Foundation Models for Chest Radiography:

Knowledge Distillation and Impacts on Performance and Bias Propagation

Fadi Zahar (MSc AI)
Supervisor – Dr Ben Glocker
BioMedIA,
Imperial College London

February 2025

IMPERIAL

Plan

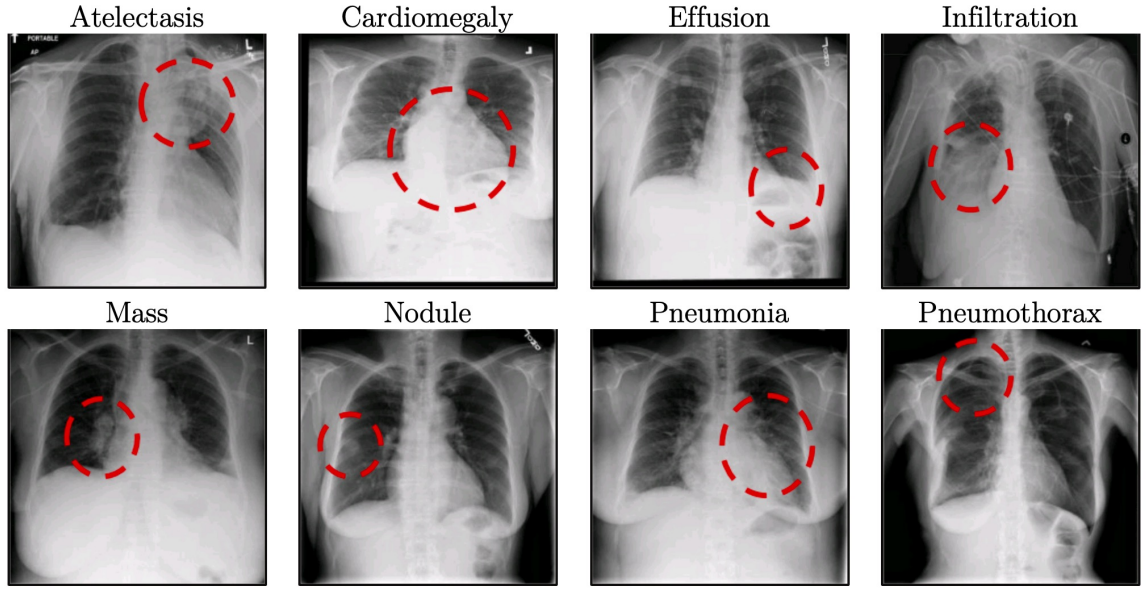
1. Introduction & Background
2. Methodology
3. Results
4. Conclusion

1.1. Deep Learning in Medical Imaging (1)

Medical Imaging → crucial for *prevention*, *diagnosis*, *prognosis*, and *treatment* of various conditions (e.g., cancer)

Deep Learning (DL) models → have shown significant success:

- Accelerate medical image analysis (automation)
- Match or exceed human expert performance



Common thoracic diseases observed in chest X-rays [1]

e.g.

Lung cancer → leading cause of cancer death globally [2]

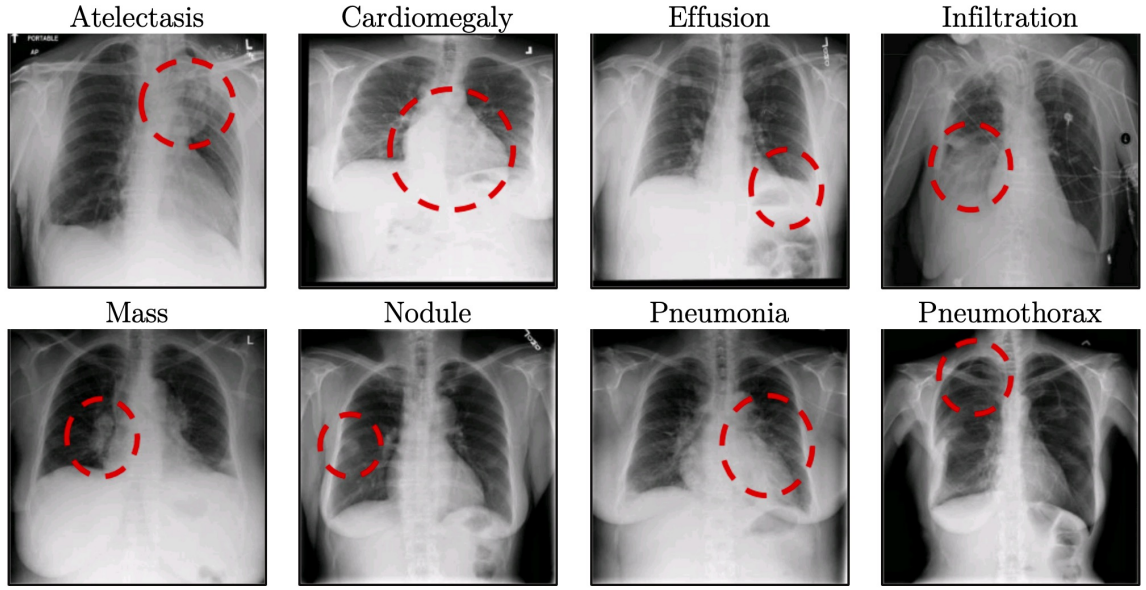
Can be identified at an earlier, more treatable stage through low-dose **CT** screening or chest X-rays (**CXRs**)

1.1. Deep Learning in Medical Imaging (2)

Medical Imaging → crucial for *prevention*, *diagnosis*, *prognosis*, and *treatment* of various conditions (e.g., cancer)

Deep Learning (DL) models → have shown significant success:

- Accelerate medical image analysis (automation)
- Match or exceed human expert performance



Common thoracic diseases observed in chest X-rays [1]

However, DL models need very large amounts of representative data for good generalisation.

Otherwise, they risk to **fail** due to data distribution shifts in different clinical settings:

- Population shifts *(demographics)*
- Acquisition shifts *(imaging modalities)*

Quality medical imaging data → **scarce** due to *expensive labelling* & *data privacy concerns*

1.2. Foundation Models

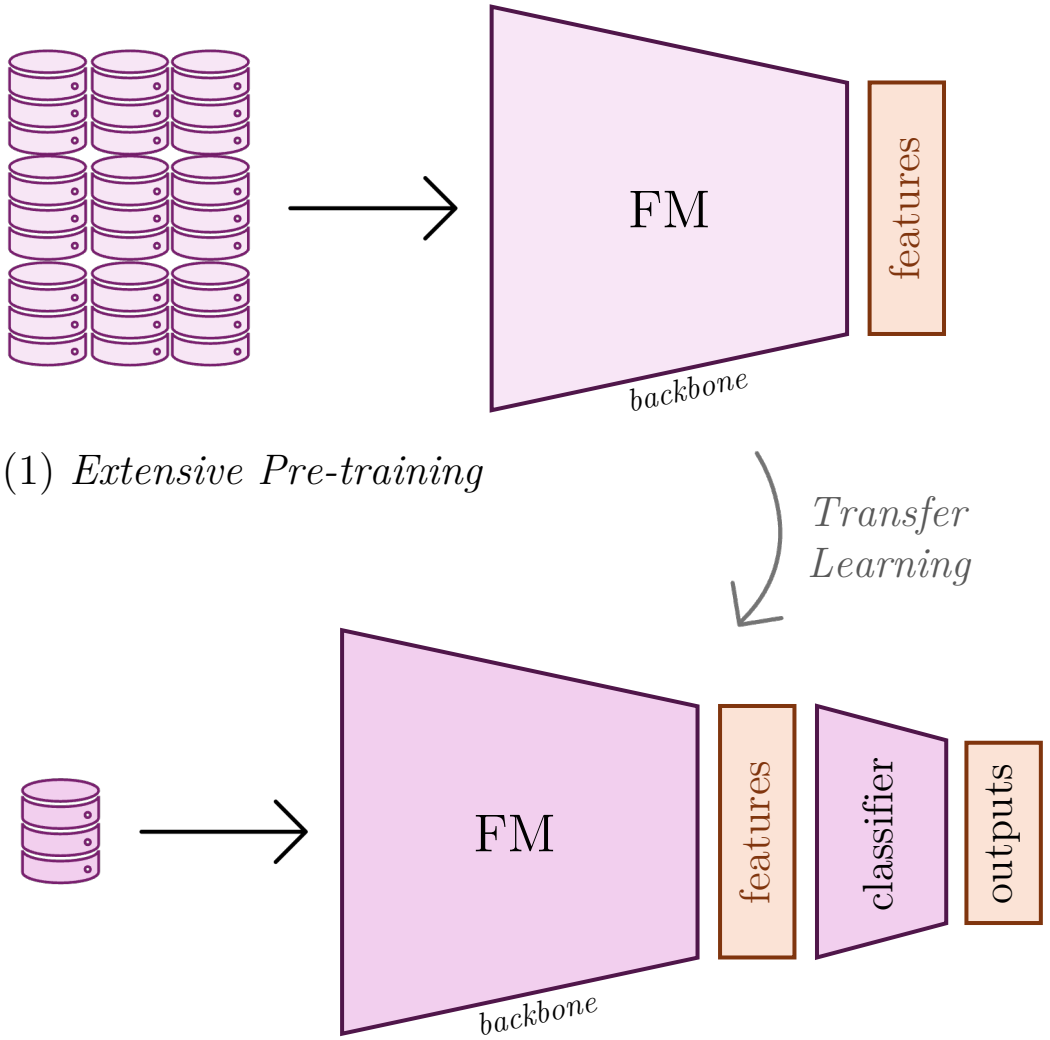
Solution: **Foundation Models (FM)** [3]

(1) Large models pretrained on **heterogeneous**, **diverse**, **large-scale datasets**; often via *Semi-Supervised* or *Self-Supervised Learning (SSL)*

(2) Learn *rich feature representations* (features) that can be used in **downstream task adaptation** with **much less training data**

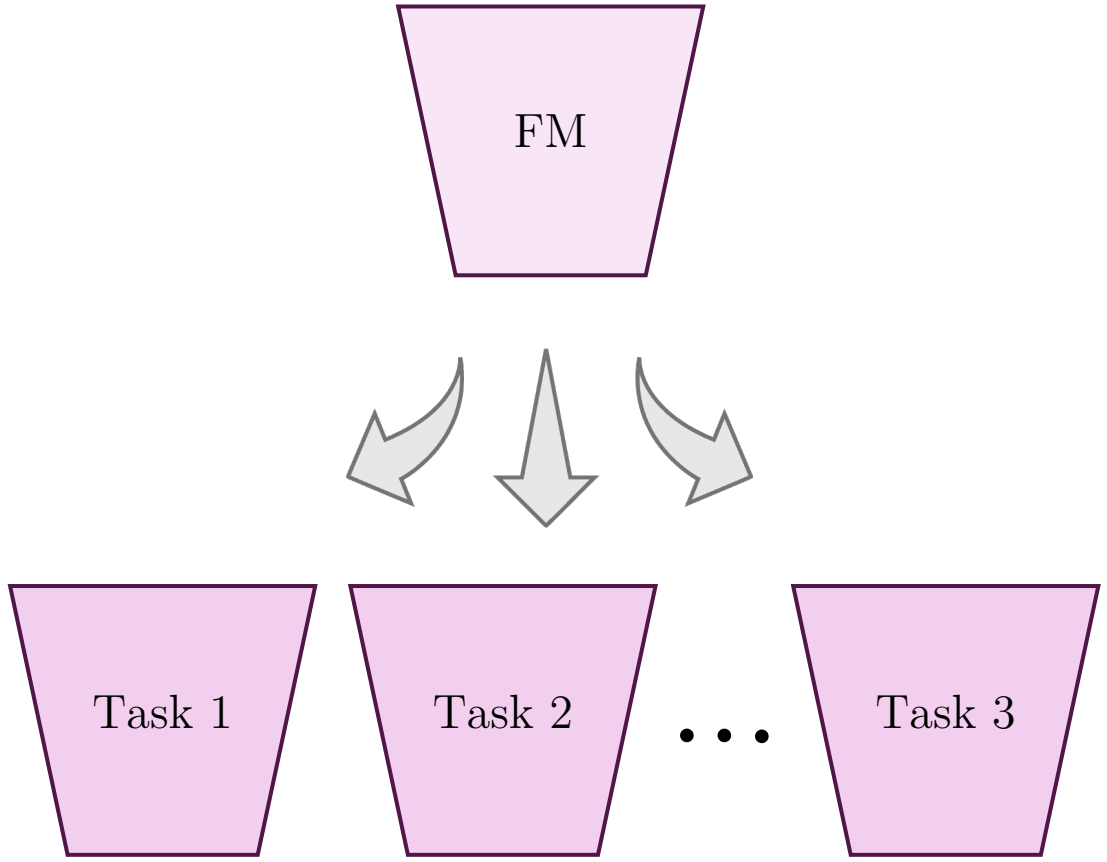
e.g., Google's chest radiography FM: **CXR-FM** [4]

developed from 821,544 CXRs sourced from India and the United States



1.3. Associated Challenges

- General lack of profound understanding of Foundation Models (FMs)' mechanisms and impacts



Downstream Task Adaptation

1.3. Associated Challenges

- General lack of profound understanding of Foundation Models (FMs)' mechanisms and impacts

- Can propagate encoded biases trough downstream tasks: any *defect* is likely '*inherited*' by all adapted models


Radiology: Artificial Intelligence

ORIGINAL RESEARCH

Risk of Bias in Chest Radiography Deep Learning Foundation Models

Ben Glocker, PhD • Charles Jones, MEng • Mélanie Roschewitz, MSc • Stefan Winzeck, PhD

Deep learning–based predictive models have found great success in medical imaging applications, such as disease detection in chest radiography (1). However, training of these models requires access to large amounts of representative data. Generalization across different clinical sites remains a major challenge for wider clinical adoption (2). Training on limited data makes models susceptible to failure whenever the data characteristics change, often caused by differences in the patient demographic characteristic (ie, population shift) and/or imaging technique (ie, acquisition shift) (3,4).

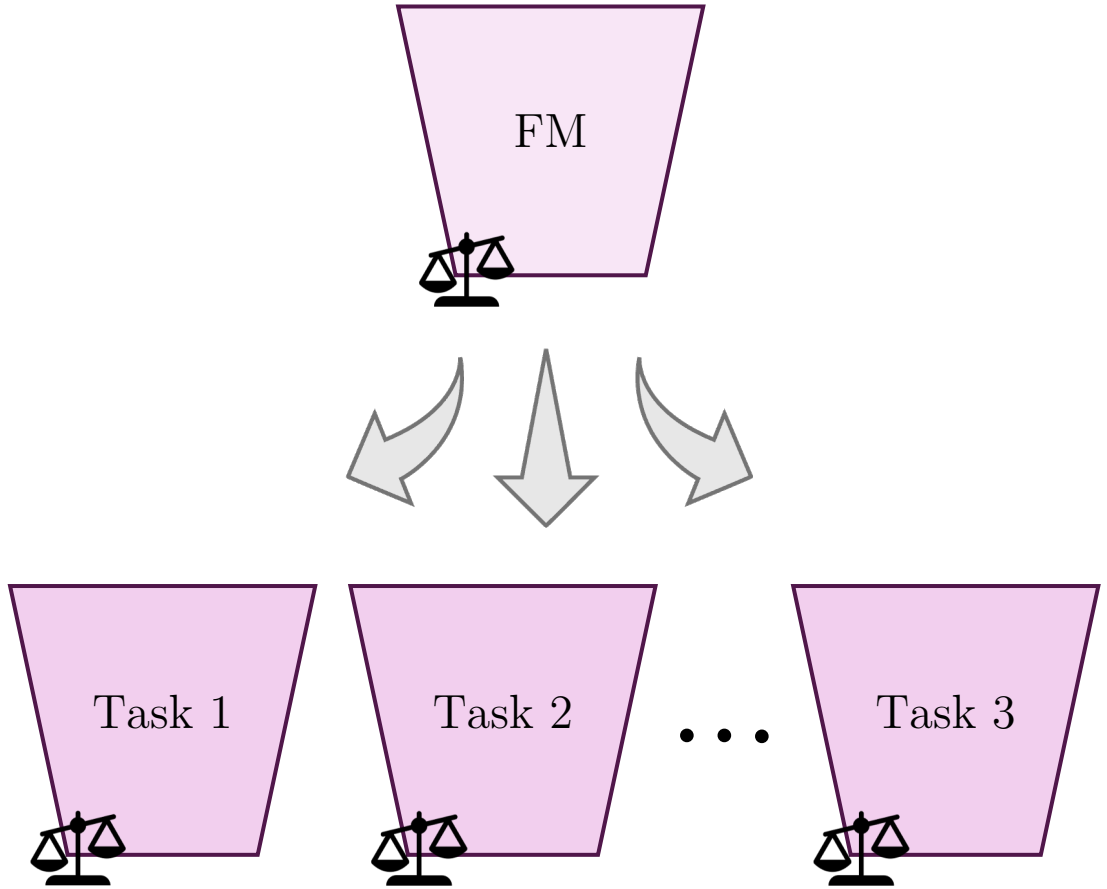
Foundation models have emerged as a promising solution to mitigate these issues (5,6). These models are pretrained on large-scale, heterogeneous, and diverse datasets, often by self-supervised or semisupervised learning strategies that do not require ground truth annotations, with the hope to provide robust backbones for task-agnostic feature extraction. These backbone features then serve as inputs for the subsequent, data-efficient training of task-specific prediction models. The term *foundation model* is now widely used to describe pretrained, versatile deep learning models that can be adapted to a wide range of downstream prediction tasks (7).

