Haoran Zhang[1,4], Judy W. Gichoya[2], Dina Katabi[1] & Marzyeh Ghassemi[1,3]
As artificial intelligence (AI) rapidly approaches human-level performance in medical imaging, it is crucial that it does not exacerbate or propagate healthcare disparities. Previous research established AI's capacity to

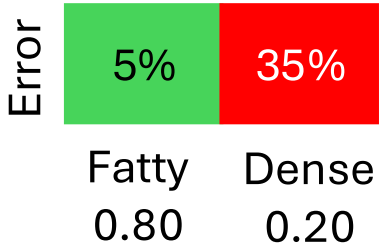
Action: Balanced training, reweighting **w.r.t age** (GroupDRO, JTT, DFR)

Conclusion: The model is struggling on younger patients

A toy example



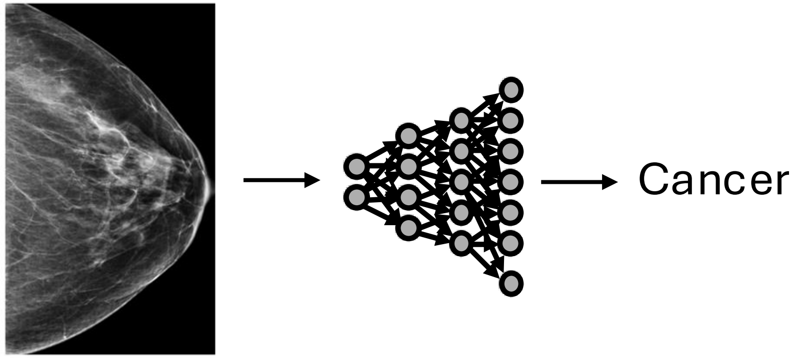
Population



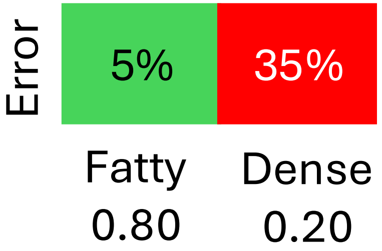
Avg Model error = (0.8 *5) + (0.2 *35)

= 11%

A toy example



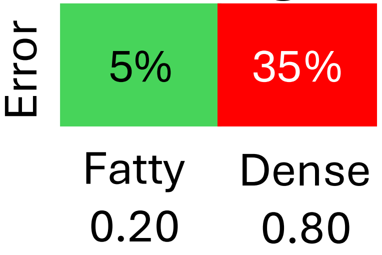
Population



Avg Model error = (0.8 *5) + (0.2 *35)

= 11%

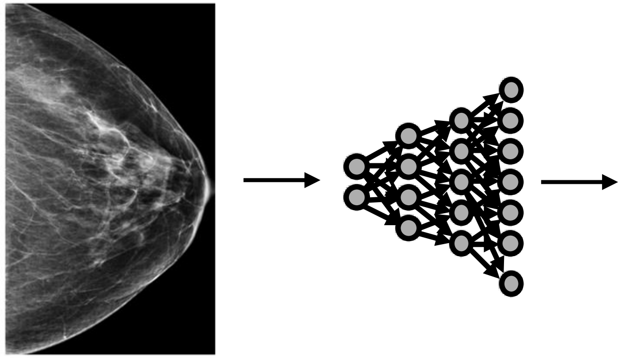
Young



Model error = (0.2 *5) + (0.8 *35)

= 29%

A toy example



Cancer

Population

Error

5%	35%
Fatty 0.80	Dense 0.20

Avg Model error = (0.8 *5) + (0.2 *35)

= 11%

Significant Gap

Young

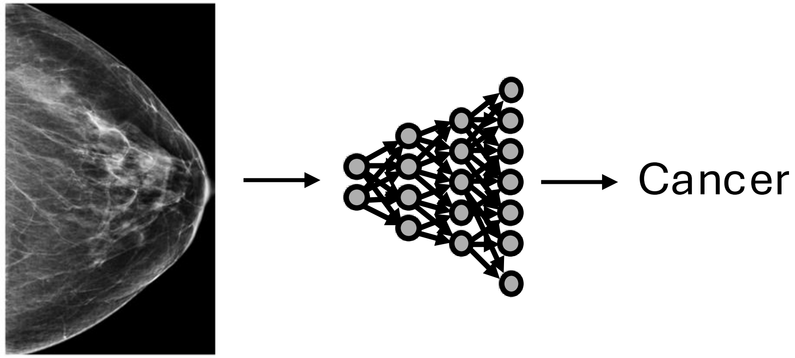
Error

5%	35%
Fatty 0.20	Dense 0.80

Model error = (0.2 *5) + (0.8 *35)

