IMPERIAL

Imperial College London

Department of Computing

MSc Individual Project

Foundation Models for Chest Radiography: Knowledge Distillation and Impacts on Performance and Bias Propagation

---

Author: Fadi Zahar
CID: 01552383

Supervisor: Dr Ben Glocker

Second Marker: Dr Bernhard Kainz

Submitted in partial fulfilment of the requirements for the MSc Degree in Artificial Intelligence of Imperial College London

January 2025

Abstract

Background: Medical imaging is integral to modern diagnostics, increasingly enhanced by Artificial Intelligence (AI) to improve accuracy and efficiency. Foundation Models (FMs) like Google's chest radiography FM (CXR-FM) offer promising adaptability to various downstream medical tasks without extensive labelled data. However, these models often lack transparency and can perpetuate biases, presenting challenges in subgroup performance that are difficult to address due to their typically restricted access and modification capabilities.

Objectives: This thesis investigates whether Knowledge Distillation (KD) can reconstruct Google's CXR-FM into accessible and tuneable student models (CXR-FMKD), aiming to inherit its strengths and surpass its performance in downstream tasks while enhancing transparency and capacity for bias mitigation.

Methods: Using a DenseNet169 architecture for the student models, this study applies various KD techniques and loss functions, including Mean Squared Error and Cosine Similarity, to transfer knowledge from CXR-FM to CXR-FMKD. Performance and bias were evaluated on the CheXpert and MIMIC datasets using metrics like AUC-ROC, AUC-PR, and Youden's J statistic. A novel bias score was introduced to quantitatively assess demographic biases related to biological sex and racial identity.

Results: The CXR-FMKD student models demonstrated significant performance improvements, with up to a 19.6% increase in Youden's J statistic at 20% FPR over CXR-FM, and showed robust generalisability across datasets. Bias analysis revealed reductions of up to 85% in bias scores, indicating a shift towards disease discrimination based on more clinically relevant features rather than protected characteristics.

Conclusion: The results validate KD as an effective method to enhance the transparency, tunability, and fairness of opaque FMs in medical imaging through the development of distilled, robust student models. Notably, this approach underscores the importance of open access to FM weights for comprehensive model enhancements and the development of equitable AI diagnostics.

Acknowledgements

I would like to thank my supervisor, Dr Ben Glocker, for his mentorship, patience, and support throughout this research project. Notably, for the extensive knowledge he shares, for always accommodating my impromptu office visits, and for those meetings that invariably extended beyond the clock. His guidance and encouraging presence made working on the project a very smooth experience, and his trust and the level of independence granted were invaluable.

I am also grateful to the entire team at the BioMedIA lab. Their demonstration of what real-world impact through research and excellence looks like has been nothing short of inspiring. Their amicability and willingness to assist me and the other MSc students played a key role in this research journey.

Special acknowledgement also goes to my second supervisor, Dr Bernhard Kainz, for his feedback and readiness to advise and help.

Outside of the lab, I would like to thank all the amazing people I have met in our MSc AI cohort, and to Dr Rob Craven, our MSc year coordinator, whose care and dedication elevated our postgraduate journey. His support and the enriching discussions we had throughout the year have greatly enhanced this experience.

Last but certainly not least, my heartfelt thanks go to my friends and family for their unwavering support and encouragement. Special mention goes to my parents, my brother, and my sister, who have been a constant source of motivation and strength. I am forever grateful for your love and support.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviations

AI	Artificial Intelligence
AUC-PR	Area Under the Precision-Recall Curve
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BCE	Binary Cross-Entropy
CE	Cross-Entropy Loss
CMR	Cardiovascular Magnetic Resonance
CNN	Convolutional Neural Network
CS	Cosine Similarity
CT	Computed Tomography
CVDs	Cardiovascular Diseases
CXR	Chest X-Ray
DL	Deep Learning
FFT[1]	Full Fine-Tuning
FM	Foundation Model
FPR	False Positive Rate
GFLOPS	Giga Floating Point Operations Per Second
GPU	Graphical Processing Unit
ID	In-Distribution
KD	Knowledge Distillation
KL	Kullback-Leibler Divergence Loss
KS	Kolmogorov-Smirnov
LP	Linear Probing
LR	Learning Rate
MAE	Mean Absolute Error
MCE	Multi-label Cross Entropy
ML	Machine Learning
MLP	Multilayer Perceptron
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NN	Neural Network

[1] Note: In the context of this paper, FFT refers to Full Fine-Tuning, not Fast Fourier Transform.

NLP	Natural Language Processing
OOD	Out-of-Distribution
PCA	Principal Component Analysis
PET	Positron Emission Tomography
SD	Standard Deviation
SPECT	Single-Photon Emission Computed Tomography
SupCon	Supervised Contrastive (Learning)
t-SNE	t-Distributed Stochastic Neighbour Embedding
TPR	True Positive Rate

Custom Terms

CXR-FM	Google's proprietary Chest X-Ray Foundation Model, employed as the 'teacher' model in the Knowledge Distillation process.
CXR-FMKD	Our newly developed 'student' model, based on the DenseNet169 architecture. This model is derived from the CXR-FM 'teacher' through the application of Knowledge Distillation techniques.
CXR-Model	Our baseline model with the same DenseNet169 architecture as CXR-FMKD, but trained independently without the application of Knowledge Distillation from CXR-FM.

Chapter 1

Introduction

In this introductory chapter, we explore the motivation behind this research study, highlighting the critical role of medical imaging and the integration of Artificial Intelligence in analysing medical images. We will discuss the significance of Foundation Models within this context and identify their current limitations, focusing specifically on a Foundation Model for chest radiography. This discussion naturally leads to the adoption of Knowledge Distillation as a strategic tool to address these issues, which is a central focus of this study. Subsequently, we will outline the contributions of this paper and the structure of the report. Finally, the chapter concludes with an ethical consideration regarding the data used in our experiments.

1.1. Motivation

1.1.1. Overview and Critical Role of Medical Imaging

Around one in every two individuals will develop some form of cancer throughout their life [1]. However, when diagnosed early, survival rates can drastically increase for the patient. Notably, figures from *Cancer Research UK* reveal that survival rates for the most common cancers are more than three times higher when diagnosed at stages one or two, with 10-year survival rates exceeding 80 per cent at these early stages compared to only around 25 to 5 per cent at stages three or four [2]. This critical role of early detection for life-threatening diseases underscores one of the many essential contributions of medical imaging in modern healthcare [3–5]. Indeed, such imaging technology has become indispensable in the prevention, diagnosis, prognosis (the likely course and outcome of a disease), and treatment of numerous conditions, including cancer [6–8]. By enabling detailed visualisation of organs and tissues within the body, medical imaging facilitates early and accurate identification of anomalies, guiding appropriate and timely interventions [7, 9].

The origins of medical imaging trace back to the accidental discovery of X-rays by Wilhelm Conrad Röntgen in 1895 [10–13]. This breakthrough revolutionised medicine by enabling the non-invasive visualisation of internal structures—procedures that do not require instruments to be inserted into the body. Röntgen's discovery earned him the Nobel Prize in Physics in 1901 [14]. Since then, the field has evolved dramatically, incorporating advanced technologies that offer detailed insights into the human body. As of 2010, medical imaging studies numbered in the five billions globally [15], reflecting its integral role in healthcare systems worldwide.

Exploring further, medical imaging encompasses a range of modalities that provide different perspectives on the body's internal structures. The following details some of these modalities

and their common uses [9, 16]. X-rays are widely used for detecting bone fractures, breast cancer, and dental issues, while computed tomography (CT) scans specifically uses a series of X-rays to offer cross-sectional images that are critical for identifying tumours and monitoring disease progression. Ultrasound employs high-frequency sound waves to visualise soft tissues, making it a staple in obstetrics (monitoring pregnancy) and cardiology. Magnetic resonance imaging (MRI) uses powerful magnets and radio waves to produce detailed images of soft tissues, aiding in the detection of neurological disorders, musculoskeletal problems, and tumours. Nuclear medical imaging techniques, such as positron emission tomography (PET) and single-photon emission computed tomography (SPECT), provide functional images that are used for detecting and evaluating cancers and heart diseases. Illustrations of these imaging modalities, specifically visualising the human brain, can be found in **Figure 1**.

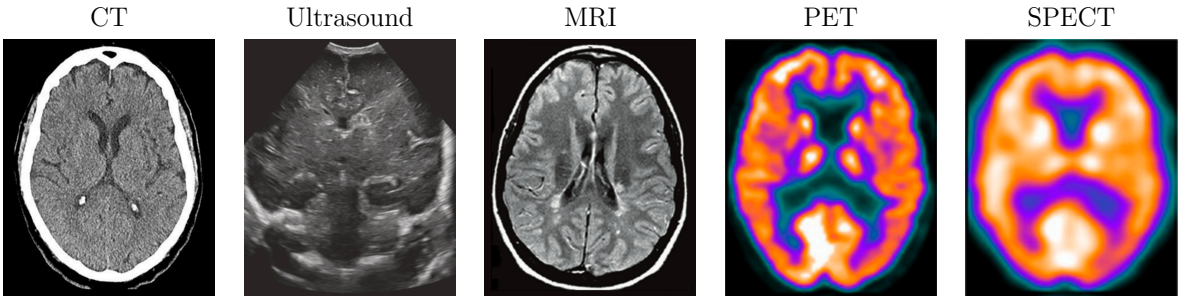

Figure 1. Comparative Visualisation of the Human Brain Using Various Medical Imaging Techniques. From left to right: Computed tomography (CT) image sourced from Figure 1 in [17]; ultrasound image sourced from Figure 3 in [18]; proton-density weighted magnetic resonance imaging (MRI) image sourced from Figure 1 in [19]; and ^{18}F-FDG positron emission tomography (PET) and single-photon emission computed tomography (SPECT) images sourced from Figure 1 in [20]. Each modality provides unique diagnostic information, highlighting different aspects of brain anatomy and function. It should be noted that the images are derived from different patients.

As discussed earlier, these imaging modalities play a crucial role in the management of serious health conditions. For instance, breast cancer, the leading cause of cancer-related death among women [21–25], can be detected early through mammography (X-ray imaging of the breast), significantly improving survival rates [26–30]. Lung cancer, the most common cause of cancer death globally [31], can be identified at an earlier, more treatable stage through low-dose CT screening or chest X-rays [32–35]. Cardiovascular diseases (CVDs), which represent the leading cause of global mortality [36, 37], benefit from imaging technologies such as cardiac CT and cardiac magnetic resonance imaging (CMR), which can detect abnormalities in the heart and blood vessels, enabling timely and potentially life-saving interventions [38–41].

Medical imaging has thus established itself as a cornerstone of modern medicine, essential in enhancing healthcare. Statistically, its impact is profound: medical images now constitute the largest portion of healthcare data, with 43.4 million imaging tests reported in England alone from February 2022 to January 2023 [42], highlighting its extensive utilisation in clinical practice. This widespread adoption has significantly improved patient outcomes by facilitating early diagnosis and treatment of diseases. Its continued advancements promise even greater improvements in disease management, driven by innovations in imaging technology and the integration of Artificial Intelligence to further refine diagnostic accuracy and efficiency, as seen in the sections below.

1.1.2. Importance of AI in Medical Imaging

The advent of Artificial Intelligence (AI), particularly through Deep Learning (DL) and Neural Networks (NNs) [43], has brought about a significant transformation in the healthcare sector, most notably in the field of medical imaging [44, 45]. Indeed, DL, a powerful subset of AI, leverages layered NNs to create models capable of analysing vast amounts of data and learning complex patterns at an unprecedented scale. This approach has revolutionised computer vision, enabling the execution of sophisticated tasks such as *image classification* [46–48], *object detection* [49, 50], and *semantic segmentation* [51, 52] with remarkable accuracy.

These advancements have naturally extended to medical imaging, where DL models excel at utilising and analysing the extensive data generated, and are now predominantly used in the field [44, 45, 53–56]. This proficiency is crucial for the effective application of computer vision tasks within the medical domain. For instance, *image classification* allows for the categorisation of scans to indicate the presence or absence of diseases, such as distinguishing between cancerous (malignant) and non-cancerous (benign) mammograms [57, 58]. Similarly, *object detection* plays a key role in identifying and localising abnormalities within an image, such as detecting and marking the presence of nodules in a chest X-ray [59, 60], which are often early signs of lung cancer, typically using bounding boxes. Moreover, *semantic segmentation* assists in the detailed mapping of organs and tissues by assigning a label to each pixel in an image [61], crucial for tasks like the automatic segmentation of the spinal canal, vertebrae, and intervertebral discs in lumbar spine MRI images [62]. *Image registration*, which aligns multiple images into a common coordinate system and is also used for tasks like motion tracking [63] and multi-modal image fusion [64], is another domain where DL has enhanced accuracy and efficiency, and significantly accelerated the process [65].

In sum, AI, particularly through DL, not only accelerates medical imaging analyses but also enhances their accuracy, often matching or surpassing the performance of human experts [44]. By processing vast datasets of medical images, DL algorithms can detect patterns and anomalies that might be overlooked by the human experts [53, 55, 66], thereby reducing misdiagnoses and ensuring timely and accurate patient care. This capability significantly advances medical imaging by automating processes, thus making them faster and more reliable.

1.1.3. Foundation Models in Medical Imaging

Given the paradigm shift it carries, AI has therefore been increasingly harnessed to assist radiologists by enhancing the processing and analysis of medical images in both clinical and research settings [67]. However, while these DL predictive models have demonstrated remarkable success, their development hinges on the availability of significant quantities of representative data. Indeed, the quality of data used for model training, ensuring it accurately represents the characteristics of the tasks at hand, is crucial for optimal performance and has been progressively viewed as more important than the sophistication of model architectures. Such data-centric approach, which prioritises high-quality data over novel algorithms, has been shown to consistently outperform model-centric strategies [68]. However, this data requirement is particularly challenging in the medical imaging field, where large datasets are often difficult to obtain due to patient privacy issues, the limited capacity of a single institution to provide sufficient data [69], and the limited availability of labelled medical images requiring expensive expert knowledge [70]. In this context, training models on relatively small datasets can lead to poor generalisation across different clinical environments. This often occurs due to distribution

shifts, where models trained on data from specific settings underperform when applied to data from different patient populations and hospitals [71–73]. Such shifts can arise from variations in patient demographics (population shifts) or the technology used in imaging (acquisition shifts). These limitations have hindered the broad adoption of these models in clinical practice [74].

Here comes Foundation Models (FMs), a term coined by the Stanford Institute for Human-Centered Artificial Intelligence (HAI) [75] in August 2021[2], offering a promising solution to these challenges [76]. These models are pretrained on extensive, heterogeneous, and diverse datasets, typically using self-supervised learning methods that do not require ground truth labels, or semi-supervised learning methods that use minimal labelled data. Through the lens of predictive tasks, which are the main focus in this study, such pretraining enables FMs to become robust feature extractors, producing outputs broadly applicable across various applications. These extracted features can then be used as inputs for subsequent training phases tailored to specific downstream tasks, such as classification or detection, but now requiring only relatively small amounts of data [77, 78].

The success of such FMs relies heavily on the concept of **transfer learning**, which involves applying knowledge gained from one task to another [75, 79]. This process capitalises on the knowledge the model has acquired from its initial training to boost its performance on new tasks, especially when labelled data is scarce. To build on the previous discussion, in DL, *pretraining* serves as a foundational method for transfer learning: a model is initially trained on a large 'surrogate' task, primarily as a stepping stone, and then *fine-tuning* is used to tailor the model to a targeted downstream application. The effectiveness of FMs is greatly enhanced by the scale of their pretraining, which is made possible by advances in computing hardware with Graphical Processing Units (GPUs), sophisticated model architectures such as Transformers that utilise hardware parallelism [80], and the availability of vast amounts of unlabelled data, which is leveraged through self-supervised and semi-supervised learning techniques as previously mentioned.

In the domain of medical imaging, FMs are therefore particularly advantageous given the challenges in assembling large, high-quality datasets. Notably, recent studies have demonstrated that self-supervised pretraining on massive, unlabelled datasets of medical images—some exceeding 100 million images—can significantly boost the models' effectiveness across various downstream tasks [81, 82]. The resulting FMs have indeed shown superior generalisation and enhanced performance even on novel, out-of-distribution (OOD) datasets containing data from sources different than those seen during training.

Chest radiography, the process of obtaining chest X-rays (CXRs), exemplifies the impact of these models, being the most frequently conducted radiological examination worldwide. In fact, statistics from industrialised countries indicate an average of 238 erect-view CXR exams per 1,000 people each year according to a 2008 report from the United Nations [83]. This prevalent use of CXRs can be linked to their affordability and minimal radiation exposure, often serving as the initial screening, diagnosis, and management of a wide array of health conditions [84–87]. **Figure 2** provides examples of common thoracic diseases detected in CXRs, while also underscoring the difficulty of automating their accurate diagnosis due to the often subtle and varied appearance of these pathologies.

[2] Also marking the establishment of Stanford's interdisciplinary Center for Research on Foundation Models (CRFM) dedicated to fundamentally advancing the research and development of FMs.

Utilising a method known as supervised contrastive (SupCon) learning [88], Google Health's proprietary CXR Foundation Model, hereafter referred to as **CXR-FM**, was developed from 821,544 chest radiographs sourced from India and the United States [89]. This model served as a base for further adaptation and was fine-tuned using smaller datasets for a range of diagnostic tasks, including detecting conditions such as tuberculosis, airspace opacity, fractures, and COVID-19 outcomes. Notably, this method demonstrated a significant reduction in the amount of labelled training data required for downstream task fine-tuning—up to a 688-fold decrease—while also improving the performance. For instance, with as few as 45 CXRs used for downstream task training, an area under the receiver operating characteristic curve (AUC) of 0.95 was achieved for detecting tuberculosis, a performance shown to be noninferior to that of radiologists. This work underscores the potential of FMs in optimising diagnostic accuracy and efficiency in medical imaging, particularly in contexts with dynamically changing patient populations and data distribution shifts.

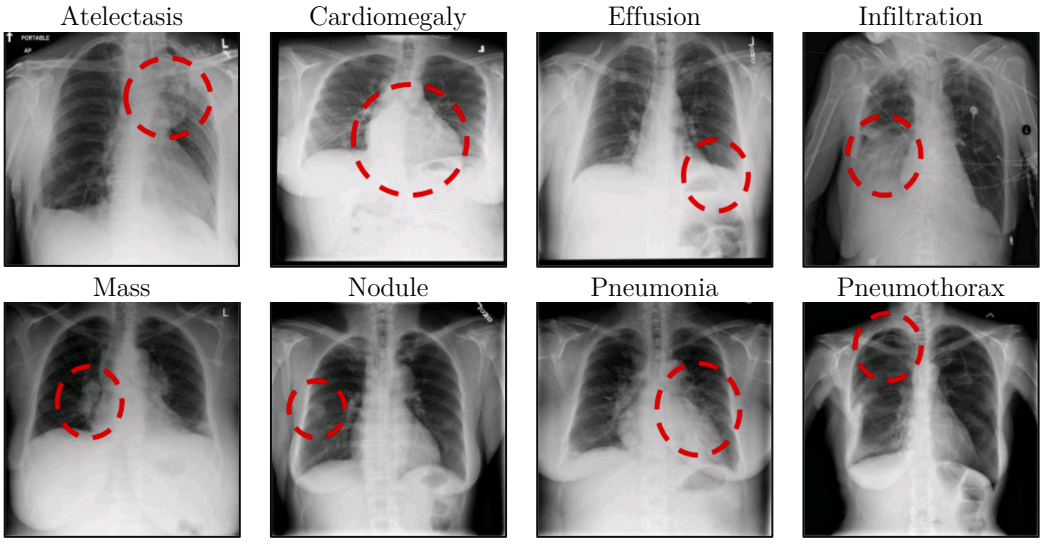

Figure 2. Examples of Common Thoracic Diseases Observed in Chest X-Rays.
Adapted from [90], which presented the development of the 'ChestX-ray8' database. This figure illustrates eight chest X-rays (CXRs), each depicting a different thoracic condition localised by a dotted red circle. The displayed conditions include atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, and pneumothorax. The variability and subtlety of these conditions exemplify the diagnostic challenges in automated CXRs analysis.

1.1.4. Current Limitations of Foundation Models

In their discussion on FMs, Bommasani et al. [75] describe *homogenisation* as the unification of methods used to develop machine learning (ML) systems for a broad spectrum of applications. While this unified approach offers substantial advantages by enabling models to handle diverse tasks efficiently, it also introduces considerable risks. A significant concern is the propagation of any inherent defects from the FM to all derivative models—those adapted for specific downstream tasks. Such issues not only create singular points of failure but also amplify the effects of these defects across multiple applications.

Transitioning to the medical field, the challenges outlined above become particularly critical and hamper the widespread clinical implementation of FMs. A major concern is that the vast scale of these models can propagate encoded biases through downstream tasks, with defects present in a FM being 'inherited' by all adapted models [75] This inherent bias, combined with a general

lack of profound understanding of how these models function and their impacts, as well as the processes by which biases are encoded, propagated, and reinforced, presents important ethical and regulatory challenges for their use in healthcare [91], and particularly in radiology [92].

A critical example of this issue is highlighted in a study by Glocker et al. [93], which demonstrated biases related to race and biological sex in the previously introduced Google's CXR-FM [89]. These biases resulted in significant performance disparities across different patient subgroups. Specifically, the model's performance in classifying the 'no finding' label dropped by approximately 6.8% to 7.8% for female patients, while the performance in classifying the 'pleural effusion' label decreased by about 10.7% to 11.6% for Black patients. These discrepancies raise substantial safety and ethical concerns regarding the clinical deployment of such models, potentially leading to the underdiagnosis of underserved populations [94]. Furthermore, the shown ability of DL models to inadvertently identify protected characteristics such as race and other demographics [95, 96] compounds these issues, necessitating a comprehensive analysis of biases in FMs used in medical imaging.

This situation is exacerbated by the opaque nature of FMs. Access to the CXR-FM, for example, was traditionally provided through a programming platform that processes input images and outputs corresponding extracted features. It is important to note that this interface does not make the network weights publicly available, preventing updates to the feature extractor parameters during training for downstream tasks. More generally, these FMs often rely heavily on API calls for feature extraction, which obscures the internal workings that could help provide insights into bias sources. This lack of transparency, including the training strategies and history that typically involves some private datasets, is compounded by the competitive landscape of AI development where sharing proprietary model insights is not incentivised, further hindering the ability to scrutinise and improve these models [75]. It should be noted that only recently, on November 25th, 2024, have the weights for CXR-FM been made public [97], setting an important precedent and further highlighting the need for open FMs.

1.1.5. Need for Knowledge Distillation

To address these challenges, this study will explore Knowledge Distillation (KD) as a means to transfer the learned knowledge from a large, complex model—the FM *teacher*, in this case, the CXR-FM studied in [89]—to a typically smaller, more manageable model—the *student*. By creating a distilled student model, we gain full access to its internal workings, parameters, and training strategies, which is crucial for exploring effective bias mitigation techniques and enabling further fine-tuning; effectively providing better control and tunability [98]. This approach allows for a deeper understanding of how any potential biases are encoded in the student model, helping to prevent the introduction of unwanted correlations and ultimately reducing bias-related performance disparities [99, 100].

From a general training perspective, adapting FMs for specific tasks typically begins with the FM serving as a backbone that extracts numerical features from input data, such as images. These output features are then processed through either linear or non-linear combinations (e.g., NN configurations) to yield task-specific outputs, such as probability scores for each class in image classification tasks. When fine-tuning these FMs to create a task-specific model, there are two main strategies: the first involves unfreezing the backbone, allowing modifications to the input-to-feature mapping which potentially overhauls the entire network's learning patterns; the second, more data-efficient method, involves freezing the backbone and solely learning and

updating the task-specific parameters. While the latter approach is often more practical due to its lower data requirements, it risks perpetuating any potential biases embedded in the original FM because the core mechanism for feature extraction does not change. This approach, necessitated by the closed nature of the initial CXR-FM model with locked backbone, could inhibit effective debiasing. Fine-tuning in this restricted framework may depend on and even reinforce biased heuristics present in the data, as the model might utilise simplified correlations or 'shortcuts' to make predictions, thereby magnifying existing disparities among different subgroups [101, 102]. As argued by [93], the performance differences related to bias among subgroups can thus be partly attributed to the fact that the CXR-FM was locked during task-specific training.

Employing KD from the CXR-FM to get a distilled student therefore allows us to circumvent the restriction of being unable to update the feature extraction mechanism during task-specific fine-tuning. This enhances our ability to apply debiasing techniques and improve both model transparency and performance. However, it is important to recognise that KD involves several challenges in effectively transferring knowledge from the larger teacher model to the smaller student model. These challenges include selecting an appropriate student model architecture that can sufficiently capture the complexity of the teacher while operating under constraints such as reduced size or computational demand. Additionally, choosing the right KD loss function to guide the student's training and carefully considering the selection of the transfer dataset used for KD are critical for ensuring effective knowledge transfer.

Ultimately, KD provides a framework in this context for creating robust, transparent, and fair models that are better suited for clinical applications, ensuring that potential biases are identified and corrected, and that the models are tuned to perform optimally across diverse patient populations and subgroups.

1.2. Contributions

This thesis aims to address the limitations and biases associated with FMs in medical imaging, with a particular focus on chest radiography and Google's proprietary CXR-FM [89]. This effort effectively builds upon the work conducted by Glocker et al. [93], which identified biases within CXR-FM. By leveraging KD, this research seeks to develop a robust distilled student model, named **CXR-FMKD**, that not only inherits the strengths of the original CXR-FM—aiming to match or exceed its performance on downstream CXR disease detection tasks—but also enhances transparency, tunability, and ultimately the capacity for bias mitigation.

While the weights of CXR-FM have recently been made public [97], as previously mentioned, our overarching goal remains to address the transparency issues that plague many FMs. Specifically, our research aims to investigate whether KD can serve as a viable strategy to reconstruct FMs (while maintaining its strengths), thus potentially enabling comprehensive mitigation of biases and performance disparities across patient subgroups.

It is also important to clarify that the main objective of this study is not merely to optimise performance on a specific task but to explore the broader implications of KD in enhancing the fairness and transparency of FMs.