In medical imaging, pretraining is particularly attractive because of the difficulty of collecting large amounts of high-quality training data. Recent work includes self-supervised pretraining on large unlabeled medical imaging datasets, which appears to improve performance on data not only from similar sources (in-distribution) but also across various downstream tasks on new, out-of-distribution data (8). These findings were corroborated by Ghesu et al (9), who proposed a foundation model trained on more than 1 million diverse medical images. Similarly, Sellergren et al (10) recently developed a chest radiography foundation model, demonstrating that it can improve performance in downstream tasks as well as drastically reduce the amount of labeled training data required for task-specific fine-tuning. The pretrained model yielded an area under the receiver operating characteristic curve (AUC) of 0.95 for detecting tuberculosis when using only 45 chest radiographs for task-specific training, which was noninferior to radiologic performance. Outcome prediction after COVID-19 was better with

This copy is for personal use only. To order printed copies, contact reprints@rsna.org

Indeed, Glocker et al. [5] showed that Google's **CXR-FM** exhibited *biases* related to race and biological sex and corresponding *subgroup performance disparities*

→ making it unsafe for clinical deployment



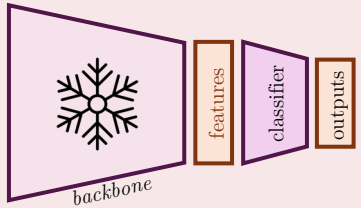
Downstream Task Adaptation

1.4. Need for Knowledge Distillation

For effective bias mitigation → **full access to the FM is needed**

However, **CXR-FM**, like many FMs, is only released through 'API calls':

its parameters are NOT publicly available
→ backbone is frozen

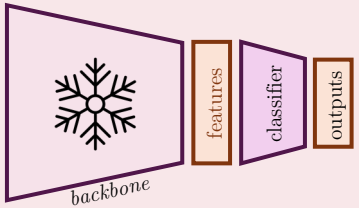


1.4. Need for Knowledge Distillation

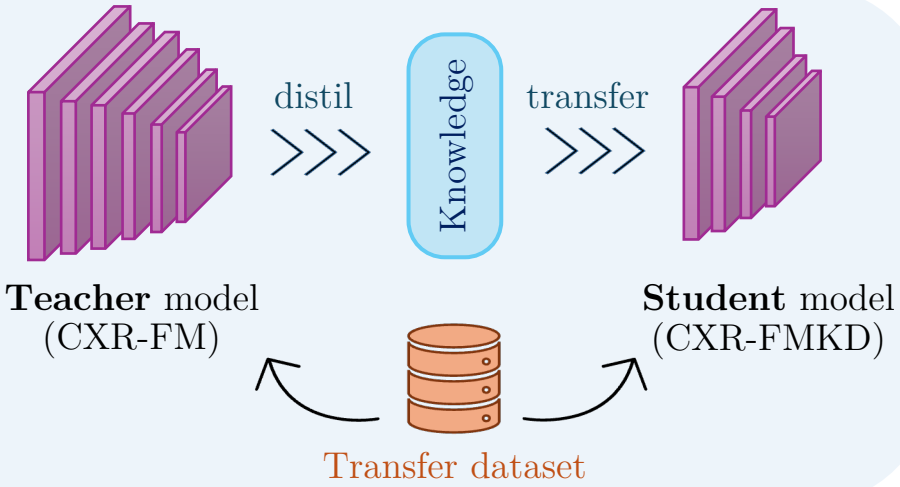
For effective bias mitigation → **full access to the FM is needed**

However, **CXR-FM**, like many FMs, is only released through 'API calls':

its parameters are NOT publicly available
→ backbone is frozen



Knowledge Distillation [6]



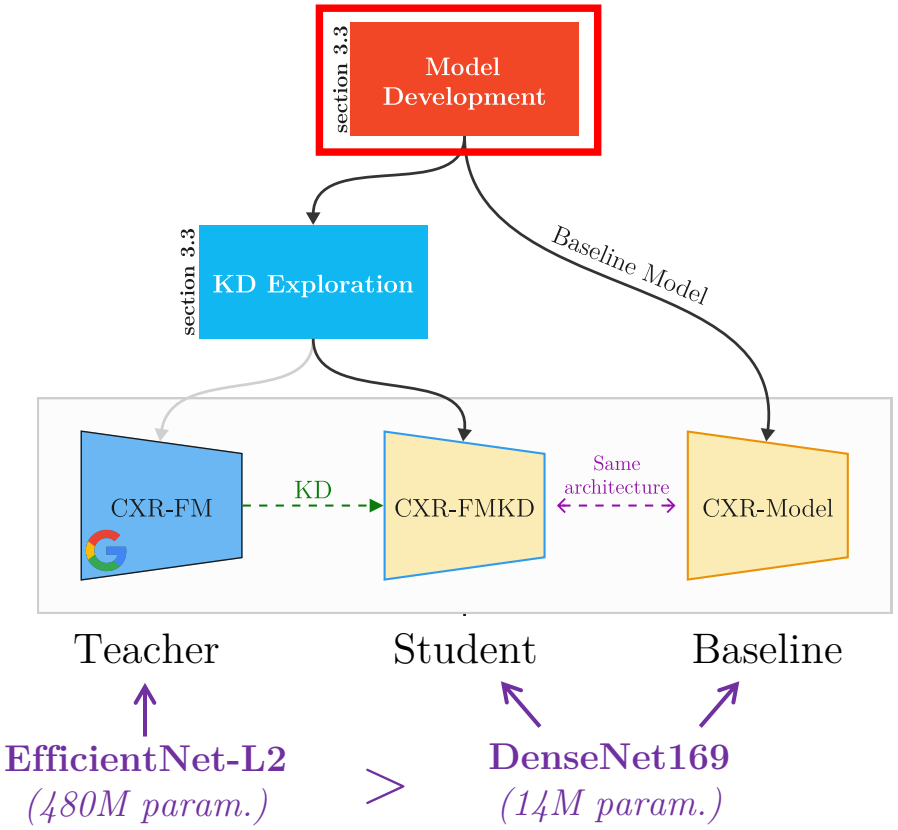
Thesis Contributions:

Assess whether **Knowledge Distillation** (KD) is a viable strategy to robustly **reconstruct FMs**, focusing on Google's **CXR-FM**

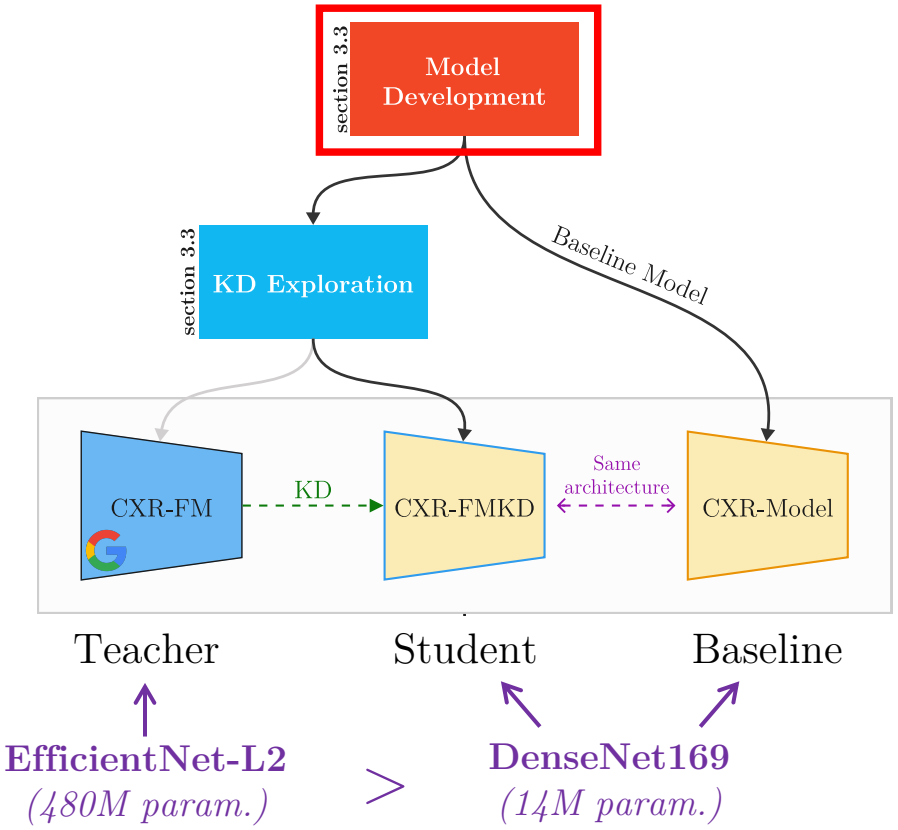
⇒ Use KD to develop robust distilled student models, **CXR-FMKD**, from CXR-FM:

- Offer enhanced *transparency*, *tunability*, and ultimately *bias mitigation capabilities* → safer for clinical adoption
- Ensure CXR-FMKD inherits CXR-FM's strengths and *matches/exceeds its performance* in downstream CXR tasks

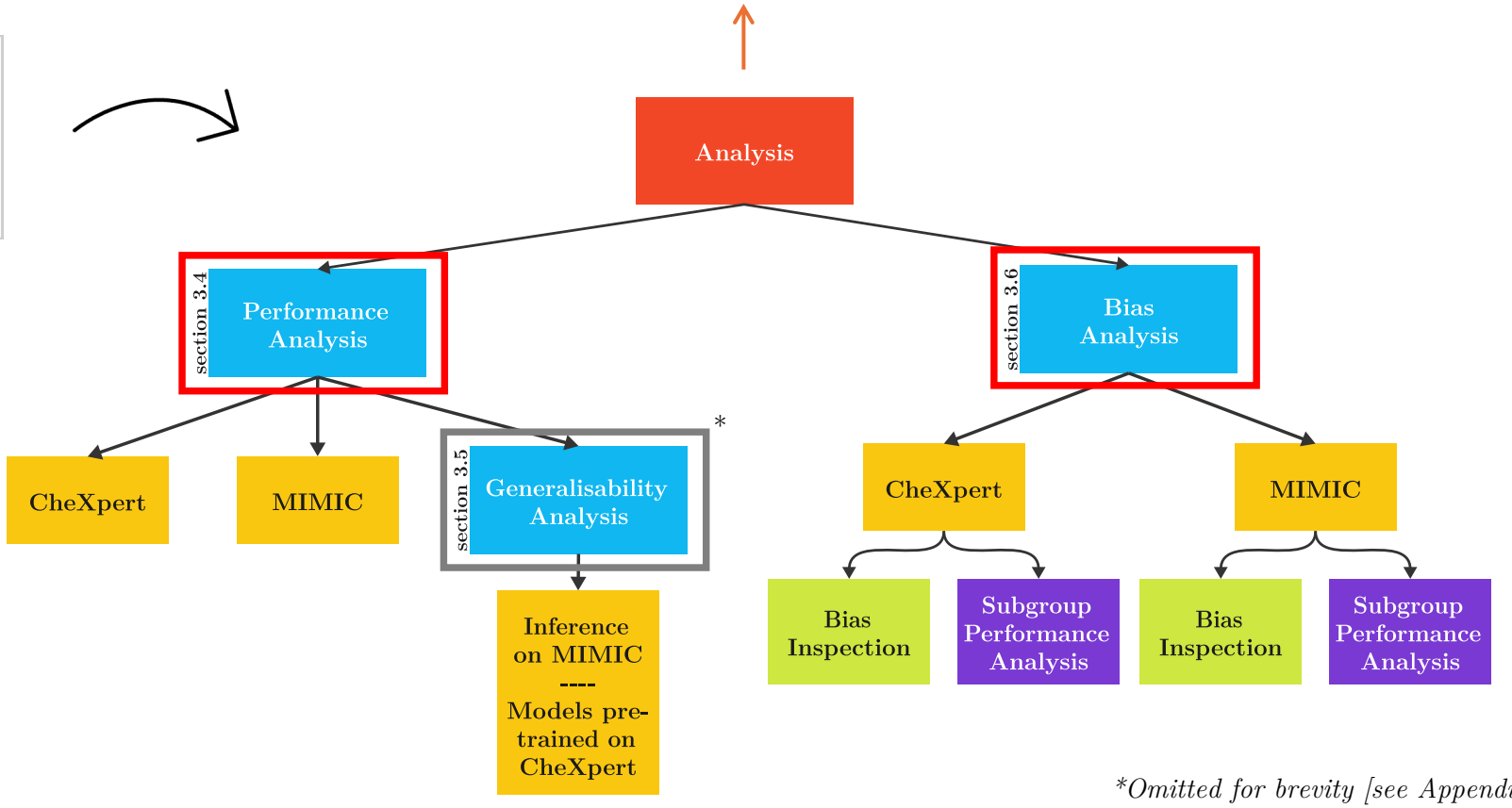
2.1. Research Design Overview



2.1. Research Design Overview



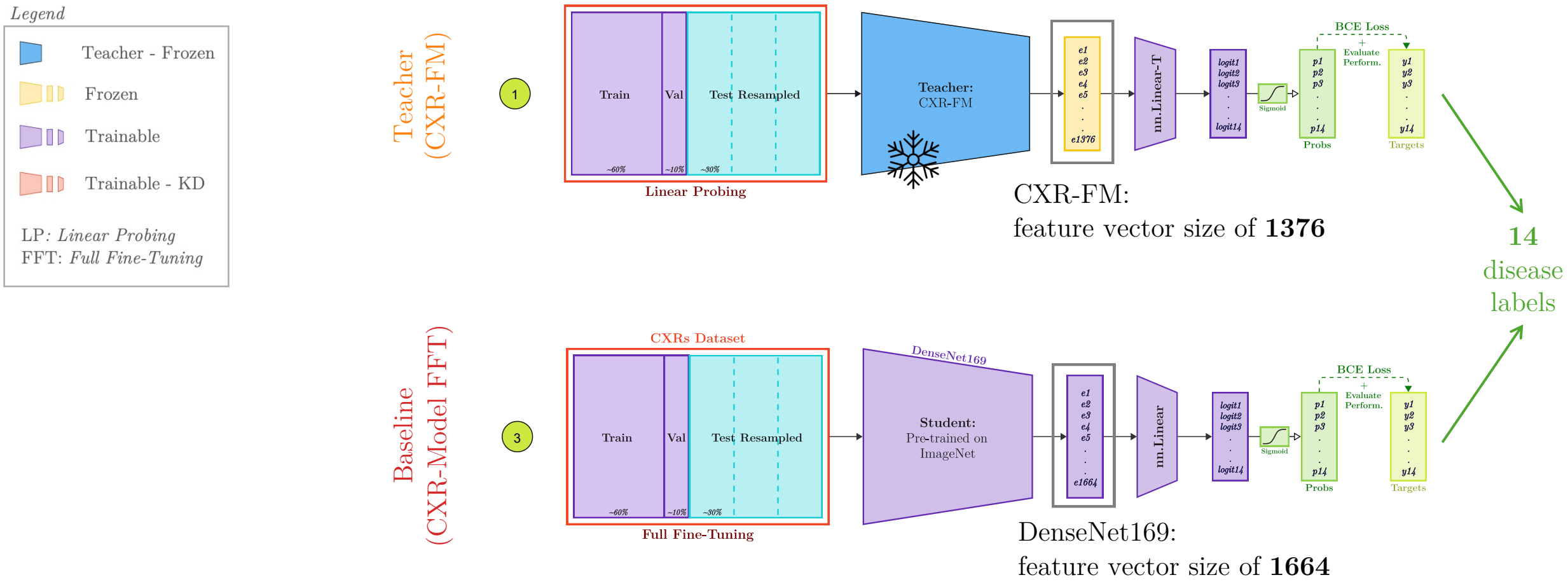
Datasets: **CheXpert** and **MIMIC** share the same **14** disease labels



Omitted for brevity [see Appendix]

2.2. Models Architecture Overview (1)

Model Development



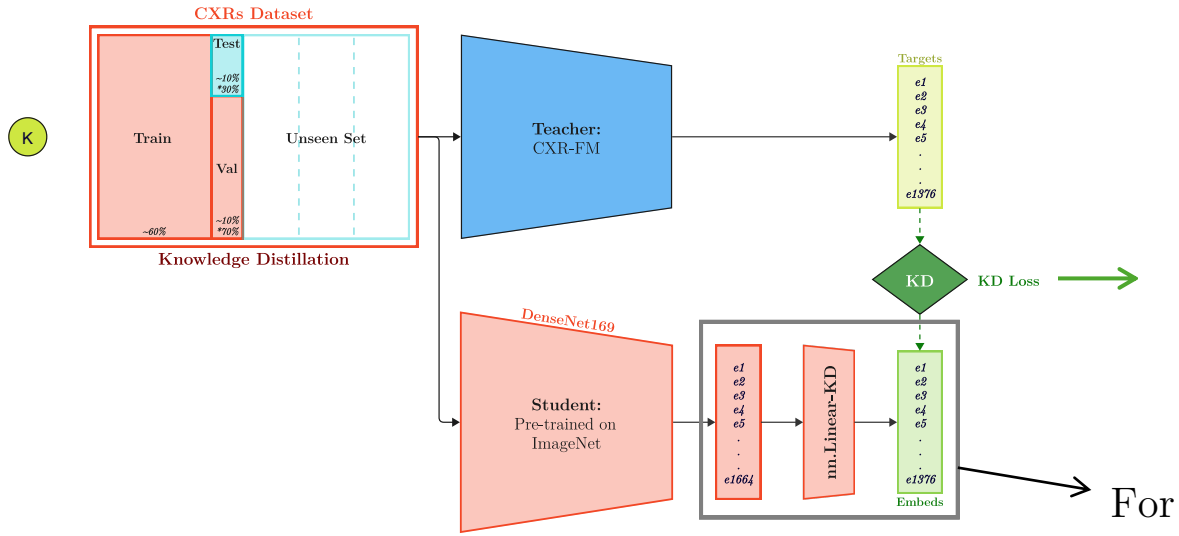
2.2. Models Architecture Overview (2)

Model Development

Legend

- Teacher - Frozen
- Frozen
- Trainable
- Trainable - KD

LP: *Linear Probing*
FFT: *Full Fine-Tuning*



KD Losses: (x11)

Inspired from KD literature [6]

KD (MSE)
KD (MAE)
KD (HuberLoss)
KD (CS)
KD (MSE-CS Naive)
KD (MSE-CS Learned)
KD (MSE-CS | 0.5-0.5)
KD (MSE-CS | 0.6-0.4)
KD (MSE-CS | 0.7-0.3)
KD (MSE-CS | 0.8-0.2)
KD (MSE-CS | 0.9-0.1)

For KD, map **1664** to **1376** to match feature sizes

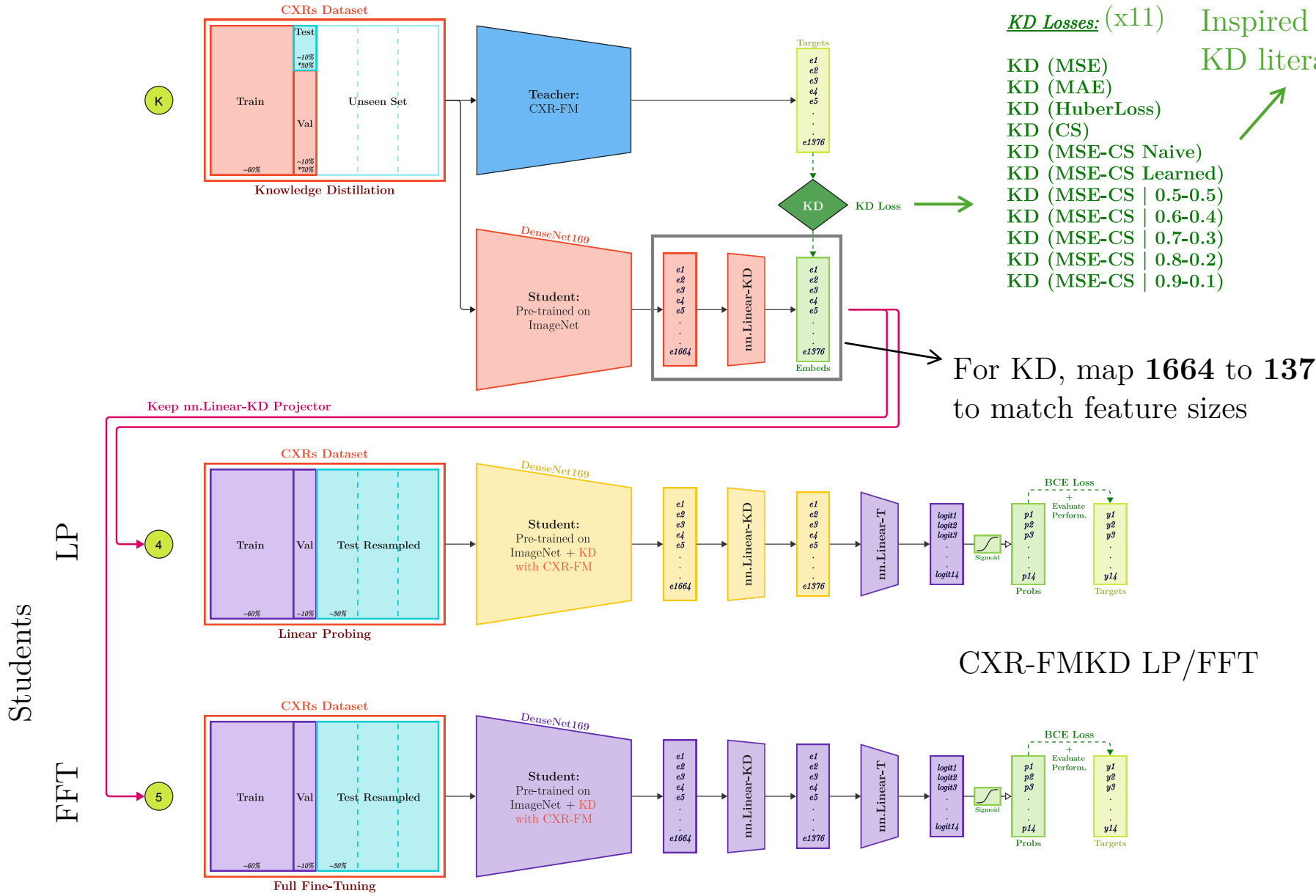
2.2. Models Architecture Overview (2)

Model Development

Legend

- Teacher - Frozen
- Frozen
- Trainable
- Trainable - KD

LP: *Linear Probing*
FFT: *Full Fine-Tuning*



KD Losses: (x11)