= 29%

A toy example

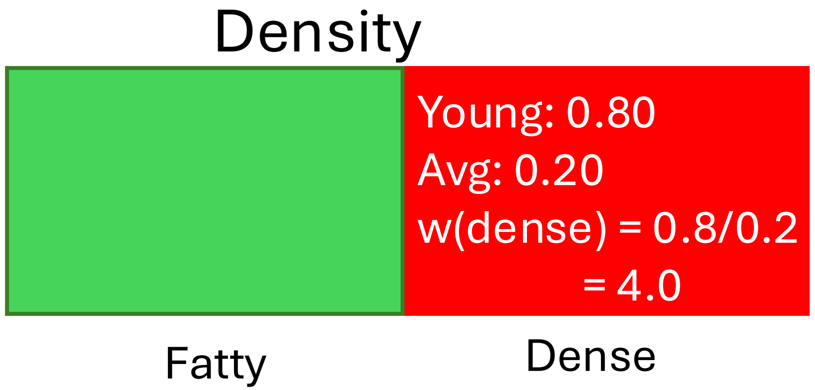


Population

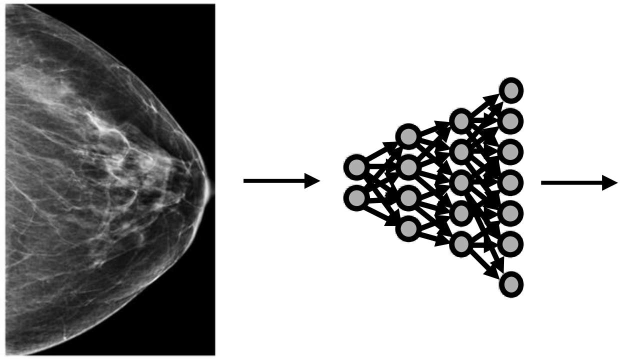
Avg Model error = (0.8 *5) + (0.2 *35)
= **11%**

Young

Model error = (0.2 *5) + (0.8 *35)
= **29%**



A toy example



Cancer

Population

Error

Fatty	Dense
	35%
0.80	0.20

Avg Model error = (0.8 *5) + (0.2 *35)

= 11%

Young

Error

Fatty	Dense
5%	35%
0.20	0.80

Model error = (0.2 *5) + (0.8 *35)

= 29%

Density

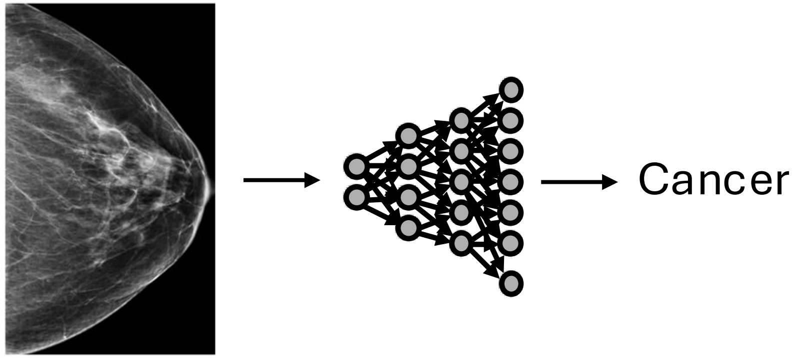
Fatty	Dense
Young: 0.20 Avg: 0.80 w(fatty) = 0.2/0.8 = 0.25	Young: 0.80 Avg: 0.20 w(dense) = 0.8/0.2 = 4.0

Model error Stratified by density

=(0.25*0.8*5 +4.0*0.2*35)

= 29%

A toy example



Population

Error

35%

Fatty 0.80 | Dense 0.20

Avg Model error = (0.8 *5) + (0.2 *35)

= 11%

Young

Error

5%	35%
Fatty 0.20	Dense 0.80

Model error = (0.2 *5) + (0.8 *35)

= 29%

Density

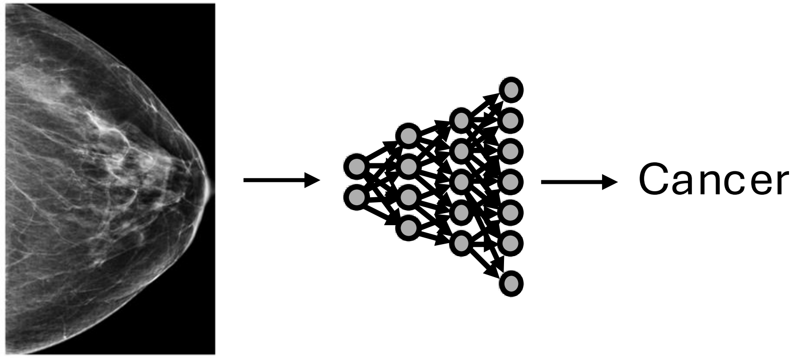
Young: 0.20 Avg: 0.80 w(fatty) = 0.2/0.8 = 0.25	Young: 0.80 Avg: 0.20 w(dense) = 0.8/0.2 = 4.0
Fatty	Dense