The key contributions of this work include:

1. **KD Exploration:** We investigated various KD techniques and loss functions to effectively transfer knowledge from the CXR-FM teacher and create our CXR-FMKD student model.
2. **Performance Analysis:** We evaluated the performance of the CXR-FMKD model using the CheXpert [103] and MIMIC [104] CXR datasets, which share the same 14 disease classes in a multilabel classification setting. Comparisons were made against a baseline model, **CXR-Model**, which shares the same architecture as CXR-FMKD but was trained independently without KD from CXR-FM, and with the original CXR-FM to benchmark improvements.
3. **Generalisability Analysis:** We assessed the generalisability of the models by training initially on CheXpert and subsequently testing on MIMIC, leveraging the shared 14 disease classifications. Various testing strategies were employed, including *Direct Transfer* without adaptation, *Linear Probing* by fine-tuning only the final classifier layer, and *Full Fine-Tuning* across all layers. Performance metrics were compared with those of CXR-Model and CXR-FM to determine relative improvements.
4. **Bias Analysis:** We adopted and adapted methods similar to those used in the study by Glocker et al. [93] to investigate biases within our models, focusing on demographic factors such as sex and race. Additionally, we developed a novel bias quantification score to systematically assess biases and compare the bias performance of the different models.
5. **Performance vs. Bias Analysis:** We analysed relationships between the models' performance and their biases, drawing conclusions on how biases impact model performance and reliability.

The findings demonstrate that our distilled CXR-FMKD model not only enhances performance over the original CXR-FM model but also reduces biases, as discussed in later chapters.

1.3. Report Structure

This report adheres to the typical structure of academic papers in the field, with a specific design choice to merge the discussion section within the results (in Chapter 4) for direct interpretation of the different contributions and experimental results. **Chapter 2**, *Background and Related Work*, provides an extensive overview of FMs and KD, setting the theoretical and practical groundwork for the methodologies used. **Chapter 3**, *Methodology*, details the experimental design and methods employed to develop and evaluate the CXR-FMKD model, emphasizing the KD techniques and strategies for performance and bias analysis highlighted in the *Contributions* section above. **Chapter 4**, *Results*, presents and discusses the findings from the performance and generalisability analysis of CXR-FMKD compared to baseline models and the original CXR-FM, including an in-depth look at bias analysis and its impact on model performance. The final **Chapter 5**, *Conclusion and Future Work*, summarizes the key findings, discusses their broader implications for medical imaging, and suggests potential future research directions.

1.4. Ethical Considerations

Ethical approval was not required for this research study as it solely utilised secondary data from the CheXpert [103] and MIMIC [104] CXR datasets, which are publicly accessible and do not necessitate permission for their use.

Chapter 2

Background and Related Work

This chapter provides a detailed overview of FMs and KD, setting the theoretical and practical groundwork that is most relevant for this research project and especially for the methodologies used in building up the models and experiments.

2.1. Foundation Models in Medical Imaging

2.1.1. Overview and Evolution

Definition and Origins

Reviewing and building on the discussions in **section 1.1**, Foundation Models (FMs) represent a significant paradigm shift in AI, characterised by their ability to be pre-trained on extensive and diverse datasets and then adapted to a wide range of downstream tasks through fine-tuning [75]. The term 'Foundation Models' was initially picked to highlight their foundational yet incomplete nature, allowing for a broad adaptability across different downstream applications. Initially, DL models in the early 2010s focused on specific tasks like image classification or language translation. However, with advancements in computational power, model architectures, and the availability of large-scale datasets, these models evolved into sophisticated systems capable of handling a broad spectrum of tasks, taking the form of robust backbones (i.e., FMs) to be used as a source of applicable 'pretrained' knowledge. Notably, current FM examples such as BERT [105] and GPT-3 [106] for Natural Language Processing (NLP), and CLIP [107] for computer vision exemplify this evolution, showcasing how a single model can achieve impressive performance across various domains.

Characteristics and Capabilities

FMs possess several key characteristics that contribute to their versatility and effectiveness, particularly in medical imaging [78]. In fact, FMs are highly scalable making them capable to digest vast amounts of data while enabling the mining of complex patterns and features that are usually missed by smaller and thinner models. Additionally, FMs are adaptable, allowing FMs to be fine-tuned for specific tasks with relatively small amounts of task-specific data, especially in medical imaging, where labelled data can be scarce and expensive, also exacerbated by distribution shifts [108]. The robust pre-training phase of FMs is responsible for the general applicability of FMs across different domains, as they learn rich representations from heterogeneous and typically unlabelled data sources.

Broader Medical Impact

FMs have revolutionised various fields by offering a 'generalist' tool, adaptable for specific tasks with minimal additional training. For instance, Google's CXR-FM which was trained on large-scale CXR datasets can identify conditions like tuberculosis and COVID-19 with high precision [89] after task-specific fine-tuning with as low as 45 CXRs in some cases. Another niche case study includes the detection of health conditions in retinal images by training a FM to learn generalisable representations from unlabelled retinal images [109]. The adapted model consistently outperformed several baseline models in the prognosis and diagnosis of eye diseases and the prediction of systemic disorders (such as heart failure and myocardial infarction). Other efforts in the field of medical imaging comprises the reconstruction of a large-scale dataset of CXRs to promote the iterative development of a vision-language FM (CheXagent) which achieved distinguished performance across eight different CXR tasks [110].

Beyond medical imaging, FMs serve as a cornerstone for Generalist Medical AI (GMAI) [76], the main purpose being to facilitate advancements in disease prediction, patient monitoring, and personalised treatment plans (precision medicine [111]) as well as multi-modal understanding.

2.1.2. Types and Modalities of Foundation Models

Here, various categories of FMs used in computer vision exists, separated between *Traditional Models* and *Prompted Models*. The latter is further divided into *Textually Prompted Models*, *Visually Prompted Models*, and *Heterogeneous Models*. Traditional FMs for computer vision typically only deal with images as input for both pretraining—using self-supervised or semi-supervised learning—and for task-specific fine-tuning. In *Textually Prompted Models*, such as CLIP [107], task adaptability is enhanced by combining visual inputs with textual prompts that define the desired classifications. In *Visually Prompted Models* like SAM [112], this concept is extended by the inclusion of visual prompts—such as points, bounding boxes, or masks—to guide segmentation tasks more specifically. Furthermore, *Heterogeneous Models* are characterised by multimodal inputs, including audio and video, expanding their applicability across varied data. Comprehensive discussions with comparisons and examples of these models are provided in the survey papers [78, 113].

2.1.3. Theoretical Underpinnings and Learning Paradigms

Pre-training Techniques

As previously mentioned, FMs rely on pretraining techniques to learn robust representations from large-scale unlabelled data. Self-supervised learning strategies, such as SimCLR [114], BYOL (Bootstrap Your Own Latent) [115], and MAE (Masked Autoencoders) [116], are core to this process. These strategies typically fall into two main categories: *contrastive* and *generative* objectives. Contrastive methods, like SimCLR, aim to bring related data points closer in the feature space while pushing unrelated ones apart. This approach helps the model learn distinctive and discriminative features. Generative techniques, such as MAE, involve reconstructing missing parts of the input data, which allows the model to understand and define complex structures within the data. Hybrid approaches, such as DINO [117], combine elements of both contrastive and generative methods to leverage the strengths of each, enhancing the model's pre-training efficiency and effectiveness.

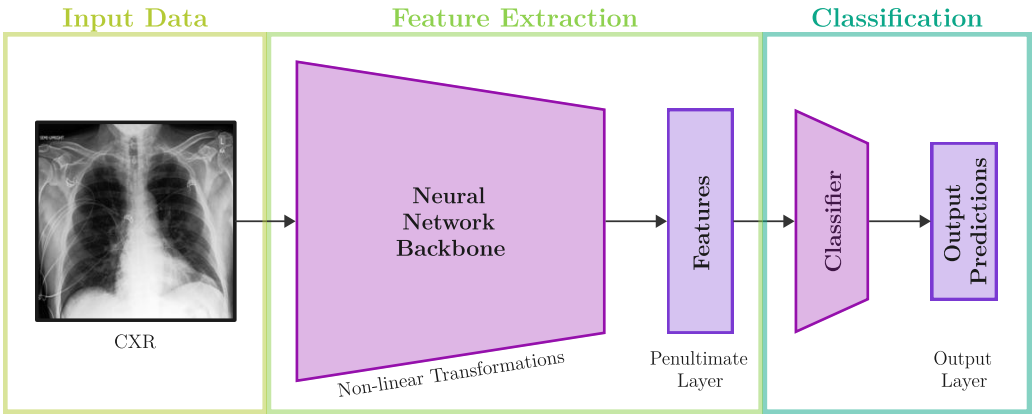

Figure 3. Schematic Representation of Typical Deep Neural Network.
The figure showcases the input data (a CXR in this case), feature extraction, and classification components of the Deep Neural Network. The feature extraction part is composed of a non-linear transformations section (the backbone) which extract the feature representations for specific inputs in the penultimate layer. The last classification layer processes the features extracted, using the information they contain to make the final predictions.

Representation Learning

Representation learning is a fundamental aspect of ML that involves the identification and extraction of meaningful patterns from raw input data. This process is crucial for transforming complex data into a structured format—structured and understandable representations—that simplifies subsequent ML tasks such as clustering, classification, and retrieval [118]. In the context of deep NNs, as illustrated in **Figure 3**, representation learning functions as the feature extraction stage, where it produces information-rich features that are crucial for understanding the data. These features, technically outputted by the backbone, are then used by the classifier—i.e., final classification layer(s)—to make the final predictions in a classification task for example. These learned features effectively serve as a foundational component for more advanced learning strategies, including transfer learning.

Transfer Learning

Transfer learning builds on the concept of representation learning by utilising the features extracted to adapt models efficiently to new tasks. This method involves pre-training a model on a large and diverse dataset, or a specific task, to capture generalisable knowledge in the features, which is then fine-tuned to perform related but distinct downstream tasks [119]. The initial phase of representation learning is pivotal in this process, as it allows the model to effectively abstract and retain useful features that can be repurposed across various applications. By leveraging these pre-learned representations, transfer learning enables models to quickly adapt to new tasks with significantly less data, circumventing the need for extensive data collection and labelling. This approach is enhanced by meta-learning techniques, where the model optimises its ability to learn new tasks rapidly, further reducing the reliance on large labelled datasets and facilitating efficient model adaptation in resource-constrained scenarios—where features extracted from initial inputs already provide useful information even before any task-specific fine-tuning.

2.1.4. Downstream Task Adaptation

Fine-tuning Techniques

Adapting pre-trained FMs to specific medical imaging tasks involves several fine-tuning techniques that modify the models to better fit the new data [75]:

Fine-tuning is a widely used technique in transfer learning, where the parameters of a pre-trained FM are adjusted using task-specific data. This method leverages the knowledge nurtured in the model from its extensive pre-training phase and refines it to improve performance on the target downstream task [120]. Fine-tuning is categorised into two main strategies: *comprehensive fine-tuning*, which updates all model parameters, and *targeted fine-tuning*, which selectively modify certain layers or parameters, and include both linear and non-linear approaches.

Comprehensive Fine-tuning: This strategy involves adjusting all layers of the pre-trained FM to improve performance on the downstream task. Contrary to the more limited *targeted* approaches, *comprehensive* fine-tuning is most effective with a substantial amount of labelled data, as it seeks to refine the entire model's input-to-feature mapping—i.e., its understanding—to specific target tasks, thereby requiring more examples to prevent overfitting and ensure robust generalisation [121].

Targeted Fine-tuning (Linear and Non-linear Probing): In this approach, adjustments are more focused, involving only certain layers such as the final classification layer, as seen in **Figure 3**. This approach is particularly useful when the available labelled data for the specific task is limited and the pre-trained model already has a significant understanding of general input image features:

- **Linear Probing:** Only the final layer of the model is trained, keeping the rest of the network parameters fixed, making it computationally efficient.
- **Non-linear Probing:** This involves training additional layers or modules, such as Multilayer Perceptrons (MLPs), added to the pre-trained model (attached to its backbone), which allow for more nuanced adjustments. However, this comes at the cost of increased computational resources and greater data requirements.

Challenges in Adaptation

One of the main issues in adapting FMs for medical imaging is the heterogeneity and multimodality of the data. Indeed, multiple medical imaging techniques exist (e.g., MRI, CT, ultrasound, X-rays), each exhibiting unique data characteristics which also vary across clinical centres [122]. This variability can lead to a model that is well trained on one data format but poorly on another [123]. To mitigate this first challenge, extensive heterogenous training datasets are essential. Historical methods also entail training different models per database or using a shared encoder with modality-specific decoders [123]. Additionally, techniques such as data augmentation and domain adaptation can help make the models more robust to different types of data [75].

Another critical downside is overfitting which occurs when a model is fine-tuned using a small sample size, where the model ends up learning the training data too well, leading to high variability and poor performance on unseen data [124]. Instead of only capturing the meaningful underlying pattern in the data, the model will tend to overfit on noise. Several techniques are used to reduce overfitting and allow the model to make more generalised predictions. These

include regularisation techniques that penalises larger weights (such as L1 and L2 regularisation), dropout, data augmentation, and early stopping [125].

As discussed previously, fine-tuning large FMs demands excessive computational and memory resources, especially problematic for smaller institutions with limited funds. Gradient checkpointing is exploited as a memory optimisation strategy [126], reducing memory usage during backpropagation through the storage of certain activations and recomputing others as needed. In terms of computational resource, specialised hardware accelerators like Field-Programmable Gate Arrays (FPGAs) can accelerate the data processing for large models [127]. Finally, algorithmic optimisation methods, equally known as low-storage adaptation methods, can accelerate the training process. For instance, algorithms like Parameter-Efficient Fine-Tuning (PEFT) [121] freezes the majority of the model's parameters, fine-tuning only a small subset of the weights. Similarly, Low-Rank Adaptation, or LoRA, freezes the pre-trained network parameters and introduces trainable rank-decomposition matrices into the network [128]which facilitates the adaptation of large ML models. In this regard, both PEFT and LoRA greatly reduce the number of trainable parameters for downstream tasks.

Lastly, fine-tuning relies on the careful selection of hyperparameters, ultimately dictating the model's performance. Key hyperparameters such as learning rate, batch size, and the extent of parameter freezing need to be optimised for each specific task. Automated hyperparameter tuning methods using Bayesian optimisation [129] and systematic approaches like grid search or random search can help identify the optimal settings.

2.1.5. Architectural Backbones

DL methods dominate large areas of medical imaging analysis. At the forefront are convolutional neural networks (CNNs) such as ResNets [48] or DenseNets [47] for classification tasks or the famous U-Net for segmentation tasks [130]. Vision Transformers (ViTs) [131] have also taken the computer vision field by storm, increasingly replacing CNNs in various medical imaging tasks [132]. It naturally follows that these architectures have been instrumental in developing the FMs. This section therefore provides a general, historical view on ML model architectures and the corresponding status-quo.

Convolutional Neural Networks

Convolutional Neural Networks (CNNs), designed primarily for image processing, have been foundational in advancing image analysis. From their inception in the 1980s, marked by the introduction of the LeNet architecture for digit recognition by [133], CNNs have evolved significantly. A pivotal moment was the success of AlexNet in 2012 [134], which dramatically reduced error rates in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [135]—a standard benchmark for large-scale object recognition (encompassing image classification and object detection) based on the ImageNet dataset [136] which comprises millions of labelled images across thousands of categories, including a diverse array of objects, scenes, and animals, in various versions. AlexNet's architecture featured deep layers, ReLU activations, and dropout to mitigate overfitting, setting a new standard in image classification and object detection. Subsequent architectures like VGGNet (2015) [137] emphasized the importance of network depth, utilising numerous weight layers and small 3x3 convolution filters. Inception (GoogLeNet) [138] introduced inception modules that facilitated deep networks without a proportional increase in parameters. ResNet (2016) [48], with its residual connections, enabled the training of exceptionally deep networks, reaching 152 layers and setting a new benchmark for performance in the ILSVRC 2015. DenseNets (2017) [47], emerging shortly after ResNet, introduced a novel

architecture that significantly differs from its predecessors by connecting each layer to every other layer in a feed-forward fashion. This unique setup encourages feature reuse throughout the network, making it very efficient both computationally and in terms of parameter usage. DenseNet's ability to alleviate the vanishing-gradient problem, strengthen feature propagation, and substantially reduce the number of parameters without compromising the depth or performance makes it particularly advantageous for complex tasks in medical imaging where preserving feature information throughout the layers is crucial.

In this context, CNNs have a rich history in medical imaging applications, beginning in the 1990s [139–141] and achieving prominence with the introduction of U-Net for biomedical image segmentation, which won the 2015 ISBI cell tracking challenge [142]. U-Net's architecture, designed for segmentation tasks, remains widely used, with its variants continuing to achieve outstanding performance in various medical imaging applications [143–146].

Vision Transformers

Transformers [80], initially developed for NLP, have significantly impacted computer vision with the advent of Vision Transformers (ViTs) [131]. ViTs partition images into sequences of patches, processing them akin to words in a sentence. While there is no definitive evidence that ViTs surpass CNNs in medical image analysis, they offer distinct advantages [132]. ViTs generally provide larger effective receptive fields, potentially allowing for better contextual understanding, crucial for medical diagnoses that require a holistic view of the image. However, ViTs are computationally intensive and demand large datasets.

From an intuitive perspective, CNNs impose a stronger inductive prior than ViTs by utilising convolution operations that effectively leverage spatial hierarchies through shared weights across pixels. This approach is particularly adept at extracting local features and reducing parameter count, making CNNs popular for detailed medical imaging analysis. In contrast, ViTs, which process images in sequences of patches similar to textual data, often require larger datasets and substantial computational power, which may pose challenges in medical settings where data privacy and availability are important concerns.

Fusion Architectures

In addition to pure CNN and ViT architectures, fusion architectures have emerged, combining elements of both to enhance performance. These architectures typically use CNNs to extract local features and ViTs to capture global dependencies. For instance, models like the TransUNet [147] integrate transformers into the U-Net framework, enhancing its ability to capture long-range dependencies and contextual information while maintaining the efficiency of CNN-based feature extraction.

2.1.6. Foundation Model in Chest Radiography

Our primary investigative model in this study is Google's proprietary CXR-FM [89], used for CXR tasks. This FM underwent a two-step pre-training process: initially, it was trained on a large and diverse dataset of natural images to develop a broad visual understanding, followed by a targeted CXR pre-training on 821,544 de-identified CXRs sourced from both India and the United States. This second phase employed a method known as supervised contrastive (SupCon) learning [88], which brings together image representations with identical labels, while distancing those with different labels. This SupCon learning represents a somewhat unconventional pre-training approach for a FM, as it leverages annotated data, contrasting with the more common

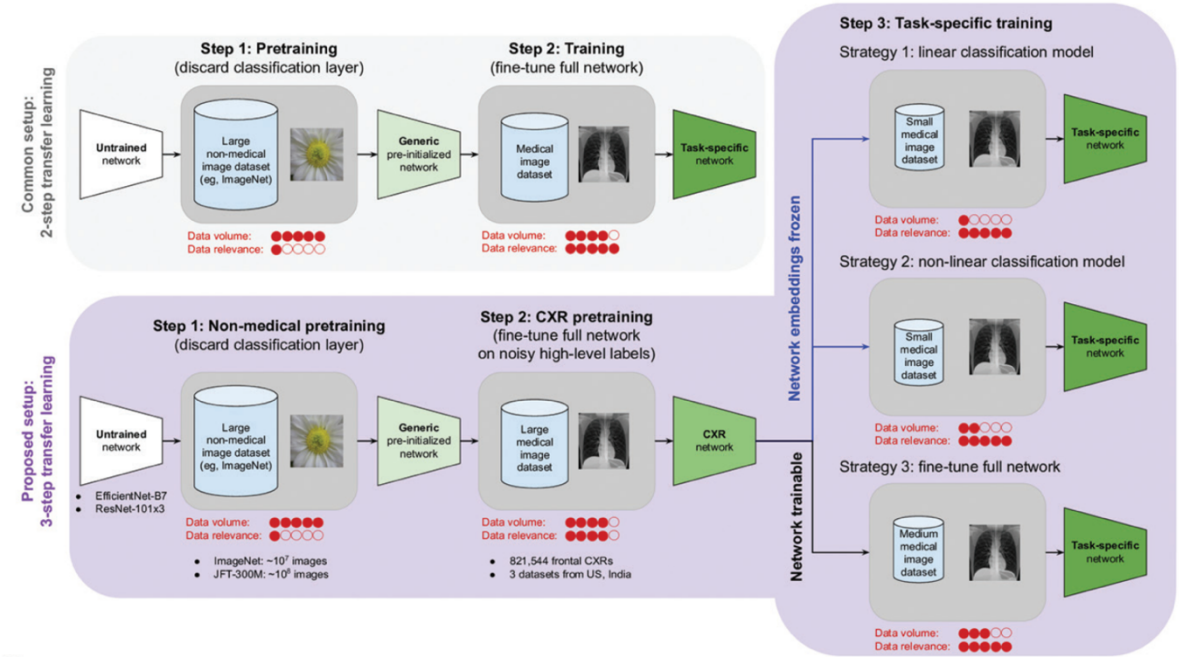

Figure 4. Pre-Training and Fine-Tuning Strategies for Enhancing Foundation Model Performance in Medical Imaging.
This figure, adapted from [89], contrasts two different training approaches for foundation models applied to chest radiography. The upper half of the diagram depicts a conventional single-phase training setup using only non-medical pretraining data like ImageNet. The lower half introduces a three-step training approach: (1) Non-Medical Pretraining on ImageNet to establish a broad base of visual features; (2) CXR Pretraining on a large dataset of chest radiographs to tailor the model's capabilities to medical imaging; and (3) Task-Specific Training, where the model undergoes fine-tuning through one of three strategies depending on data availability and task complexity. Strategy 1 involves Targeted Linear Probing with minimal data, Strategy 2 utilises Targeted Non-linear Probing for deeper model adjustments with a moderate amount of data, and Strategy 3 comprises Comprehensive Fine-tuning across all layers for optimal performance, requiring the most data.

practice of using unlabelled data in self-supervised learning. Such specialised pre-training utilised noisy disease labels, extracted via NLP from radiology reports, to focus on distinguishing images exhibiting abnormalities. Ultimately, this unique two-phase pre-training has proven highly effective, enabling the development of robust models with up to 688-fold less data than traditional methods that only utilises pre-training on generic non-medical datasets, across various architectures.

Going into more details, **Figure 4**—adapted from [89]—exemplifies the effectiveness of the fine-tuning techniques discussed earlier in **section 2.1.4**, comparing a conventional pre-training setup—which typically involves only non-medical pre-training using natural images datasets such as ImageNet—with CXR-FM's novel approach mentioned above that also includes an additional pre-training phase on a large medical (CXR) dataset with noisy high-level labels before task-specific fine-tuning. The proposed three-step development approach is detailed below and clearly delineates the three different fine-tuning strategies employed for the network (step 3) after the pre-training phase (steps 1-2):

Step 1: Non-Medical Pretraining — Initially, the FM is pre-trained on a large, non-medical dataset such as ImageNet to develop a broad base (knowledge) of visual features.

Step 2: CXR Pretraining — In the novel approach, after initial pre-training, the model undergoes a second phase of pre-training on a large CXR dataset. This tailors the model's capabilities to high-level medical features, enhancing its relevance to medical imaging.

Step 3: Task-Specific Training — The final step involves fine-tuning the now specialised CXR network (FM) using one of three strategies, reflecting their varying data requirements and complexity:

- *Strategy 1 (Targeted Linear Probing):* Focuses on training a small linear model atop the pre-trained layers, requiring the least amount of labelled data.
- *Strategy 2 (Targeted Non-linear Probing):* Employs a more complex non-linear model (Multilayer Perceptron) for deeper adjustments, needing a moderate amount of data.
- *Strategy 3 (Comprehensive Fine-tuning):* Adjusts all layers of the network, effectively utilising the model's extensive pre-training but requiring the most labelled data to refine and optimise performance across all layers.

2.1.7. Foundation Models Limitations

As mentioned in the Introduction Chapter, Glocker et al. [93] highlighted inherent biases in CXR-FM when tested on CheXpert. This issue is compounded by the model's limited accessibility, characterised by a frozen backbone that only outputs feature embeddings via API calls in response to CXR inputs. Below, we further elaborate on these limitations of FMs.

Transparency

FMs have rapidly penetrated several consumer-facing applications and industries, steadily increasing its societal impact. However, transparency is recognised as one of its main pain points, undermining the ethical development of FMs [148]. A paper introduced in 2023 by Bommasani et al. [149] from the Stanford Institute for Human-Centered Artificial Intelligence (HAI), implemented a FM specific transparency Index that integrates multiple indicators to assess the transparency of models. Applying this index across the industry revealed a low average transparency score of 37%, stressing on the need to have better documentation of data sources, disclosure of downstream impact of models, and supporting the community of consumers. This opacity is problematic in a field where transparency is critical for patient safety and regulatory compliance. In effect, the healthcare industry presents several technical and ethical challenges, directly related to interpretability, ethical implications, data privacy concerns, and unintended bias [75]. The inherent opacity of FMs often lead to a lack of interpretability, making it difficult for healthcare professionals to understand and trust the decision-making processes of these AI systems [150].

Bias Challenges

As discussed in **section 1.1**, FMs can inadvertently propagate existing biases from their training data, leading to skewed outcomes and unequal treatment across different demographic groups [75, 93]. This can become particularly troubling in clinical environments where biased decisions could result in disparate healthcare outcomes [151]. To address this, enhancing dataset diversity, implementing bias correction techniques, and continuous tracking of model performance are primordial to ensure fairness and mitigate biases [152]. The mentioned biases related to race and biological sex for Google's CXR-FM, as demonstrated by Glocker et al., therefore underscores the necessity of robust bias mitigation strategies before clinical deployment [94].

Catastrophic Forgetting

Catastrophic forgetting is a widespread issue in DL where models lose previously acquired knowledge upon learning new information. This challenge is particularly pronounced when adapting FMs to new tasks, as the fine-tuning process can inadvertently overwrite existing knowledge, requiring the re-training of the model from scratch each time a new task is added [153]. In the healthcare sector, where FMs may be required to adapt continuously to various medical conditions and contexts, catastrophic forgetting can severely impact the reliability of these models. An effective strategy known as elastic weight consolidation (EWC), localises and preserves important parameters from previously learned tasks [154]. This approach slows down the re-training of the model, enabling it to retain important information while accommodating new knowledge. Additionally, memory-augmentation strategies, like dual control memory augmentation [155] can be utilised, allowing the model to access and recall past information when needed. Such model relies on a dual-controller mechanism with an external memory module, allowing it to conserve its ability to function effectively across diverse (medical) applications.