KD (MSE)
KD (MAE)
KD (HuberLoss)
KD (CS)
KD (MSE-CS Naive)
KD (MSE-CS Learned)
KD (MSE-CS | 0.5-0.5)
KD (MSE-CS | 0.6-0.4)
KD (MSE-CS | 0.7-0.3)
KD (MSE-CS | 0.8-0.2)
KD (MSE-CS | 0.9-0.1)

Inspired from KD literature [6]

For KD, map **1664** to **1376** to match feature sizes

CXR-FMKD LP/FFT

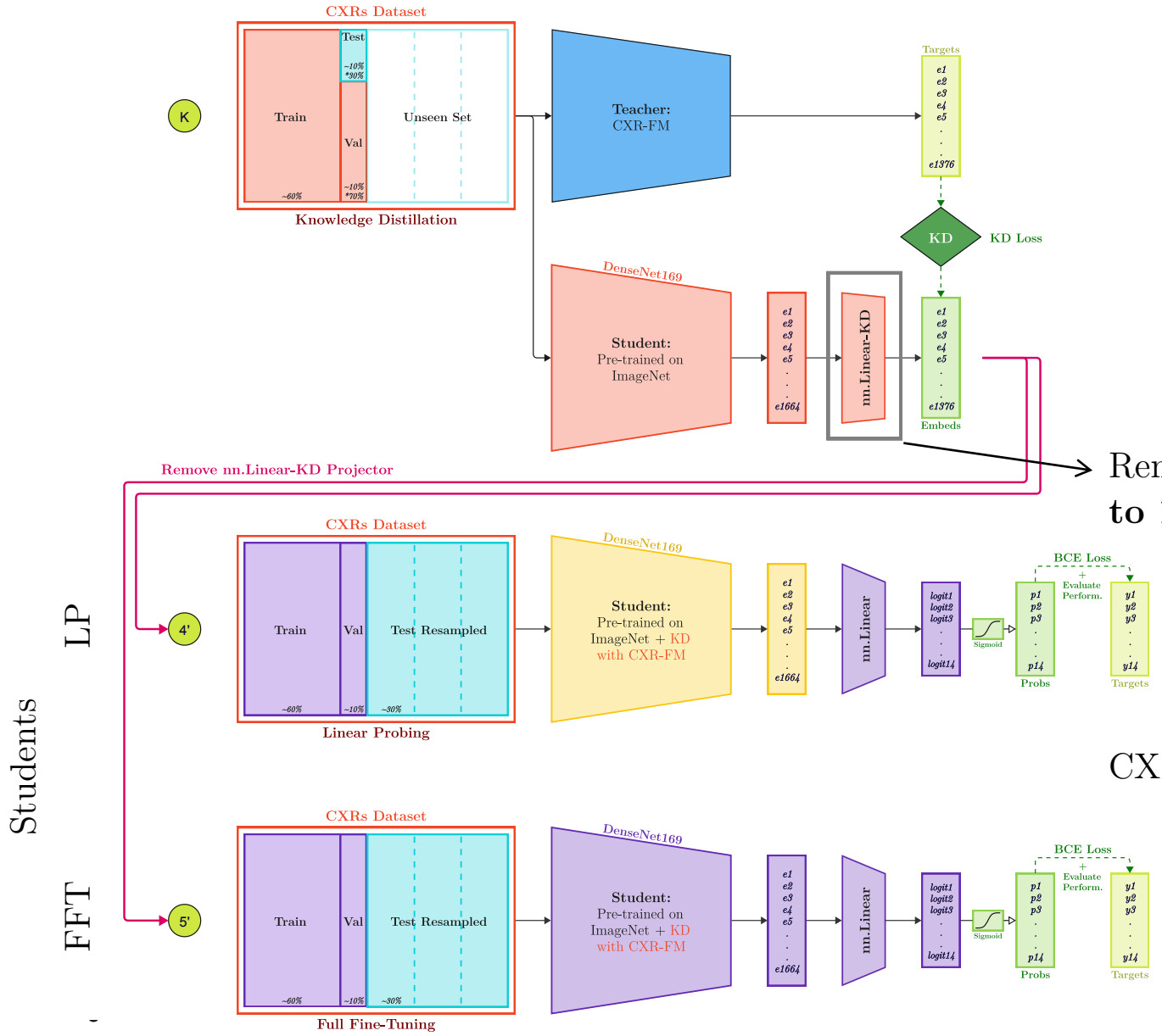
2.2. Models Architecture Overview (3)

Model Development

Legend

- Teacher - Frozen
- Frozen
- Trainable
- Trainable - KD

LP: *Linear Probing*
FFT: *Full Fine-Tuning*



KD Losses:

KD (MSE)
KD (MAE)
KD (HuberLoss)
KD (CS)
KD (MSE-CS Naive)
KD (MSE-CS Learned)
KD (MSE-CS | 0.5-0.5)
KD (MSE-CS | 0.6-0.4)
KD (MSE-CS | 0.7-0.3)
KD (MSE-CS | 0.8-0.2)
KD (MSE-CS | 0.9-0.1)

Remove projector, **go back to 1664** feature size

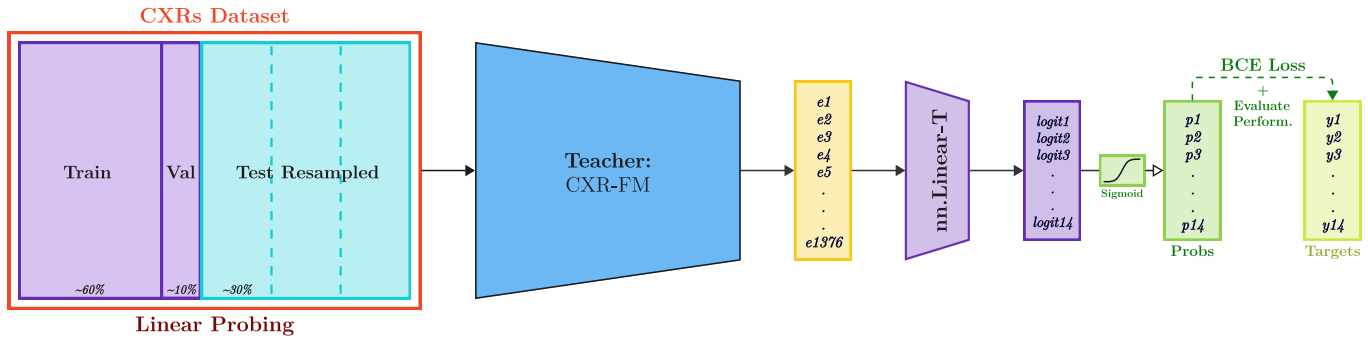
CXR-FMKD-**Direct** LP/FFT

2.3. Selected Models

Model Development

Teacher (CXR-FM)

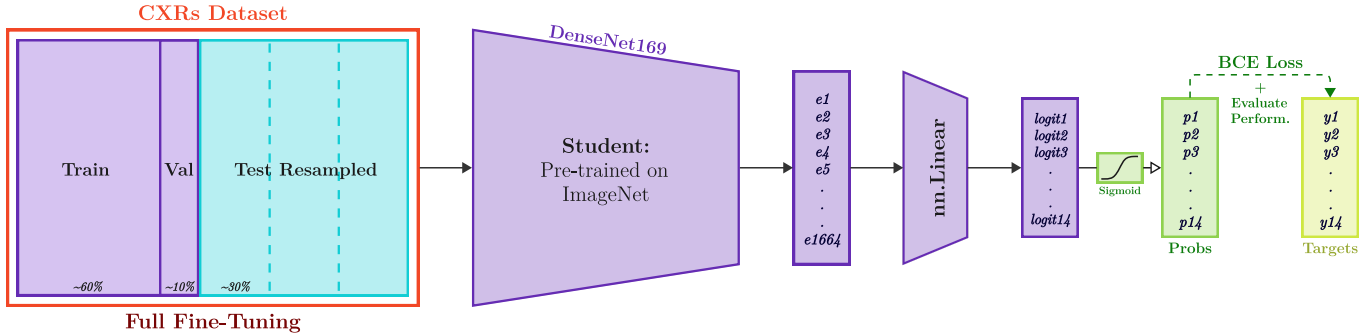
1



(x1)

Baseline (CXR-Model FFT)

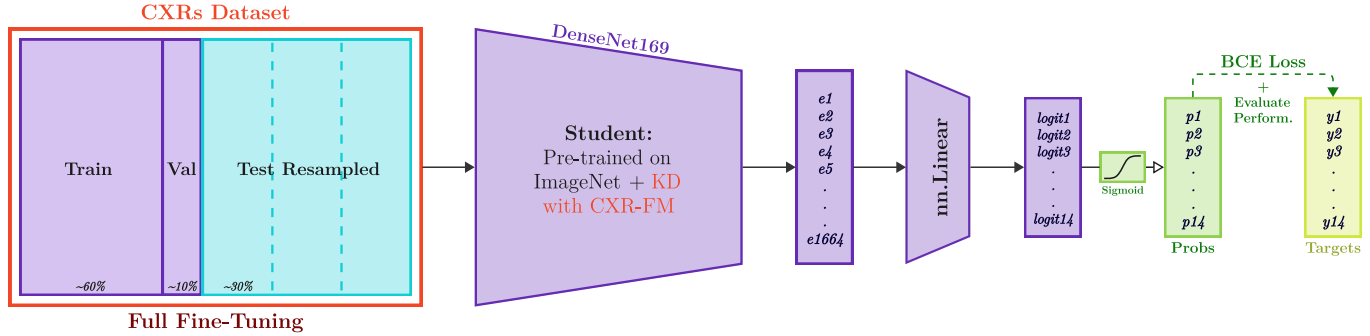
3



(x1)

Student (CXR-FMKD-**Direct** FFT)

5'

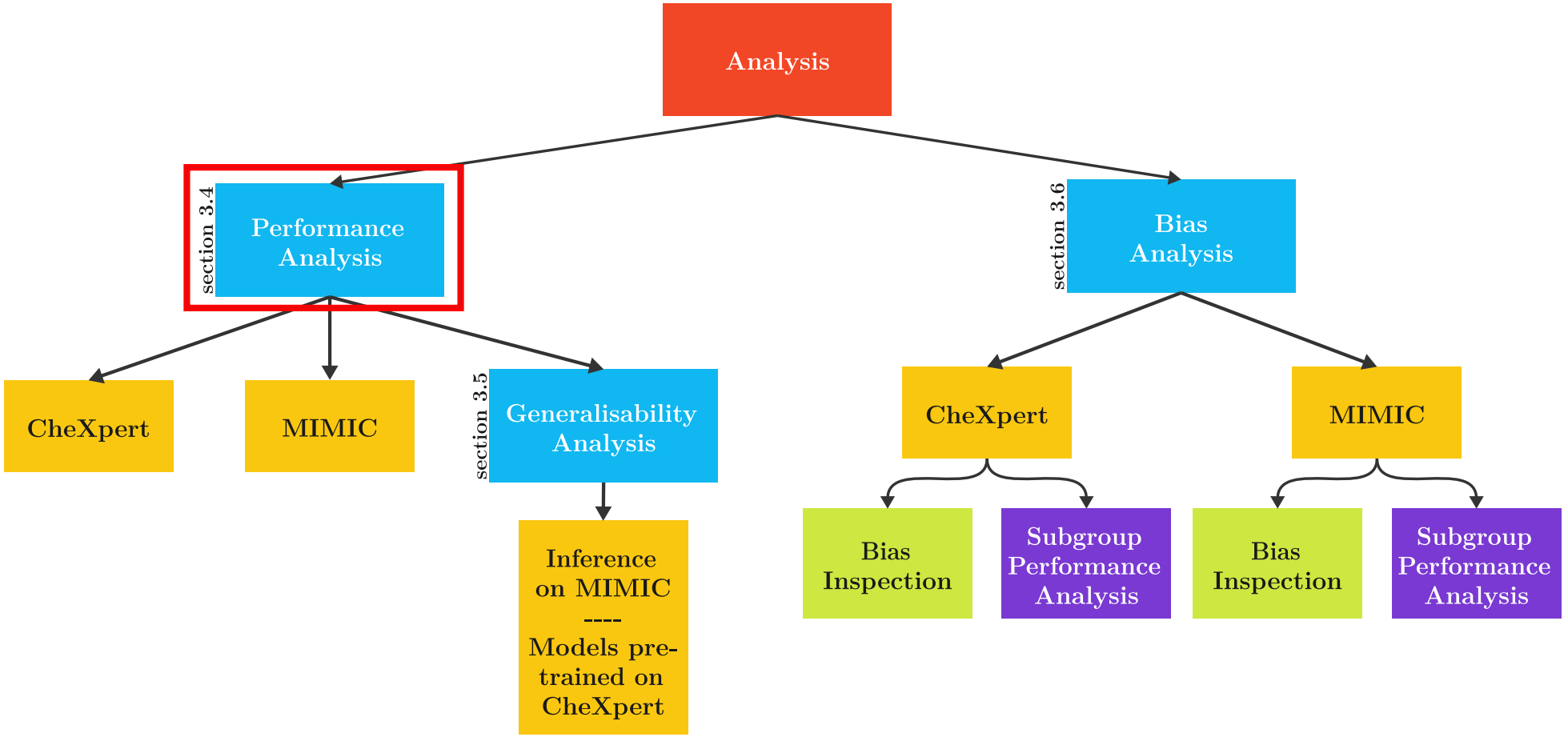


(x3):

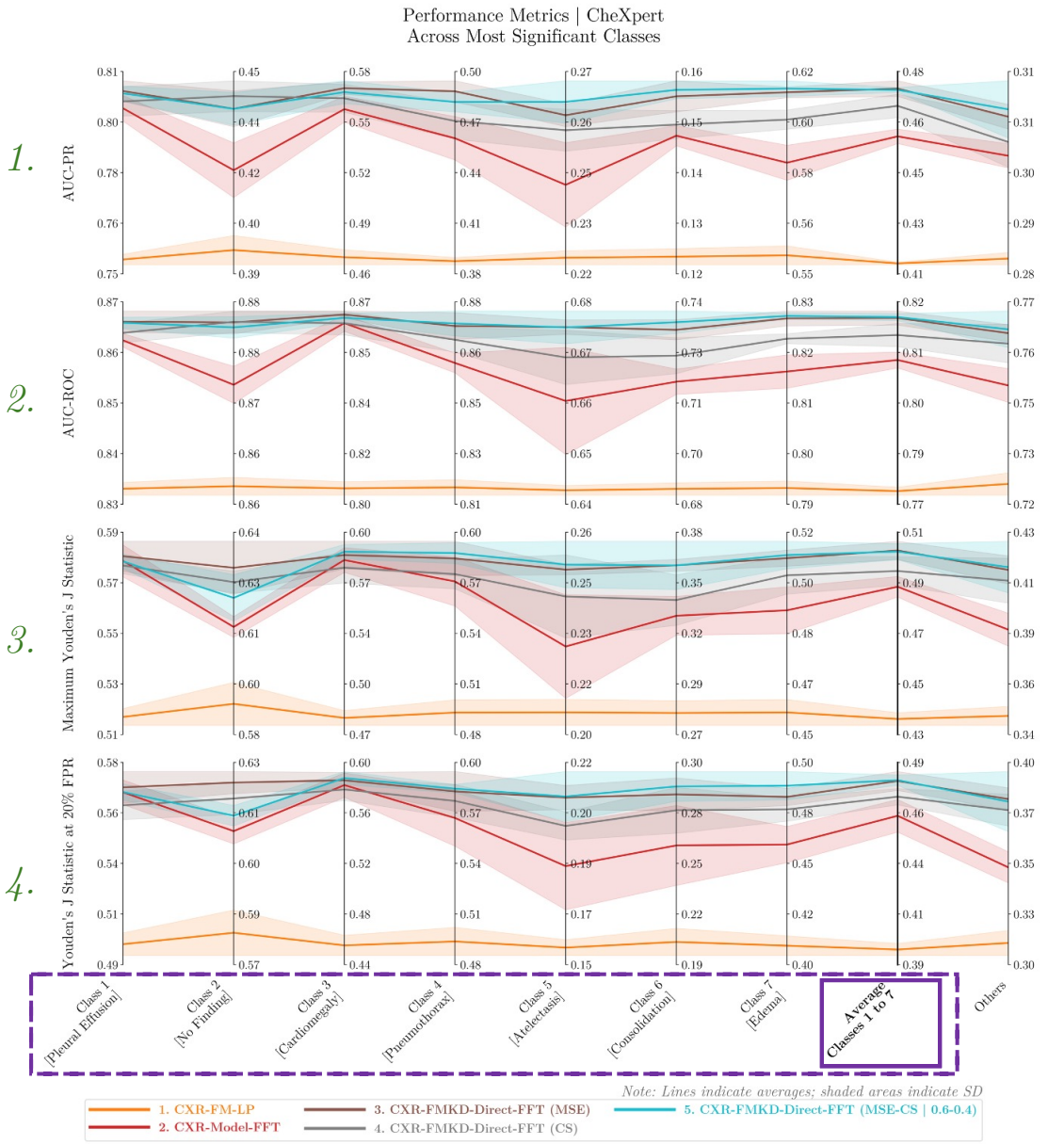
(x1) **KD-MSE**
(x1) **KD-CS**
(x1) **KD-MSE&CS**