Model error Stratified by density

= 29%

Matches

A toy example



Population

Error

Fatty 0.80	Dense 0.20
5%	35%

Avg Model error = (0.8 *5) + (0.2 *35)

= 11%

Density

Fatty	Dense
Young: 0.20 Avg: 0.80 w(fatty) = 0.2/0.8 = 0.25	Young: 0.80 Avg: 0.20 w(dense) = 0.8/0.2 = 4.0

Down weight easy one

up weight hard one

Model error Stratified by density

= 29%

Young

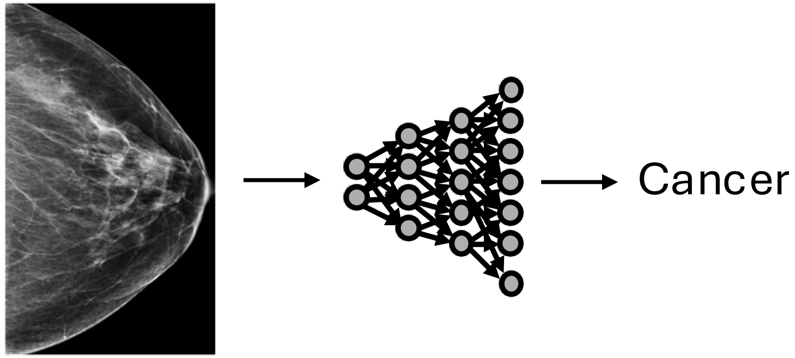
Error

Fatty 0.20	Dense 0.80
5%	35%

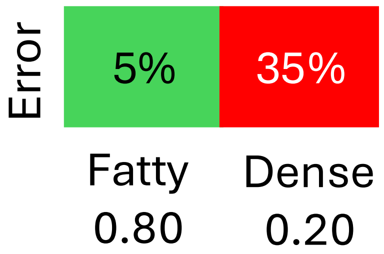
Model error = (0.2 *5) + (0.8 *35)

= 29%

A toy example

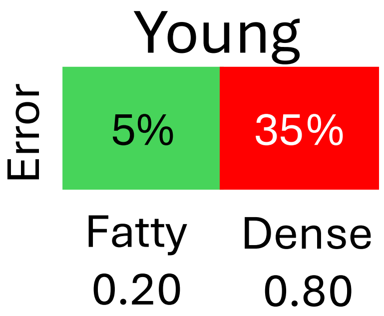


Population



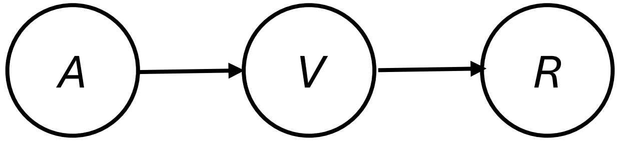
Avg Model error = (0.8 *5) + (0.2 *35)

= 11%



Model error = (0.2 *5) + (0.8 *35)

= 29%



Pseudo young Population

Density (V)

Fatty	Dense
Young: 0.20 Avg: 0.80 w(fatty) = 0.2/0.8 = 0.25	Young: 0.80 Avg: 0.20 w(dense) = 0.8/0.2 = 4.0

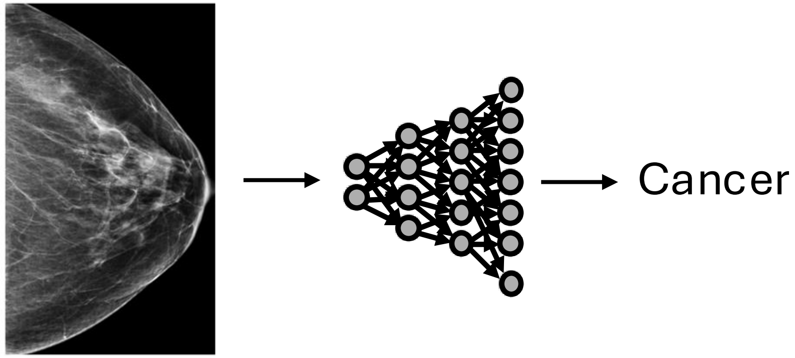
Down weight easy one

up weight hard one

Model error Stratified by density

= 29%

A toy example



Population

	Fatty 0.80	Dense 0.20
Error	5%	35%

Avg Model error = (0.8 *5) + (0.2 *35)
= 11%

Pseudo young Population

Density (V)