2.2. Knowledge Distillation

2.2.1. Overview and Motivation

Introduction and Importance

Knowledge Distillation (KD) is part of a broader set of model compression and acceleration techniques, which include parameter pruning and sharing, low-rank factorisation, and transferred compact convolutional filters [156]. These techniques focus on facilitating the deployment of large-scale models in resource-constrained environments by reducing the computational complexity and storage requirements of deep neural networks.

In this context, KD enables a large, cumbersome model (often referred to as the 'teacher') to transfer its learned knowledge to a smaller, more efficient model (the 'student') [157]. This process is essential for scenarios where computational resources are limited, such as in mobile devices, embedded systems, and real-time applications. By distilling the knowledge from the teacher model, which is typically trained on extensive datasets with significant computation, the student model aims to achieve comparable performance while being more suitable for deployment in hardware-constrained settings.

In this study, the application of KD diverges from traditional objectives such as preparing models for resource-constrained settings. Instead, the focus is on reconstructing the teacher model, here the CXR-FM [89], to allow greater control and fine-tuning capabilities. This approach is driven by the need to address the challenges related to encoded biases in the CXR-FM and in FMs generally, as discussed in [93].

Historical Context and Relevance

The concept of KD has evolved significantly since its inception. Initially, Bucilă et al. [158] pioneered the idea of transferring knowledge from a large ensemble of models to a single, smaller model, demonstrating the feasibility and benefits of this approach. Subsequently, Hinton et al. formalised KD in their seminal paper [159], highlighting its potential to leverage the strengths

of large models to improve the performance of smaller models. Over time, KD has gained traction, especially in fields such as Computer Vision, NLP, and Reinforcement Learning, where the need for efficient models is paramount. The technique has been refined to include various methods for transferring knowledge, such as using soft targets, logits, and intermediate layer representations which will be discussed in the following sections.

2.2.2. Teacher-Student Architecture

In KD, the design of the teacher-student architecture plays a pivotal role in the effective transfer of knowledge. **Figure 5** provides a visual representation of the generic framework for KD within teacher-student models, illustrating how knowledge is transferred from a usually larger, more complex teacher model to a smaller, more efficient student model. The structures of the teacher and student models need to be carefully selected to optimise knowledge transfer and minimise the capacity gap between the two.

KD was originally conceptualised to compress ensembles of deep NNs, focusing on transferring knowledge from deeper and wider networks to shallower and thinner ones [159]. The student model can take various forms: a simplified version of the teacher with fewer layers and channels, a quantised version maintaining the original structure but with lower-bit parameter precision, a small network with efficient operations, or even the same network as the teacher. However, significant differences in model capacity can hinder the transfer process.

To bridge the gap, several methods have been developed, including introducing teacher assistants to ease the transition, leveraging residual learning, and employing structure compression techniques that transfer knowledge across layers. Additionally, advanced methods like network quantisation and block-wise knowledge transfer have been proposed to align the structures of teacher and student models more closely.

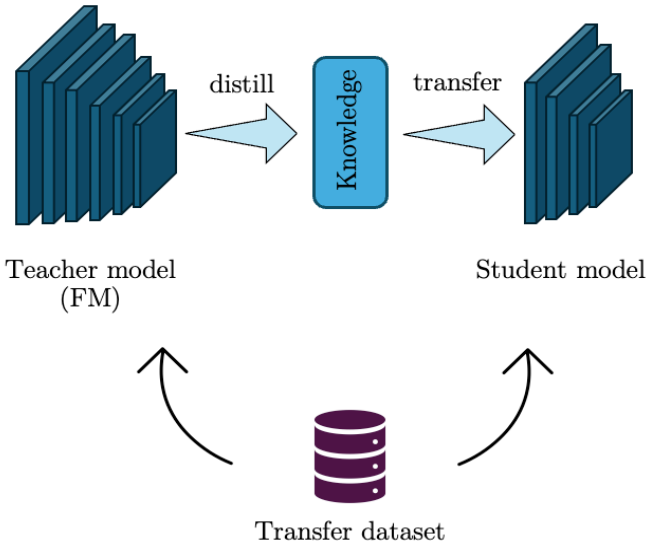

Figure 5. Framework for Knowledge Distillation in Teacher-Student Models.
Adapted from [157]. This figure illustrates the generic framework for knowledge distillation, showing a larger teacher model and a smaller student model with arrows indicating the direction of knowledge transfer. The transfer dataset, which is related to both models, is used for the distillation process.

2.2.3. Matching Knowledge

KD employs various techniques to efficiently transfer knowledge from a larger teacher model to the typically smaller student model. The knowledge transferred can be categorised into three main types: response-based, feature-based, and relation-based knowledge, as detailed by Gou et al. [157]. **Figure 6** specifically illustrates the different sources and components from which each knowledge type is derived within the teacher-student framework, enabling the transfer of knowledge across different layers of the teacher model. The subsequent section will follow the terminology and equations used in the KD survey by Gou et al. [157].

Response-Based Knowledge

Response-based KD is a straightforward yet impactful method for transferring knowledge. As depicted in **Figure 6**, this approach focuses solely on utilising the outputs from the teacher model's final layer, known as logits. The goal of response-based approaches is to effectively compress the complex knowledge encapsulated in the larger teacher model into a more efficient

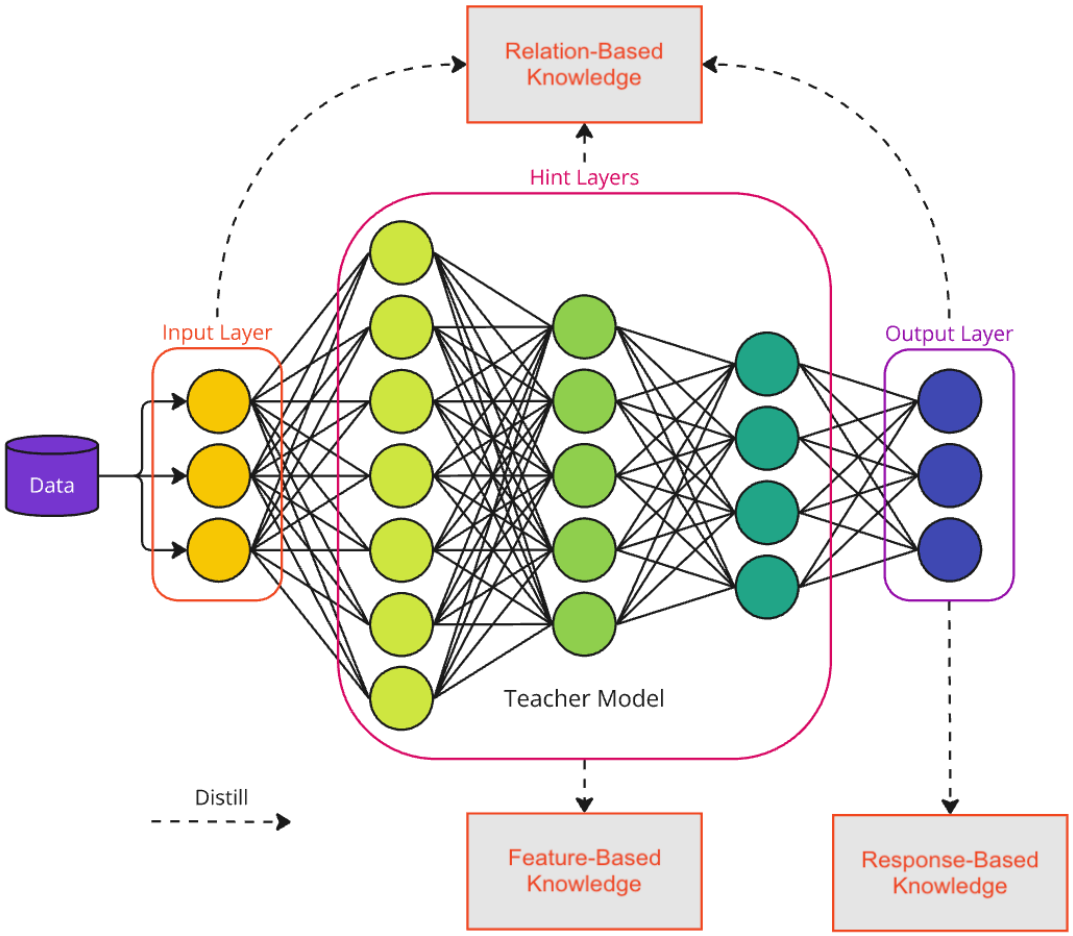

Figure 6. Visual Representation of Knowledge Types in Knowledge Distillation.
Adapted from [157]. This diagram categorises and illustrates the three main types of knowledge utilised in Knowledge Distillation: response-based, feature-based, and relation-based. Each category is visually differentiated and linked to its specific source within the teacher model: response-based knowledge from the output layer, feature-based knowledge from hint (intermediate) layers, and relation-based knowledge encompassing input, output, and hint layers. This representation shows how each knowledge type flows from the teacher to the student model, enhancing the student's ability to replicate and build upon the teacher's expertise, providing an understanding of the depth and complexity involved in effective knowledge transfer.

student model by aligning the student's predictions with the teacher's outputs, taken as 'ground truth'. Both the effectiveness and simplicity of this framework have established it as a foundational approach in various applications, including image classification, object detection, and semantic landmark localisation considered below.

Fundamentally, response-based KD focuses on minimising the divergence between the logits of the teacher and student models, which can be mathematically expressed as:

$$L_{\mathrm{RespKD}}(z_{\mathrm{t}},z_{\mathrm{s}}) = L_{\mathrm{R}}(z_{\mathrm{t}},z_{\mathrm{s}})$$

where z_{t} and z_{s} represent the logits of the teacher and student, respectively, and $L_{\mathrm{R}}(.)$ being the divergence loss function.

One of the most influential contributions to this field is the concept of 'soft targets', pioneered by Hinton et al. in their 2015 paper [159]. Soft targets refer the teacher model's predicted class probabilities for an input, refined beyond simple 'hard' class labels. These probabilities are computed using a Softmax function moderated by a temperature parameter T. The introduction of T by Hinton et al. addresses the issue of overconfident predictions typical of Softmax functions, where the probability mass is disproportionally concentrated on the class with the highest logit. This squashing effect can obscure important cues about the model's generalisation behaviour, as the relative probabilities of incorrect classes are neglected by being squashed towards zero. The modified Softmax operation can be expressed as follows:

$$p(z_i, T) = \frac{\exp\,(z_i/T)}{\sum_j \exp\,(z_j/T)}$$

Where z_i is the logit corresponding to class i and T controls the smoothness of the probability distribution. In a typical training phase, the hard label of an input is treated as a one-hot vector, focusing only on the correct class and discarding valuable information about the relative probabilities assigned to incorrect classes. Higher temperatures in the softmax operation produces softer predictions, revealing nuanced patterns, like correlations and hierarchical relationships between classes in the teacher's predictions as referred to by Hinton et al. as *dark knowledge*. By relying on matching the teacher's output distribution, these soft targets seem to be particularly valuable in regularising the student model and implicitly learning the collective knowledge of multiple models in the case of ensemble learning.

To facilitate the alignment of the student's probability distribution with the teacher's when combined with soft targets, the distillation loss can be reformulated as:

$$L_{ResD}\big(p(z_t, T), p(z_s, T)\big) = L_R\big(p(z_t, T), p(z_s, T)\big)$$

In terms of the loss function in this case, a common choice is the Kullback-Leibler (KL) divergence.

Taking one step further, this approach can be combined with a cross-entropy loss between the student's predictions and the ground-truth labels. This approach forms the basis of vanilla KD, as illustrated in frameworks like those proposed by Müller et al. [160], by combining label smoothing to adjust the target distribution and reduce overconfidence. Other papers have also noted similarities between the use of soft targets and techniques like label smoothing and regularisation [161].

Applications of this technique have extended beyond simple classification tasks. Namely, in semantic landmark localisation tasks like human pose estimation for example, Zhang et al. [162] introduces a Fast Pose Distillation (FPD) strategy to train lightweight neural networks, by including heatmaps representing the likelihood of specific landmark positions in the teacher model output (**Figure 7**).

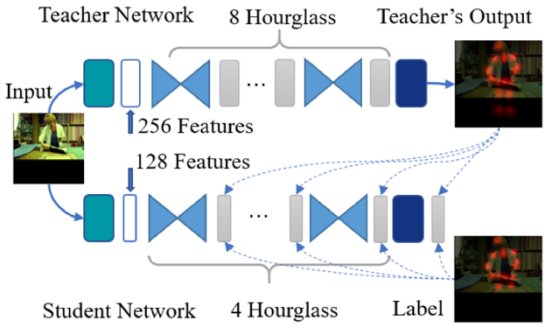

Figure 7. Fast Pose Distillation Framework. [162]

Similarly, Chen et al. [163] demonstrated its utility in object detection, where the teacher's response includes both teacher-bounded regression loss and a weighted cross-entropy, to achieve a balance between speed and accuracy (**Figure 8**).

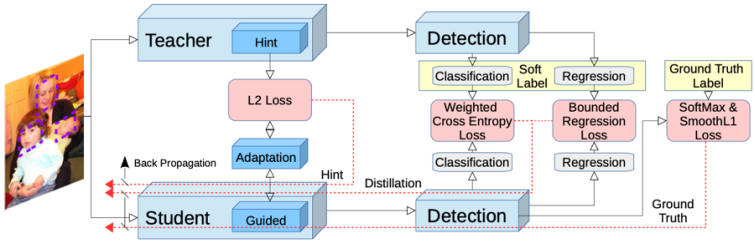

Figure 8. Teacher-Bounded Regression Loss and Weighted Cross-Entropy Used for Knowledge Distillation. [163]

Feature-Based Knowledge

To address the challenges of response-based knowledge generally limited by the insights enclosed within the teacher model's final layer, researchers have been leveraging intermediate layers of deep neural networks. This idea results from the high proficiency of deep learning architecture to hold rich hierarchical feature representations where lower layers capture basic features, and higher layers capture more abstract, task-relevant features (increasing levels of abstraction as one moves through the network layers). This capability, also known as representation learning [118], enables networks to extract meaningful intermediate representations, also known as feature maps, from the raw data to guide the student model's learning process.

The concept of using intermediate representations for knowledge transfer was first introduced in FitNets [164], where the authors proposed using 'hints' to improve the training of student models by matching the feature activations/intermediate weights of the teacher and the student. This

mechanism was proven particularly efficient to extend KD when the student network has deeper layers and fewer parameters (**Figure 9**, (a)).

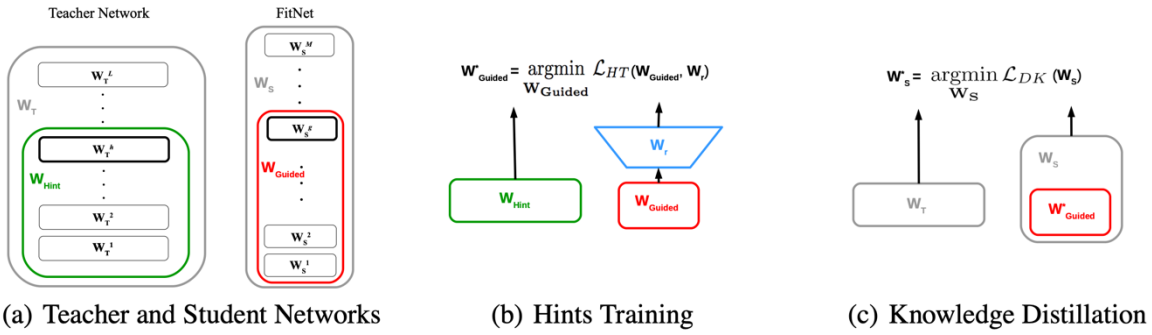

Figure 9. Introduction of Hints for Knowledge Distillation. [164]

Hints supervision is enabled at intermediate layers of the student network, wherein a specific intermediate layer of the teacher network, referred to as the hint layer, is selected to provide rich feature representations capturing mid-level abstractions. The student network goes through a similar selection process to identify its intermediate layer that will mimic the feature representations of the teacher's hint layer during training. The corresponding intermediate layer of the student model is also known as the guided layer. Once the intermediate representations are aligned, the student is fine-tuned using standard KD.
To bypass the challenges of matching the intermediate weights of models with different dimensions (student vs teacher), the weights of the guided student network are passed through a fully connected layer known as a projector/linear transformation (W_r), making the networks dimensions compatible (**Figure 9**, (b)).

Similar feature representation architectures papers include Wang et al. [165], and Xu et al. [166]. Building on this foundation, subsequent methods have explored indirect matching of features to enhance flexibility and effectiveness of KD.

One notable approach is the use of attention maps derived from feature maps to represent knowledge [167]. The authors define attention in CNNs as spatial maps that indicate the focus areas during decision-making processes. The attention of a CNN feature map $\boldsymbol{x}$ is defined as $\frac{\partial L}{\partial x}$, where L is the learning objective. As the attention becomes larger, any small perturbation at any location of the feature map (e.g. significant change in pixels in a picture) will have a more significant impact to the final output, hence adding more attention on this specific perturbation

Figure 10. Knowledge Distillation Through Attention Maps. [167]

location. Similarly, feature maps will require an additional transformation layer that matches the weights dimensions between models.
These attention maps emphasize regions of interest within the feature space as explained above, making knowledge transfer more interpretable and focused (**Figure 10**). It was also shown that performant models will generally have similar attention map generated during training and knowledge transfer.

This concept was later adapted by Huang and Wang [168], to introduce neuron selectivity transfer, a method that focuses on selective activations of neurons to extract knowledge. This adapted technique refers to the preference of neurons to pick certain input patterns, which holds valuable information about the task-specific features learned by the network. As an example, in a convolutional neural network (CNN) trained for image classification, some neurons may activate strongly for edges, textures, or specific object parts. By sticking to this principle, the goal is to align the selectivity distribution of neurons in the student network with that of the teacher network (**Figure 11**). This is done by applying a feature map reduction operation to spatially summarise each channel in a layer and obtain the activation map of a convolutional layer, reflecting how neurons in a specific layer respond to input data.
To dig more in the technicality of the paper, the alignment of the selectivity distributions is compared using the Maximum Mean Discrepancy (MMD). MMD measures the distance between two probability distributions using the cosine of angle between teacher/student feature vectors in a Reproducing Kernel Hilbert Space (RKHS).

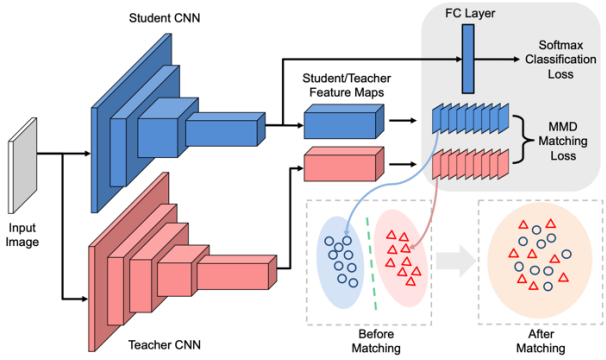

Figure 11. Neuron Selectivity Transfer Architecture. [168]

Additionally, Heo et al. [169] explored the activation boundaries of hidden neurons to enhance knowledge transfer. This method aims to add more emphasis on decision-critical regions within the feature space, concentrating more on the activation of neurons, rather than the magnitude of neuron responses. This methodology is analogous to matching the sparsity patterns of the teacher and student networks through the lens of a ReLU function. In effect, for the models to match, they should have similar sparsity patterns after the ReLU activation, denoted by the following indicator function which activates a neuron if its value is larger than 0:

$$\rho(x) = 1[x > 0]$$

To adhere to the above principle, Heo et al. [169] implemented a new loss function to transfer the neuron activation that penalises when the activations of the teacher and student networks are not the same:

$$\mathcal{L}(I) = \left\| \rho\big(T(I)\big) - \rho\big(S(I)\big) \right\|_1$$

Since the model is trained to keep the activation of the teacher neurons, the activation boundaries are accurately transferred. This concentration provides more accurate transfer of the activation boundaries.

On another note, to simplify intermediate representations, Kim et al. [170] introduced the concept of factor transfer (FT). The FT method utilises two main convolutional modules: a paraphraser and a translator. The paraphraser, trained in an unsupervised fashion by a reconstruction loss, converts the teacher model feature maps to the 'teacher factors' which learns the main information of the teacher network. The translator then transforms the feature maps of the student network to the 'student factors' with the same dimension as that of the teacher factor. The factor transfer loss is used to mimic the teacher factors by minimising the difference between factors in the training of the translator that generates student factors (**Figure 12**).

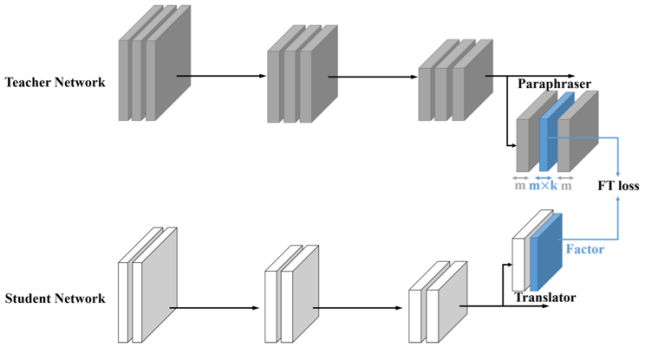

Figure 12. Overview of the Factor Transfer. [170]

Another innovative method brought up by Passalis and Tefas [171] transfers knowledge by aligning the probability distributions of feature spaces. In this paper, the authors propose modelling the interactions between data samples in the feature space as a probability distribution that expresses the similarities between samples. This method enables the student model to learn by matching its feature space distribution to that of the teacher model and closely capturing the underlying data structure (**Figure 13**).

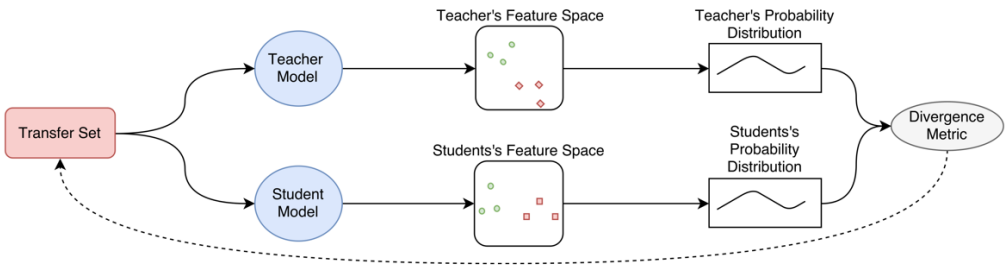

Figure 13. Probabilistic Knowledge Transfer. [171]

More recently, Chen et al. [172] introduced a cross-layer KD architecture to address semantic alignment challenges between teacher and student models, by adaptively assigning teacher layers to student layers using attention mechanisms. In fact, layer semantics usually vary in networks

while semantic mismatch in manual layer associations could lead to performance degeneration caused by negative regularisation. Chen et al. addresses this issue using a learned attention distribution, where each student layer distils knowledge from multiple teacher layers rather than a specific hinted layer (**Figure 14**).

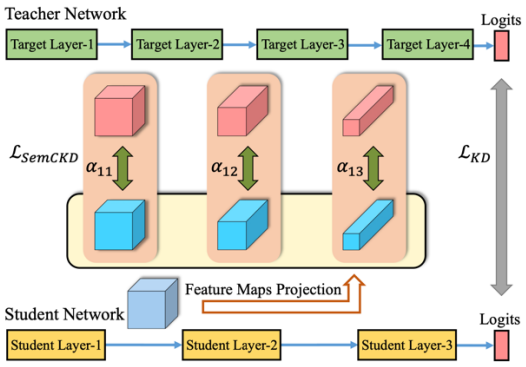

Figure 14. Semantic Calibration Knowledge Distillation. [172]

The general formulation of feature-based KD loss can be expressed as:

$$L_{FeatKD}\big(f_t(x), f_s(x)\big) = L_F(\Phi_t\big(f_t(x)\big), \Phi_s\big(f_s(x)\big)$$

where $f_t(x)$ and $f_s(x)$ are the feature maps from the intermediate layers of the teacher and student models, respectively. Transformation functions $\Phi_t(.)$ and $\Phi_s(.)$ are often applied when the feature maps of the teacher and student differ in shape or dimensionality, as referred above as a projector/linear transformation through a fully connected layer. The similarity function $L_F(.)$ measures the alignment between these transformed feature maps and can take various forms, such as $L_2(.)$, $L_1(.)$, cross-entropy loss $L_{CE}(.)$, or maximum mean discrepancy ($L_{MMD}(.)$) loss, depending on the specific distillation objective discussed in some of the previous papers.

Relation-Based Knowledge

As discussed in the previous sections, response-based and feature-based KD utilise the outputs of specific layers in the teacher network, either by matching the soft labels for one input or mimicking intermediate layers. Relational-based KD goes one step further by looking at the relations between intermediate features from multiple inputs, also known as relative differences Park et al. [173] (**Figure 15**).

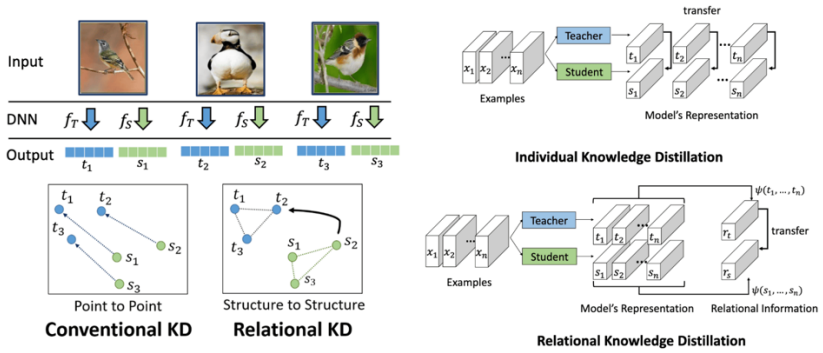

Figure 15. Conventional (Response-Based) Knowledge Distillation Compared to Relational Knowledge Distillation. [173]

The first paper released on relation-based knowledge is the flow of solution process (FSP) matrix, first proposed by Yim et al. [174] to investigate the relationships between different feature maps. FSP is inspired by the step-by-step approach or flow generally used by a genuine teacher to intuitively guide its student in solving a complex problem. Techniques like Gram matrices, which measure the correlations and directionality between different sets of features, or graph-based methods are used to compute the loss in this case. The FSP matrix is generated by taking the inner products between features from two layers and is denoted as follows:

$$G_{i,j}(x;W) = \sum_{s=1}^{h}\sum_{t=1}^{w} \frac{F^1_{s,t,i}(x;W) \times F^2_{s,t,i}(x;W)}{h \times w}$$

where x and W represent the input image and the parameters of the network.