⚠ *Removing projector → improved results → akin to SSL literature*

3.1. Performance Analysis (1)

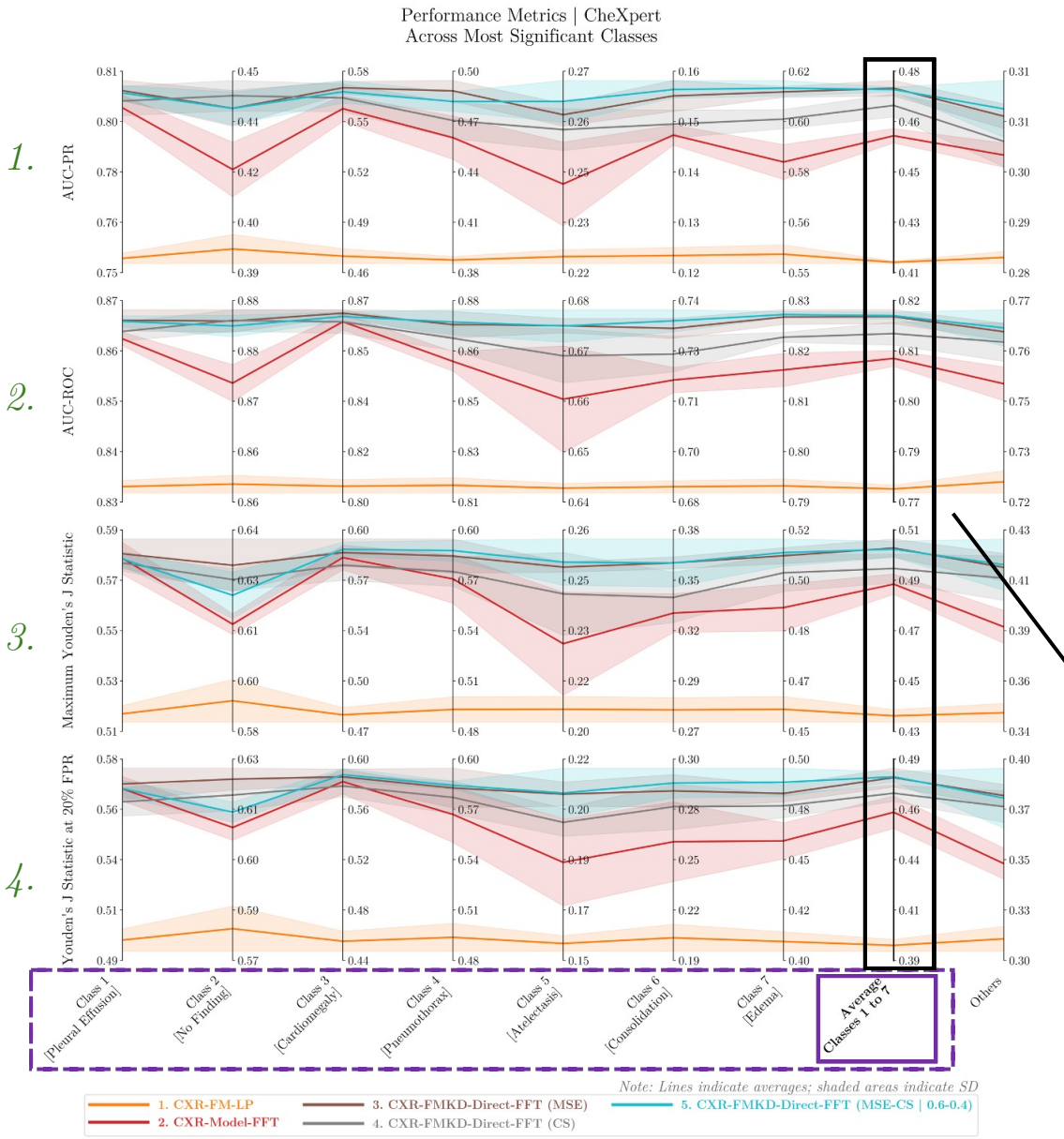


3.1. Performance Analysis (2)



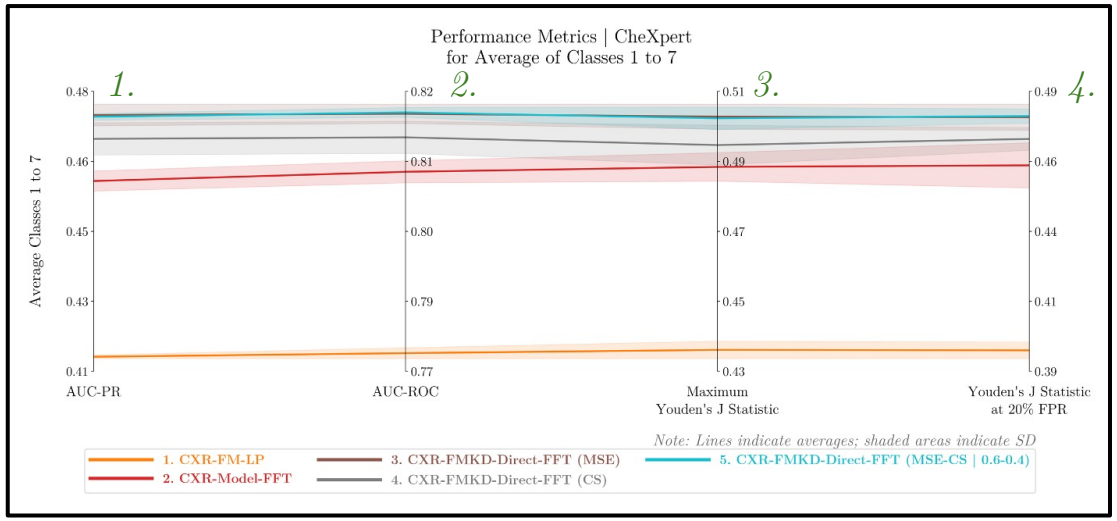
- Focus on **Average of the 7 most clinically significant classes** [5, 8]
- Use 4 metrics [5, 8] :
  1. *AUC-PR*
  2. *AUC-ROC*
  3. *Maximum Youden's J Statistic*
  4. *Youden's J Statistic at 20% FPR*

3.1. Performance Analysis (2)



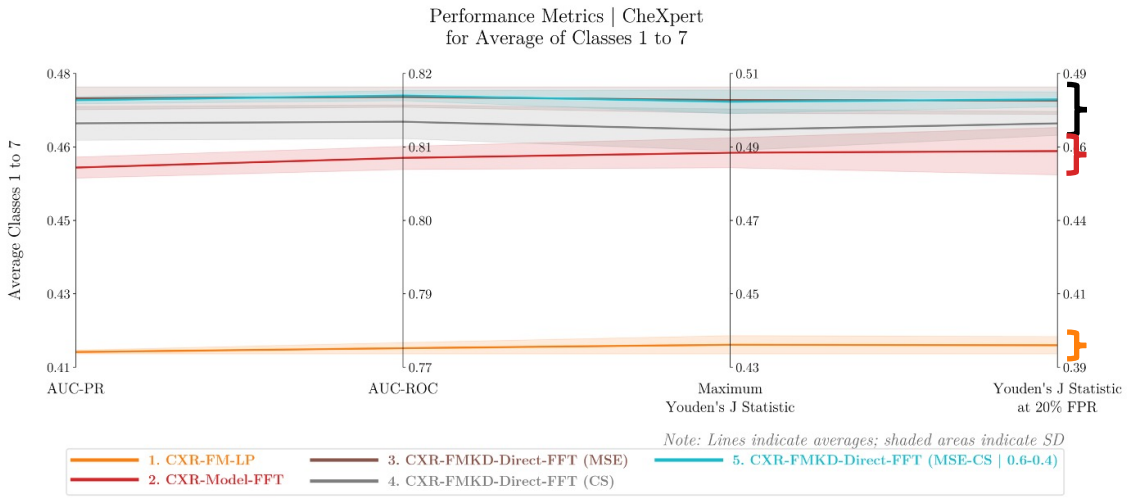
- Focus on **Average of the 7 most clinically significant classes** [5, 8]
- Use 4 metrics [5, 8] :
  1. *AUC-PR*
  2. *AUC-ROC*
  3. *Maximum Youden's J Statistic*
  4. *Youden's J Statistic at 20% FPR*

Summary Plot



3.1. Performance Analysis (3)

CheXpert

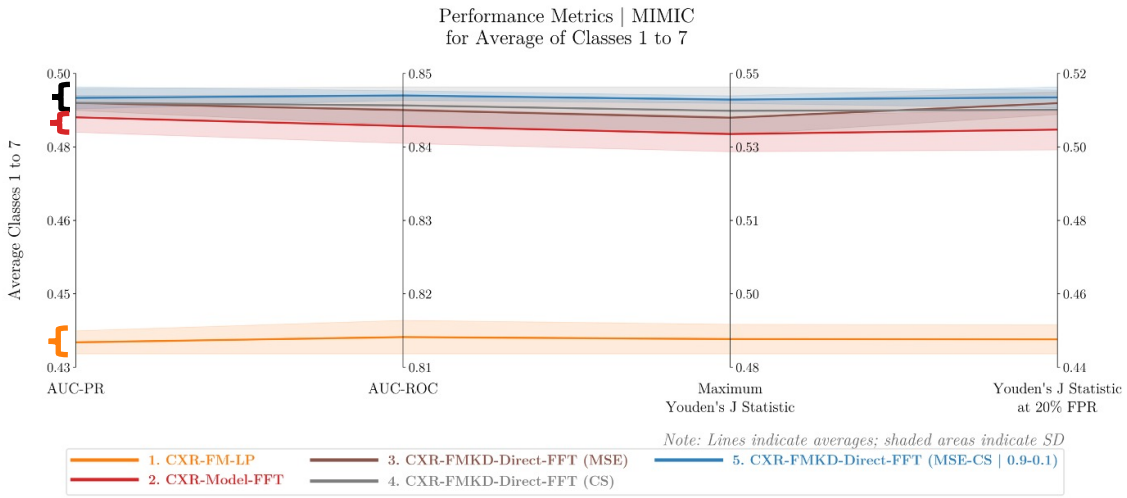


Students

Baseline

Teacher

MIMIC



CheXpert

	Teacher	Baseline	Students		
			MSE	CS	MSE-CS | 0.6-0.4
	Avg ± SD	Avg ± SD	Avg ± SD	Avg ± SD	Avg ± SD
Epoch at Lowest Val Loss	**8.40 ± 2.70**	**13.40 ± 1.82**	**3.00 ± 0.71**	**1.80 ± 0.45**	**3.20 ± 0.84**

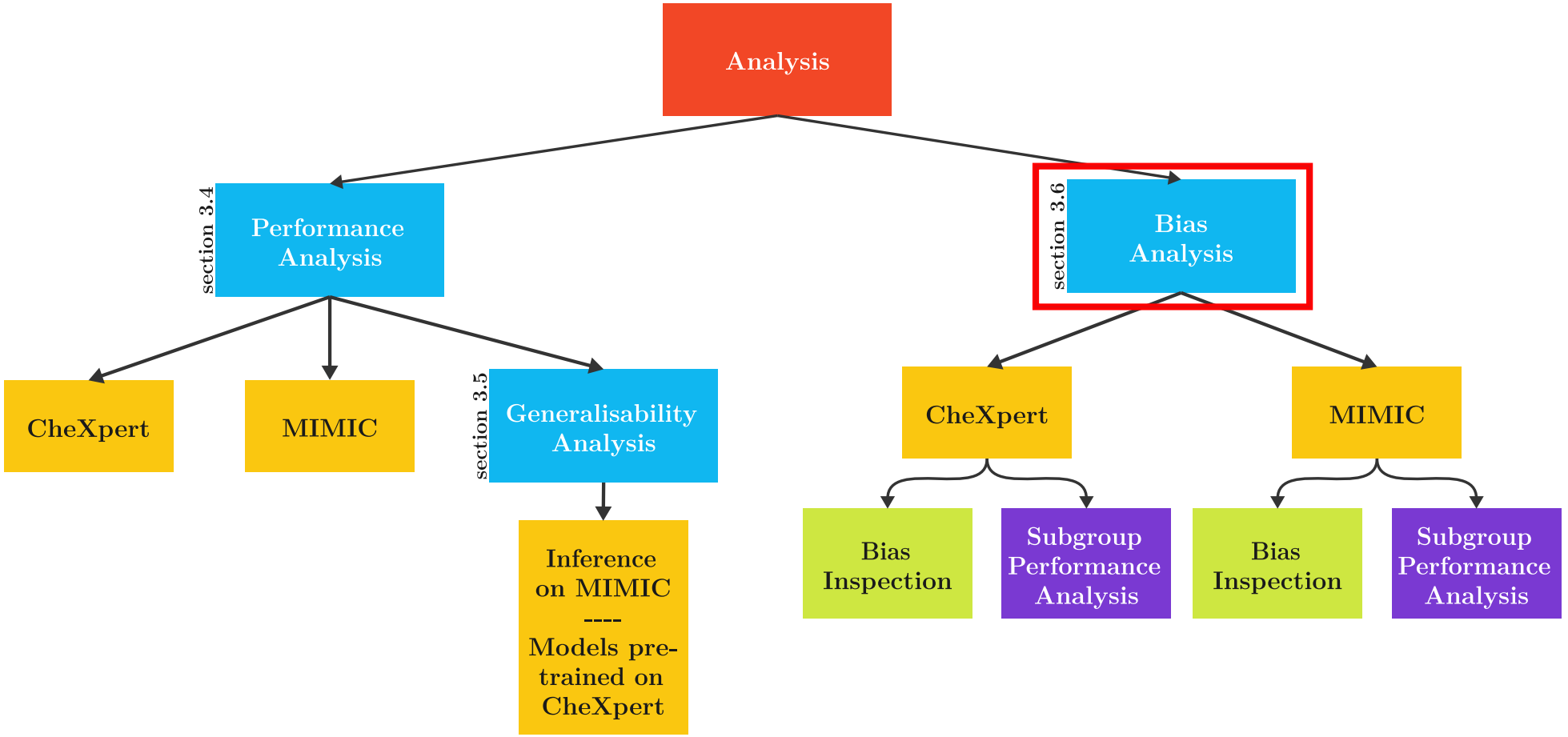
MIMIC

	Teacher	Baseline	Students		
			MSE	CS	MSE-CS | 0.9-0.1
	Avg ± SD	Avg ± SD	Avg ± SD	Avg ± SD	Avg ± SD
Epoch at Lowest Val Loss	**8.60 ± 5.37**	**9.60 ± 1.14**	**4.20 ± 1.30**	**2.60 ± 0.55**	**3.20 ± 1.10**

Students:

- Achieve **highest performance**, surpassing both teacher (significantly) and baseline
- Exhibit **faster convergence** rates than both baseline (significantly) and teacher
- Benefit from *enhanced knowledge* via KD with teacher → rich features leading to faster convergence (strength of FM)

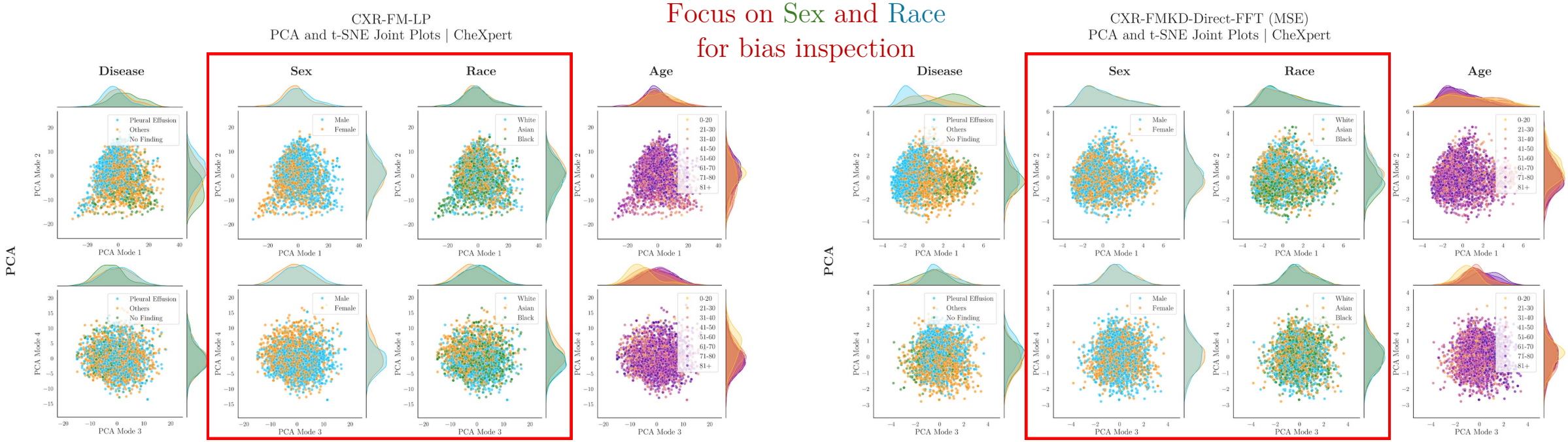
3.3. Bias Analysis (1)



3.3. Bias Analysis (2)

⇒ Has the model learned to discriminate based on *protected characteristics*?

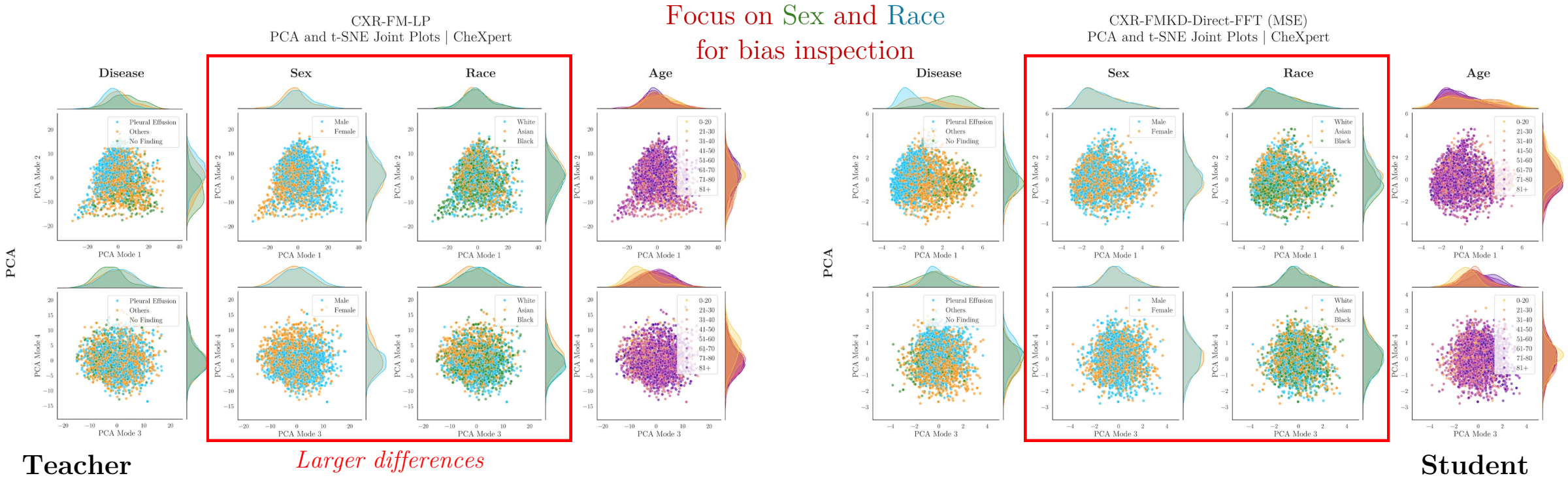
Focus on Sex and Race for bias inspection



Teacher *Larger differences* **Student**

3.3. Bias Analysis (2)

⇒ Has the model learned to discriminate based on *protected characteristics*?