Fatty	Dense
Young: 0.20 Avg: 0.80 w(fatty) = 0.2/0.8 = 0.25	Young: 0.80 Avg: 0.20 w(dense) = 0.8/0.2 = 4.0

Down weight easy one

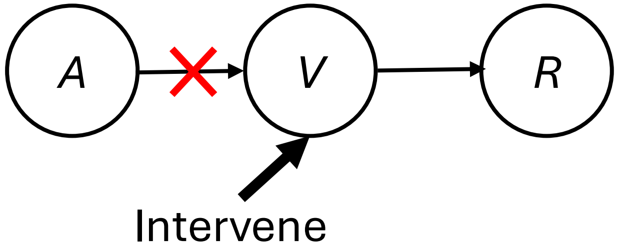
up weight hard one

Model error Stratified by density
= 29%

Young

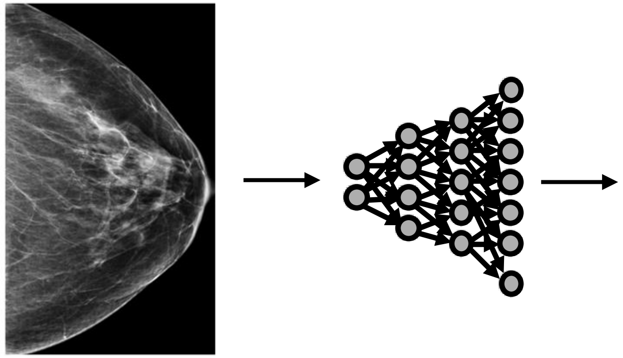
	Fatty 0.20	Dense 0.80
Error	5%	35%

Model error = (0.2 *5) + (0.8 *35)
= 29%



What would happen if population is same as the subgroup?

A toy example



Cancer

Population

Error

5%	35%
Fatty 0.80	Dense 0.20

Avg Model error = (0.8 *5) + (0.2 *35)

= 11%

Pseudo young Population

Density (V)

Young: 0.20 Avg: 0.80 w(fatty) = 0.2/0.8 = 0.25	Young: 0.80 Avg: 0.20 w(dense) = 0.8/0.2 = 4.0
Fatty	Dense

Down weight easy one

up weight hard one

Model error Stratified by density

= 29%

Young

Error

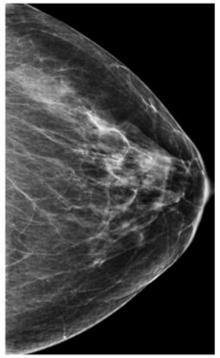
5%	35%
Fatty 0.20	Dense 0.80

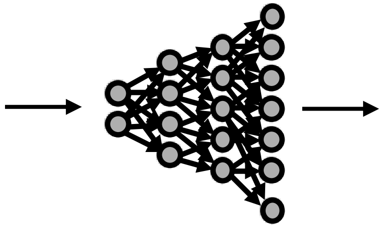
Model error = (0.2 *5) + (0.8 *35)

= 29%

Conclusion: The model is struggling on ~~younger patients~~ **dense breasts**

A toy example





Cancer

$$\mathbb{E}[m(R,Y) \mid A = a]$$

Young

Error	5%	35%
	Fatty 0.20	Dense 0.80

Model error = (0.2 *5) + (0.8 *35)

= 29%

$$M_a(V) = \mathbb{E}[w_a(V)m(R,Y)]$$

Density (V)

Fatty	Dense
Young: 0.20 Avg: 0.80 w(fatty) = 0.2/0.8 = 0.25	Young: 0.80 Avg: 0.20 w(dense) = 0.8/0.2 = 4.0

Down weight easy one

up weight hard one

Model error Stratified by density

= 29%

$$T_a(V) = \mathbb{E}[m(R,Y) \mid A = a] - \underbrace{\mathbb{E}[w_a(V)m(R,Y)]}_{M_a(V)}$$

If $T_a \approx 0$ => V explains the performance gap

$$\implies (R,Y) \perp A \mid V$$

So, out of many hypotheses proposed by Ladder,
we reject them by choosing appropriate V's and doing this conditional
Independence test.