Zhang and Peng [175] introduced a novel graph-based distillation framework, leveraging multiple self-supervised tasks by constructing a dual logits and representation graph. This infrastructure enables KD from multiple teacher models into a single student model (**Figure 16**) as follows:

The logits graph component addresses the multi-distribution joint matching challenge by transferring classifier-level knowledge and employing the Earth Mover (EM) distance to measure information among the logits of different teachers.

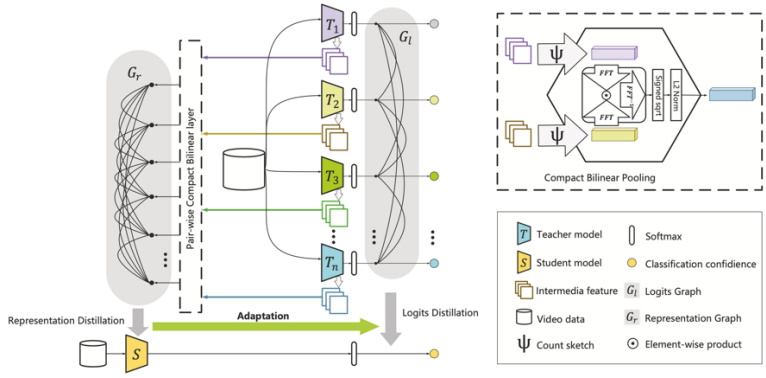

Figure 16. Graph-Based Distillation Framework. [175]

The representation graph component handles the heterogeneity among features from various self-supervised tasks and is mainly used to distil internal feature knowledge. The mechanism

behind it is achieved through pairwise ensembled representations resulting from compact bilinear pooling.

A study released by Lee et al. [176] proposes an innovative way to extract key information in the feature maps. The authors suggested using Singular Value Decomposition (SVD) to break up feature maps from the teacher network. This technique revealed enhanced capture of essential information while reducing spatial redundancy, facilitating the efficient transfer of core knowledge to the student network.

An important aspect of relation-based KD explores pairwise interactions between hint layers to enhance the knowledge transfer process. The information flow generated through these pairwise interactions record the dynamic behaviour of neural networks during different learning phases. Additionally, to facilitate the transfer of knowledge between heterogeneous architectures, an auxiliary teacher model can be introduced to link the teacher and student networks [177].

Another study introduced by Lee and Song [178] significantly advances relation-based KD by introducing a graph-based approach that captures and transfers relational knowledge within datasets. This multi-head approach captures the embedding procedures of the teacher model through the representation of intra-data relations as graphs, consequently transferred to the student model (**Figure 17**). The transfer of this representation induce the student network to gain a relational bias supporting its understanding and processing of the data structure.

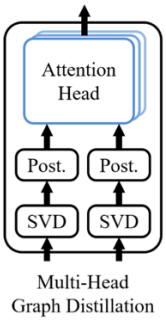

Figure 17. Multi-Head Graph Distillation. [178]

The general formulation of relation-based KD loss based on the relations of feature maps can be formulated as:

$$L_{RelaKD}(f_t, f_s) = \mathcal{L}_{R^1}\left(\Psi_t(\widehat{f}_t, \breve{f}_t), \Psi_s(\widehat{f}_s, \breve{f}_s)\right)$$

where f_t and f_s represents the feature maps of the teacher and student networks, respectively. Feature maps pairs are taken from both the teacher and student model respectively $(\widehat{f}_t, \breve{f}_t)$, $(\widehat{f}_s, \breve{f}_s)$ whereas Ψ represents the similarity functions of these pairs. All these mathematical functions are then encapsulated within a correlation function between the teacher and student feature maps denoted as $\mathcal{L}_{R^1}(.)$.

2.2.4. Distillation Methods

To conclude, it is interesting to mention that KD methods can be broadly divided into offline, online, and self-distillation. For a visual representation of these distillation methods, refer to **Figure 18** below.

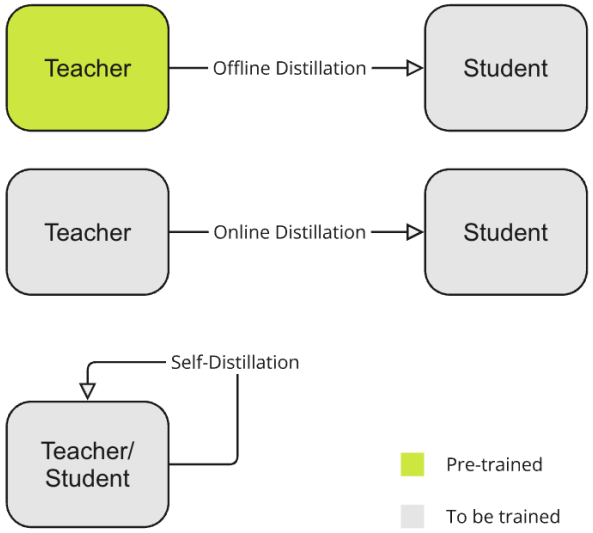

Figure 18. Diverse Approaches to Knowledge Distillation.
Adapted from [157]. This diagram illustrates three principal methods of knowledge distillation—offline, online, and self-distillation—highlighting how each strategy manages the transfer of knowledge from teacher to student models. Each section is color-coded with green indicating models pre-trained prior to the distillation process and grey indicating models to be trained during the distillation process.

First and foremost, offline distillation is the traditional/most common method of KD, where a pre-trained teacher model transfers its knowledge to a student model in order to replicate the teacher's outputs Hinton et al. [159]. This process can be broken down into two distinct steps, starting with the large teacher model following a standard training stage; followed by a KD step (any of the types discussed in the previous section), to train the student network.

Next, online distillation was introduced to overcome the challenges introduced by vanilla KD Mirzadeh et al. [179]. As an alternative, this method initially involves untrained teacher and student networks. Both models are updated in parallel making the whole KD architecture a one-phase end-to-end trainable framework. This form of distillation is generally leveraged in the case of low capacity/performance teacher model Zhang et al. [180].

Lastly, the framework behind self-distillation is made up of a unique component also viewed as a special case of online distillation. Meaning that the same initially untrained network is used for both the teacher and student models Hou et al. [181].

2.3. Foundation Model Distillation

This section delves into recent research on FM distillation and advancements that have significantly contributed to the field of medical imaging, and is especially relevant to this paper.

In recent years, a new research study was released by Daniel et al. [182], with the goal of enhancing pneumonia differentiation using human KD. In this study, a dataset composed of 1,082 CXR images from a university hospital was used to train a DL model while leveraging human expert image annotations. To overcome the issue of a limited and complex training dataset, the researchers implemented a novel KD process, referred as Human Knowledge Distillation. The process starts like any KD process with training a teacher model on annotated images that include complete localisation information from experts (Stage 1, **Figure 19**). The student model is then trained on raw images, while incorporating additional consistency regularisation based on the teacher model's predictions for the corresponding annotated images (Stage 2, **Figure 19**). A combined loss function using a weighted sum of the consistency loss (the Mean Squared Error (MSE) between the feature maps of the last convolutional layer) and the classification loss was utilised in the study. The final stage includes fine tuning of the student model on raw images only, in the absence of the consistency regularisation component (Stage 3, **Figure 19**). This method allows the student model to indirectly leverage localisation information through the teacher model, providing additional guidance and supervision to the model, improving its convergence and performance. The methodology implemented in the study was evaluated through the lens of several model types, with a significant increase in performance for all the students, with the best-performing model achieving a +2.3% improvement in overall accuracy compared to a baseline model.

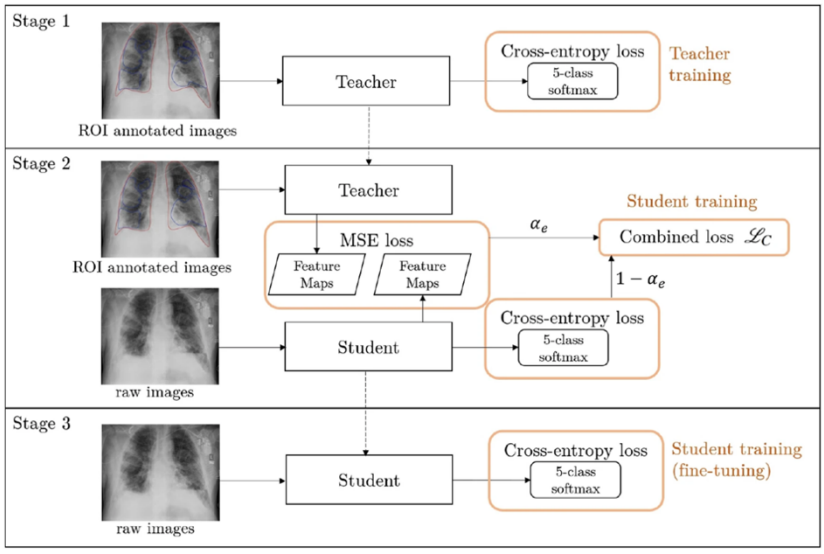

Figure 19. Human Knowledge Distillation Framework. [182]

Another recent study implemented a new KD framework aimed at detecting medial and lateral meniscus tear by Ying et al. [183]. The training dataset includes 199 paired knee Arthroscopy-MRI exams utilised to train a multimodal teacher model and a single-modality, MRI based

student model. The methodology used was the standard offline distillation process, coupled with a response-based framework where the teacher model outputted logits (final layer) and soft targets guide the supervision of the student model. The main goal of the study was to transfer the information from missing modality to the student model by minimising both the loss functions of Mean Squared Error (MSE) and Cross-Entropy (CE). As a result, the distilled model successfully benefited from the multimodal teacher model to achieve an enhanced meniscus tear detection performance. Some of the performance metrics used include a significantly higher accuracy of (0.764/0.734), sensitivity of (0.838/0.661), and F1-score of (0.680/0.754) for both medial and lateral tear detection.

2.3.1. Challenges in Knowledge Distillation

Despite its success, KD faces several challenges that could decrease its applicability and effectiveness.

Methodological Challenges

Response-based KD, which primarily relies on the outputs of the final layer, ignores any intermediate representations learned by the teacher model. These intermediate features often capture essential hierarchical information that is critical for deep networks, as highlighted by Romero et al. [164], which is overlooked by this method. On the other hand, response-based methods are also constrained to supervised learning settings, as it depends on labelled data to compute both soft and hard targets.
While feature-based knowledge transfer capture meaningful intermediate representations for the learning of student models, identifying the most relevant layers from both the teacher network and the corresponding layers in the student network remains one of the main challenges as highlighted by Yen-Chang et al. [184]. Furthermore, the substantial differences in size and dimensionality between the hint layers of the teacher and the guided layers of the student introduce additional complexity. Another challenge, raised by Rijk et al. [185], concerns the balancing of the feature alignment loss with the primary task loss. Improper balance can lead to conflicting gradients, which can destabilise the training process.

Complexity and Compatibility Issues (Teacher-Student Gaps)

The capacity gap between teacher and student models is another important challenge since teacher networks typically have more complex architectures. Addressing this issue requires innovative strategies, such as implementing intermediate robust models (teacher assistant models) [186] or leveraging advanced transfer learning techniques [187].

Quality of Knowledge Transfer

Another notable challenge is balancing model compression to a student network with its own performance as discussed by Malihi et al. [188]. Some techniques that could degrade the compressed model performance include over-regularisation and insufficient supervision during the transfer process. Effective combinations of KD techniques are needed to ensure high-fidelity knowledge transfer without sacrificing accuracy as discussed by Sarfraz et al. [189].

Transfer Datasets

A fundamental factor impacting the successful transfer of knowledge to a student model lies in the choice of relevant transfer datasets [190]. These datasets are generally sampled from the original training sample set used for training the teacher in order to match the target distribution. The reliance on the original training data raises several issues especially when large pre-trained models are released with restricted access to the desired data. As stated by Nayak et al. [191], this restriction emerges from practical constraints, including privacy concerns from sensitive and personal data (medical records and personally identifiable information). Proprietary restrictions are becoming more common, with large corporation moving away from opensource to hold their competitive advantage, as these datasets represent a significant investment in time and money. Further, some ML use cases such as reinforcement learning employs agents that explore and interact with an environment in real-time. In such scenarios, observations are often processed dynamically without being stored, leading to data transience - impossible to reproduce.

Chapter 3

Methodology

This chapter outlines the experimental and methodological framework employed in this study. It details the research design and execution, with a focus on using KD to develop and evaluate the CXR-FMKD student model, distilled from the CXR-FM teacher. The chapter also discusses the methodologies adopted for performance, generalisability, and bias analysis, explaining how these experiments contribute to understanding the impact of KD on the CXR-FMKD model and its comparative performance against baseline models and the original CXR-FM. It covers the datasets used, the experimental setup, and the specific KD techniques and metrics employed to assess model performance and bias.

3.1. Research Design

Following on from the *Contributions* **section 1.2** in Chapter 1, *Introduction*, this research investigates whether KD can serve as a viable strategy to reconstruct FMs to enhance their transparency, tunability, and ultimately their fairness. We focus on chest radiography, specifically through Google's proprietary CXR-FM, and explore the development of the CXR-FMKD student model, which is derived from CXR-FM through KD.

This investigation is prompted by the biases recently identified in the CXR-FM model by Glocker et al. [93], a study that serves as a key reference throughout this research. To establish clear benchmarks for improvements in our models and to build upon their findings, we engage in the same predictive task of CXR disease detection, which they explored using the CheXpert [103] dataset. We employ similar performance metrics and bias analysis methodologies as used by Glocker et al. Throughout our experiments, we compare our **CXR-FMKD** model to both the original **CXR-FM** and a baseline model, **CXR-Model**, which shares the same core backbone architecture as CXR-FMKD but was trained independently without KD from CXR-FM, representing a 'traditional' approach to the disease classification task.
As previously mentioned in the *Introduction*, optimising the performance of our models on the detection task is not the primary concern of our study. Instead, we aim to demonstrate the potential of KD to address and mitigate the limitations and biases inherent in the CXR-FM and, by extension, other similar FM models used in healthcare.

Taking a deeper look at the design and workflow of this research, as illustrated in **Figure 20**:

(1) We begin by exploring various KD techniques suitable for constructing multiple versions of our CXR-FMKD model, each employing different strategies and KD losses. These explorations

and the development of our baseline models will be detailed in **section 3.3,** *Model Development and Knowledge Distillation Strategy.*

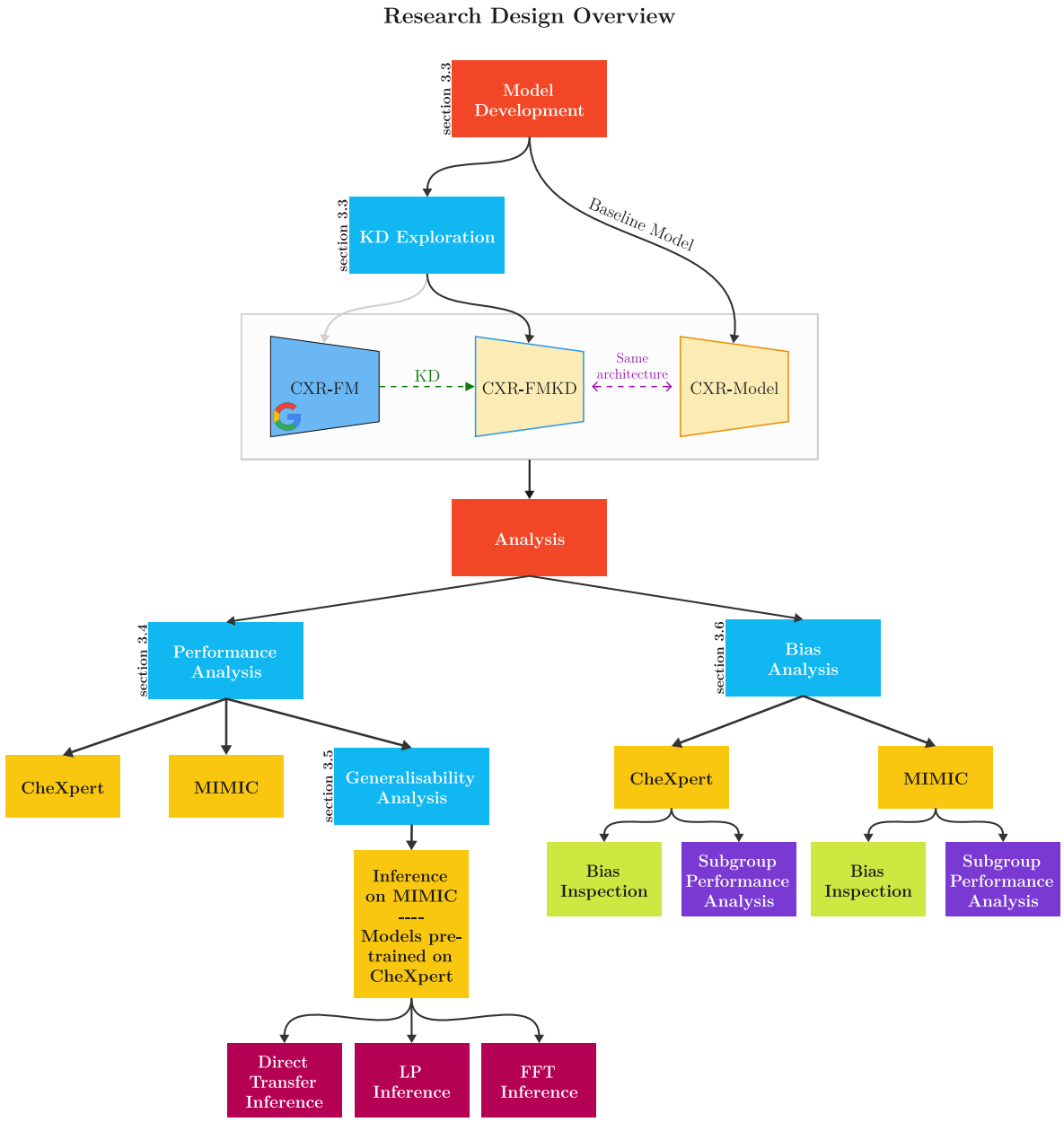

Figure 20. Overview Diagram of the Thesis Research Design and Workflow.
This diagram depicts the workflow and experimental structure of the thesis. It outlines the development of the CXR-FMKD via Knowledge Distillation (KD) from Google's proprietary CXR-FM and the baseline CXR-Model that shares the same core architecture as CXR-FMKD. Key experiments include KD Exploration, Performance Analysis, Generalisability Analysis, and Bias Analysis using the CheXpert and MIMIC datasets. Each step is linked to specific sections in Chapter 3. LP refers to Linear Probing and FFT to Full Fine-Tuning.

(2) Next, we assess the performance of our newly developed CXR-FMKD model to determine if it achieves our goal of matching or exceeding the performance of its teacher, the CXR-FM, in disease detection tasks, and how it stands against the baseline CXR-Model. This performance analysis is conducted using the CheXpert dataset [103], similarly to Glocker et al., and is repeated on the MIMIC dataset [104], which shares the same 14 CXR disease classes. This dual-

dataset approach provides a comprehensive view of the robustness of our results. Details of this analysis will be presented in **section 3.4**, *Performance Analysis*.

(3) To evaluate the generalisability of our CXR-FMKD—a key attribute of FMs—we conduct a series of inference testing experiments. Generalisability, in this context, reflects the model's ability to maintain performance when applied to out-of-distribution (OOD) data, significantly different from the data seen during training. This analysis starts by training the models on the CheXpert dataset, followed by testing on MIMIC using varying levels of fine-tuning, leveraging the shared 14 disease classes to assess performance across datasets. The testing process is stratified into three strategies of increasing levels of fine-tuning to progressively gauge adaptability: 'Direct Transfer' without adaptation, 'Linear Probing' (LP) by fine-tuning only the final classifier layer, and 'Full Fine-Tuning' (FFT) of all layers. These tests, detailed in in **section 3.5**, *Generalisability Analysis*, help illustrate how well the distilled model, CXR-FMKD, inherits and possibly enhances the robustness and transfer (learning) capabilities of the original CXR-FM, and how it compares to the baseline CXR-Model trained without KD.

(4) Further, **section 3.6**, *Bias Analysis*, addresses the core motivation of the study by examining the biases in our models, employing methodologies similar to those used by Glocker et al. in their paper. This section includes a detailed bias inspection of the CXR-Model, CXR-FM, and CXR-FMKD models by analysing features extracted before task-specific adaptations—the 'representation' each model generates from the CXR images before applying this information for disease detection. This analysis aims to identify if the models have inappropriately learned to rely on protected characteristics such as race and sex for disease classification, potentially leading to biased decisions. To provide a more systematic approach to investigating bias than previously used, we have developed a novel bias score to quantify such biases, facilitating clearer comparisons between our models. These aspects are detailed in **section 3.6.1**, *Bias Inspection*. Complementing this, we link these biases in the generated features to classification performance across patient subgroups and any revealed disparities, discussed in **section 3.6.2**, *Subgroup Performance Analysis*.

Details about the hyperparameters used in the experiments as well as the implementation, including the software and tools used, the hardware specifications, and environment setup, are also documented in this chapter.

3.2. Study Datasets

3.2.1. CheXpert and MIMIC Datasets

As part of our study, we utilise two publicly available CXR datasets: CheXpert [103] and MIMIC-CXR [104], hereafter referred to as MIMIC. Specifically, we use the MIMIC-CXR-JPG [192] version, derived from the original MIMIC-CXR dataset and processed to include CXR labels and standard references for data splits. Both datasets are enriched with detailed patient demographic information—including biological sex, self-reported racial identity, and age—and feature the same set of 14 disease labels.

To obtain these labels, Irvin et al. [103] developed the CheXpert labeller, which automatically identifies the presence of 14 common observations in radiology reports using NLP. It also addresses the inherent uncertainties in radiographic interpretations by assigning uncertainty

labels. Thus, for each CXR, each disease label is categorised as *unmentioned* (**blank**), *negative* (**0**), *uncertain* (**-1**), or *positive* (**1**).

Here, the choice of these 14 labels was based on clinical relevance and prevalence of the conditions, adhering to the standards suggested by the Fleischner Society's recommended glossary [193]. The labels are:

1. *Pleural Effusion*
2. *No Finding*
3. *Cardiomegaly*
4. *Pneumothorax*
5. *Atelectasis*
6. *Consolidation*
7. *Edema*
8. *Pleural Other*
9. *Enlarged Cardiomediastinum*
10. *Pneumonia*
11. *Lung Lesion*
12. *Lung Opacity*
13. *Fracture*
14. *Support Devices*

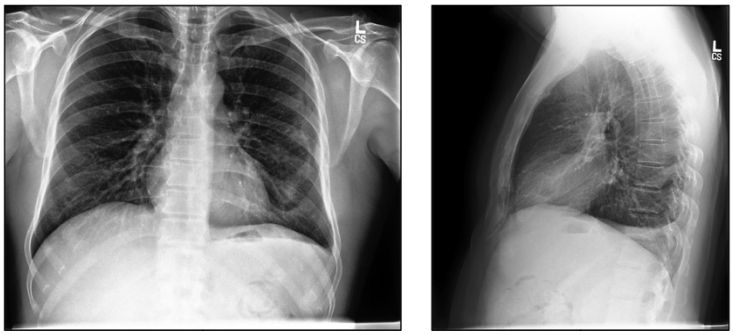

Figure 21. Example of Frontal and Lateral Chest X-Rays from the CheXpert Dataset.
Adapted from [103]. This figure displays two types of chest X-rays: the left image shows a frontal view, and the right image shows a lateral view, both extracted from the CheXpert dataset.

A specialised CheXpert documentation [194] provides additional details about these labels. Notably, '*Pneumonia*' label, is included to represent images suggestive of primary infection, despite being a clinical diagnosis. The labels '*No Finding*' and '*Support Devices*' are categorised as observations rather than pathologies: '*No Finding*' is assigned (classified as positive) when no pathology[3] is marked as positive or uncertain. Additionally, '*Support Devices*' is the only other label that can co-occur with '*No Finding*', indicating the presence of medical support devices like pacemakers, valves, or tubes visible in the CXR.

Both CheXpert and MIMIC utilise the CheXpert labeller to ensure consistency in the labelling process, enhancing the comparability and utility of these datasets for our experiments, particularly for the *Generalisability Analysis* experiment.

In terms of dataset specifics:

- CheXpert consists of 224,316 CXRs from 65,240 patients, collected between 2002 and 2017 in both inpatient (where patients are admitted and stay overnight) and outpatient (where patients are treated and return home on the same day) settings at a large hospital in Palo Alto, CA, USA.
- MIMIC consists of 377,110 CXRs from 65,379 patients, collected between 2011 and 2016 in the emergency department of a major hospital in Boston, MA, USA.

Both datasets provide a standardised reference for data splits, ensuring structured and consistent use in research.

[3] The extended list of pathologies reviewed to determine the '*No Finding*' label includes more than the ones in the principal 14 labels. This detailed list is accessible via the following link: *https://github.com/stanfordmlgroup/chexpert-labeler/blob/master/phrases/mention/no_finding.txt.*

3.2.2. Study Data Generation

Initial pre-processing and resampling were applied to both the CheXpert and MIMIC datasets before their use in our study. We employed the same approach as that used by Glocker et al. [93, 99], adopting identical study samples and respective training/validation/test splits as those from a previous work by Gichoya et al. [96]. The latter focused on evaluating the ability of AI to recognise racial identity from medical images and provides an important discussion on the definitions of race and racial identity.

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 22. Demographic Distribution of Patient Data in CheXpert and MIMIC Datasets.
This table, taken from Glocker et al. [99], 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.

To prepare the final study samples, several modifications and constraints were applied, including the removal of inconsistent or incomplete patient information, such as inconsistently documented races in MIMIC or unknown ethnicities in CheXpert. Only frontal CXRs were retained, discarding all lateral viewpoint images—both views are illustrated in **Figure 21**. For frontal CXRs, both posteroanterior (PA) and anteroposterior (AP) views were included, differing in the direction of the X-ray beam travel—from the back (posterior) to the front (anterior) or the

inverse, respectively. The race and sex demographics were standardised to maintain consistency across datasets, retaining only those CXRs classified with races as *Black*, *Asian*, or *White*, and biological sex as *male* or *female*, which are pivotal for subsequent bias analysis. Additionally, all CXR images were resized to **224x224 pixels**, partly to align with computational limitations typical in research settings, which often utilise smaller image sizes than those used in clinical practice. The *training/validation/test* splits were set at **60%/10%/30%**.