- Observed *pairwise subgroup comparisons*: Male vs. Female – White vs. Asian – White vs. Black – Asian vs. Black
- **Novel Bias Score** developed to quantify *marginal distribution differences*, based on Kolmogorov-Smirnov statistical tests and bootstrapping with 5000 simulations → **ranges from 0 (*unbiased*) to 150 (*most biased*)**. [see Appendix]

3.3.1. Bias Inspection

CheXpert

	Race				Sex		Overall	
	Aggregate P-Value Significance	Bias Score	Aggregate P-Value Significance	Combined Bias Score	Aggregate P-Value Significance	Combined Bias Score	Aggregate P-Value Significance	Combined Bias Score
Teacher	*White vs Asian* FALSE : 59.43% TRUE : 8.65% TRUE+ : 31.93%	56.54	Race Attribute		Sex Attribute		Attributes Average	
	White vs Black FALSE : 65.00% TRUE : 30.49% TRUE+ : 4.51%	37.26	FALSE : 46.81% TRUE : 22.44% TRUE+ : 30.75%	68.56	*Male vs Female* FALSE : 0.04% TRUE : 1.41% TRUE+ : 98.55%	149.24	FALSE : 23.42% TRUE : 11.93% TRUE+ : 64.65%	108.90
	Asian vs Black FALSE : 16.01% TRUE : 28.18% TRUE+ : 55.81%	111.90						
Baseline	*White vs Asian* FALSE : 95.18% TRUE : 4.60% TRUE+ : 0.22%	4.93	Race Attribute		Sex Attribute		Attributes Average	
	White vs Black FALSE : 61.75% TRUE : 8.85% TRUE+ : 29.39%	52.94	FALSE : 74.43% TRUE : 7.74% TRUE+ : 17.83%	34.48	*Male vs Female* FALSE : 82.46% TRUE : 12.63% TRUE+ : 4.91%	20.00	FALSE : 78.45% TRUE : 10.19% TRUE+ : 11.37%	27.24
	Asian vs Black FALSE : 66.37% TRUE : 9.77% TRUE+ : 23.87%	45.56						
Student (MSE)	*White vs Asian* FALSE : 96.92% TRUE : 2.92% TRUE+ : 0.15%	3.15	Race Attribute		Sex Attribute		Attributes Average	
	White vs Black FALSE : 71.35% TRUE : 4.50% TRUE+ : 24.15%	40.73	FALSE : 80.31% TRUE : 3.57% TRUE+ : 16.13%	27.76	*Male vs Female* FALSE : 94.82% TRUE : 4.95% TRUE+ : 0.23%	5.30	FALSE : 87.56% TRUE : 4.26% TRUE+ : 8.18%	16.53
	Asian vs Black FALSE : 72.64% TRUE : 3.28% TRUE+ : 24.08%	39.40						
Student (CS)	*White vs Asian* FALSE : 71.31% TRUE : 9.56% TRUE+ : 19.13%	38.26	Race Attribute		Sex Attribute		Attributes Average	
	White vs Black FALSE : 72.54% TRUE : 6.53% TRUE+ : 20.93%	37.93	FALSE : 65.09% TRUE : 10.86% TRUE+ : 24.05%	46.94	*Male vs Female* FALSE : 61.84% TRUE : 13.07% TRUE+ : 25.09%	50.70	FALSE : 63.47% TRUE : 11.97% TRUE+ : 24.57%	48.82
	Asian vs Black FALSE : 51.41% TRUE : 16.50% TRUE+ : 32.08%	64.63						
Student (MSE-CS)	*White vs Asian* FALSE : 95.25% TRUE : 4.53% TRUE+ : 0.21%	4.85	Race Attribute		Sex Attribute		Attributes Average	
	White vs Black FALSE : 61.09% TRUE : 13.48% TRUE+ : 25.43%	51.63	FALSE : 75.98% TRUE : 7.40% TRUE+ : 16.61%	32.32	*Male vs Female* FALSE : 75.99% TRUE : 18.45% TRUE+ : 5.56%	26.79	FALSE : 75.99% TRUE : 12.92% TRUE+ : 11.09%	29.56
	Asian vs Black FALSE : 71.61% TRUE : 4.19% TRUE+ : 24.20%	40.49						

CheXpert

MIMIC

	Race				Sex		Overall	
	Aggregate P-Value Significance	Bias Score	Aggregate P-Value Significance	Combined Bias Score	Aggregate P-Value Significance	Combined Bias Score	Aggregate P-Value Significance	Combined Bias Score
Teacher	*White vs Asian* FALSE : 71.44% TRUE : 25.46% TRUE+ : 3.10%	30.11	Race Attribute		Sex Attribute		Attributes Average	
	White vs Black FALSE : 82.01% TRUE : 13.03% TRUE+ : 4.95%	20.46	FALSE : 61.74% TRUE : 24.42% TRUE+ : 13.85%	45.19	*Male vs Female* FALSE : 19.17% TRUE : 10.99% TRUE+ : 69.84%	115.75	FALSE : 40.45% TRUE : 17.70% TRUE+ : 41.84%	80.47
	Asian vs Black FALSE : 31.75% TRUE : 34.75% TRUE+ : 33.49%	84.99						
Baseline	*White vs Asian* FALSE : 99.08% TRUE : 0.90% TRUE+ : 0.01%	0.92	Race Attribute		Sex Attribute		Attributes Average	
	White vs Black FALSE : 63.80% TRUE : 23.93% TRUE+ : 12.26%	42.33	FALSE : 74.74% TRUE : 17.46% TRUE+ : 7.80%	29.16	*Male vs Female* FALSE : 19.13% TRUE : 48.28% TRUE+ : 32.59%	97.16	FALSE : 46.93% TRUE : 32.87% TRUE+ : 20.19%	63.16
	Asian vs Black FALSE : 61.33% TRUE : 27.54% TRUE+ : 11.13%	44.23						
Student (MSE)	*White vs Asian* FALSE : 94.69% TRUE : 5.12% TRUE+ : 0.19%	5.40	Race Attribute		Sex Attribute		Attributes Average	
	White vs Black FALSE : 60.22% TRUE : 18.76% TRUE+ : 21.01%	50.28	FALSE : 64.69% TRUE : 18.52% TRUE+ : 16.79%	43.71	*Male vs Female* FALSE : 22.71% TRUE : 42.42% TRUE+ : 34.87%	94.72	FALSE : 43.70% TRUE : 30.47% TRUE+ : 25.83%	69.22
	Asian vs Black FALSE : 39.16% TRUE : 31.66% TRUE+ : 29.18%	75.43						
Student (CS)	*White vs Asian* FALSE : 88.34% TRUE : 10.41% TRUE+ : 1.25%	12.28	Race Attribute		Sex Attribute		Attributes Average	
	White vs Black FALSE : 46.80% TRUE : 23.79% TRUE+ : 29.41%	67.90	FALSE : 63.22% TRUE : 20.70% TRUE+ : 16.09%	44.83	*Male vs Female* FALSE : 20.73% TRUE : 12.67% TRUE+ : 66.60%	112.57	FALSE : 41.97% TRUE : 16.68% TRUE+ : 41.34%	78.70
	Asian vs Black FALSE : 54.51% TRUE : 27.89% TRUE+ : 17.60%	54.29						
Student (MSE-CS)	*White vs Asian* FALSE : 97.44% TRUE : 2.50% TRUE+ : 0.07%	2.60	Race Attribute		Sex Attribute		Attributes Average	
	White vs Black FALSE : 38.39% TRUE : 31.23% TRUE+ : 30.37%	76.79	FALSE : 55.37% TRUE : 22.43% TRUE+ : 22.20%	55.72	*Male vs Female* FALSE : 16.73% TRUE : 19.83% TRUE+ : 63.43%	114.98	FALSE : 36.05% TRUE : 21.13% TRUE+ : 42.82%	85.35
	Asian vs Black FALSE : 30.29% TRUE : 33.55% TRUE+ : 36.16%	87.78						

Bias Scores:
0 – 150
(unbiased) – (most biased)

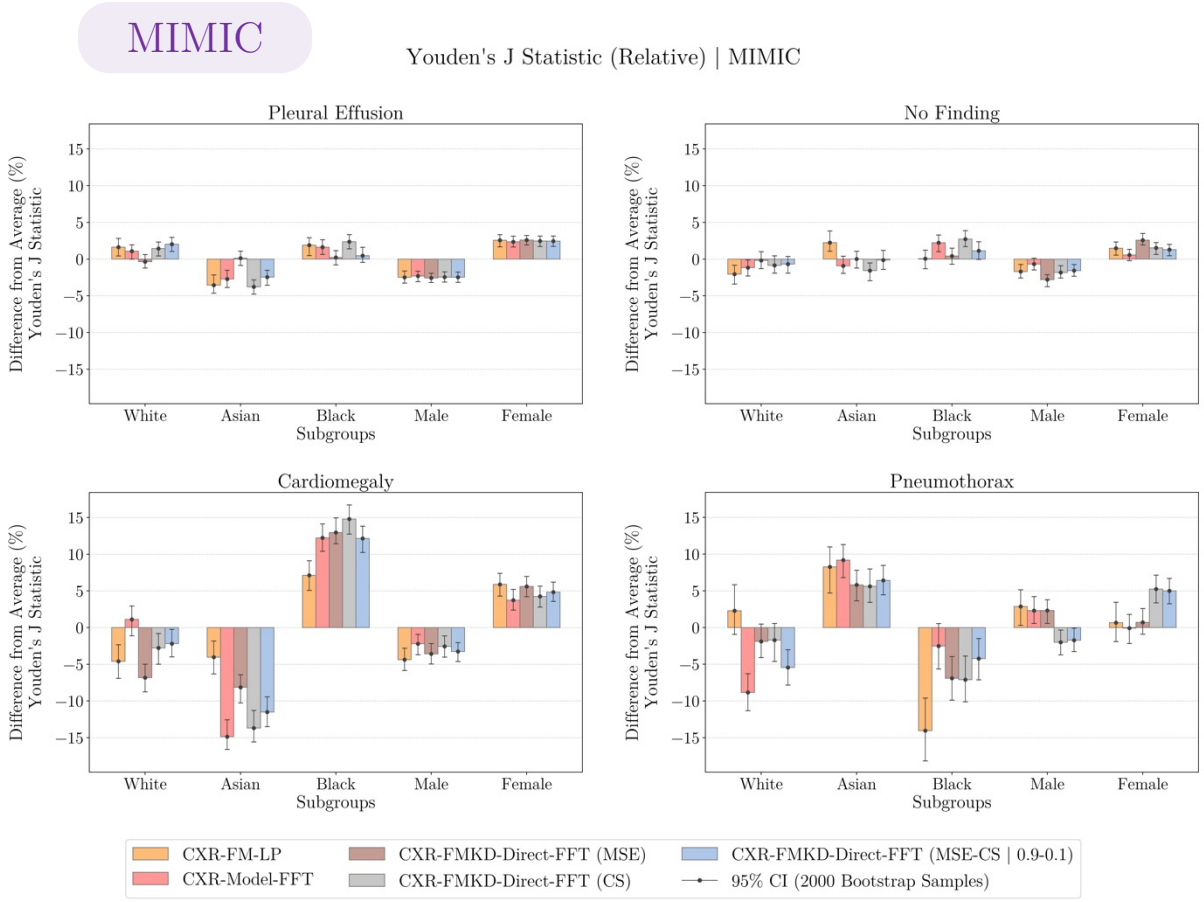
- *Student (MSE)* and baseline show **lowest bias**; teacher among the most biased
- *Consistent bias trends* across attributes, but nuances within race
- Models in MIMIC *more biased* overall than CheXpert → **more nuanced** results for MIMIC

3.3.2. Subgroup Performance Analysis

CheXpert

- **Significant reduction** in subgroup performance disparities for *students* compared to teacher
- Lower bias scores correlate with reduced disparities

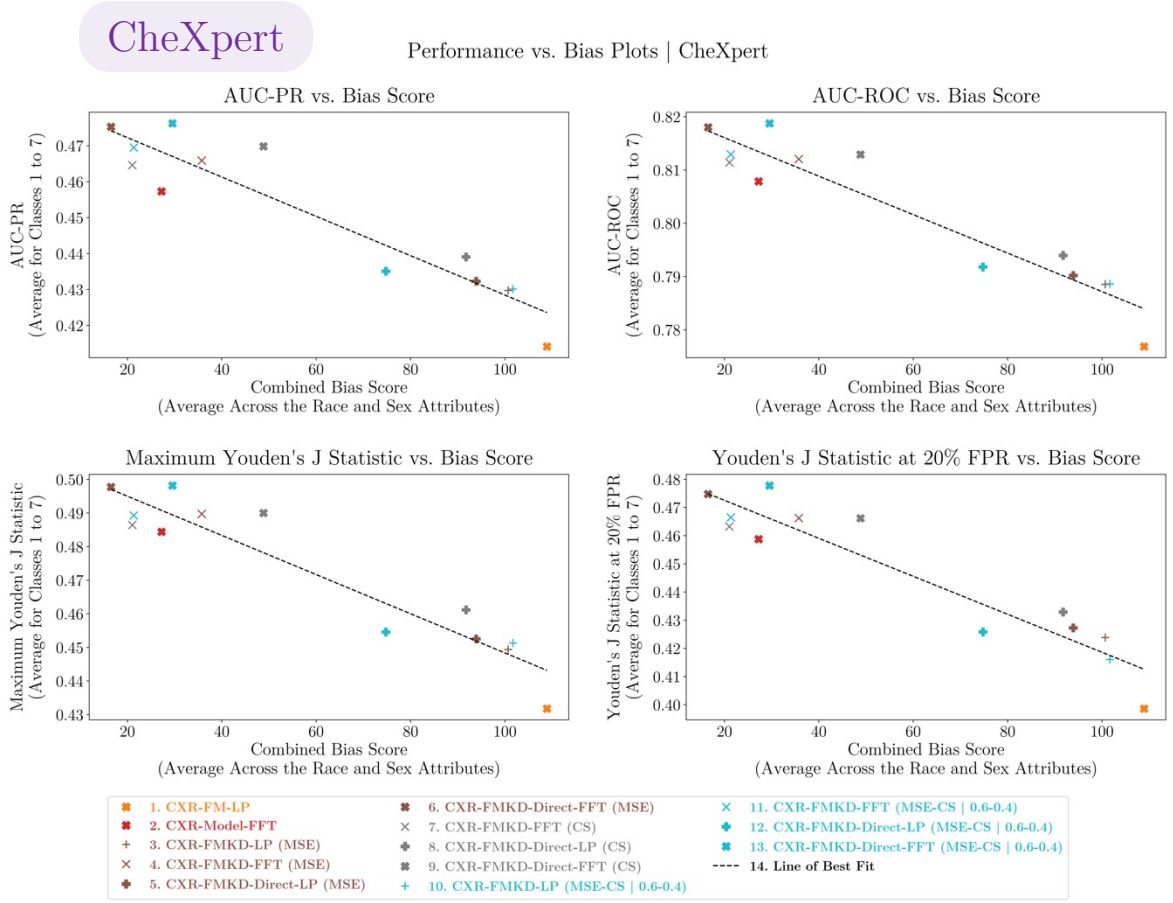
MIMIC



- **No clear reduction** in subgroup performance disparities for *students* compared to teacher or baseline → high bias scores across all models

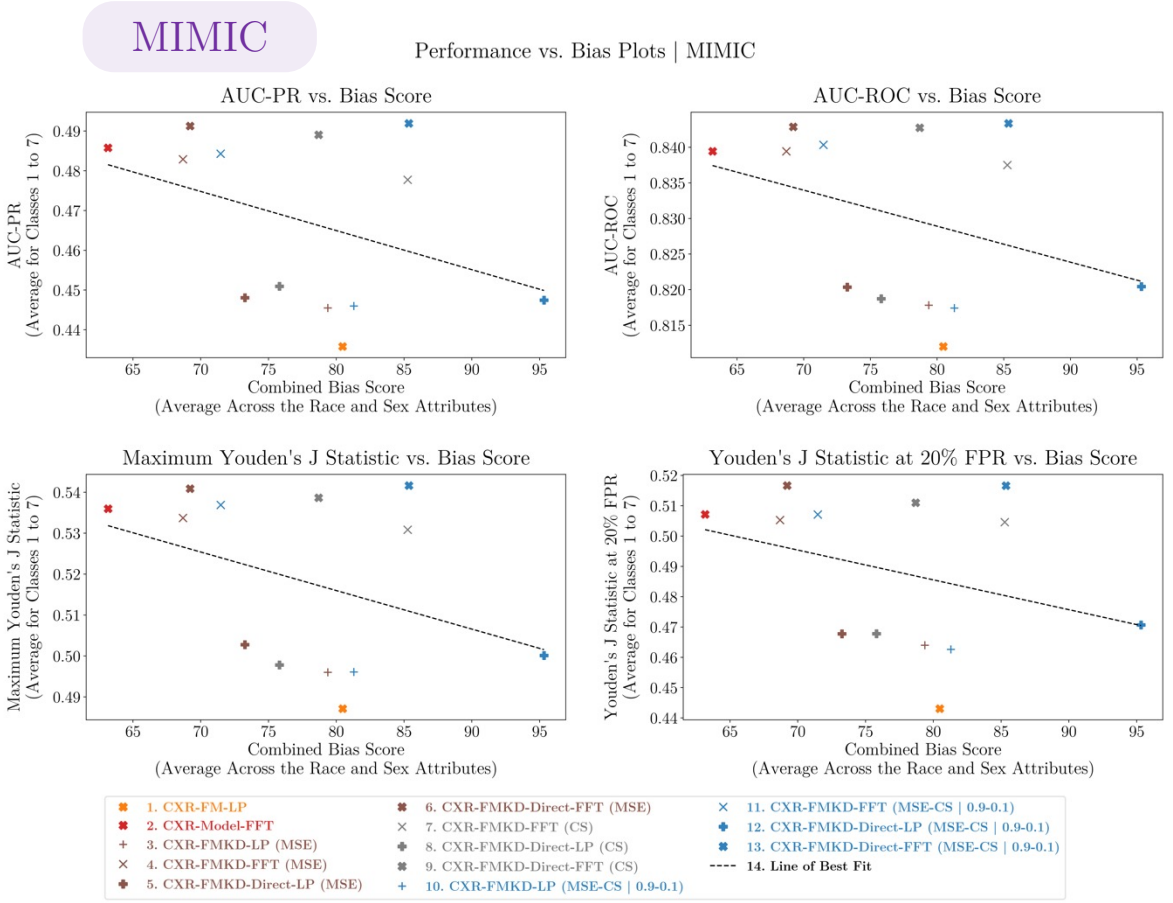
3.3.3. Performance vs. Bias Analysis

CheXpert



- **Lower bias** correlates with **higher performance**
- Superior CheXpert performance → involves *minimising reliance* on protected characteristics

MIMIC



- **Top-performance** achieved across **varying bias levels** → we should favour less biased models
- Models seem to *leverage* protected characteristics

4. Conclusion (1)

This paper successfully **demonstrates the potential of KD** as a strategy to **robustly reconstruct FMs**, effectively addressing issues related to their inherent *lack of transparency* and the *propagation of biases*.

Results

- **Performance Analysis:** CXR-FMKD-Direct FFT student models show up to **20%** improvement in performance metrics over CXR-FM teacher and **4%** over CXR-Model FFT baseline.
- **Convergence**: Students chieve minimum validation loss up to **87%** faster than baseline and **79%** faster than teacher.
- **Bias Analysis (CheXpert):** Students exhibit up to an **85%** reduction in bias scores, correlating with performance gains and a shift towards disease discrimination based on clinically relevant features rather than protected characteristics. Notable reductions in subgroup performance disparities compared to teacher.
- **Bias Analysis (MIMIC):** High bias scores across all models with close clustering, limiting clear trends in subgroup performance disparities. Top performance achieved across varying bias levels.

Student (MSE) stood out as the **most balance**, achieving high performance with low bias → whereas student (MSE-CS), despite its best performance overall, registered higher bias scores and variability.

4. Conclusion (2)

Implications

- **Dataset-specific strategies** → crucial for model selection and bias mitigation due to the diverse impacts of protected characteristics across datasets like CheXpert and MIMIC.
- **Essential open access to FMs** → needed for clinical deployment and advancing medical AI.
- **Recommendations for using KD** → include starting with MSE as KD loss for simplicity and effectiveness, removing projectors post-KD akin to SSL practices, and applying FFT.
- **Proposed Novel Bias Score** → simplifies detailed bias assessment across models

Future Work

- **Diverse Datasets:** Evaluate models on more varied datasets beyond CheXpert and MIMIC.
- **KD Enhancements:** Further explore KD loss functions, transfer datasets, and model architectures.
- **Expanded Subgroup Analysis:** Extend subgroup performance analysis to include all (14) disease labels for a more comprehensive assessment across diverse patient subgroups.

References

[1] Louati, H., Louati, A., Bechikh, S. et al. Topology optimization search of deep convolution neural networks for CT and X-ray image classification. BMC Med Imaging 22, 120 (2022). https://doi.org/10.1186/s12880-022-00847-w

[2] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A (2024) Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 74:229–263

[3] Bommasani R, Hudson DA, Adeli E, et al (2021) On the Opportunities and Risks of Foundation Models.