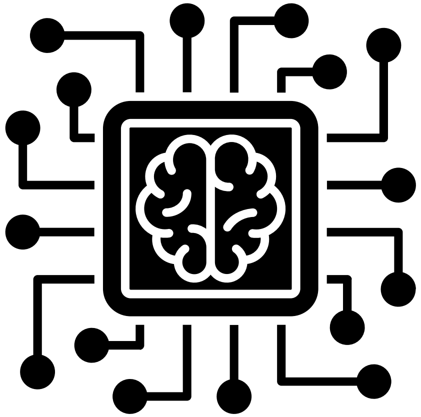
$$T_a(V) = \mathbb{E}[m(R,Y) \mid A = a] - \underbrace{\mathbb{E}[w_a(V)m(R,Y)]}_{M_a(V)}$$

If $T_a \approx 0$ => V explains the performance gap

$$\implies (R,Y) \perp A \mid V$$

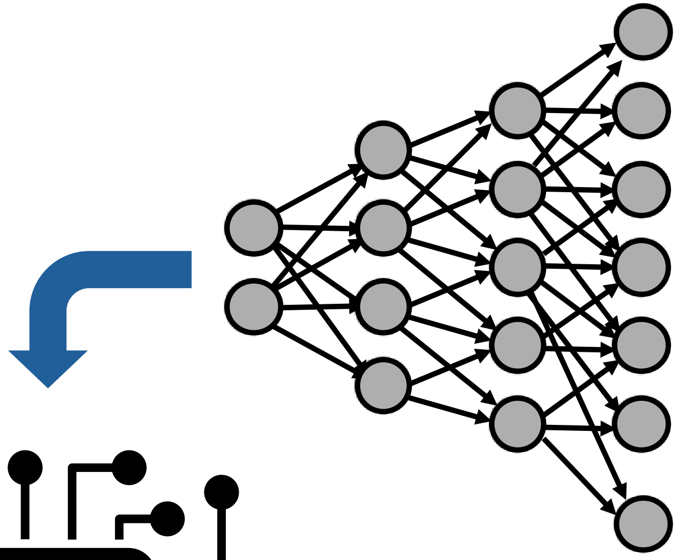
How to find the V ?

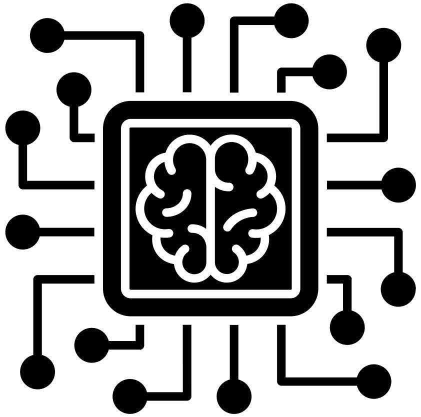
Proposal: Where we are going?



An AI Agent

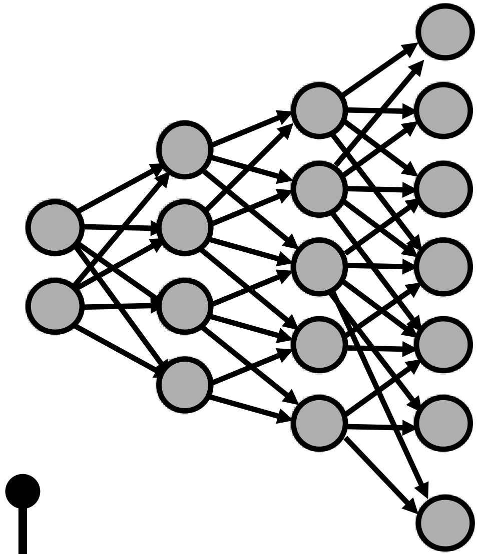
Proposal: Where we are going?