The dataset specifics for the refined study data are as follows:

- CheXpert with 127,118 CXRs from 42,884 patients, divided into the following splits: 76,205 CXRs for training; 12,673 CXRs for validation; 38,240 CXRs for testing.
- MIMIC with 183,207 CXRs from 43,209 patients, divided into the following splits: 110,280 CXRs for training; 17,665 CXRs for validation; 55,262 CXRs for testing.

The processing steps to generate these study samples from the original datasets are detailed in the released code repository by Glocker et al. [99], available at: *https://github.com/biomedia-mira/chexploration.*

As a general note, for model development, the training set is used for training the models, the validation set is used for model selection, and the test set is held-out (unseen during training) and used for assessing the performance of the models on the disease detection task as well as for inspecting any potential biases.

Regarding the study samples demographics, as detailed in Glocker et al. [99] and illustrated in their corresponding demographic data table in **Figure 22**, both CheXpert and MIMIC exhibit significant imbalances and skewness across demographic subgroups. The majority of scans are from patients identifying as White, comprising 78% in CheXpert and 77% in MIMIC. Patients identifying as Asian and Black are underrepresented, with Asian patients making up 15% in CheXpert and 4% in MIMIC, and Black patients accounting for 7% in CheXpert and 19% in MIMIC. Notably, the proportion of female patients varies by race, with 40%, 43%, and 49% for White, Asian, and Black patients, respectively, in CheXpert, and 43%, 44%, and 60% in MIMIC. We can also observe that Black patients in CheXpert are on average 5-8 years younger than their Asian and White counterparts, while in MIMIC, the age difference ranges from 2-5 years.

3.2.3. Test Set Resampling

To ensure a more accurate evaluation of subgroup performance and model bias, especially relevant for **section 3.6** (*Bias Analysis*), we implement the same test set resampling approach with replacement as described by Glocker et al. [93, 99], effectively expanding the test sets for CheXpert and MIMIC threefold.

As outlined in [99], the need for this arises from the limitations of standard random dataset splitting techniques, which can inadvertently preserve biases from the training data within the test set. This mirroring effect distorts the analysis of performance disparities across demographic subgroups, making it difficult to isolate and assess sources of bias in the model [195]. In fact, studies have demonstrated that differences in prevalence rates and demographic distributions among subgroups contribute to discrepancies in predictive performance [72, 196]. To mitigate this, it is therefore crucial to work with a test set that offers a more balanced representation of the population.

Since curating such unbiased datasets is often impractical, we apply the previously mentioned strategic resampling with replacement to construct test sets that counteract these imbalances. This technique adjusts for variations in subgroup characteristics like race, age, and disease prevalence. By doing so, we create a more controlled testing environment that facilitates the faithful detection of performance disparities across subgroups.

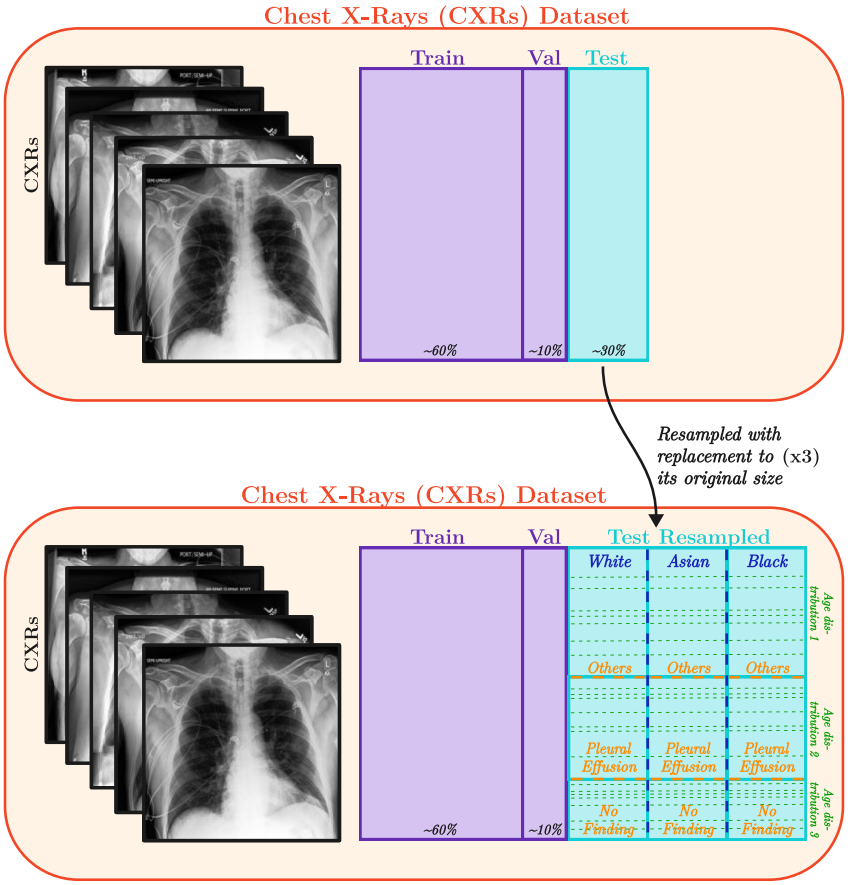

Figure 23. Stratified Resampling for Demographic and Clinical Balance in Test Dataset. The resampling process triples the test set size and adjusts to compensate for differences across subgroups, including variations in race, age, and disease prevalence. 'CXRs' refers to chest X-rays.

As depicted in **Figure 23**, this resampling process begins by equalizing the racial distributions within the test set to ensure proportional representation of White, Asian, and Black patients. At this stage, the test set has effectively tripled in size, with each racial subgroup expanded to match the size of the original complete test set. Concurrently, we document the overall disease prevalence and age distributions per disease label as observed in the original test set.
Following this, each racial subgroup undergoes a further resampling with replacement, maintaining its enlarged size to align disease prevalences with those recorded for the overall dataset. Age distributions are similarly balanced within each race-disease subgroup, employing replacement to achieve these demographic adjustments based on the documented age proportions for each disease label. It is important to note that this process leads to duplicates within the newly formed test set.

In our resampling process, we concentrate on two primary disease labels: '*Pleural Effusion*' and '*No Finding*', grouping all other conditions under the label '*Others*'. This choice mirrors the approach from Glocker et al., where '*Pleural Effusion*' is viewed as indicating the presence of thoracic pathology and '*No Finding*' signifies the absence of detected pathology. These labels are therefore mutually exclusive, which simplifies the analysis of bias inspection and subgroup performance. Notably, this choice simplifies the subsequent bias analysis, making it more focused, and aids in examining the model's behaviour under conditions of high clinical relevance, as '*Pleural Effusion*' is among the most prevalent and significant conditions [103]. For a more in-depth understanding of the reasons behind the focus on these specific labels and the interesting demographic and prevalence shifts they encapsulate, we refer readers to the extensive discussions in the study by Glocker et al. [99], under 'Methods'.

To conclude this section on resampling, while it is necessary to have a balanced test set for **section 3.6** (*Bias Analysis*) to accurately assess model subgroup performances and potential biases, we use the same resampled test set across all experiments, including for **section 3.4** (*Performance Analysis*) and **section 3.5** (*Generalisability Analysis*). Although this might slightly affect absolute performance metrics, it does not impact our study's primary objective, which is to compare relative performances between our models—CXR-FM, CXR-FMKD, and the baseline CXR-Model. Moreover, the *Bias Analysis* section utilises the results from the *Performance Analysis*, which includes examining the mentioned subgroup performances and analysing the feature representations generated for bias inspection. This integrated approach, by using the resampled test sets consistently throughout our experiments, ensures a seamless and coherent analysis process, enabling a detailed inspection of how model performance correlates with potential biases within the same controlled testing framework.

Consequently, the final dataset splits used in all of our experiments are structured as follows:

- CheXpert: 76,205 CXRs for training; 12,673 CXRs for validation; 114,720 CXRs for testing (3 times the original test set size of 38,240).
- MIMIC: 110,280 CXRs for training; 17,665 CXRs for validation; 165,786 CXRs for testing (3 times the original test set size of 55,262).

In the context of these dataset configurations, it is important to highlight findings from Glocker et al. [99], which revealed that using this test set resampling for CheXpert, for example, significantly reduced performance disparities for the *'No Finding'* label across subgroups. This suggests that such previous disparities were predominantly driven by statistical imbalances in the test data, rather than by inherent model biases. Conversely, for the *'Pleural Effusion'* label, disparities persisted even after resampling, suggesting that these may be due to intrinsic biases within the model in question. Therefore, applying test set resampling helps eliminate disparities that result from population and prevalence shifts across subgroups, and expose those due to actual model bias.

3.3. Model Development and Knowledge Distillation Strategy

3.3.1. Multi-Label Classification for Disease Detection

As discussed in **section 3.2**, both datasets used in our study, CheXpert and MIMIC, share the same set of 14 disease labels. Given that each patient may exhibit multiple thoracic pathologies,

we effectively operate within a multi-label classification framework for our disease detection task [103]. Indeed, each CXR can represent one or more pathologies, as defined by these 14 labels. Consequently, the (one-dimensional) output layer of a model trained for this task includes 14 output neurons, one for each disease class. In such multi-label classification scenarios, a sigmoid function is applied to each of the logits ($logit_c$) from the output layer to generate probability scores (p_c) for each class c—this approach is appropriate as the sigmoid function maps the input values between 0 and 1. Taking an input x, the sigmoid function $\sigma(.)$ applied to this input is defined as:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

This leads to the following expression for the probability score p_c given a logit $logit_c$:

$$p_c = \sigma(logit_c) = \frac{1}{1 + e^{-logit_c}}$$

Figure 24 illustrates this application of the sigmoid function on the logits produced from the feature representation generated by the model's NN backbone following an input CXR. For generality, this illustration assumes a disease detection scenario encompassing C classes.

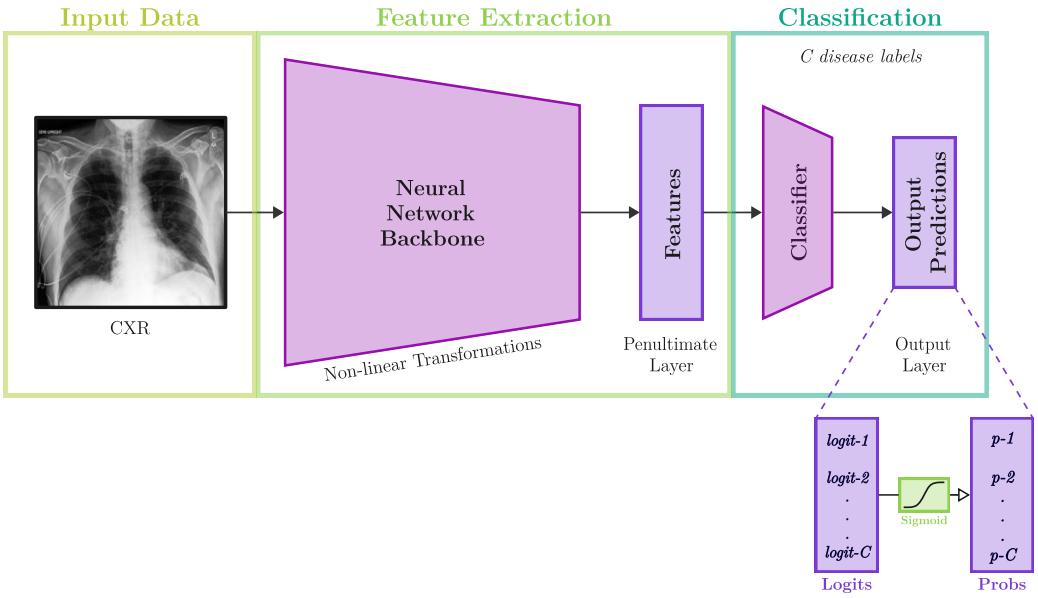

Figure 24. Application of the Sigmoid Function in Multi-Label Chest X-ray Classification. This figure illustrates the application of the sigmoid function to the logits from the output layer of the model, generating probability scores for each of the C disease classes in a general multi-label chest X-ray classification scenario for disease detection.

For training a model on this multi-label classification task, the Binary Cross-Entropy (BCE) loss is employed to optimise the model using the ground truth labels (y_c) for each class c. Typically, such BCE loss is used in binary classification tasks with only two possible classes signalling, for example, the presence of absence of a specific disease. For a single training example i, with a predicted probability $p^{(i)}$ and a ground truth label $y^{(i)}$ (either 0 or 1), where the superscript $^{(i)}$ denotes the specific training input i (i.e., the i^{th} CXR), the BCE loss is calculated as follows:

$$\text{BCE}(p^{(i)},\ y^{(i)}) = -\left(y^{(i)}\ log(p^{(i)}) + (1 - y^{(i)})\ log(1 - p^{(i)})\right)$$

This approach can be easily extended to the multi-label classification setting, where each class c is treated as a separate binary classification task. Here, the BCE loss is computed for each class and then summed across all classes to derive the final Multi-label Cross Entropy (MCE) loss. For a single training example i, with a predicted probability $p_c^{(i)}$ and a ground truth label $y_c^{(i)}$ for each class c, the MCE loss is defined as follows:

$$\text{MCE}(p^{(i)},\, y^{(i)}) = -\sum_{c=1}^{C} \left(y_c^{(i)}\ log(p_c^{(i)}) + (1 - y_c^{(i)})\ log(1 - p_c^{(i)})\right)$$

For all N training examples in a batch or set, the training Loss $\mathcal{L}$ is the sum of MCE losses across all examples:

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{c=1}^{C} \left(y_c^{(i)}\ log(p_c^{(i)}) + (1 - y_c^{(i)})\ log(1 - p_c^{(i)})\right)$$

$$\mathcal{L} = \sum_{i=1}^{N} \text{MCE}(p^{(i)},\, y^{(i)})$$

This discussion leads to an important point relating to the 'ground truth' labels $y_c^{(i)}$, which are used in both training—through the application of the BCE loss as seen above—and in computing the corresponding performance metrics for our disease detection task using the CheXpert and MIMIC datasets. As detailed in **section 3.2**, each disease label is categorised as *unmentioned* (**blank**), *negative* (**0**), *uncertain* (**-1**), or *positive* (**1**). However, to apply the training loss $\mathcal{L}$ mentioned previously, a binary categorisation of the ground labels (either 0 or 1) is required. In their original CheXpert paper [103], Irvin et al. explore and discuss several approaches to deal with the uncertainty labels, including simply ignoring them or applying a binary mapping where all *uncertain* labels are assigned as either *positive* (**1**) or *negative* (**0**). Following and building upon the methodologies employed by Glocker et al. [93, 99], the decision was made to set labels as 1 for *positive* (**1**) findings, and 0 for all other cases—namely the *negative* (**0**), *unmentioned* (**blank**), and *uncertain* (**-1**) labels. This approach ensures that all instances are retained, which is necessary for effective training.

3.3.2. Teacher Model: CXR-FM

Our primary investigative model is Google's proprietary CXR-FM [89], which has recently been made publicly available [97]. This move sets an important precedent and highlights the necessity for transparent FMs. During the development phases of CXR-FM [89], various architectures were explored depending on the nonmedical dataset used for its generic pre-training: EfficientNet-B7 was utilised for ImageNet [136], while larger ResNet architectures (ResNet-101 × 3 and ResNet-152 × 4) were employed for the JFT-300M dataset. However, the latest iteration of CXR-FM, as detailed in their GitHub page and the recent release [97], adopts the EfficientNet-L2 architecture, a scaled up version of EfficientNet-B7 but uses a lower resolution [197].

Returning to the main objective of this section as outlined in the *Research Design*, CXR-FM serves as the teacher in our KD process to develop the distilled student model, CXR-FMKD. **Figure 25** illustrate how this model is applied in our multi-label disease classification problem.

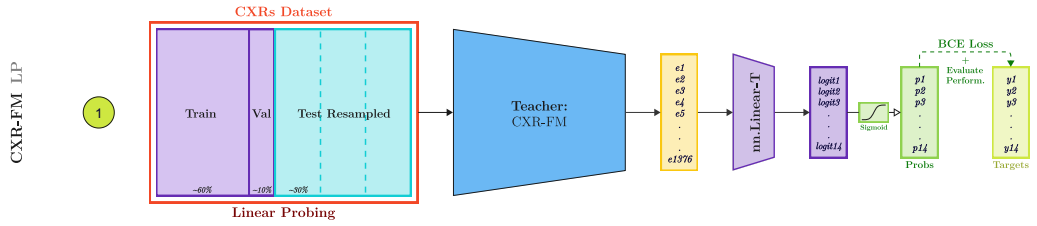

Figure 25. Architecture and Integration of CXR-FM in the CXR Disease Detection Task.
This diagram illustrates the integration of Google's proprietary CXR-FM into our multi-label disease classification framework, applicable to both the CheXpert and MIMIC datasets. Please refer to **Figure 26** for a detailed legend explaining the technical terms and components used in this and related diagrams.

As previously noted, access to CXR-FM was traditionally provided through a programming interface that processes input CXR images to output extracted features. Notably, this platform does not expose the network weights, thus precluding any updates to the feature extractor

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 26. 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.

parameters during downstream training tasks. As illustrated in **Figure 25**, the generated feature is an embedding vector of size 1376, depicted in yellow, indicating that both the embedding and the CXR-FM's backbone, shown in blue, are frozen—non-trainable and not subject to fine-tuning. This embedding serves as a rich representation extracted from the CXR, which is then inputted into a 'submodel', specifically the *nn.Linear-T* layer. This is a single, fully-connected classification layer that outputs a logit for each of the 14 disease labels, as described in **section 3.3.1**. These logits are subsequently passed through a sigmoid function to derive the final predictions in the form of probability scores.

This configuration, initially the intended and sole method of employing CXR-FM when it was not publicly accessible, presents limitations concerning bias and tunability, echoing concerns highlighted in earlier discussions. It is worth noting that this setup follows the architectural integration described by Glocker et al. [93] under their 'CXR-linear' model. While they also developed more complex models such as the 'CXR-MLP-3' and 'CXR-MLP-5', featuring multilayer perceptrons (MLPs) as the 'submodels' with three and five hidden layers respectively, these showed comparable performance and bias profiles to the simpler 'CXR-linear'. Therefore, to streamline our experiments and maintain a manageable number of models for analysis, we continue with the single-layer approach across all models, including CXR-FMKD, to map extracted features to the 14 disease classes.

The *nn.Linear-T* layer, depicted in purple to highlight its training and fine-tuning phases influenced by the training and validation datasets also in purple, is the only component actively trained and fine-tuned for the disease detection task. For a detailed understanding of the diagram's components, colour codes, and their significance, please refer to **Figure 26**. Additionally, this approach aligns with the Linear Probing (LP) strategy discussed in Chapter 2, where only the classifier layer is trainable while the remainder of the network remains frozen. Consequently, we have further designated this model as CXR-FM LP, setting a baseline for comparing with other models in our study that will explore both Linear Probing and Full Fine-Tuning (FFT).

3.3.3. Student Model: CXR-FMKD

Having explored how the teacher model, CXR-FM, integrates into our disease detection task—with its architecture frozen and access limited only to extracted features—we now have a clearer understanding of how the knowledge distillation (KD) process, as described in **Chapter 2**, *Background and Related Work*, can be implemented. This insight helps in crafting the distilled student model, CXR-FMKD, fitting within the broader teacher-student KD framework.

Initially, it is important to note that with only the 1376-sized feature embeddings from the CXR-FM teacher accessible, our KD options are somewhat limited. This confines the teacher model to the category of a 'pre-trained' teacher, where the NN backbone is established prior to KD. This scenario categorically falls under the **Offline Distillation** scheme, as described by [157]. In terms of knowledge matching, the only data we can utilise are the input CXRs and their corresponding feature embeddings generated by CXR-FM. The 'output' of this teacher model, therefore, is not directly related to the disease detection task but instead pertains to the feature representation from the penultimate layer. As previously mentioned, to utilise this output, we implement a single, fully-connected classification layer (*nn.Linear-T*) for generating the final outputs. This structure predominantly aligns with **Feature-based KD**, where intermediate layer(s) are used for distillation. One could argue that in our scenario, this is also comparable to

Response-based KD, which primarily focuses on the outputs—specifically, the logits—of a teacher model's final layer. In our case, the 1376-sized embedding vector essentially represents the sole output we are examining. Therefore, the extensive embedding vector essentially serves as the primary output, guiding our focus on KD methods pertinent to both **Feature-based** and **Response-based KD**. For instance, references such as [166] demonstrate the use of the penultimate layer's feature representation for developing a feature-normalised KD approach in image classification. Others have employed features from intermediate hint layers to align student and teacher embeddings [164, 198]. In our work, we consider employing KD from the 1376-sized feature embedding of the teacher to a corresponding feature embedding in the penultimate layer of our student model.

> *Note:* **Relation-based KD**, which further examines the relationships between different layers as well as between inputs and outputs across data samples, could also be adapted—limited to the input CXRs and output embeddings without hidden layer hints. However, given our project's success with simpler matching techniques as found in **Feature-based** and **Response-based KD**, and their satisfactory performance outcomes, we opted to continue with these methods. This choice aligns with our overarching goal of assessing the viability of KD to reconstruct FMs and address transparency issues, rather than optimising for performance. This approach provides a clear and straightforward interpretation and application of basic KD techniques.

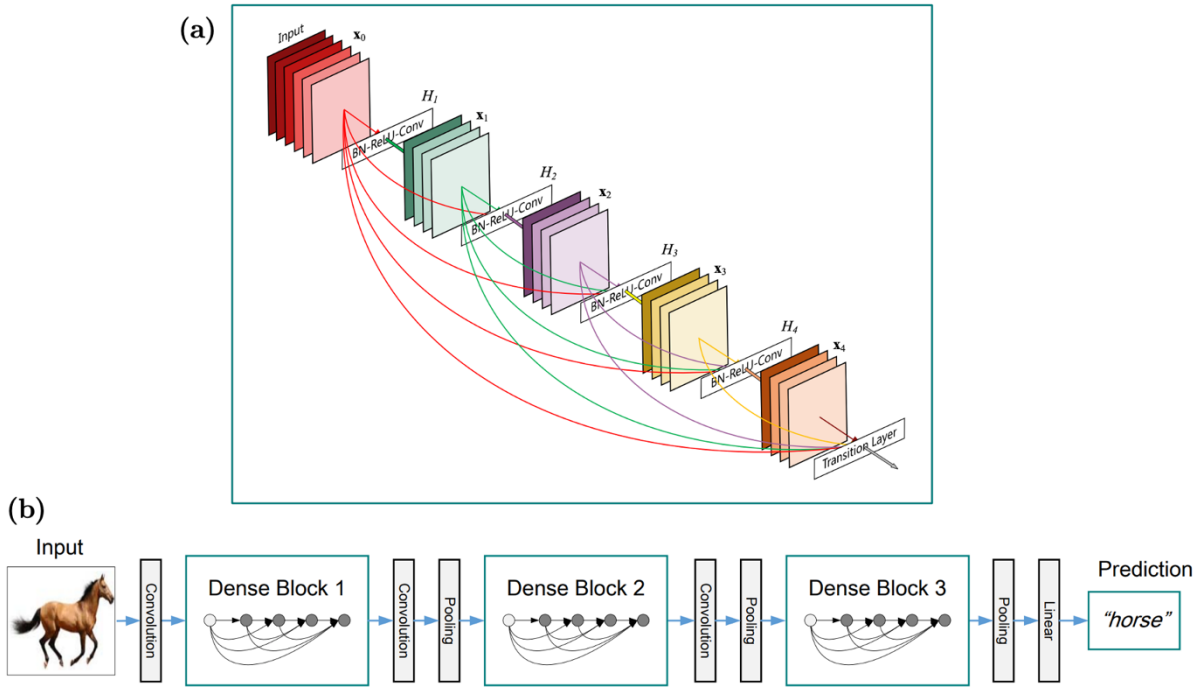

Figure 27. Architectural Overview of DenseNet with Dense Blocks and Transition Layers. Adapted from the seminal paper on DenseNets [47]. **(a)** Illustrates a 5-layer dense block, showcasing how each layer incorporates all preceding feature-maps as inputs, encouraging feature reuse throughout the network. **(b)** Diagram of a DenseNet structured for a simple classification task, featuring three dense blocks. Positioned between these blocks, the transition layers are responsible for resizing the feature-map dimensions via convolution and pooling processes.

Student Model Architecture

Here comes the selection of the student model, which defines the penultimate layer size and any necessary projections to match the teacher and student embedding sizes. Glocker et al. [93, 99]

used the DenseNet121 as a baseline model in their comparative analysis of bias and performance with CXR-FM. This model's weights were initialised from a pre-training on ImageNet-1K, a large dataset of natural images covering 1000 object classes. It was then fine-tined on our disease detection task using the CheXpert and MIMIC datasets. A similar model has been used in the studies by Seyyed-Kalantari et al. [94, 199], showcasing its state of the art performance for such CXR disease detection tasks. Additionally, in their CheXpert paper, Irvin et al. [103] also employed DenseNet121 in their experiments and found that it performed better than the other CNNs they tested, namely ResNet152, Inception-v4, and SE-ResNeXt101. For a visual representation of the DenseNet architecture, refer to **Figure 27**, which illustrates the complex structure of the DenseNet with its dense blocks and transition layers, taken from the original paper [47].

However, DenseNet121 has a penultimate layer size of 1024 before the classification head, which is task-specific and will be replaced in our application with *nn.Linear*, a single fully-connected layer similar to *nn.Linear-T*. In implementing KD between the student and teacher models, we aimed to preserve all information from the teacher, CXR-FM, which maintains its 1376 features. Using DenseNet121 would therefore require an upward projection of its 1024-sized embedding to match the 1376 dimensions of the teacher, which is needed to perform KD. This smaller embedding size led us to opt for a different DenseNet version. Assuming extrapolation of their enhanced performance in this disease classification task—linked to their overall powerful architecture that encourages feature reuse throughout the network [47] as detailed in chapter 2—we chose a DenseNet-169. The latter, positioned just above DenseNet121, has 1664 feature maps at its penultimate layer, larger than the 1376 of CXR-FM, thus allowing for a downward projection to match the teacher's embedding.

The decision to opt for DenseNet169 over DenseNet121 was motivated by the fact that DenseNet-169's penultimate layer, with 1664 features, is closer in size to CXR-FM's 1376 features compared to DenseNet121's 1024 features. This proximity is arguably preferred, ensuring that the student model can effectively learn from the teacher without excessive dimensionality manipulations. Furthermore, the main rationale behind this choice is that reducing the dimensions from a larger number of features (from 1664 to 1376) is generally more information-preserving than expanding a smaller number (from 1024 to 1376) upwards, thus making DenseNet169 our preferred option.