[4] Sellergren AB, Chen C, Nabulsi Z, et al (2022) Simplified Transfer Learning for Chest Radiography Models Using Less Data. Radiology 305:454–465

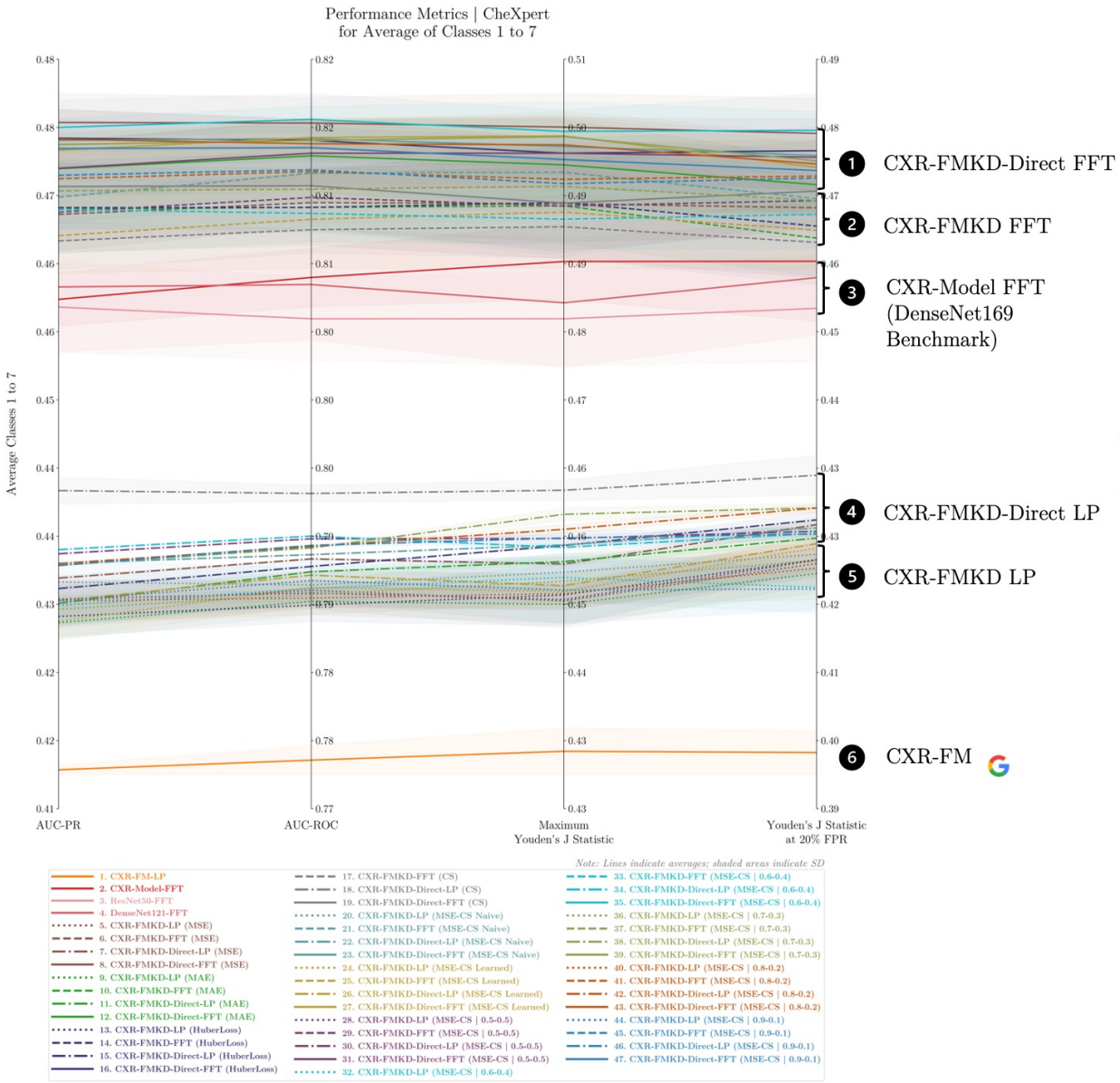
[5] Glocker B, Jones C, Roschewitz M, Winzeck S (2023) Risk of Bias in Chest Radiography Deep Learning Foundation Models. Radiol Artif Intell 5:230060

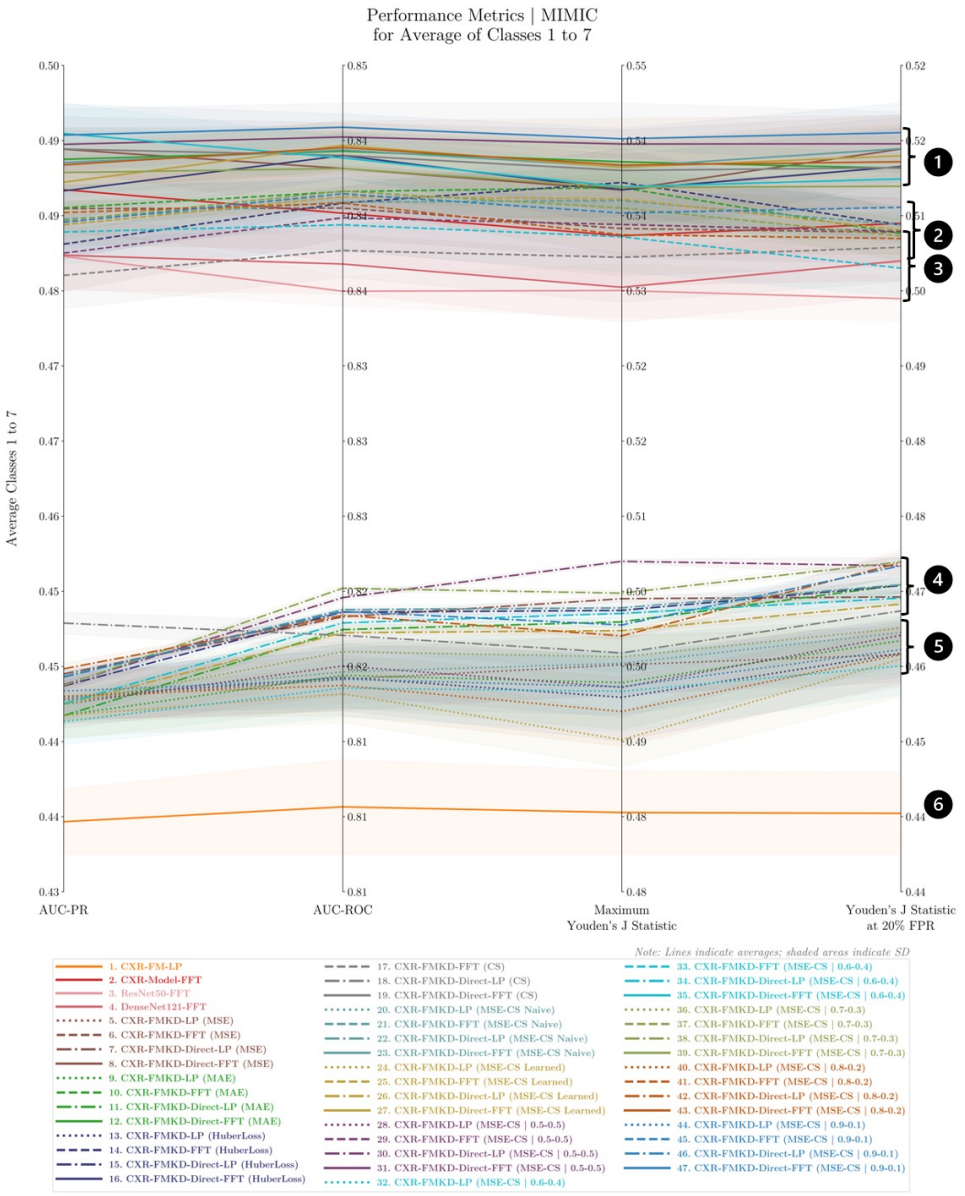
[6] Gou J, Yu B, Maybank SJ, Tao D (2020) Knowledge Distillation: A Survey. Int J Comput Vis 129:1789–1819

[7] Glocker B, Jones C, Bernhardt M, Winzeck S (2023) Algorithmic encoding of protected characteristics in chest X-ray disease detection models. EBioMedicine 89:104467

[8] Irvin J, Rajpurkar P, Ko M, et al (2019) CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In: 33rd AAAI Conference on Artificial Intelligence 2019. AAAI Press, Honolulu, Hawaii, USA, pp 590–597

Appendix


Performance Metrics | CheXpert
for Average of Classes 1 to 7
Average Classes 1 to 7
AUC-PR
AUC-ROC
Maximum
Youden's J Statistic
Youden's J Statistic
at 20% FPR
1 CXR-FMKD-Direct FFT
2 CXR-FMKD FFT
3 CXR-Model FFT
(DenseNet169
Benchmark)
4 CXR-FMKD-Direct LP
5 CXR-FMKD LP
6 CXR-FM
Note: Lines indicate averages; shaded areas indicate SD
1. CXR-FM-LP
2. CXR-Model-FFT
3. ResNet50-FFT
4. DenseNet121-FFT
5. CXR-FMKD-LP (MSE)
6. CXR-FMKD-FFT (MSE)
7. CXR-FMKD-Direct-LP (MSE)
8. CXR-FMKD-Direct-FFT (MSE)
9. CXR-FMKD-LP (MAE)
10. CXR-FMKD-FFT (MAE)
11. CXR-FMKD-Direct-LP (MAE)
12. CXR-FMKD-Direct-FFT (MAE)
13. CXR-FMKD-LP (HuberLoss)
14. CXR-FMKD-FFT (HuberLoss)
15. CXR-FMKD-Direct-LP (HuberLoss)
16. CXR-FMKD-Direct-FFT (HuberLoss)
17. CXR-FMKD-FFT (CS)
18. CXR-FMKD-Direct-LP (CS)
19. CXR-FMKD-Direct-FFT (CS)
20. CXR-FMKD-LP (MSE-CS Naive)
21. CXR-FMKD-FFT (MSE-CS Naive)
22. CXR-FMKD-Direct-LP (MSE-CS Naive)
23. CXR-FMKD-Direct-FFT (MSE-CS Naive)
24. CXR-FMKD-LP (MSE-CS Learned)
25. CXR-FMKD-FFT (MSE-CS Learned)
26. CXR-FMKD-Direct-LP (MSE-CS Learned)
27. CXR-FMKD-Direct-FFT (MSE-CS Learned)
28. CXR-FMKD-LP (MSE-CS | 0.5-0.5)
29. CXR-FMKD-FFT (MSE-CS | 0.5-0.5)
30. CXR-FMKD-Direct-LP (MSE-CS | 0.5-0.5)
31. CXR-FMKD-Direct-FFT (MSE-CS | 0.5-0.5)
32. CXR-FMKD-LP (MSE-CS | 0.6-0.4)
33. CXR-FMKD-FFT (MSE-CS | 0.6-0.4)
34. CXR-FMKD-Direct-LP (MSE-CS | 0.6-0.4)
35. CXR-FMKD-Direct-FFT (MSE-CS | 0.6-0.4)
36. CXR-FMKD-LP (MSE-CS | 0.7-0.3)
37. CXR-FMKD-FFT (MSE-CS | 0.7-0.3)
38. CXR-FMKD-Direct-LP (MSE-CS | 0.7-0.3)
39. CXR-FMKD-Direct-FFT (MSE-CS | 0.7-0.3)
40. CXR-FMKD-LP (MSE-CS | 0.8-0.2)
41. CXR-FMKD-FFT (MSE-CS | 0.8-0.2)
42. CXR-FMKD-Direct-LP (MSE-CS | 0.8-0.2)
43. CXR-FMKD-Direct-FFT (MSE-CS | 0.8-0.2)
44. CXR-FMKD-LP (MSE-CS | 0.9-0.1)
45. CXR-FMKD-FFT (MSE-CS | 0.9-0.1)
46. CXR-FMKD-Direct-LP (MSE-CS | 0.9-0.1)
47. CXR-FMKD-Direct-FFT (MSE-CS | 0.9-0.1)


Performance Metrics | MIMIC
for Average of Classes 1 to 7
Average Classes 1 to 7
AUC-PR
AUC-ROC
Maximum
Youden's J Statistic
Youden's J Statistic
at 20% FPR
Note: Lines indicate averages; shaded areas indicate SD
1. CXR-FM-LP
2. CXR-Model-FFT
3. ResNet50-FFT
4. DenseNet121-FFT
5. CXR-FMKD-LP (MSE)
6. CXR-FMKD-FFT (MSE)
7. CXR-FMKD-Direct-LP (MSE)
8. CXR-FMKD-Direct-FFT (MSE)
9. CXR-FMKD-LP (MAE)
10. CXR-FMKD-FFT (MAE)
11. CXR-FMKD-Direct-LP (MAE)
12. CXR-FMKD-Direct-FFT (MAE)
13. CXR-FMKD-LP (HuberLoss)
14. CXR-FMKD-FFT (HuberLoss)
15. CXR-FMKD-Direct-LP (HuberLoss)
16. CXR-FMKD-Direct-FFT (HuberLoss)
17. CXR-FMKD-FFT (CS)
18. CXR-FMKD-Direct-LP (CS)
19. CXR-FMKD-Direct-FFT (CS)
20. CXR-FMKD-LP (MSE-CS Naive)
21. CXR-FMKD-FFT (MSE-CS Naive)
22. CXR-FMKD-Direct-LP (MSE-CS Naive)
23. CXR-FMKD-Direct-FFT (MSE-CS Naive)
24. CXR-FMKD-LP (MSE-CS Learned)
25. CXR-FMKD-FFT (MSE-CS Learned)
26. CXR-FMKD-Direct-LP (MSE-CS Learned)
27. CXR-FMKD-Direct-FFT (MSE-CS Learned)
28. CXR-FMKD-LP (MSE-CS | 0.5-0.5)
29. CXR-FMKD-FFT (MSE-CS | 0.5-0.5)
30. CXR-FMKD-Direct-LP (MSE-CS | 0.5-0.5)
31. CXR-FMKD-Direct-FFT (MSE-CS | 0.5-0.5)
32. CXR-FMKD-LP (MSE-CS | 0.6-0.4)
33. CXR-FMKD-FFT (MSE-CS | 0.6-0.4)
34. CXR-FMKD-Direct-LP (MSE-CS | 0.6-0.4)
35. CXR-FMKD-Direct-FFT (MSE-CS | 0.6-0.4)
36. CXR-FMKD-LP (MSE-CS | 0.7-0.3)
37. CXR-FMKD-FFT (MSE-CS | 0.7-0.3)
38. CXR-FMKD-Direct-LP (MSE-CS | 0.7-0.3)
39. CXR-FMKD-Direct-FFT (MSE-CS | 0.7-0.3)
40. CXR-FMKD-LP (MSE-CS | 0.8-0.2)
41. CXR-FMKD-FFT (MSE-CS | 0.8-0.2)
42. CXR-FMKD-Direct-LP (MSE-CS | 0.8-0.2)
43. CXR-FMKD-Direct-FFT (MSE-CS | 0.8-0.2)
44. CXR-FMKD-LP (MSE-CS | 0.9-0.1)
45. CXR-FMKD-FFT (MSE-CS | 0.9-0.1)
46. CXR-FMKD-Direct-LP (MSE-CS | 0.9-0.1)
47. CXR-FMKD-Direct-FFT (MSE-CS | 0.9-0.1)

Noval Bias Score Calculation

				Disease Detection	Race Attribute			Sex Attribute	Overall
				#	#	#	#	#	
Model Name	**Mode**	**Explained Variance**		**Pleural Eff. vs No Finding**	**White vs Asian**	**White vs Black**	**Asian vs Black**	**Male vs Female**	STEP 1
CXR-FMKD-Direct-FFT (MSE)	PCA Mode 1	19.47%	FALSE	0	4961	4938	4875	4938	Perform 5000 iterations of balanced stratified sampling to create a 3000-patient set, conduct pairwise subgroup statistical tests, then tabulate the occurrences of outcomes, categorised by p-value significance into 'FALSE', 'TRUE', and 'TRUE+'.
			TRUE	0	39	61	121	61	
			TRUE+	5000	0	1	4	1	
	PCA Mode 2	9.95%	FALSE	0	4761	0	0	4474	
			TRUE	0	231	0	0	499	
			TRUE+	5000	8	5000	5000	27	
	PCA Mode 3	7.16%	FALSE	0	4974	3956	4367	4792	
			TRUE	41	26	996	609	202	
			TRUE+	4959	0	48	24	6	
	PCA Mode 4	4.96%	FALSE	0	4382	4786	4982	4428	
			TRUE	5	570	207	18	542	
			TRUE+	4995	48	7	0	30	
				%	%	%	%	%	
Model Name	**Mode**	**Explained Variance**		**Pleural Eff. vs No Finding**	**White vs Asian**	**White vs Black**	**Asian vs Black**	**Male vs Female**	STEP 2
CXR-FMKD-Direct-FFT (MSE)	PCA Mode 1	19.47%	FALSE	0%	99%	99%	98%	99%	Convert the occurrences of 'FALSE', 'TRUE', and 'TRUE+' into percentages.
			TRUE	0%	1%	1%	2%	1%	
			TRUE+	100%	0%	0%	0%	0%	
	PCA Mode 2	9.95%	FALSE	0%	95%	0%	0%	89%	
			TRUE	0%	5%	0%	0%	10%	
			TRUE+	100%	0%	100%	100%	1%	
	PCA Mode 3	7.16%	FALSE	0%	99%	79%	87%	96%	
			TRUE	1%	1%	20%	12%	4%	
			TRUE+	99%	0%	1%	0%	0%	
	PCA Mode 4	4.96%	FALSE	0%	88%	96%	100%	89%	
			TRUE	0%	11%	4%	0%	11%	
			TRUE+	100%	1%	0%	0%	1%	
				%	%	%	%	%	
Model Name	**Mode**	**Normalised Exp. Var.**		**Pleural Eff. vs No Finding**	**White vs Asian**	**White vs Black**	**Asian vs Black**	**Male vs Female**	STEP 3
CXR-FMKD-Direct-FFT (MSE)	PCA Mode 1	47%	FALSE	0%	47%	46%	46%	46%	Calculate the Normalised Explained Variance for the first four PCA modes and use them to adjust the percentages in each corresponding row. For each demographic comparison, sum these weighted percentages for 'FALSE', 'TRUE', and 'TRUE+' outcomes, and combine with assigned p-scores to calculate a bias score per column. Average these for Race and Sex, then further average to get the overall bias score.
			TRUE	0%	0%	1%	1%	1%	
			TRUE+	47%	0%	0%	0%	0%	
	PCA Mode 2	24%	FALSE	0%	23%	0%	0%	21%	
			TRUE	0%	1%	0%	0%	2%	
			TRUE+	24%	0%	24%	24%	0%	
	PCA Mode 3	17%	FALSE	0%	17%	14%	15%	17%	
			TRUE	0%	0%	3%	2%	1%	
			TRUE+	17%	0%	0%	0%	0%	
	PCA Mode 4	12%	FALSE	0%	10%	11%	12%	11%	
			TRUE	0%	1%	0%	0%	1%	
			TRUE+	12%	0%	0%	0%	0%	
			FALSE	**0%**	97%	71%	73%	95%	
		100%	**TRUE**	**0%**	3%	4%	3%	5%	
			TRUE+	**100%**	0%	24%	24%	0%	
	P-VALUES RANGE		P-SCORES		3.15	40.73	39.40	5.30	
FALSE	(p > 0.05)		0			80%		95%	88% FALSE
TRUE	(0.001 < p < 0.05)		100			4%		5%	4% TRUE
TRUE+	(p < 0.001)		150			16%		0%	8% TRUE+
						27.76		5.30	16.53
						Race Attribute Bias Score		**Sex Attribute** Bias Score	**Average** Bias Score

Figure A1. Calculation Process of the Novel Bias Score.
This table illustrates the calculation of our novel bias scores for the CXR-FMKD-Direct FFT (MSE) model based on 5000 simulations. It details the categorisation of p-values, the conversion of these categories into percentages, and the final bias score calculation weighted by the explained variance of the PCA modes.