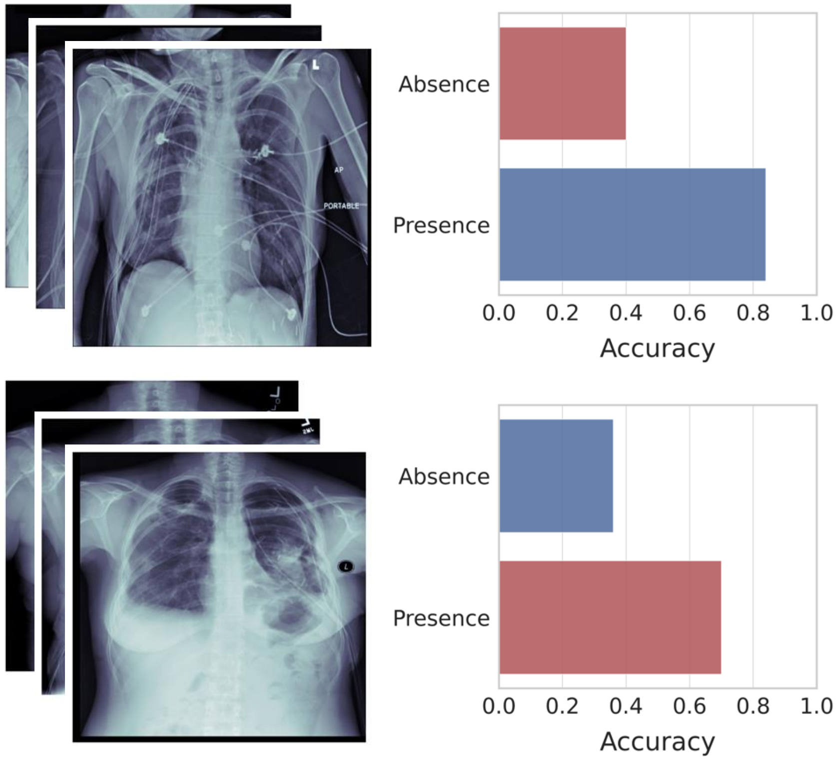




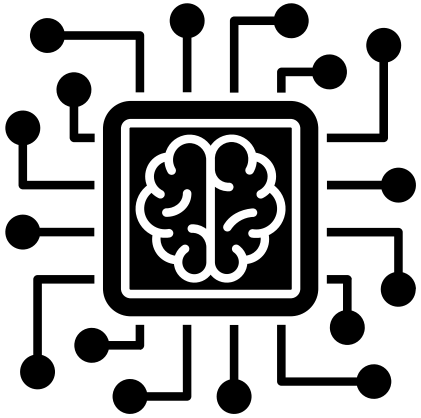
An AI Agent

Proposal: Where we are going?





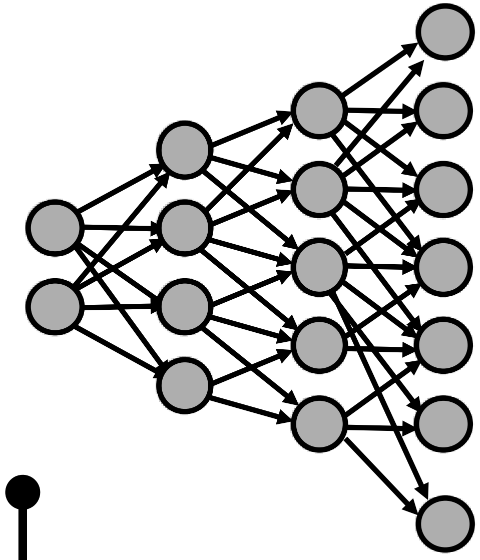
Sub groups

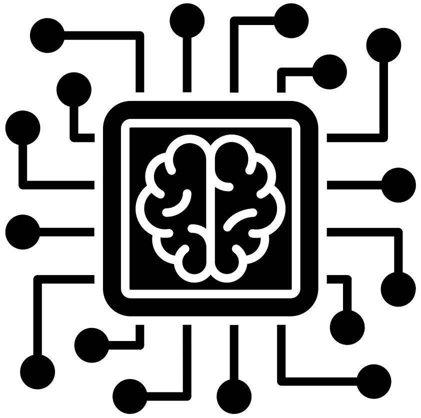


An AI Agent

Finds subgroups using Ladder or from metadata attributes

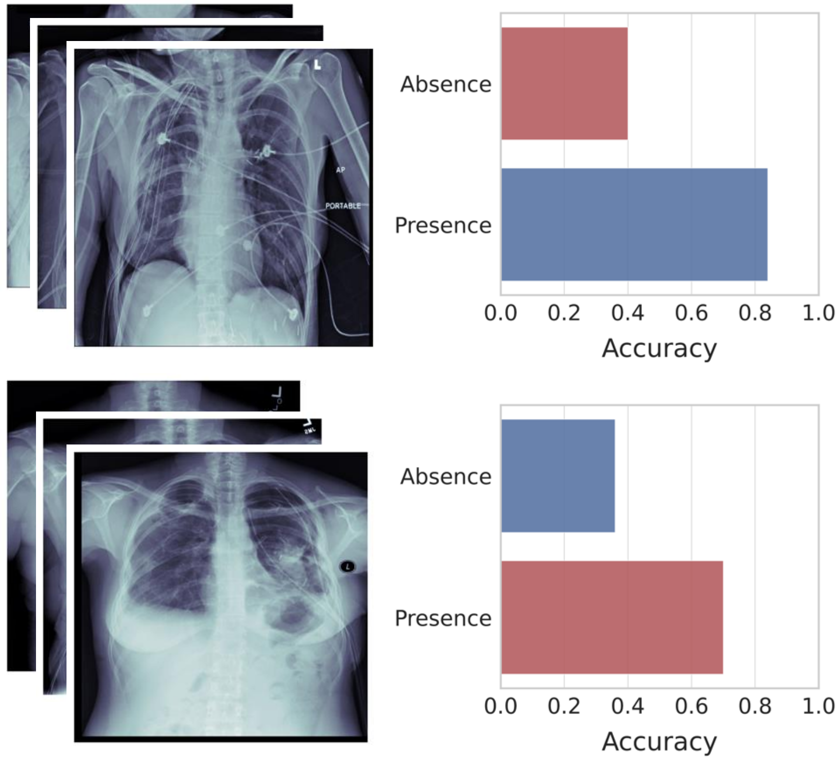
Proposal: Where we are going?