It should be noted that, more broadly in the KD field, the selection of which layers in the teacher serve as the hint layers and which layers in the student should be treated as the guided layers remains an area of active research with no definitive guidelines [164]. This uncertainty also applies to how we match the sizes of the hint and guided layers' feature representations.

KD Losses

Upon selecting DenseNet169 and adjusting its penultimate layer from 1664 to 1376 features using a single fully-connected layer, termed *nn.Linear-KD*, we next needed to choose an appropriate KD method. This involves applying a KD loss that, when minimised during training in our offline distillation setting, aligns the student model's features (predictions) closely with those of the teacher ('ground truth'), which are frozen and serve as a static target.

As outlined in the KD survey by Gou et al. [157], particularly in section 2 and tables 1 - 2, the L_2-norm has been predominantly employed in KD settings similar to ours [182, 183]. The L_2-norm and Mean Squared Error (MSE) are practically identical, differing only in that one takes

the square root of the sum of squared errors, while the other averages them. Other losses considered include the L_1-norm, which is analogous to Mean Absolute Error (MAE) differing only by a constant factor, Huber Loss, Cross-Entropy Loss (CE) , Maximum Mean Discrepancy Loss, and Kullback-Leibler Divergence Loss (KL). Cosine Similarity loss (CS), though rarer and more commonly used in NLP for word vector embeddings, presents a unique approach.

Consequently, we decided to experiment with the following losses: **MSE**, **MAE**, **Huber Loss**—which strategically combines MSE and MAE, CS for its distinctive properties, and **combinations of MSE and CS** to explore the effects of their contrasting characteristics. We opted not to pursue more complex losses such as KL and CE, which are more suited to scenarios involving probability distributions—*typically more relevant in response-based KD scenarios where we have logits from the final output layer which are converted into probabilities via Softmax, often incorporating a temperature parameter to produce soft targets, as discussed in the seminal paper by Hinton et al. on KD [159], which also introduces the concept of 'dark knowledge'.* These losses did not align with our configuration of matching large embedding vectors. Instead, the simpler losses we selected are more compatible with our setup and facilitate a clearer understanding of the knowledge transfer process.

In total, we explored eleven different KD losses. The typical losses such as MSE, MAE, and Huber loss are applied as defined in PyTorch's documentation[4]. Below, we detail the specific implementations and equations for the more nuanced losses only, particularly where combinations or modifications are involved. The eleven losses, with their **names** as used in our experiments, are:

1) **MSE**: Commonly used to calculate the average of the squares of the differences between teacher (target) and student (output) embeddings; termed as L_{MSE}.

2) **MAE**: Calculates the average of the absolute differences between teacher and student embeddings; termed as L_{MAE}.

3) **HuberLoss**: Huber loss, termed as $L_{\text{HuberLoss}}$, combines MSE and MAE strategically. This hybrid loss penalises small errors in a quadratic fashion similar to MSE, while being less sensitive to large errors (outlier) like MAE, offering a more balanced approach.

4) **CS**: Measures the cosine of the angle between the student's and the teacher's embeddings, enhancing alignment. This loss is termed as L_{CS} and defined as follows:

$$L_{\text{CS}} = 1 - \frac{\mathbf{f}_t \cdot \mathbf{f}_s}{\|\mathbf{f}_t\| \, \|\mathbf{f}_s\|}$$

Here, $\mathbf{f}_t$ represents the feature embedding of the teacher model, with a size of 1376, and $\mathbf{f}_s$ represents the corresponding feature embedding of the DenseNet169 student model, which has been projected down from 1664 to 1376. The term $\mathbf{f}_t \cdot \mathbf{f}_s$ is the dot product of these two embedding vectors, and $\|\mathbf{f}_t\|$ and $\|\mathbf{f}_s\|$ are their norms. $\frac{\mathbf{f}_t \cdot \mathbf{f}_s}{\|\mathbf{f}_t\| \, \|\mathbf{f}_s\|}$ is effectively the cosine similarity. To align with our goal of maximising similarity (i.e., minimising the angle between embeddings), we modify the traditional cosine similarity by subtracting it from one, thus framing it as a loss where zero indicates perfect alignment.

[4] *https://pytorch.org/docs/stable/nn.html#loss-functions*

5) MSE-CS Naive: Naively averages MSE and CS, treating both equally without normalising influence based on their scale. This loss is termed as $L_{\text{MSE-CS Naive}}$ and defined as follows:

$$L_{\text{MSE-CS Naive}} = 0.5 \times L_{\text{MSE}} + 0.5 \times L_{\text{CS}}$$

6) MSE-CS Learned: This is an attempt to dynamically balance the contributions of MSE and CS based on learned precision parameters—this is inspired by the concepts of precision and variance in Bayesian statistics. It adjusts the weight of each loss component based on their respective uncertainties, represented by log variance terms. This loss is termed as $L_{\text{MSE-CS Learned}}$ and defined as follows:

$$L_{\text{MSE-CS Learned}} = (e^{-\alpha} L_{\text{MSE}} + \alpha) + (e^{-\beta} L_{\text{CS}} + \beta)$$

Here, α and β are learnable parameters and can be seen as the log variance of L_{MSE} and L_{CS}, respectively. $e^{-\alpha}$ and $e^{-\beta}$ denote the precisions, where higher precision indicates lower uncertainty or variance. Such combination encourages the model to not only minimise the loss but also become more confident (reduce variance).

> *Note*: This approach was explored as a secondary consideration in our study. We primarily focused on fixed combinations of MSE and CS, described in options **7-11**, due to their greater interpretability and clarity. These fixed combinations align more closely with our study's goal as a 'feasibility study' to assess the viability of reconstructing FMs through KD as mentioned previously.

For losses **7-11**: **MSE-CS | α-β**
Fixed weight combinations with specific weights for α and β respectively, as follows: **7)** 0.5-0.5, **8)** 0.6-0.4, **9)** 0.7-0.3, **10)** 0.8-0.2, and **11)** 0.9-0.1.

These combinations integrate MSE and CS with predetermined weights to explore how their distinct characteristics impact the model's downstream performance. Initial explorations indicated that the best performances for the student model, CXR-FMKD, were achieved using KD with MSE. This finding led us to investigate combinations of MSE and CS, with a bias towards favouring MSE, thus assigning it a consistently higher weight relative to CS. This approach is designed to capitalise on the strengths of MSE while still leveraging the unique directional insights provided by CS. The contrast between these two methods—MSE focusing on error magnitude and CS on the alignment of embedding vectors—makes it particularly interesting to see how their combination influences model performance. The combined loss for each configuration is defined as follows:

$$L_{\text{MSE-CS | α-β}} = \alpha \times \overline{L_{\text{MSE}}} + \beta \times \overline{L_{\text{CS}}}$$

$$\text{with } (\alpha, \beta) \in \{(0.5, 0.5), (0.6, 0.4), (0.7, 0.3), (0.8, 0.2), (0.9, 0.1)\}$$

Here, $\overline{L_{\text{MSE}}}$ and $\overline{L_{\text{CS}}}$ represent the MSE and CS losses normalised by their respective average values at the training plateau. This normalisation facilitates a balanced integration of the losses, ensuring that neither dominates due to scale differences during the training process.

KD Process

Now that we have outlined the key initial design choices, we can look into the details of the KD process used to develop the CXR-FMKD models. Refer to **Figure 28** and **Figure 29**, which

illustrate the development of four distinct versions of the CXR-FMKD models based on the KD loss applied. The versions are categorised as follows:

1. CXR-FMKD LP and CXR-FMKD FFT (**Figure 28**): In these models, the *nn.Linear-KD* projector is retained following KD. Additionally, another *nn.Linear-T* single fully-connected layer is integrated to map the 1376 features to the 14 disease labels. In the CXR-FMKD LP variant, both the DenseNet169 backbone and *nn.Linear-KD* are frozen, and only the weights of *nn.Linear-T* are trained. In contrast, the CXR-FMKD FFT variant involves training the full model, updating all weights including those of the backbone and *nn.Linear-KD*.
2. CXR-FMKD-Direct LP and CXR-FMKD-Direct FFT (**Figure 29**): In these variants, the *nn.Linear-KD* projector is removed post-KD, and a 'direct' *nn.Linear* layer is added, mapping from 1664 directly to the 14 disease labels. For CXR-FMKD-Direct LP, the DenseNet169 backbone remains frozen, and only the *nn.Linear* layer's weights are updated. Conversely, in CXR-FMKD-Direct FFT, the entire model, including the backbone, is fine-tuned for the CXR disease detection task.

With a total of eleven KD losses, we end up with 44 CXR-FMKD models (for each dataset) to analyse and compare against the CXR-FM teacher model and the CXR-Model baseline.

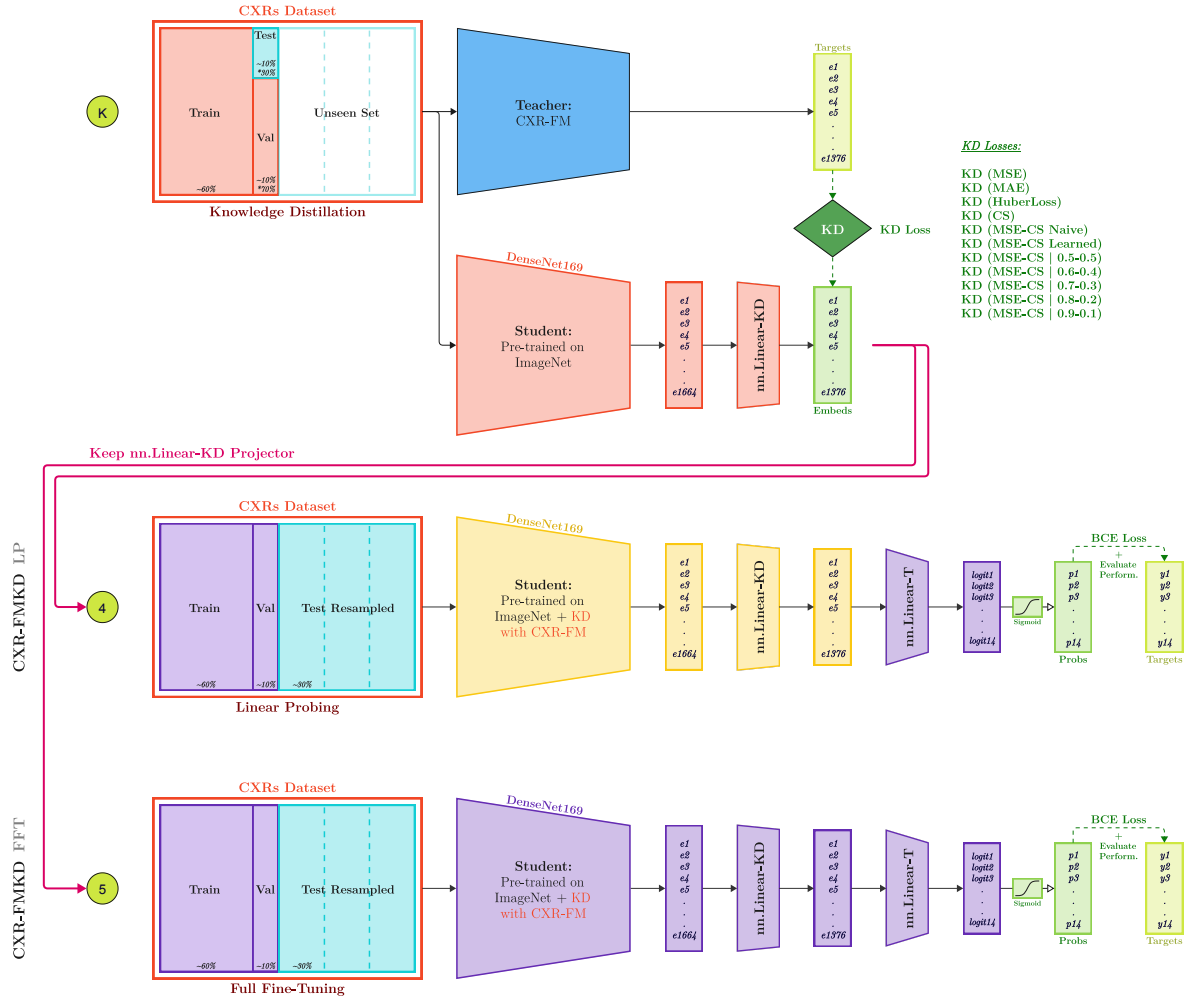

Figure 28. Development and Integration of CXR-FMKD in the CXR Disease Detection Task.
This diagram illustrates the development of our CXR-FMKD student models and their integration into our multi-label disease classification framework, applicable to both the CheXpert and MIMIC datasets. Please refer to **Figure 26** for a detailed legend explaining the technical terms and components used in this and related diagrams.

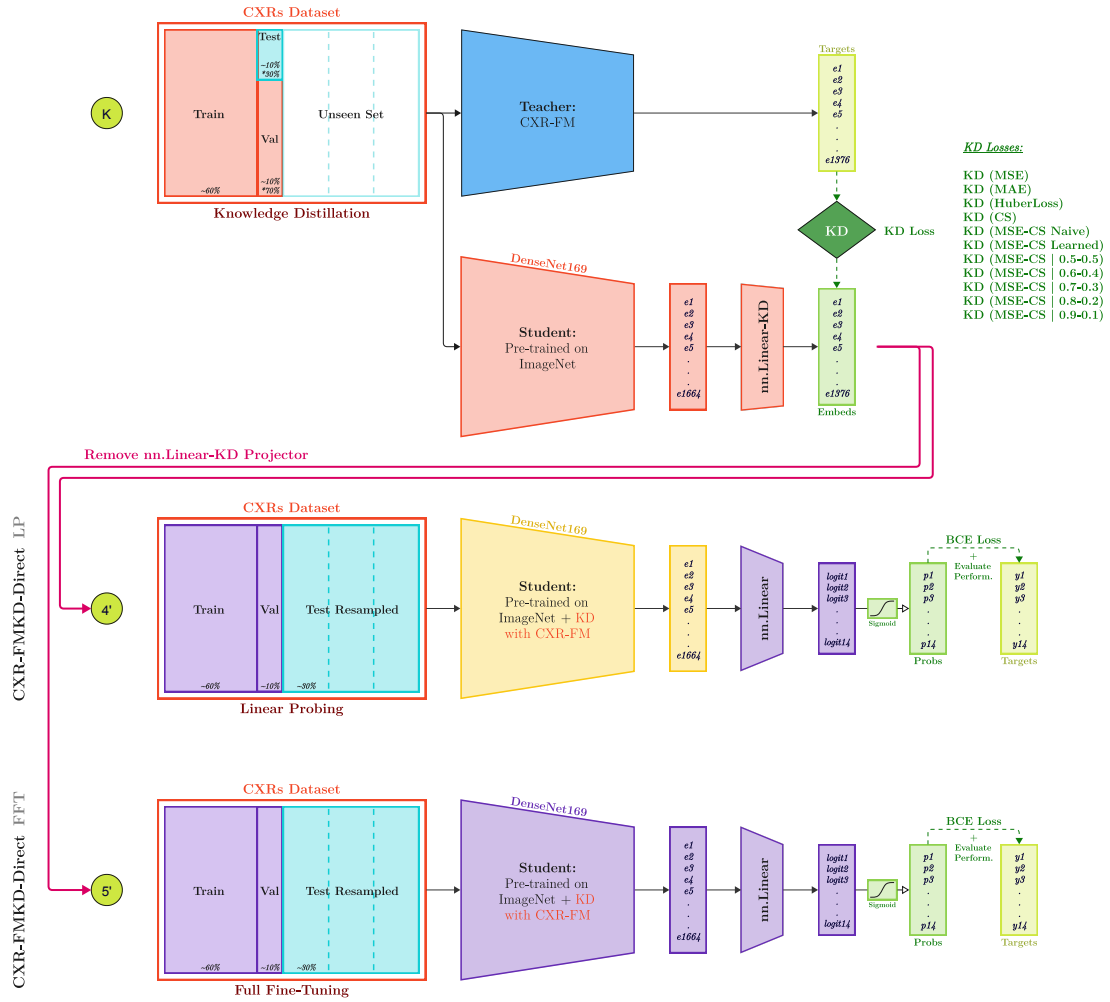

Figure 29. Development and Integration of CXR-FMKD-Direct in the CXR Disease Detection Task. This diagram illustrates the development of our CXR-FMKD-Direct student models and their integration into our multi-label disease classification framework, applicable to both the CheXpert and MIMIC datasets. Please refer to **Figure 26** for a detailed legend explaining the technical terms and components used in this and related diagrams.

Now, in terms of the KD process steps and other pertinent considerations, referring back to **Figure 28** and **Figure 29**, the KD process begins by selecting a transfer set used for transferring knowledge from the CXR-FM to the DenseNet169 student model. We chose to keep the training set unchanged (see **section 3.2.3**), and for the initial validation set, we further divided it into a test-KD subset (30%) and validation-KD subset (70%). The initial test set remains unseen during the KD process, which is crucial as it will be used post-KD for the disease detection task, ensuring the student model's training is not influenced by this data, and its classification performance evaluated on the final test set is faithful.

After adapting the student model to match the teacher's feature size, KD can start. It is important to note that during the experiments, the validation-KD set is utilised for model selection. Although no specific task is being trained for at this stage, the performance of the model is monitored through the relative increase or decrease in the KD loss. The purpose of preserving the test-KD set is to validate the final model's performance and ensure it maintains efficacy on unseen data and behaves as expected. Keeping the training set unchanged aligns with the priority to retain as much data (training instances) as possible within the training phase to facilitate effective KD.

Once KD is complete, the DenseNet169 student model, initially pretrained on ImageNet, now has its weights further refined with the newly acquired 'knowledge' from the CXR-FM

teacher. The model is then adapted for the CXR disease detection task, following the procedures outlined at the beginning of this subsection and as illustrated in the corresponding figures. This post-KD step utilises the original training/validation/test set splits. Ultimately, this process results in the production of our four distinct models: CXR-FMKD LP, CXR-FMKD FFT, CXR-FMKD-Direct LP, and CXR-FMKD-Direct FFT.

In this regard, it is essential to emphasize the rationale behind creating these four models:

- **Creating the 'Direct' CXR-FMKD variants:**
  By removing the *nn.Linear-KD* projector, the model reverts to its original form—the DenseNet169 backbone. Subsequently, the newly added *nn.Linear* classification layer maps this backbone's 1664 features to the 14 disease labels. This approach mirrors common practices in Self-Supervised Learning (SSL), notably discussed in section 3.2, *Role of the Projector*, of the SSL Cookbook by Balestriero et al. [200]. In SSL, a similar layout is often employed where an encoder's output is projected to a smaller embedding using a 2- or 3-layer MLP with ReLU activations in joint embedding methods for contrastive learning. The SSL loss is then applied to these smaller embedding outputs from the projector, and post-training, the projector is typically disregarded. This technique, pioneered by Chen et al. in their seminal SimCLR paper [114], has proven to significantly boost top-1 accuracy on ImageNet by about 20% during a 100-epoch training phase. These SSL dynamics resonate with our approach in KD, and removing the projector to create the CXR-FMKD-Direct variants aims to enhance performance in a similar manner.
- **LP versus FFT:**
  The variations between LP and FFT setups enable comprehensive comparisons against the CXR-FM teacher model. As previously discussed, CXR-FM, which is also referred to as CXR-FM LP, can only be linearly probed due to its frozen backbone when applied to the disease detection tasks. In contrast, the CXR-FMKD (including CXR-FMKD-Direct) models are designed as robust reconstructions of the CXR-FM to address and potentially circumvent transparency issues central to this thesis. The overarching goal of such KD efforts is to produce a reconstructed CXR-FM, the CXR-FMKD, which can then be fully fine-tuned for the disease detection tasks (CheXpert and MIMIC)—its intended application. We anticipate that this FFT approach will match or surpass the performance of the original CXR-FM. Here, also presenting results from the LP approach serves not only to reflect and mirror the operational integration of the CXR-FM but also provides a direct comparison of how this reconstructed, albeit imperfect, replica of the CXR-FM fares when trained in the same manner for the specified disease detection tasks.

All in all, these four versions of the CXR-FMKD student model facilitate a comprehensive overview of performance, allowing for detailed analysis and comparison across the different configurations.

Finally, for completeness,

Table 1 provides a comparative size overview of the DenseNet169 student model against the CXR-FM teacher, which utilises the EfficientNet-L2 architecture [197]. The teacher model is indeed significantly larger, with approximately 34 times more parameters than the student model.

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

Table 1. Comparative Overview of the Teacher and Student Model Sizes.

3.3.4. Baseline Model: CXR-Model

As mentioned earlier, the student model is based on the Densenet169 architecture, with a rationale for this choice detailed in **section 3.3.3**. Accordingly, we utilised the same architecture for our CXR-Model baseline. Its development also involved implementing both LP and FFT setups, resulting in two variants: CXR-Model LP and CXR-Model FFT. These latter mirror the full architecture of the CXR-FMKD-Direct models when adapted for the disease detection task, yet without the prior enrichment of weights via KD from the CXR-FM teacher model. Instead, the CXR-Model is simply a DenseNet169 backbone pretrained on ImageNet and subsequently fine-tuned directly for the CXR disease detection task, with a *nn.Linear* classifier mapping the backbone's 1664 features to the 14 CXR disease labels. Figure X illustrates the development and integration of CXR-Model LP and CXR-Model FFT.

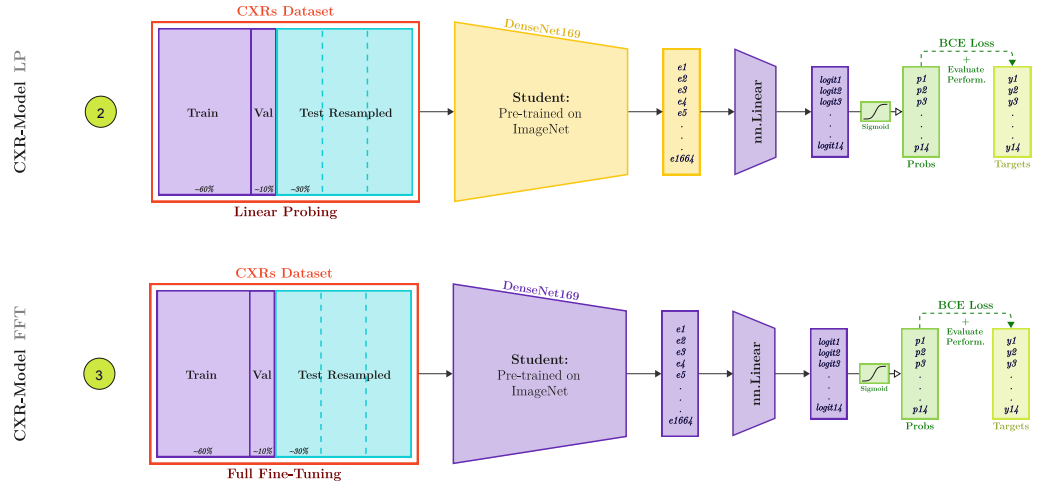

Figure 30. Development and Integration of CXR-Model baselines in the CXR Disease Detection Task. This diagram illustrates the development of our CXR-Model student models and their integration into our multi-label disease classification framework, applicable to both the CheXpert and MIMIC datasets. Please refer to **Figure 26** for a detailed legend explaining the technical terms and components used in this and related diagrams.

The CXR-Model FFT serves as a direct baseline to compare with the CXR-FMKD FFT, and even more directly with the CXR-FMKD-Direct FFT, which shares the same full architecture. This comparison aims to evaluate the effects of having undergone KD versus not, and, in a way, also helps assess how FFT impacts the models, particularly in terms of how extensively FFT overwrites and modifies the CXR input-to-feature mapping. Specifically, it explores whether having passed through KD still makes a difference (positive or negative) or provides an advantage to CXR-FMKD.

On the other hand, CXR-Model LP acts as the baseline for comparison with CXR-FMKD LP and CXR-FMKD-Direct LP, as well as with CXR-FM (also known as CXR-FM LP). This CXR-Model LP is expected to show the poorest performance, as it consists of a frozen backbone pretrained on ImageNet connected to a trainable single fully connected layer *nn.Linear*, mapping the 1664 features to 14 disease classes. This feature embedding represents what a model, which has only learned from natural non-medical images and has no context of CXR, can extract and represent—which is expected to be relatively poor and not very informative or useful for CXR disease classification. Indeed, given such 'information-poor' feature vector as input, this *nn.Linear* is likely not to have sufficient 'capacity' to learn effective classification/discrimination capabilities for the disease detection task, compared to a more complex projector like an MLP. Therefore, CXR-Model LP is a valuable baseline for comparing the utility of the features

generated by the CXR-FM backbone, which has undergone a second pretraining step with CXRs (through SupConv learning) atop its initial non-medical pretraining on ImageNet and JFT-300M datasets [89]. By extension, this comparison also applies to CXR-Model LP versus CXR-FMKD LP and CXR-FMKD-Direct LP, demonstrating how effectively the features generated post-knowledge transfer from the teacher have equipped the student to perform well in CXR disease detection.

Lastly, for reference, we have also trained and evaluated two other versions of CXR-Model FFT using the ResNet50 (named ResNet50 FFT) and DenseNet121 (named DenseNet121 FFT) architectures respectively. ResNet50 is a standard model often employed as a student model in KD literature [157], while DenseNet121 was specifically utilised in the studies by Glocker et al. For ResNet50, we used a *nn.Linear* layer to map from 2048 features to the 14 disease labels, and for DenseNet121, from 1024 features to the 14 labels. This approach allowed us to compare the performance of these backbones directly with DenseNet169, providing a basis for architectural comparison. It was later revealed that our DenseNet169 model achieved similar or better performance compared to DenseNet121 and ResNet50.

3.3.5. Training Setup and Hyperparameters

We adopted a training setup and hyperparameters similar to those used by Glocker et al. [93], with specific adaptations for the KD process, including doubling the number of training epochs and implementing a learning rate (LR) scheduler to enhance convergence and optimisation. Below, **Table 2** summarises the primary hyperparameters employed in our study.

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 2. Summary of the Main Training Hyperparameters.

We chose batch sizes of 128, diverging from Glocker et al.'s use of 150, to adhere to the conventional 'power of 2' rule for optimised GPU memory alignment, although the impact on performance is arguably minimal. For the KD training phase, we observed better convergence and performance with the use of LR schedulers, specifically with the *'OneCycle LR'* [201]. We also experimented with various LRs including 0.1, 0.01, 0.001, and 0.0005, finding that a base LR of 0.001 provided the best results. For the 'OneCycle' LR, we also choose a max LR of 0.01 up to which the initial base LR of 0.001 is increased (annealed), and is then decreased (decayed) down to a rate much lower than the initial one.