Appendix

Noval Bias Score Calculation

				Disease Detection	Race Attribute			Sex Attribute	Overall
				#	#	#	#	#	
Model Name	**Mode**	**Explained Variance**		**Pleural Eff. vs No Finding**	**White vs Asian**	**White vs Black**	**Asian vs Black**	**Male vs Female**	STEP 1
CXR-FMKD-Direct-FFT (MSE)	PCA Mode 1	19.47%	FALSE	0	4961	4938	4875	4938	Perform 5000 iterations of balanced stratified sampling to create a 3000-patient set, conduct pairwise subgroup statistical tests, then tabulate the occurrences of outcomes, categorised by p-value significance into 'FALSE', 'TRUE', and 'TRUE+'.
			TRUE	0	39	61	121	61	
			TRUE+	5000	0	1	4	1	
	PCA Mode 2	9.95%	FALSE	0	4761	0	0	4474	
			TRUE	0	231	0	0	499	
			TRUE+	5000	8	5000	5000	27	
	PCA Mode 3	7.16%	FALSE	0	4974	3956	4367	4792	
			TRUE	41	26	996	609	202	
			TRUE+	4959	0	48	24	6	
	PCA Mode 4	4.96%	FALSE	0	4382	4786	4982	4428	
			TRUE	5	570	207	18	542	
			TRUE+	4995	48	7	0	30	

Appendix

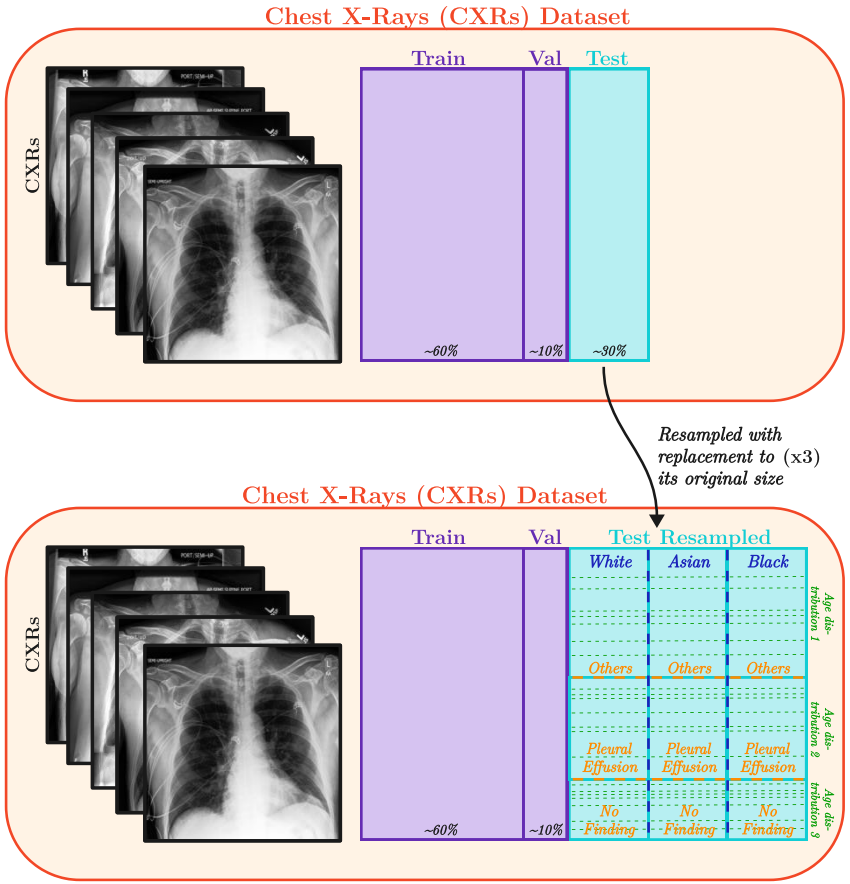
				%	%	%	%	%	
Model Name	**Mode**	**Explained Variance**		**Pleural Eff. vs No Finding**	**White vs Asian**	**White vs Black**	**Asian vs Black**	**Male vs Female**	STEP 2
CXR-FMKD-Direct-FFT (MSE)	PCA Mode 1	19.47%	FALSE	0%	99%	99%	98%	99%	Convert the occurrences of 'FALSE', 'TRUE', and 'TRUE+' into percentages.
			TRUE	0%	1%	1%	2%	1%	
			TRUE+	100%	0%	0%	0%	0%	
	PCA Mode 2	9.95%	FALSE	0%	95%	0%	0%	89%	
			TRUE	0%	5%	0%	0%	10%	
			TRUE+	100%	0%	100%	100%	1%	
	PCA Mode 3	7.16%	FALSE	0%	99%	79%	87%	96%	
			TRUE	1%	1%	20%	12%	4%	
			TRUE+	99%	0%	1%	0%	0%	
	PCA Mode 4	4.96%	FALSE	0%	88%	96%	100%	89%	
			TRUE	0%	11%	4%	0%	11%	
			TRUE+	100%	1%	0%	0%	1%	

Appendix

Model Name	Mode	Normalised Exp. Var.		% Pleural Eff. vs No Finding	% White vs Asian	% White vs Black	% Asian vs Black	% Male vs Female	STEP 3
CXR-FMKD-Direct-FFT (MSE)	PCA Mode 1	47%	FALSE	0%	47%	46%	46%	46%	Calculate the Normalised Explained Variance for the first four PCA modes and use them to adjust the percentages in each corresponding row. For each demographic comparison, sum these weighted percentages for 'FALSE', 'TRUE', and 'TRUE+' outcomes, and combine with assigned p-scores to calculate a bias score per column. Average these for Race and Sex, then further average to get the overall bias score.
			TRUE	0%	0%	1%	1%	1%	
			TRUE+	47%	0%	0%	0%	0%	
	PCA Mode 2	24%	FALSE	0%	23%	0%	0%	21%	
			TRUE	0%	1%	0%	0%	2%	
			TRUE+	24%	0%	24%	24%	0%	
	PCA Mode 3	17%	FALSE	0%	17%	14%	15%	17%	
			TRUE	0%	0%	3%	2%	1%	
			TRUE+	17%	0%	0%	0%	0%	
	PCA Mode 4	12%	FALSE	0%	10%	11%	12%	11%	
			TRUE	0%	1%	0%	0%	1%	
			TRUE+	12%	0%	0%	0%	0%	
			FALSE	**0%**	97%	71%	73%	95%	
		100%	**TRUE**	**0%**	3%	4%	3%	5%	
			TRUE+	**100%**	0%	24%	24%	0%	

	P-VALUES RANGE	P-SCORES	3.15	40.73	39.40	5.30		
FALSE	(p > 0.05)	0		80%		95%	88%	FALSE
TRUE	(0.001 < p < 0.05)	100		4%		5%	4%	TRUE
TRUE+	(p < 0.001)	150		16%		0%	8%	TRUE+
				27.76		5.30	16.53	
				Race Attribute Bias Score		**Sex Attribute** Bias Score	**Average** Bias Score	

Appendix



Test set has been resampled to ensure approximately:

1. Equal *Race* proportions—of White, Asian, and Black
2. Equal *Disease* prevalence within each (race)-subgroup
3. Equal *Age-bin* distribution within each (race, disease)-subsubgroup

Figure A2. Stratified Resampling for Demographic and Clinical Balance in Test Dataset
The resampling process triples the test set size and adjusts to compensate for differences across subgroups, including variations in race, age, and disease prevalence. 'CXRs' refers to chest X-rays.

Appendix

Attribute	CheXpert				MIMIC-CXR			
	All	White	Asian	Black	All	White	Asian	Black
All data								
Patients	42,884	33,338	6642	2904	43,209	32,756	1881	8572
Scans	127,118	99,027 (78)	18,830 (15)	9261 (7)	183,207	14,1865 (77)	7106 (4)	34,236 (19)
Age (years)	63 ± 17	64 ± 17	61 ± 17	56 ± 17	65 ± 17	66 ± 16	63 ± 18	61 ± 17
Female	52,436 (41)	39,735 (40)	8132 (43)	4569 (49)	85,193 (47)	61,626 (43)	31,22 (44)	20,445 (60)
No finding	10,916 (9)	8236 (8)	1716 (9)	964 (10)	56,615 (31)	41,215 (29)	22,21 (31)	13,179 (38)
Pleural effusion	51,574 (41)	40,545 (41)	7953 (42)	3076 (33)	46,224 (25)	38,693 (27)	19,16 (27)	5615 (16)
Training data								
Patients	25,730	20,034	3945	1751	25,925	19,613 (76)	1110 (4)	5202 (20)
Scans	76,205	59,238 (78)	11,371 (15)	5596 (7)	110,280	86,098 (78)	4248 (4)	19,934 (18)
Age (years)	63 ± 17	64 ± 17	62 ± 17	56 ± 17	65 ± 17	66 ± 16	63 ± 18	60 ± 17
Female	31,432 (41)	23,715 (40)	4976 (44)	2741 (49)	51,138 (46)	37,518 (44)	1897 (45)	11,723 (59)
No finding	6514 (9)	4910 (8)	1046 (9)	558 (10)	34,530 (31)	25,170 (29)	1330 (31)	8030 (40)
Pleural effusion	31,015 (41)	24,405 (41)	4754 (42)	1856 (33)	27,806 (25)	23,526 (27)	11,08 (26)	3172 (16)
Validation data								
Patients	4288	3348	666	274	4321	3242 (75)	209 (5)	870 (20)
Scans	12,673	9945 (79)	1809 (14)	919 (7)	17,665	13,369 (76)	776 (4)	3520 (20)
Age (years)	62 ± 17	63 ± 17	62 ± 17	55 ± 16	65 ± 17	67 ± 16	60 ± 22	62 ± 17
Female	5030 (40)	3933 (40)	667 (37)	430 (47)	8245 (47)	5755 (43)	336 (43)	2154 (61)
No finding	1086 (9)	817 (8)	175 (10)	94 (10)	5393 (31)	3903 (29)	232 (30)	1258 (36)
Pleural effusion	5049 (40)	3988 (40)	738 (41)	323 (35)	4575 (26)	3721 (28)	230 (30)	624 (18)
Test data								
Patients	12,866	9956	2031	879	12,963	9901 (76)	562 (5)	2500 (19)
Scans	38,240	29,844 (78)	5650 (15)	2746 (7)	55,262	42,398 (77)	2082 (4)	10,782 (19)
Age (years)	63 ± 17	64 ± 17	61 ± 17	57 ± 16	65 ± 17	66 ± 16	65 ± 17	61 ± 17
Female	15,974 (42)	12,087 (41)	2489 (44)	1348 (49)	25,810 (47)	18,353 (43)	889 (43)	6568 (61)
No finding	3316 (9)	2509 (8)	495 (9)	312 (11)	16,692 (30)	12,142 (29)	659 (32)	3891 (36)
Pleural effusion	15,510 (41)	12,152 (41)	2461 (44)	897 (33)	13,843 (25)	11,446 (27)	578 (28)	1819 (17)

Breakdown of demographics over the set of patient scans by racial groups and training, validation and test splits. Percentages in brackets are with respect to the number of scans. We also report the number of unique patients for each group.

Table 1: **Characteristics of the study population.**

Figure A3. Demographic Distribution of Patient Data in CheXpert and MIMIC Datasets. This table, taken from Glocker et al. [7], presents comprehensive demographic and diagnostic data across the CheXpert and MIMIC datasets. It details the number of patients, the number of chest X-ray (CXR) scans, average age ± standard deviation, gender distribution, and prevalence of 'No Finding' and 'Pleural Effusion' labels. Data are segmented for all data, training, validation, and test sets, and further stratified by racial groups (White, Asian, Black) within each dataset.

Appendix

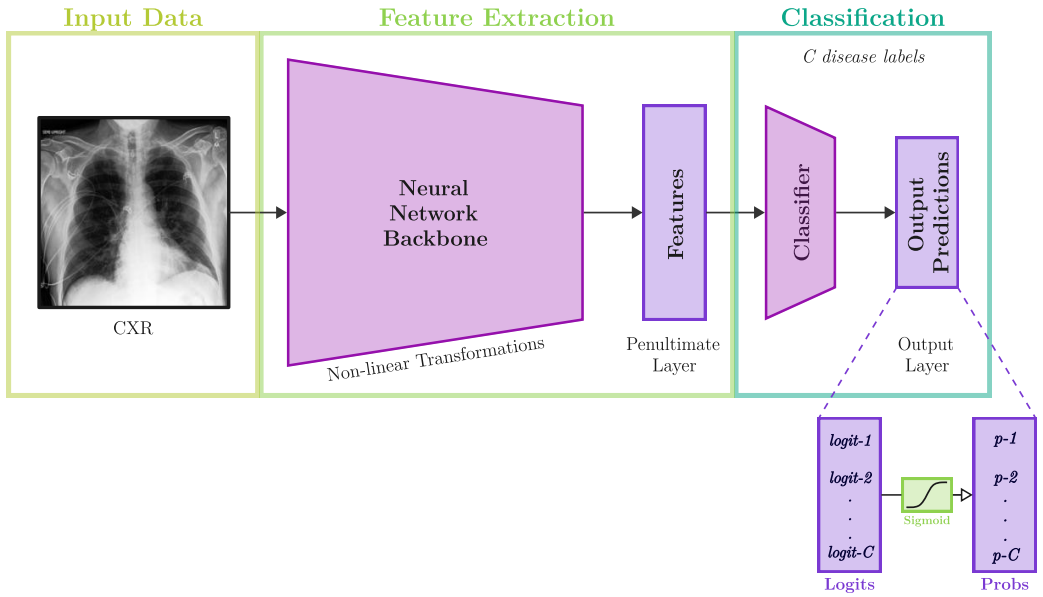


Figure A4. Application of Sigmoid Function in Multi-Label Chest X-Ray Classification. This figure illustrates the application of the sigmoid function to the logits from the output layer of the model, generating probability scores for each of the C disease classes in a general multi-label chest X-ray classification scenario for disease detection.

Appendix

General Legend for Model Development Diagrams

Abbreviations:
- **CXR**: Chest X-Ray
- **FM**: Foundation Model
- **CXR-FM**: Google's proprietary Chest X-Ray Foundation Model
- **KD**: Knowledge Distillation
- **Val**: Validation (set)
- **ImageNet**: A large dataset of natural images.

Model Development Techniques:
- **LP**: Linear Probing
- **FFT**: Full Fine-Tuning

Technical Terms:
- **nn.Linear**: PyTorch Linear Layer—applies an affine linear transformation to the incoming data.
  Used in three main configurations:
  1. **nn.Linear**: Maps DenseNet169 backbone feature embedding (1664 dimensions) to 14 outputs (disease labels).
  2. **nn.Linear-T**: Maps CXR-FM backbone feature embedding (1376 dimensions) to 14 outputs (disease labels).
  3. **nn.Linear-KD**: Maps DenseNet169 backbone feature embedding (1664 dimensions) to match the CXR-FM embedding size (1376 dimensions) for KD purposes.

Vectors:
- ***e1, e2, e3, ...*** : 1st, 2nd, 3rd, ... components (values) of the **embedding** vector—the output from the penultimate layer.
- ***logit1, logit2, logit3, ...*** : 1st, 2nd, 3rd, ... components (values) of the **logits** vector—the raw output from the final output layer.
- ***p1, p2, p3, ...*** : 1st, 2nd, 3rd, ... components (values) of the **predicted probabilities** vector (**Probs**)—the result of applying the sigmoid function to the logits vector for *multi-label classification*.
- ***y1, y2, y3, ...*** : 1st, 2nd, 3rd, ... components (values) of the **ground truth/target labels** vector (Targets)—the true labels.

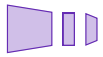
Model Development and Evaluation Tools:
- **BCE**: Binary Cross-Entropy Loss
- **Evaluate Perform.**: Evaluate Performance. This involves applying selected performance metrics to compare the model's predictions (**Probs**) against the actual ground truth (Targets).
- **MSE**: Mean Squared Error
- **MAE**: Mean Absolute Error
- **CS**: Cosine Similarity

Diagrams Colour Codes:

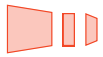
CXR-FM backbone, which is frozen (non-trainable and cannot be fine-tuned).

Sigmoid activation function.

From left to right: DenseNet169 backbone, intermediate embedding vectors, and other neural network components like nn.Linear, are frozen (non-trainable and cannot be fine-tuned).



From left to right: DenseNet169 backbone, intermediate embedding vectors, and additional neural network components such as nn.Linear are trained and fine-tuned for disease detection. The training set is used for model training, while the validation set guides model selection and checkpoints.



From left to right: DenseNet169 backbone, intermediate embedding vectors, and additional neural network components such as nn.Linear are trained and fine-tuned for KD with CXR-FM. The training set is used for model training, while the validation set guides model selection and checkpoints.

Figure A5. General Legend for Model Development Diagrams.
This detailed legend accompanies the series of figures illustrating the architecture, development, and integration of our models in the multi-label disease classification task. It provides definitions and explanations of technical terms and components used throughout these diagrams.

Appendix

Model Architecture	KD Role	# Parameters
DenseNet169	Student	14M
EfficientNet-L2	Teacher	480M

Table A1. Comparative Overview of the Teacher and Student Model Sizes.

Scenario	KD Student Training	Standard CXR Disease Detection Training
# Epochs	40	20
Batch Size	128	128
Optimizer	Adam	Adam
LR Scheduler	*'OneCycle LR'* [201]	—
Base LR	0.001	0.001*
Max LR	0.01	—

*Constant Learning Rate (LR)

Table A2. Summary of the Main Training Hyperparameters.

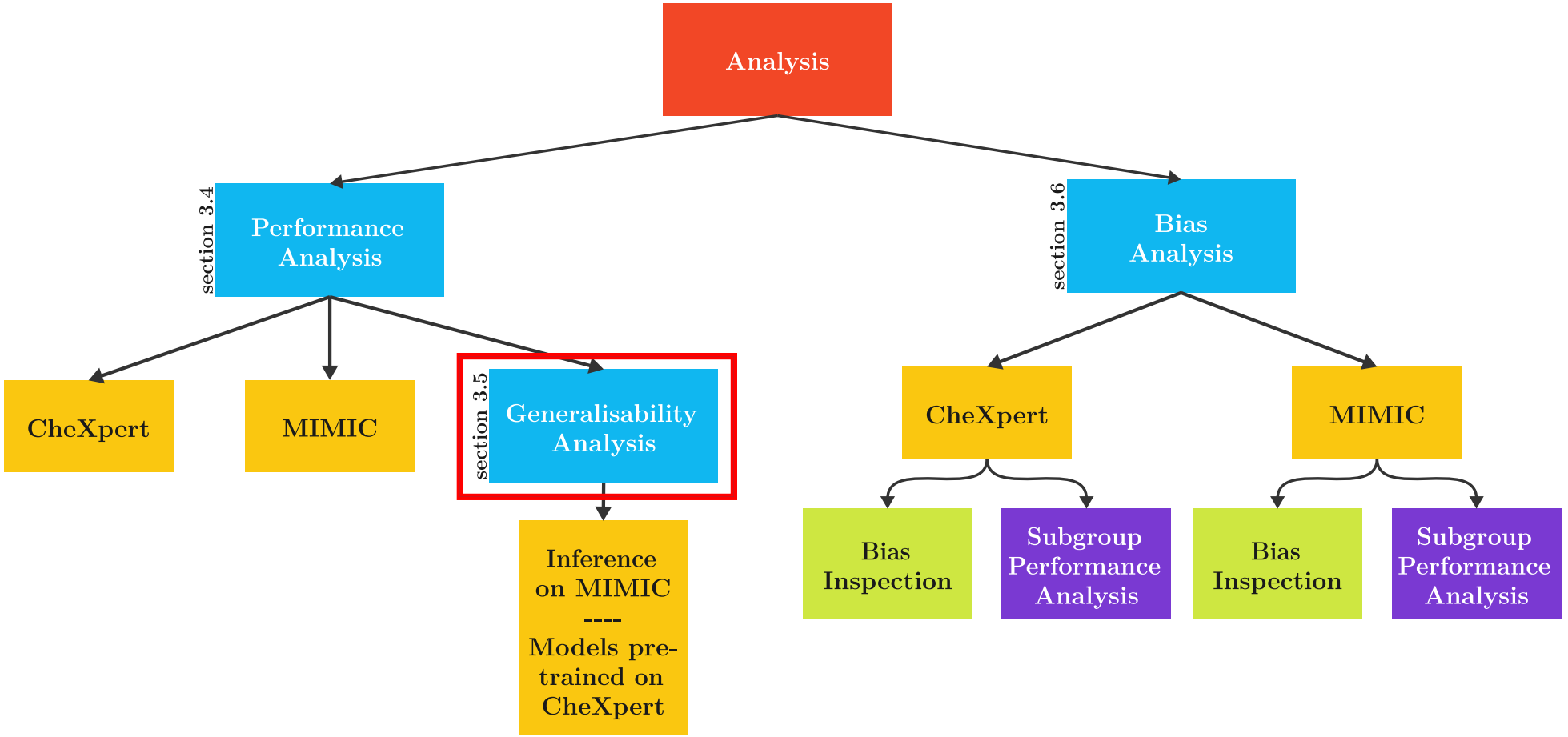
Performance Analysis

For both the CheXpert [103] and MIMIC [104] datasets, the student models achieving the best overall performances were those developed using fixed weighted combinations of Mean Squared Error (MSE) and Cosine Similarity (CS) losses during KD. Our analysis conclusively demonstrates that KD can effectively reconstruct student models that not only match but significantly outperform their teacher model. This is particularly evident with our CXR-FMKD-Direct FFT student models, which achieve the highest performance across all tested models by removing the projector used during KD to match teacher and student embeddings—echoing similar practices in the Self-Supervised Learning (SSL) literature [114, 200]—and subsequently applying comprehensive fine-tuning. These models surpass the teacher with increases up to **15.1%** in *AUC-PR*, **5.3%** in *AUC-ROC*, and **19.6%** in *Youden's J Statistic at 20% FPR* for CheXpert; and up to **12.8%** in *AUC-PR*, **4.0%** in *AUC-ROC*, and **16.1%** in *Youden's J Statistic at 20% FPR* for MIMIC. Furthermore, enriched by the teacher's knowledge, these students also outperform a traditional baseline model, CXR-Model FFT, which shares the same architecture but was trained independently without KD from CXR-FM: surpassing it by up to **3.7%** in *AUC-PR*, **1.3%** in *AUC-ROC*, and **3.6%** in *Youden's J Statistic at 20% FPR* for CheXpert; and up to **0.9%** in *AUC-PR*, **0.5%** in *AUC-ROC*, and **1.9%** in *Youden's J Statistic at 20% FPR* for MIMIC.