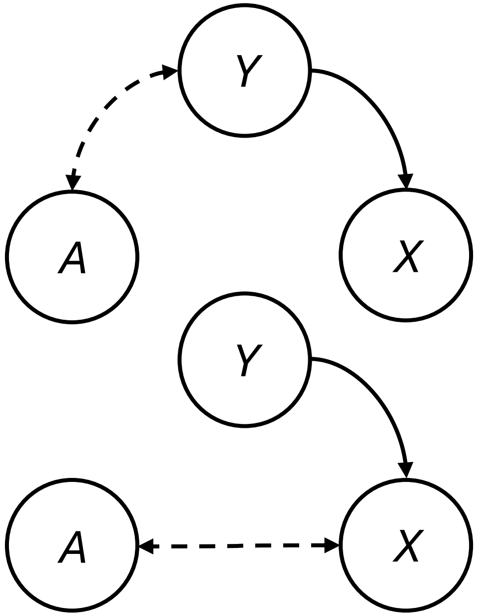


An AI Agent

Finds subgroups using Ladder or from metadata attributes



Sub groups



Data shift

Explains the performance gap

Timeline

2021
PhD Started at Pitt

2023
- Aim1 (ICML+MICCAI 2023)
- Transferred to BU

2024
- Aim3 (Ladder. ACL 2024)
- Amazon internship

2025
- Aim2 (Mammo-CLIP. MICCAI 2024)
- Amazon internship

2026

Jan-Mar
- Proposal
- Aim2 (Mammo-FM. Nat BME rebuttal)

Aug
Aim3 (Agent Ladder)

Dec
Dissertation

Papers

- Anatomy-Guided Weakly-Supervised Abnormality Localization in Chest X-rays. Ke Yu, **Shantanu Ghosh**, Zhexiong Liu, Christopher Deible, Kayhan Batmangehlich. **MICCAI 2022**
- Anatomy-specific Progression Classification in Chest Radiographs via Weakly Supervised Learning. Ke Yu, **Shantanu Ghosh**, Zhexiong Liu, Christopher Deible, Clare B. Poynton, Kayhan Batmangehlich. **RAD: AI**
- Dividing and Conquering a BlackBox to a Mixture of Interpretable Models: Route, Interpret, Repeat. **Shantanu Ghosh**, Ke Yu, Forough Arabshahi, Kayhan Batmanghelich. **ICML 2023**
- Tackling Shortcut Learning in Deep Neural Networks: An Iterative Approach with Interpretable Models. **Shantanu Ghosh**, Ke Yu, Forough Arabshahi, Kayhan Batmanghelich. **SCIS (w)@ICML 2023**
- Distilling BlackBox to Interpretable models for Efficient Transfer Learning. **Shantanu Ghosh**, Ke Yu, Kayhan Batmanghelich. **MICCAI 2023 (Top 14%)**
- Bridging the Gap: From Post Hoc Explanations to Inherently Interpretable Models for Medical Imaging. **Shantanu Ghosh**, Ke Yu, Forough Arabshahi, Kayhan Batmanghelich. **IMLH (w)@ICML 2023**
- Mammo-CLIP: A Vision Language Foundation Model to Enhance Data Efficiency and Robustness in Mammography. **Shantanu Ghosh**, Clare B. Poynton, Shyam Visweswaran, Kayhan Batmanghelich. **MICCAI 2024 (Top 11%)**
- Mammo-FM: Breast-specific foundational model for Integrated Mammographic Diagnosis, Prognosis, and Reporting. **Shantanu Ghosh**, Vedant Parthesh Joshi, Rayan Syed, Aya Kassem, Abhishek Varshney, Payel Basak, Weicheng Dai, Judy Wawira Gichoya, Hari M. Trivedi, Imon Banerjee, Shyam Visweswaran, Clare B. Poynton, Kayhan Batmanghelich. **ArXiv 2025**
- Distributionally robust self-supervised learning for tabular data. **Shantanu Ghosh**, Tiankang Xie, Mikhail Kuznetsov. **TRL (w)@NeurIPS 2024**
- LADDER: Language-Driven Slice Discovery and Error Rectification in Vision Classifiers. **Shantanu Ghosh**, Rayan Syed, Chenyu Wang, Vaibhav Choudhary, Binxu Li, Clare B. Poynton, Shyam Visweswaran, Kayhan Batmanghelich. **ACL 2025**
- Semantic Consistency-Based Uncertainty Quantification for Factuality in Radiology Report Generation. Chenyu Wang, Weichao Zhou, **Shantanu Ghosh**, Kayhan Batmanghelich, Wenchao Li. **NAACL 2025**
- PhyDiCT: Training-Free 3D CT Reconstruction from Sparse X-Rays via Differentiable Rendering and Strong Priors. Weicheng Dai, **Shantanu Ghosh**, Kayhan Batmanghelich. **CVPR 2025 submission**