Validation loss calculations at the end of each epoch track the model's progress and help define checkpoints for potential early stopping—effectively used for model selection. This strategy allows us to select the model with the lowest epoch validation loss, avoiding models where loss may start to increase again with more epochs—indicative of potential overfitting.

To assess the stability of the models, and to ensure that the results are robust and not dependent on specific initial conditions, we **repeated each experiment five times using seeds 41 through 45** for reproducibility (seeding everything). This process helps control for variability introduced by factors such as dataset shuffling and the random initialisation of added

components like *nn.Linear*, which could otherwise lead to different outcomes in each run. During the KD phase, we selected the student model with the lowest validation-KD and test-KD losses at the end of the designated checkpoint epoch. The 'best' student model from these runs is then used for subsequent CXR disease detection training which was also repeated five times. At the end of this training, and for each of the runs, the best model checkpoint—i.e., the one with the minimum validation loss—is used for final validation and testing to report model performance. We also store the penultimate feature embeddings from these models for later bias analysis.

> *Note:* As previously mentioned, our primary goal is not performance optimisation but to demonstrate the viability of using KD for FM reconstruction to address transparency issues.

Data processing

The data processing steps for the CXR images were designed to both standardise the input size and augment the dataset to improve the generalisability and robustness of our models. Here are the details of the process:

1. **Image Resizing:**
   - All CXR images were resized to 224×224 pixels, also aligning with the default input size for DenseNet169. This ensures that the network receives uniformly sized inputs.
2. **Data Augmentation:**
   - *Horizontal Flipping:* Images are randomly flipped horizontally to simulate different orientations, reflecting the variability that can occur in real-world imaging scenarios.
   - *Affine Transformations:* Images undergo random rotations of up to 15 degrees and scaling between 90% and 110%. These transformations help the model to recognise radiological features under various conditions and from different angles, enhancing its ability to handle geometric variations in clinical settings.
3. **Pseudo-RGB Conversion:**
   - We convert our grayscale CXR images to pseudo-RGB by replicating the single channel across all three RGB channels. More generally, this approach allows the utilisation of architectures pre-trained on RGB datasets.
4. **Data Loading:**
   - The training set is shuffled to ensure diverse batches of data during training, which aids in effective learning. In contrast, the validation and test sets are not shuffled, maintaining a consistent order for performance evaluation.
   - Efficient parallel data processing is enabled by setting the number of worker threads to four to optimise the loading speed.

These processing steps are crucial for preparing the imaging data effectively, ensuring that the models are not only trained on well-standardized data but also exposed to a range of simulated clinical variations through augmentation, enhancing their robustness.

3.3.6. Implementation

Software Frameworks and Libraries

Our project was implemented using *Python 3.11.4*, with *PyTorch Lightning* facilitating the management of the model training, evaluation lifecycle, and the loading and handling of our data. Additionally, models were constructed using *PyTorch*'s modular components. These include, but are not limited to, the *nn.Linear* which was used to model the single fully-connected

classification layer, the sigmoid function to convert logits to probability scores for our multi-label classification problem, the various losses used during standard and KD training phases, and the *DenseNet169* architecture that comes with weights pre-trained on ImageNet[5]. Additional libraries such as *skimage* for image resizing and manipulation, and *pandas* for dataset manipulation and querying, were also used. The full project code is available at: *https://github.com/FadiZahar/CXR-Foundation-Model-Knowledge-Distillation*

Monitoring and Logging

Training progress and key metrics such as losses were monitored using the *Weights & Biases* (WandB)[6] platform. The latter provides comprehensive logging capabilities, essential for managing outputs, performance, and experimental artifacts.

Hardware and Computational Resources

Model training and testing required high-end GPU workstations. For our experiments, two main GPU partitions were used:

1. gpus24 Partition:
   - GPUs: Workstation-grade RTX 3090 / 4090 cards.
   - GPU Memory: Each GPU has 24 GiB of VRAM.
   - System Specs: Accompanied by 62 GiB RAM and 12 CPU cores.
   - Usage: Used for various tasks in the project, predominantly when lesser memory was sufficient.
2. gpus48 Partition:
   - GPUs: Server-grade Ada A6000 / L40 cards.
   - GPU Memory: Each GPU has 48 GiB of VRAM.
   - System Specs: Accompanied by 125 GiB RAM and 16 CPU cores.
   - Usage: Used for running the KD codes (due to the high-dimensional embeddings being handled) and other demanding high-memory tasks including standard training for the CXR disease detection task.

3.4. Performance Analysis

Performance Metrics

In this section, we focus on key metrics that provide a comprehensive understanding of our models' performance, specifically the *Area Under the Receiver Operating Characteristic Curve (AUC-ROC)*, the *Area Under the Precision-Recall Curve (AUC-PR)*, *Max Youden's J Statistic*, and *Youden's J Statistic at a decision threshold of 20% FPR*. These are standard metrics used to evaluate the performance of classification models, offering important insights into their effectiveness:

- **AUC-ROC:** This metric is derived from the ROC curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. The AUC-ROC quantifies the overall ability of the model to distinguish between classes across all possible thresholds. It is particularly valuable in situations with imbalanced

[5] *https://pytorch.org/vision/main/models/generated/torchvision.models.densenet169.html*
[6] *https://wandb.ai/site/*

classes like CheXpert and MIMIC. A perfect model has an AUC-ROC of 1, indicating perfect discrimination between classes, while an AUC-ROC of 0.5 suggests performance no better than random chance.

- **AUC-PR:** This metric focuses on the precision and recall for the positive class, offering a more detailed view in scenarios where the prediction of the positive class is crucial. Like AUC-ROC, AUC-PR ranges from 0 to 1, with higher values indicating better model performance. It should also be noted that AUC-PR is more sensitive to class imbalance than AUC-ROC.

  AUC-ROC and AUC-PR are usually applied to binary classification tasks; however, they can be easily adapted to our multi-label classification setting by treating each label as a separate binary classification problem. Therefore, in our case, we end up with AUC-ROC and AUC-PR scores for each of the 14 CXR disease labels.

- **Youden's J Statistic** is calculated as TPR−FPR, offering a single-number summary of the ROC curve. The **Max Youden's J Statistic** refers to the maximum value of this statistic across all possible decision thresholds, highlighting the best possible performance of the model. It is useful for identifying the optimal decision threshold that balances true positives and false positives most effectively.

- **Youden's J Statistic at 20% FPR** specifically sets the False Positive Rate at 20%, serving as a fixed decision threshold. This metric is particularly useful for directly identifying performance disparities across patient subgroups, facilitating targeted analyses in our bias analysis section. It also reflects a strategic choice in clinical settings where maintaining a moderate level of FPR is crucial.

These metrics align with those used in the relevant studies for our research: Glocker et al. [93] utilised AUC-ROC and Youden's J Statistic @ 20% FPR for subgroup performance analysis, while the CheXpert study [103] employed AUC-ROC and AUC-PR. Additionally, we incorporate Max Youden's J Statistic to highlight the maximum potential model performance.

Most Significant Classes

In our analysis, we focus on a subset of the 14 CXR disease labels to simplify evaluation and draw clearer conclusions. Glocker et al. [93] concentrated their analyses on the metrics for '*No Finding'*, '*Pleural Effusion'*, '*Cardiomegaly'*, and '*Pneumothorax'*, labels under which models typically perform better (higher scores), facilitating clearer detection of performance disparities. Additionally, from the CheXpert study, Irvin et al. focused on the evaluation of the five most clinically significant labels due to their prevalence, namely '*Atelectasis'*, '*Cardiomegaly'*, '*Consolidation'*, '*Edema'*, and 'Pleural Effusion'. To maintain alignment with Glocker et al. while broadening our monitoring scope based on relevance, we combine insights from both studies and select seven disease labels deemed most significant for detailed analysis. These labels are (c1) '*Pleural Effusion'*, (c2) '*No Finding'*, (c3) '*Cardiomegaly'*, (c4) '*Pneumothorax'*, (c5) '*Atelectasis'*, (c6) '*Consolidation'*, and (c7) '*Edema'*. We will present performance metrics scores for each of these in our results section. The remaining seven less critical labels, such as fractures and support devices, are grouped into an '*Others'* category and averaged for a consolidated score. Critical decisions and interpretations in terms of model performance are based on the average scores of these seven prioritised diseases (given equal weighting). For completeness, we also calculated the macro-average performance score across all 14 classes.

Performance Outcomes

We perform this detailed performance analysis separately for the CheXpert and MIMIC datasets. Consequently, we examine a total of 49 models for each dataset, encompassing CXR-FM (×1), CXR-Model LP (×1), CXR-Model FFT (×1), ResNet50 FFT (×1), DenseNet121 FFT (×1), CXR-FMKD LP (×11), CXR-FMKD FFT (×11), CXR-FMKD-Direct LP (×11), and CXR-FMKD-Direct FFT (×11)—recall the CXR-FMKD being variants subjected to the 11 different KD losses. This results in 98 model analyses across the two datasets. As mentioned in **section 3.3.5**, to ensure the robustness and reproducibility of our results, each model's training was conducted five times using different seeds, leading to a total of 490 individual model outputs for this performance analysis section. This repetition of experiments enables the calculation of both average performances and standard deviations for each model, offering a statistically more robust summary that aids in assessing the consistency and reliability of our findings and patterns.

3.5. Generalisability Analysis

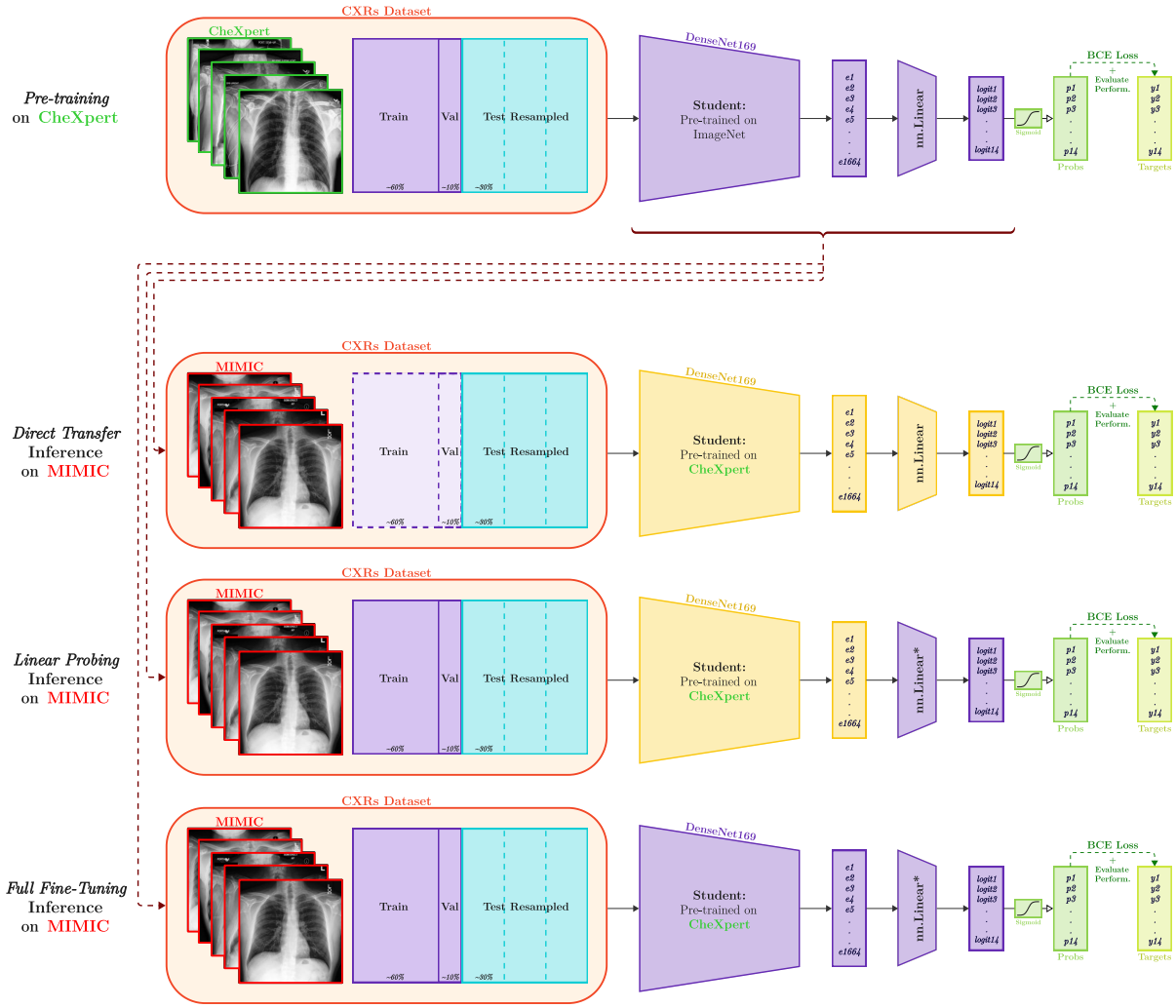

Figure 31. Inference Testing Framework for Evaluating Model Generalisability.
Diagram representing the approach to assess generalisability of models, illustrating the CXR-Model FFT pre-trained on CheXpert and tested on MIMIC. The process stratifies testing into three fine-tuning strategies: 'Direct Transfer' without adaptation, 'Linear Probing' (LP) involving fine-tuning of the final classifier layer, and 'Full Fine-Tuning' (FFT) across all layers. Please refer to **Figure 26** for a detailed legend explaining the technical terms and components used in this and related diagrams.

As laid out in **section 3.1**, to evaluate the generalisability of our CXR-FMKD models, we conducted inference testing experiments to assess their performance on OOD data. For this analysis, models initially trained on the CheXpert dataset were subsequently tested on the MIMIC dataset using various levels of fine-tuning to probe their adaptability and robustness. These datasets share 14 disease labels, facilitating a direct comparison across the three different testing strategies. **Figure 31** illustrate this inference testing framework, which we further describe below:

1. **Direct Transfer**: Models pre-trained on CheXpert were directly tested on MIMIC's test set without any adaptation. Only the test set of MIMIC is used in this case, leaving the training and validation sets untouched. This strategy evaluates the raw ability of the models to perform under new conditions without any adjustments, made possible by the shared 14 CXR disease classes between the two datasets.

2. **Linear Probing (LP)**: Here, the final classification layer (*nn.Linear*) originally trained on CheXpert is replaced with a new single fully-connected *nn.Linear** layer with randomly initialised weights. This new layer is then fine-tuned using the training and validation sets of MIMIC. This method simulates how a model might be adapted to a new dataset, where direct transfer is not possible due to different output classes between the two datasets/tasks.

3. **Full Fine-Tuning (FFT)**: Similar to LP, the final classification layer is replaced, but unlike LP, the entire model, including all layers, is now fine-tuned on MIMIC. This extensive fine-tuning process is meant to fully adapt the model to the specifics of the new dataset.

For each of these strategies, we implemented a systematic selection process to choose the representative model from CheXpert training to carry forward into MIMIC testing as part of the generalisability analysis. Given that each model configuration was trained five times with different seeds, we selected the model iteration that showed the closest performance to the mean outcomes of these five runs, focusing on the average for the seven most significant disease labels (classes 1-7) as previously defined. The selection was based on the model with the smallest sum of absolute differences between its individual performance metrics (AUC-ROC, AUC-PR, Max Youden's J, and Youden's J @ 20% FPR) for the averages of classes 1-7, and the corresponding average metrics across all runs. This selection ensures that the model used for generalisability testing represents the typical performance output, rather than an outlier.

To ensure robustness in our generalisability analysis, each experiment was also repeated five times using different random seeds. We utilised the same four performance metrics defined in the previous section to evaluate the models across varying datasets. In terms of benchmark comparisons, for each of the three inference frameworks—Direct Transfer, LP, and FFT—we compared the resulting models against their counterparts that were exclusively trained on the MIMIC dataset, as identified in the *Performance Analysis* section. However, not all models from the performance analysis were included in this comparison. We selected a subset of the most relevant and best-performing models for a focused analysis. These include:

- CXR-FM (×1)
- CXR-Model LP (×1) and CXR-Model FFT (×1)

- CXR-FMKD models: We selected three instances (×3) each from CXR-FMKD LP, CXR-FMKD FFT, CXR-FMKD-Direct LP, and CXR-FMKD-Direct FFT. These selections were based on the KD losses that derived the best performing student models.
- We excluded ResNet50 FFT and DenseNet121 FFT models from this comparison as the CXR-Model FFT showed similar or superior performance.

It should be noted that in the LP scenario, models undergo only minimal adjustments as the backbone is frozen, making the CXR-FM effectively equivalent to its benchmark counterpart. For the FFT scenario, CXR-FM also remains similar to the benchmark as its backbone is frozen and cannot be fine-tuned.

3.6. Bias Analysis

This section builds on the work by Glocker et al. [93], which demonstrated bias in Google's CXR-FM. We will employ similar techniques to analyse potential biases in our newly constructed CXR-FMKD models, allowing us to draw direct parallels between our findings.

3.6.1. Bias Inspection

Overview

For the bias analysis, we will separately analyse the models derived from the CheXpert and MIMIC datasets, which were previously discussed in the performance analysis section. It is important to note that the test set resampling, designed to correct prevalence variations across patient subgroups by race, age, and disease, plays a crucial role in this bias analysis.

We adopt the feature exploration method proposed by Glocker et al. [93, 99] to examine potential biases in the feature representations extracted from the CXRs by our models. Specifically, we inspect the feature embeddings from the penultimate layer—the output from the backbone—saved during the *performance analysis* after testing on both CheXpert and MIMIC test datasets. These embeddings are crucial as they represent the information the model uses for disease classification. The decision-making process is conducted by the classifier, in this case, a *nn.Linear* single fully-connected layer that maps these features to 14 disease classes. Biases in these feature representations could inadvertently influence the classifier's decisions, leading it to rely on unfavourable correlations related to patient sex or race as shortcuts for disease classification, rather than focusing on relevant medical indicators.

In their original form, the feature embeddings from our models are high-dimensional—1376 dimensions for CXR-FM and 1664 dimensions for the CXR-Model and CXR-FMKD models—making direct inspection challenging. To facilitate a more manageable analysis, we employ dimensionality reduction techniques such as principal component analysis (PCA) [202] and t-distributed stochastic neighbour embedding (t-SNE) [203], which project such the high-dimensional embedding to lower dimensional feature spaces. In PCA, the principal components, or modes, represent the axes along which the data's variance is maximised. The first mode captures the largest variance, explaining the most significant differences among the data points, and each subsequent mode captures progressively less variance. For a model developed for CXR disease detection, which is trained to discriminate between the presence and absence of disease, these initial modes are critical as we expect them to show the strongest discrimination between data sample with and without disease. After identifying these key PCA modes, set up to

collectively capture 99% of the variance, t-SNE is applied on top of them. This strategy helps preserve the integral structure and relationships within the original high-dimensional feature space, ensuring a detailed and accurate representation in a reduced dimensional space.

In inspecting for bias, Glocker et al. randomly selected a cohort of 3000 CXRs, ensuring an even distribution with 1000 samples from each racial group. This sample set was used to explore whether the PCA dimensions, primarily distinguishing disease presence, might also inadvertently segregate patient characteristics unrelated to the disease, such as biological sex (male or female) or racial identity (White, Black, or Asian). Furthermore, they extended this examination to t-SNE projections to ascertain if any noticeable groupings or distributional disparities emerged among patient subgroups, hinting at similar discrimination based on protected characteristics. Observations of differences in the marginal distributions for race and sex subgroups in PCA and t-SNE analyses using this sample set could suggest that the features derived by the model encapsulate variations not solely confined to disease conditions but potentially encompassing demographic attributes as well. This could suggest that the models, while effective in disease detection, might also be learning and perpetuating existing biases present in the training data.

Referring back to this balanced sample set to be examined, it is pertinent to consider the implications of test set resampling. Indeed, extracting feature embeddings for bias analysis from a well-balanced test set ensures a faithful assessment of any inherent biases. For example, consider an unbalanced sample set where 50% of the patients are male and 50% are female; if 100% of the males have '*No Finding'* and 100% of the females have '*Pleural Effusion'*, a strong separation between male and female samples might be observed. However, this separation would be attributable not to sex-based bias but to the disease distribution within the sample set. Such disparities can similarly affect other demographic factors like age or racial groups, leading to potentially misleading conclusions about bias. The example underscores the necessity of balancing for demographics and disease prevalences to ensure that the PCA modes or other analytical methods reveal true biases related to model behaviour rather than artifacts introduced by skewed prevalences.

Statistical Analysis

To systematically investigate potential biases in the PCA-reduced feature space, Glocker et al. examined the distributional differences between subgroups. This analysis helps discerning whether the disease detection model discriminates based on characteristics unrelated to the disease, such as race or sex. To quantify these distributional differences, Glocker et al. employed two-sample Kolmogorov-Smirnov (KS) tests. These tests generate p-values for the null hypothesis that the distributions of two compared subgroups are identical across the first four PCA modes, which capture the most significant variance and are thus most indicative of the model's behaviour. These statistical tests are applied across various subgroup pairings, including racial comparisons (White vs. Black, Black vs. Asian, and Asian vs. White) and sex (male vs. female), as well as across disease prevalences ('Pleural Effusion' vs. 'No Finding'). To correct for multiple testing in these comparisons, p-values were adjusted using the Benjamini-Yekutieli procedure, ensuring that findings are statistically significant at a 95% confidence level ($p < 0.05$).

Novel Bias Score

In our approach, we extended the bias inspection framework to mitigate variations in results depending on the specific 3000-patient set inspected (sampled randomly) and to automate the

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%		
		100%	**FALSE**	**0%**	97%	71%	73%	95%		
			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 32. 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.

analysis by introducing a single bias score. This score replaces reliance on visual assessments of marginal distributions and tables of p-values across the first four PCA modes for all subgroups [93], a necessity given the extensive number of models we are analysing in this study.

Initially, the presence of duplicates in the resampled test set—introduced by the sampling with replacement—could distort the analysis of marginal distributions and explain some variances in the p-values results observed. These duplicates often create artifacts in the PCA visualisations, appearing as unnatural 'bumps' in what should be smoother distribution curves, misleadingly suggesting significant differences where there might be none. Statistical tests like KS, sensitive to such discrepancies, might incorrectly deem similar distributions as significantly

different ($p < 0.05$) as a result. To overcome this, we removed duplicates from the resampled test set, termed test-set-cleaned, which reverted to being imbalanced. Subsequently, a stratified sampling approach without replacement was used to extract a set for examination from test-set-cleaned. This approach was guided by predefined prevalences based on factors such as sex, disease status, and age group, ensuring that each subgroup was proportionally represented according to the average distributions across all races in test-set-cleaned. Specifically, we ensured a balanced representation by sampling 1000 individuals from each racial group, similar to what was done by Glocker et al., and resulting in a sample set of 3000 patients as well. This approach resulted in more consistent p-values when comparing marginal distributions across different subgroups, thus enhancing the reliability of our bias analysis. However, some variability still remained.

To further enhance the robustness of our bias analysis, we adopted a (statistical)-bootstrapping-like approach by repeating the balanced stratified sampling of the 3000-patient set from the test-set-cleaned 5000 times, using seeds from 1 to 5000.
In line with creating a single bias score, we categorised the p-values from the statistical tests into three groups:

- **FALSE ($p > 0.05$)**: Indicates no significant differences in marginal distributions, suggesting an absence of bias.
- **TRUE ($0.001 < p < 0.05$)**: Indicates significant differences in marginal distributions, suggesting the presence of bias.
- **TRUE+ ($p < 0.001$)**: Indicates very significant differences in marginal distributions, suggesting a strong presence of bias.

Scores are assigned to these categories to facilitate a quantifiable analysis: 0 for FALSE, 100 for TRUE, and 150 for TRUE+, with the aim to capture a score range from 0 (no bias) to 150 (significant bias). We refer to these as 'p-scores'.

The novel bias score calculation process is illustrated in **Figure 32**, which shows the results for the CXR-FMKD-Direct FFT model trained with an MSE loss during KD. Note that for bias analysis, we are only interested in the columns for the race and sex attributes, and not for disease labels. This procedure involves several steps:

1. **Recording Results**: We document the count of FALSE, TRUE, and TRUE+ outcomes from the 5000 simulations (STEP 1 in **Figure 32**).
2. **Calculating Percentages**: We convert these counts into percentages (STEP 2).
3. **Weighting by PCA Explained Variance**: We normalise the explained variance for the first four PCA modes so their sum equals 1. Each PCA mode's contribution to the bias score is weighted by its relative importance in explaining the variance—equivalent to multiplying the percentages by the normalised explained variance in the same row (STEP 3).
4. **Combining Results**: For each demographic comparison (e.g., White vs Asian), we sum the weighted percentages for FALSE, TRUE, and TRUE+ outcomes. These are then combined with the assigned p-scores to calculate a bias score for each column. We average these scores by demographic attribute categories (e.g., *Race*, *Sex*) to yield comprehensive bias scores for each. The *Overall* bias score is further computed as the average of these category scores.

For instance, as shown to **Figure 32**, the results for the 'White vs Asian' comparison are 97%, 3%, and 0% for FALSE, TRUE, and TRUE+ respectively across the PCA modes. Applying the assigned p-scores results in a bias score of 3.15 for this comparison. Similarly, we obtain a score of 40.73 for 'White vs Black', 39.40 for 'Asian vs Black', and 5.30 for 'Male vs Female'. These

scores are then aggregated by demographic category, yielding a bias score of 27.76 for *Race* and 5.30 for *Sex*. By averaging these, we get an *Overall* bias score of 16.53.

As a brief interpretation to put into context, the relatively low scores on the **0-150 scale** suggest that the model does not exhibit significant biases in the demographic comparisons studied, as observed through the first four PCA modes. This indicates that the model does not unduly discriminate based on protected characteristics, such as race and sex. Ideally, this suggests that the model's feature representations are focused on medically relevant cues rather than demographic biases that could arise from an unbalanced training dataset. Furthermore, while the bias levels are generally low, the analysis does reveal a higher tendency for racial bias than for sex-based distinctions. Particularly, the comparisons 'White vs. Black' and 'Asian vs. Black' show more pronounced biases than 'White vs. Asian', indicating a specific area where bias mitigation efforts could be focused. In contrast, for disease detection labels like 'Pleural Effusion vs No Finding', we observe a 100% TRUE+ rating, leading to the maximum bias score of 150. This result is expected and validates the model's ability to distinguish effectively between different medical conditions, which is its primary function. However, it also highlights the importance of ensuring that this strong discriminative capability does not extend inappropriately to non-clinical attributes, underscoring the ongoing need to monitor and address potential biases.