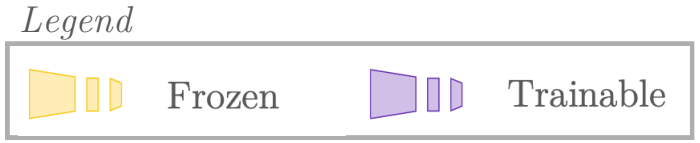
The superior performance of our CXR-FMKD-Direct FFT student models is also evidenced by their faster convergence rates, suggesting an inherited advantage from the CXR-FM teacher in swiftly adapting to downstream tasks. Compared to the CXR-Model FFT baseline, these student models achieve minimum validation loss in up to **87%** fewer epochs for CheXpert and up to **73%** fewer epochs for MIMIC. When compared to the CXR-FM teacher, they reach this benchmark in up to **79%** fewer epochs for CheXpert and up to **70%** fewer epochs for MIMIC. These results underscore the efficacy of KD as a strategy to enhance model performance by leveraging the inherent strengths and accumulated knowledge of the original CXR-FM.

Hidden Slides

3.2. Generalisability Analysis (1)



3.2. Generalisability Analysis (2)

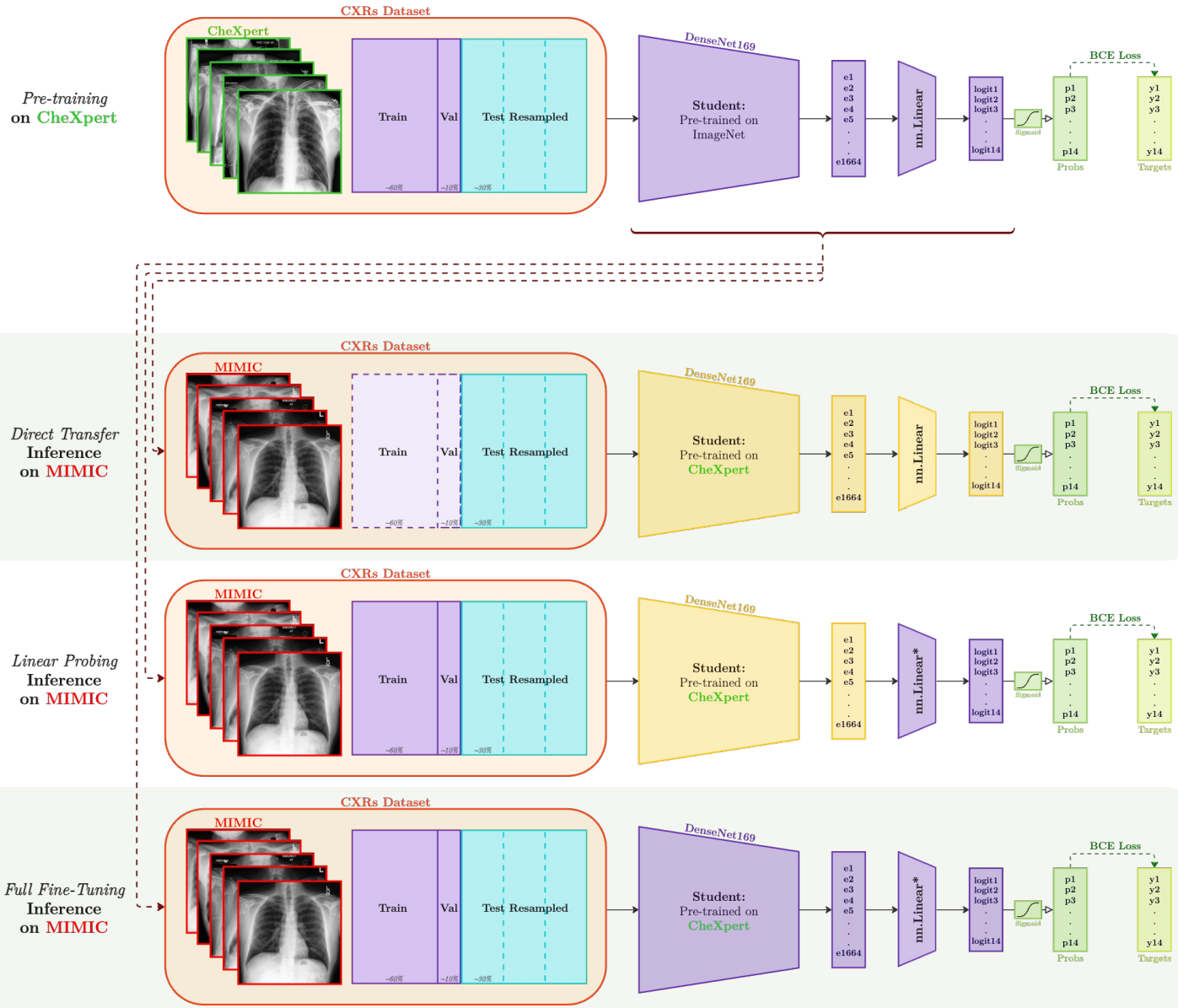


Models pre-trained on CheXpert, then evaluated on MIMIC:

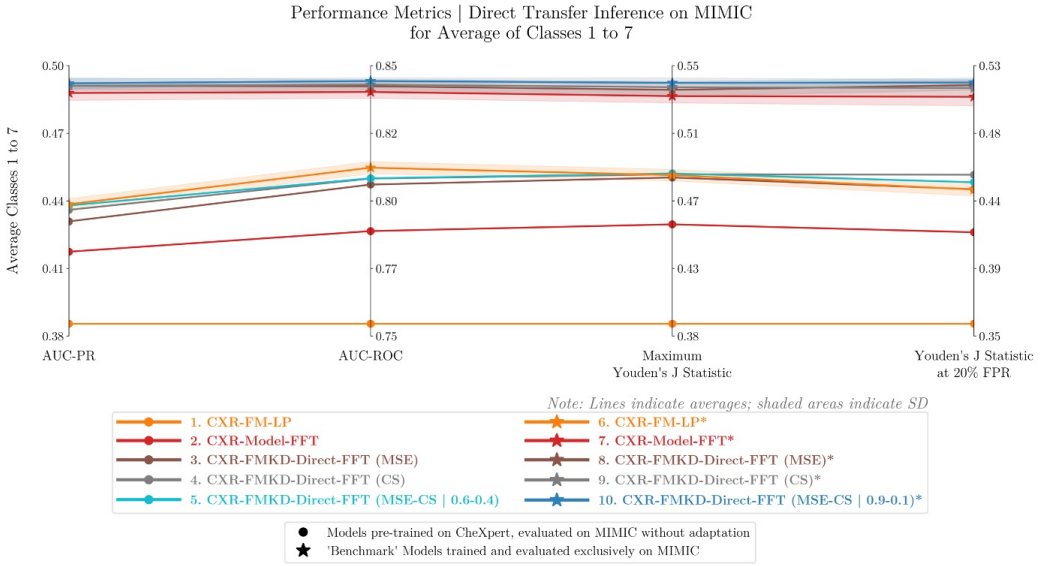
1. Direct Transfer:
without adaptation

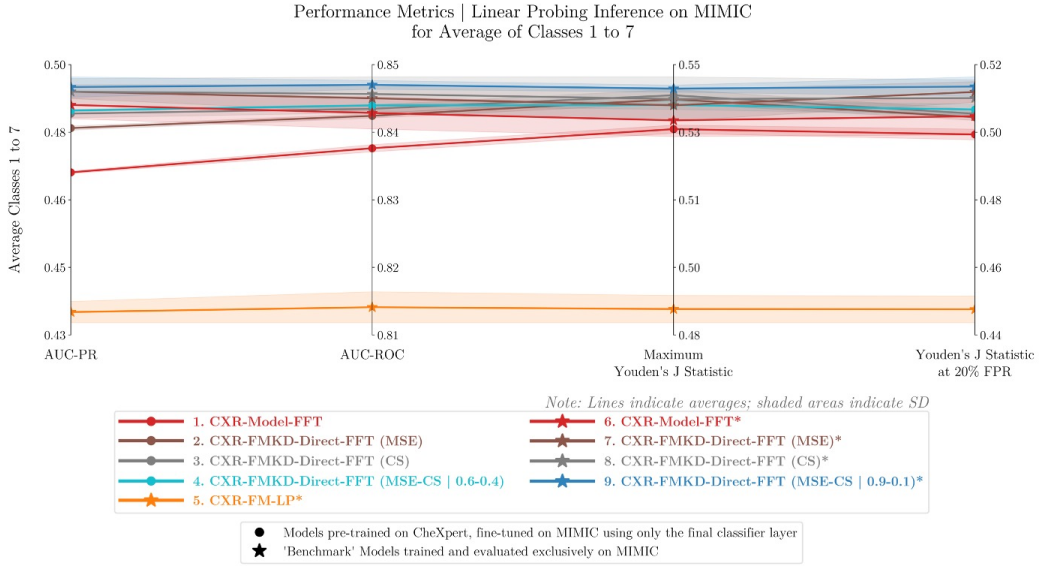
2. Linear Probing (LP):
after *fine-tuning* on MIMIC using *only the classifier*

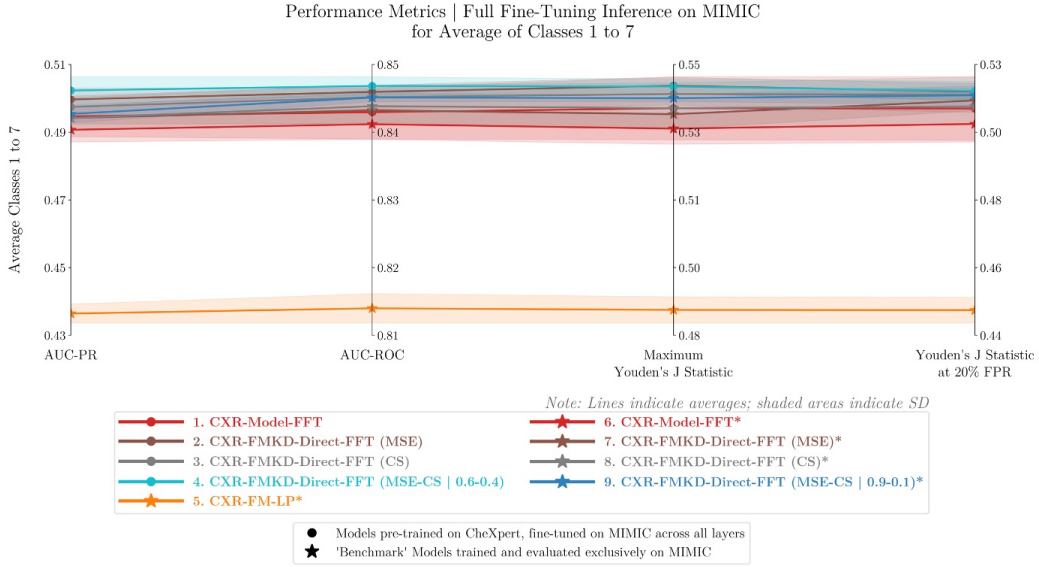
3. Full Fine-Tuning (FFT):
after *fine-tuning* on MIMIC *across all layers*



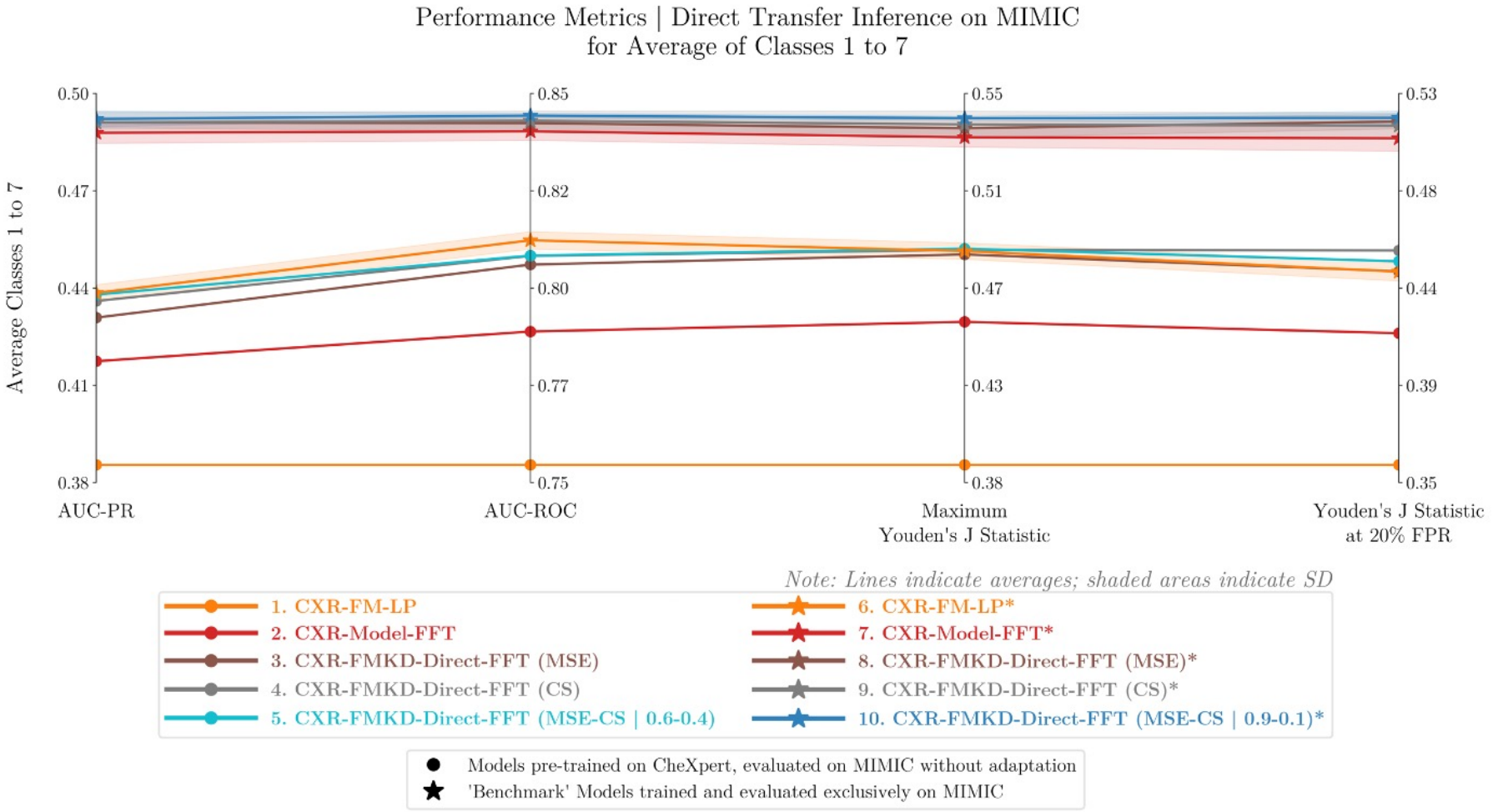
3.2. Generalisability Analysis (3)



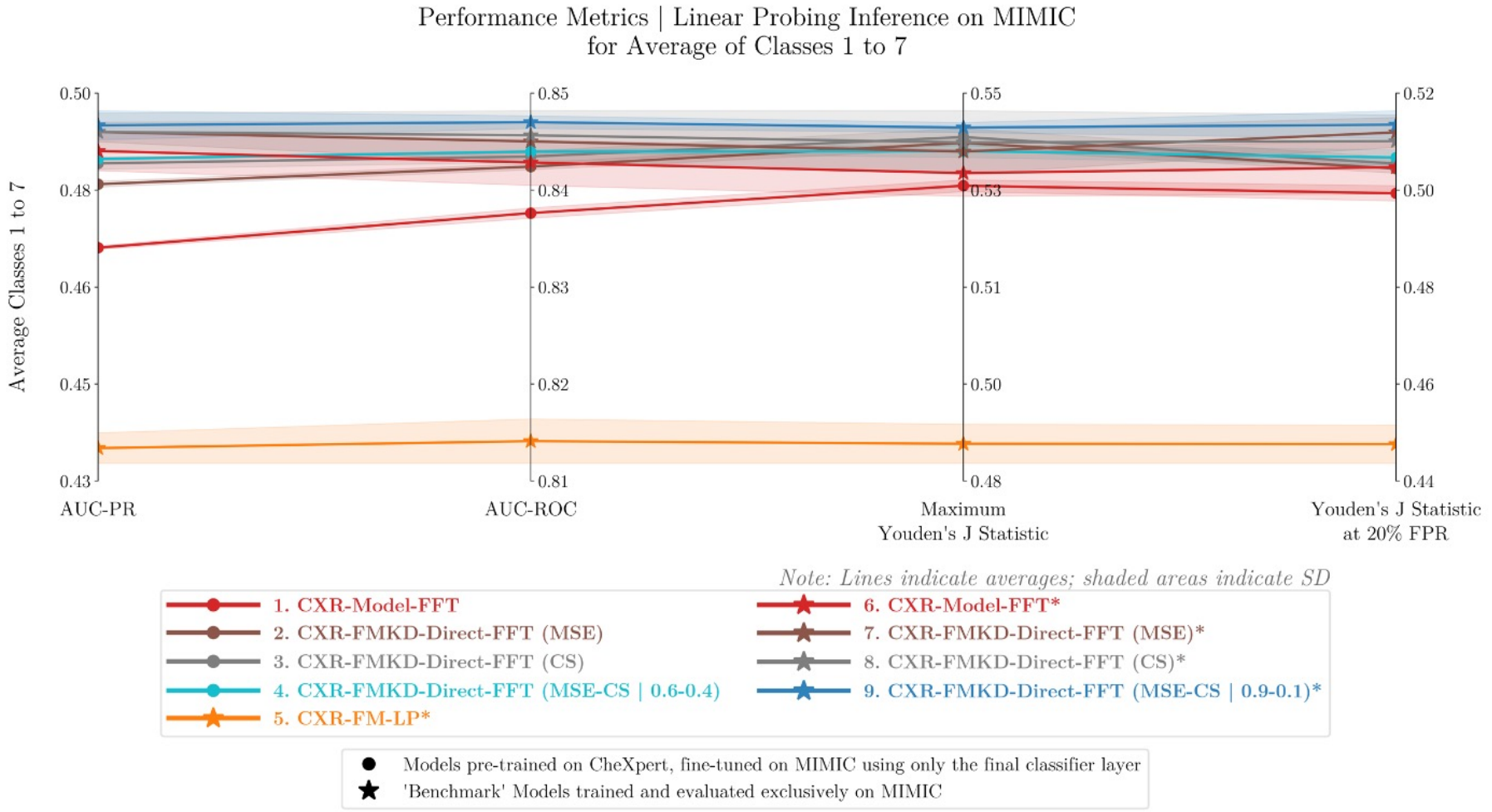




3.2. Generalisability Analysis (4)



3.2. Generalisability Analysis (5)



3.2. Generalisability Analysis (6)