Acknowledgments

University of Pittsburgh

Boston University

Boston University School of Medicine

Advisor

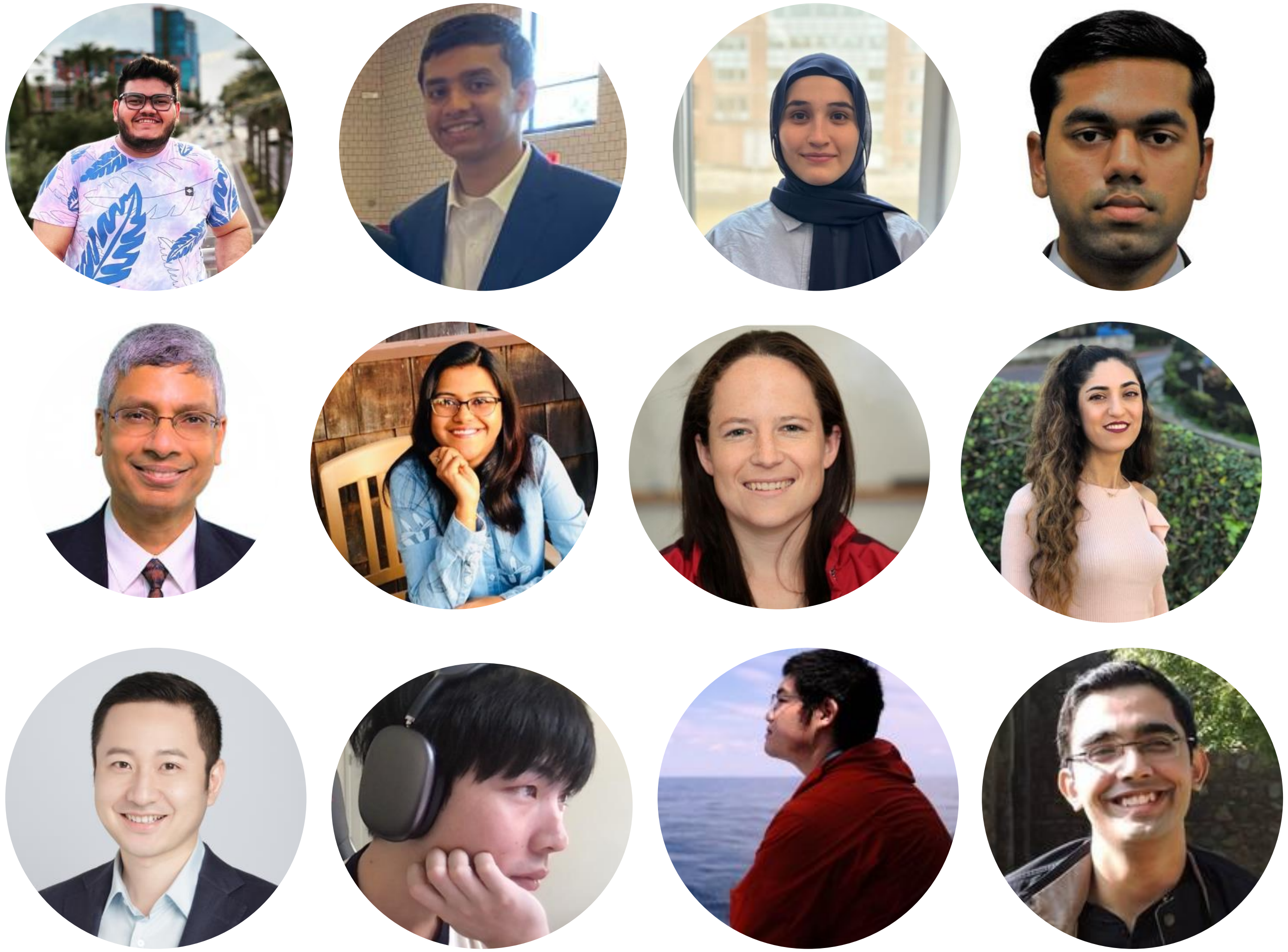
Arizona State University

Emory University School of Medicine

Department of Radiology and Imaging Sciences

Mayo Clinic

Collaborators



Thank you!

Questions?