Concluding notes:

- We focused solely on PCA modes for calculating bias scores, and generally, the results from t-SNE are consistent with those from PCA.
- The rationale behind weighting the p-value results by the explained variance of PCA modes is to prioritise the influence of modes that explain more variance, thereby reflecting their greater importance in the feature space.
- The advantage of our single bias score is that it provides a concise, unified measure of bias that can also be dissected to reflect specific demographic subgroup comparisons and attributes, offering separate scores for race and sex. This facilitates a nuanced understanding of where biases may be more pronounced within the model.
- The decision to employ repeated sampling (5000 simulations) for our bias inspection was guided by its conceptual similarity to *statistical bootstrapping*, which we believed would enhance the robustness of our findings by mitigating variations in p-value outputs. However, further validation with statisticians is recommended to solidify the statistical underpinnings and guarantees of this approach. It should be noted that our focus on relative performance comparisons among our models in this study helps ensure that our conclusions remain robust, despite potential limitations in absolute terms.

3.6.2. Subgroup Performance Analysis

As discussed in the *Bias Inspection* section, our methodology, akin to that used by Glocker et al. [93], assesses disease detection capabilities across different demographic subgroups. Our primary metrics for evaluation are *AUC-ROC* and *Youden's J statistic* determined at a fixed decision threshold, optimised to achieve a 20% FPR on the entire patient sample. This consistent criterion allows for direct observation of performance variations across subgroups.

To provide a more granular and statistically robust analysis, we implement bootstrapping with 2000 samples to compute 95% confidence intervals. This approach is crucial for our bias analysis where we select the best-performing model out of five, each trained with different seeds. Since the unique feature embeddings of these models characterise their performance and cannot be

averaged across models, the bootstrapping method enhances the statistical robustness of our results, enabling us to draw more reliable conclusions about each model's effectiveness in detecting diseases across varied demographic subgroups.

Chapter 4

Results

In this chapter, we will present the outcomes from the experiments outlined in Chapter 3, namely the performance analysis, generalisability analysis, and bias analysis, which includes both bias inspection and subgroup performance analysis. Each section will be accompanied by a discussion, where we interpret the findings in context.

4.1. Performance Analysis

4.1.1. Knowledge Distillation Exploration

The development of our CXR-FMKD models started with an exploration of various KD losses to effectively transfer knowledge from Google's proprietary CXR-FM model to the student models. We explored eleven different losses including MSE, MAE, HuberLoss, CS, and combinations of MSE and CS, namely: MSE-CS Naive, MSE-CS Learned, MSE-CS | 0.5-0.5, MSE-CS | 0.6-0.4, MSE-CS | 0.7-0.3, MSE-CS | 0.8-0.2, MSE-CS | 0.9-0.1. For detailed descriptions for each of the *KD Losses*, please refer to **section 3.3.3**.

Notations: It is important to note that KD losses specified in brackets next to a CXR-FMKD model's name indicate the specific loss used during the KD process with CXR-FM. Additionally, for clarity in discussions throughout this section: when referring to '*LP-variants*' for the CXR-FMKD models, we include both CXR-FMKD LP and CXR-FMKD-Direct LP models. Similarly, '*FFT-variants*' encompasses both CXR-FMKD FFT and CXR-FMKD-Direct FFT models. The term '*Direct-variants*' refers collectively to CXR-FMKD-Direct LP and CXR-FMKD-Direct FFT models, whereas '*non-Direct-variants*' refers to CXR-FMKD LP and CXR-FMKD FFT models. This grouping helps in the analysis by categorising models based on their adaptation strategies for the disease detection task.

CheXpert

Figure 33 showcases a custom parallel coordinate plot that illustrates the performance metrics—*AUC-ROC*, *AUC-PR*, *Maximum Youden's J Statistic*, and *Youden's J Statistic at 20% FPR*—for our models tested on CheXpert, focusing on the average results for the seven *Most Significant Classes* (1 to 7) as defined in **section 3.4**. Each line in the figure represents the average performance results from testing five distinct instances of the same model type, each developed and trained using a different seed to ensure robustness and repeatability. The variability in performance is illustrated by the shaded areas around each line, indicating the standard deviation (SD). Each CXR-FMKD plot line is colour-coded according to the type of

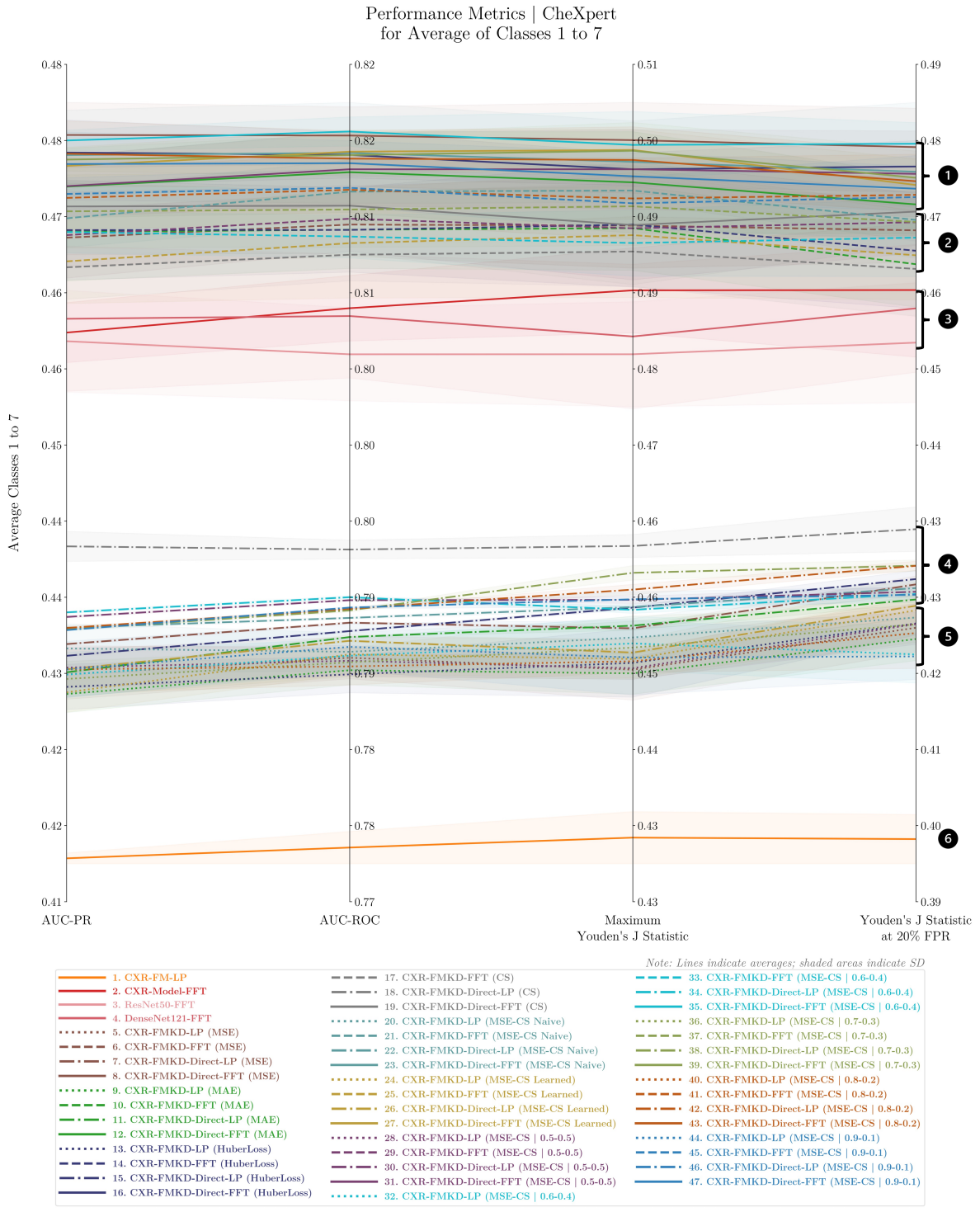

Figure 33. Comparative Analysis of Performance Metrics Across 47 Models for CheXpert Dataset. This custom parallel coordinate plot visualises the performance metrics—AUC-ROC, AUC-PR, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR—for 47 models tested on the CheXpert dataset, focusing on the average results for the most significant disease labels (Classes 1 to 7). Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. We observe a clear stratification in model performance, grouped by type from top (best performers) to bottom (least performers): CXR-FMKD-Direct FFT models ❶, CXR-FMKD FFT models ❷, Benchmark Models ❸ (i.e., CXR-Model FFT, ResNet50 FFT, DenseNet121 FFT), CXR-FMKD-Direct LP models ❹, CXR-FMKD LP models ❺, and lastly the original CXR-FM ❻.

KD loss used during the model's development, aiding in visual differentiation. The plot also employs unique visual markers to identify the CXR-FMKD model types: CXR-FMKD LP models are depicted with dotted lines, CXR-FMKD-Direct LP models with dash-dotted lines, CXR-FMKD FFT models with dashed lines, and CXR-FMKD-Direct FFT models with solid lines. More generally, except for the original CXR-FM—which can only be integrated via LP due to its frozen backbone—*LP-variants* models features dots in their lines, while *FFT-variants* models do not. In this regard, models with full solid lines can effectively be seen as the best possible (and intended) integration, in terms of fine-tuning for the downstream detection task, of each model type. Lastly, it is important to emphasize that the parallel coordinate plot displays each performance metric on its unique scale. The vertical axes are scaled to encompass the full range of observed values—from the minimum to the maximum—for each metric across all models.

A distinct stratification in model performance, grouped by type from top (best performers with higher recorded values for each performance metric) to bottom (worst performers with lower recorded values), can be observed (❶>❷>❸>❹>❺>❻):

1. **CXR-FMKD-Direct FFT** models ❶
2. **CXR-FMKD FFT** models ❷
3. **Benchmark** models ❸ (CXR-Model FFT, ResNet50 FFT, and DenseNet121 FFT)
4. **CXR-FMKD-Direct LP** models ❹
5. **CXR-FMKD LP** models ❺
6. Original **CXR-FM** ❻

Although we tested 49 model types in total, only 47 are displayed in **Figure 33**. We excluded the CXR-Model LP and CXR-FMKD LP (CS) because they significantly underperformed compared to the others, which distorted the visual representation of the data (as seen in **Figure 51**). Specifically, CXR-FMKD LP (CS) showed the lowest performance among all models tested, with considerable variability indicated by the large SD in the shaded area around its line. Conversely, CXR-FMKD-Direct LP (CS) exhibited the best performance among the CXR-FMKD-Direct LP models, as seen in **Figure 33**. This suggests a nuanced impact of the KD CS loss applied, which will be discussed later.

We now dig deeper into the performance stratification noted earlier. We observe that the CXR-FMKD FFT-variants (❶,❷) achieved higher performance metrics than their LP-variants counterparts (❹,❺). Within each of these categories, the Direct-variants outperformed the non-Direct-variants (❶>❷, ❹>❺). Regarding the benchmark models ❸—CXR-Model FFT (Densenet169), ResNet50 FFT, and DenseNet121 FFT, which were all developed using FFT—they performed better than the CXR-FMKD LP-variants, but came just below our FFT-variants. Notably, we observe two main (pronounced) performance gaps separating our model types: one between the FFT models (CXR-FMKD FFT-variants + benchmark models) and the CXR-FMKD LP-variants; and another between these LP-variants and the original CXR-FM (❶,❷,❸, gap-1, ❹,❺, gap-2, ❻). Here, the CXR-FM ❻ sits at the lower end of the performance spectrum in the parallel coordinate plot, performing worse than the other 46 models, but it still outperforms the omitted CXR-Model LP and CXR-FMKD LP (CS) models.

Looking within each stratum category:

- **CXR-FMKD-Direct FFT** ❶:
  - The models derived from KD with MSE and MSE-CS | 0.6-0.4 losses are at the top—and are effectively the best among all models tested on CheXpert. Here, CXR-FMKD-

Direct FFT (MSE-CS | 0.6-0.4) performs slightly better in terms of *AUC-ROC* and *Youden's J Statistic at 20% FPR* compared to CXR-FMKD-Direct FFT (MSE), while the latter achieves slightly better results in terms of *AUC-PR* and *Maximum Youden's J Statistic*.

- The least performing models within this group are CXR-FMKD-Direct FFT (CS), which noticeably falls into the performance range of the CXR-FMKD FFT category ❷ below, followed by CXR-FMKD-Direct FFT (MAE). Additionally, CXR-FMKD-Direct FFT (MSE-CS | 0.5-0.5) ranks low for *AUC-ROC* and *AUC-PR*, and CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1) ranks low for *Maximum Youden's J Statistic* and *Youden's J Statistic at 20% FPR*.
- Intermediate performers include CXR-FMKD-Direct FFT (HuberLoss), CXR-FMKD-Direct FFT (MSE-CS Naive), CXR-FMKD-Direct FFT (MSE-CS Learned), CXR-FMKD-Direct FFT (MSE-CS | 0.7-0.3), and CXR-FMKD-Direct FFT (MSE-CS | 0.8-0.2).

- **CXR-FMKD FFT** ❷:
  - The CS variant (CXR-FMKD FFT (CS)) remains the least performant within this category. CXR-FMKD FFT (MSE-CS Learned) also sits in the lower end. Additionally, CXR-FMKD FFT (MSE-CS | 0.6-0.4) ranks low for *AUC-ROC* and *Maximum Youden's J Statistic,* and CXR-FMKD FFT (MAE) ranks low for *Youden's J Statistic at 20% FPR*.
  - CXR-FMKD FFT (MSE-CS | 0.9-0.1), CXR-FMKD FFT (MSE-CS | 0.8-0.2), and CXR-FMKD FFT (MSE-CS Naive) now position themselves at the top.
- **Benchmark** models ❸:
  - Within this category, CXR-Model FFT leads in overall performance, followed by DenseNet121 FFT, with ResNet50 FFT trailing as the least performant. In terms of *AUC-PR* only, DenseNet121 FFT marginally outperforms CXR-Model FFT.
- **CXR-FMKD-Direct LP** ❹:
  - Surprisingly, the CS variant (CXR-FMKD-Direct LP (CS)) is now at the top across all performance metrics.
  - At the lower end, we find CXR-FMKD-Direct LP (MSE-CS Learned), which falls into the performance range of the CXR-FMKD LP category ❺ below, followed by CXR-FMKD-Direct LP (MAE). CXR-FMKD-Direct LP (MSE) also ranks low for *Maximum Youden's J Statistic* for instance.
  - More generally, there is a small overlap between CXR-FMKD-Direct LP ❹ and CXR-FMKD LP ❺ for *AUC-PR*.
- **CXR-FMKD LP** ❺:
  - CXR-FMKD LP (MSE-CS Naive) consistently ranks high across all metrics. CXR-FMKD LP (MSE-CS Learned) leads for *Youden's J Statistic at 20% FPR* and CXR-FMKD LP (MSE-CS | 0.9-0.1) leads for *AUC-ROC*.
  - CXR-FMKD LP (MAE) ranks low for *AUC-ROC*, *AUC-PR*, and *Maximum Youden's J Statistic*. Additionally, CXR-FMKD LP (HuberLoss) ranks at bottom for *AUC-PR* and *AUC-ROC*. CXR-FMKD LP (MSE-CS | 0.6-0.4) and CXR-FMKD LP (MSE-CS | 0.9-0.1) performs the worst for *Youden's J Statistic at 20% FPR*. CXR-FMKD LP (MSE-CS Learned) also shows low performance for *AUC-PR*.
  - It should be noted that this stratum category, like the others, covers a narrow performance range, with the included models exhibiting close performances.
- **Original CXR-FM** ❻:
  - CXR-FM, limited to LP integration due to its frozen backbone, ranks as the least performant compared to the other 46 models featured in the plot.

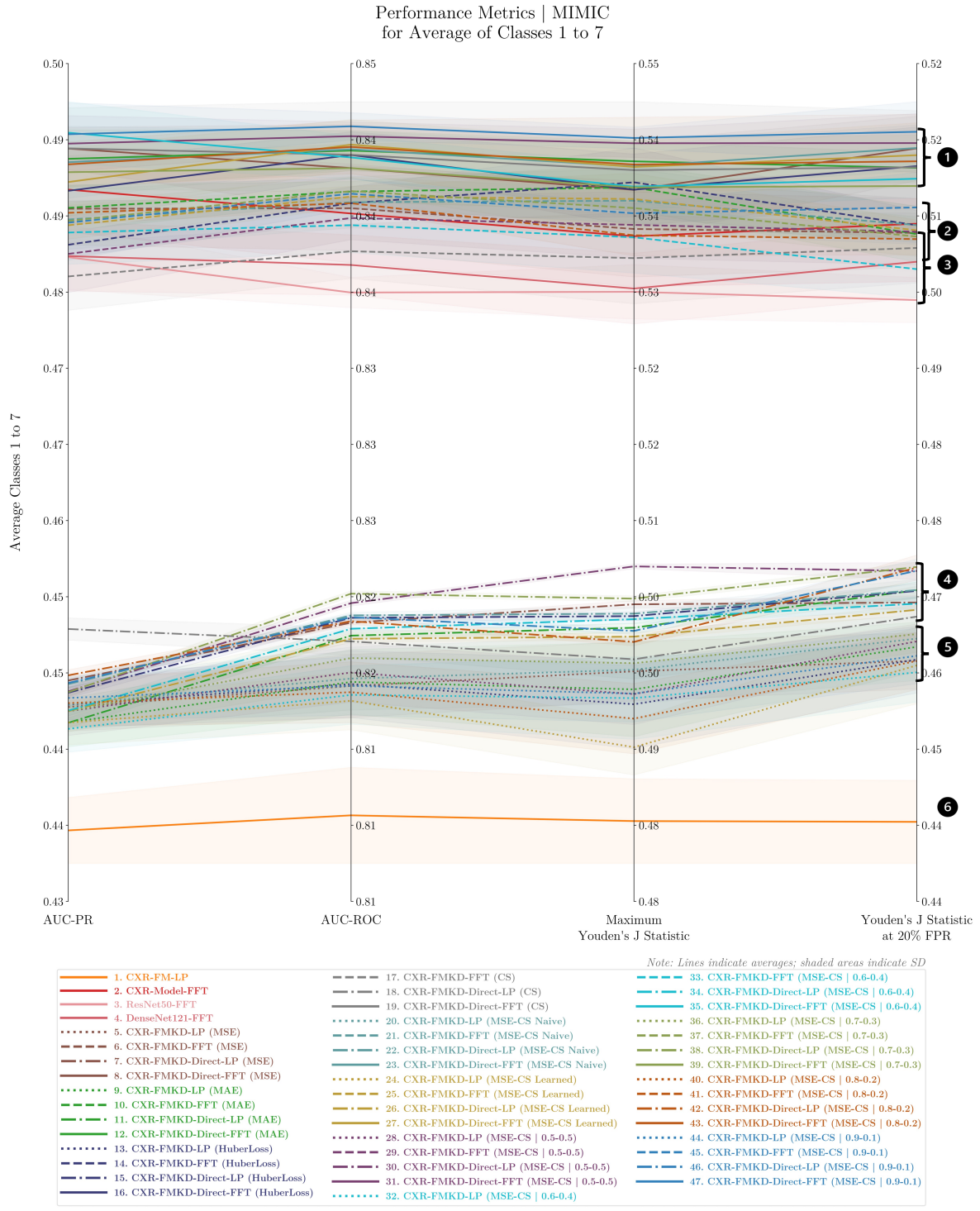

Figure 34. Comparative Analysis of Performance Metrics Across 47 Models for MIMIC Dataset. This custom parallel coordinate plot visualises the performance metrics—AUC-ROC, AUC-PR, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR—for 47 models tested on the MIMIC dataset, focusing on the average results for the most significant disease labels (Classes 1 to 7). Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. We observe a clear stratification in model performance, grouped by type from top (best performers) to bottom (least performers): CXR-FMKD-Direct FFT models ❶, CXR-FMKD FFT models ❷, Benchmark Models ❸ (i.e., CXR-Model FFT, ResNet50 FFT, DenseNet121 FFT), CXR-FMKD-Direct LP models ❹, CXR-FMKD LP models ❺, and lastly the original CXR-FM ❻.

Additional observations include:

- The parallel alignment of lines between models suggests that these models perform similarly and consistently across the four performance metrics. Overall, there are relatively few intersections of lines, indicating stable performance trends across metrics.
- Within each category, models exhibit closely clustered performances. Indeed, the distinctions highlighted earlier in terms of best and least performers pertain to models with still relatively similar performance levels.
- All categories (❶,❷,❸,❹,❺) show approximately equal performance ranges. However, FFT models (CXR-FMKD FFT-variants and benchmarks models) display the most variability, as evidenced by the larger shaded areas around their respective lines, representing higher standard deviations. In contrast, the LP-variants and CXR-FM show less variability across these repeated tests—which were conducted with five different instances of each model type.

MIMIC

Similar to **Figure 33**, **Figure 34** showcases a custom parallel coordinate plot illustrating the performance metrics—*AUC-ROC*, *AUC-PR*, *Maximum Youden's J Statistic*, and *Youden's J Statistic at 20% FPR*—this time for our models tested on the MIMIC dataset, focusing on the average results for the seven *Most Significant Classes* (1 to 7). The same stratification in model performance is observed, grouped by type from best to worst: **CXR-FMKD-Direct FFT** models ❶, **CXR-FMKD FFT** models ❷, **Benchmark** models ❸ (CXR-Model FFT, ResNet50 FFT, and DenseNet121 FFT), **CXR-FMKD-Direct LP** models ❹, **CXR-FMKD LP** models ❺, and the original **CXR-FM** ❻.
Like before, the CXR-Model LP and CXR-FMKD LP (CS) were omitted from **Figure 34** due to their significant underperformance compared to the other models, which skewed the visual representation of the data (as seen in **Figure 64**).

Compared to the CheXpert results, the performance gap between the FFT models (CXR-FMKD FFT-variants + benchmark models) and the CXR-FMKD LP-variants is wider on the MIMIC dataset, while the gap between these LP-variants and the original CXR-FM is narrower. Overall, models on MIMIC exhibit higher performance metrics than those on CheXpert. For instance, the performance range of the LP-variants on MIMIC is similar to that of the FFT models on CheXpert for *AUC-PR*, *Maximum Youden's J Statistic*, and *Youden's J Statistic at 20% FPR*.

We observe similar performance variations among model types on MIMIC as seen with CheXpert, with varied rankings within each category (❶,❷,❹,❺). As a brief overview within each 'stratum' category:

- **CXR-FMKD-Direct FFT** ❶:
  - The models derived from KD with MSE-CS | 0.5-0.5 and MSE-CS | 0.9-0.1 losses are at the top—and are effectively the best overall among all models tested on MIMIC—with CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1) slightly outperforming CXR-FMKD-Direct FFT (MSE-CS | 0.5-0.5) across the evaluated metrics. Interestingly, CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4) achieves the best performance for *AUC-PR* but then sits at the lower end for *AUC-ROC*, *Maximum Youden's J Statistic*, and *Youden's J Statistic at 20% FPR*.
  - CXR-FMKD-Direct FFT (CS) does not rank as the least performant for MIMIC. CXR-FMKD-Direct FFT (HuberLoss) has the lowest scores for *AUC-PR* and *Maximum Youden's J Statistic*. Additionally, CXR-FMKD-Direct FFT (MSE) shows poor

performance for *AUC-ROC* and *Maximum Youden's J Statistic*, and CXR-FMKD-Direct FFT (MSE-CS | 0.7-0.3) ranks low across all metrics and is the least performant overall.

- **CXR-FMKD FFT** ❷:
  - CXR-FMKD FFT (MAE) shows the strongest overall performance within this category, while CXR-FMKD FFT (HuberLoss) ranks highest for *Maximum Youden's J Statistic* and CXR-FMKD FFT (MSE-CS | 0.9-0.1) is at the top for *Youden's J Statistic at 20% FPR*.
  - The CS variant (CXR-FMKD FFT (CS)) is the least performant overall, exhibiting the worst scores for *AUC-ROC*, *AUC-PR*, and *Maximum Youden's J Statistic*. Additionally, CXR-FMKD FFT (MSE-CS | 0.6-0.4) ranks the lowest for *Youden's J Statistic at 20% FPR*.
  - There is slight overlap between the lower end of the CXR-FMKD-Direct FFT ❶ category and the upper end of the CXR-FMKD FFT ❷ category for *Maximum Youden's J Statistic*. More notably, there is significant overlap between CXR-FMKD FFT ❷ and the benchmark models category ❸, particularly evident for *AUC-PR*, as will be discussed in the point below.
- **Benchmark** models ❸:
  - Within this category, aligning with the observations from CheXpert, CXR-Model FFT leads in overall performance, followed by DenseNet121 FFT, while ResNet50 FFT ranks as the least performant.
  - As mentioned above, there is a significant overlap with the CXR-FMKD FFT ❷ category. Indeed, in terms of *AUC-PR*, CXR-Model FFT outperforms even the top-performing CXR-FMKD FFT, namely CXR-FMKD FFT (MAE), and ResNet50 FFT performs better than the least performant CXR-FMKD FFT, specifically CXR-FMKD FFT (CS). Across the other metrics, CXR-Model FFT shows performance on par with the other CXR-FMKD FFT models, while DenseNet121 FFT and ResNet50 FFT remains at the lower end.
- **CXR-FMKD-Direct LP** ❹:
  - The CS variant (CXR-FMKD-Direct LP (CS)) now leads in *AUC-PR* but then ranks lowest for *AUC-ROC*, *Maximum Youden's J Statistic*, and *Youden's J Statistic at 20% FPR*. In contrast, CXR-FMKD-Direct LP (MSE-CS | 0.7-0.3) and CXR-FMKD-Direct LP (MSE-CS | 0.5-0.5) emerges as top performers for *AUC-ROC*, *Maximum Youden's J Statistic*, and *Youden's J Statistic at 20% FPR*.
  - CXR-FMKD-Direct LP (MSE-CS Learned) consistently ranks low across all metrics, while CXR-FMKD-Direct LP (MAE) and CXR-FMKD-Direct LP (MSE-CS | 0.6-0.4) are at the lower end for *AUC-PR*. CXR-FMKD-Direct LP (MSE) ranks low for *Maximum Youden's J Statistic*.
  - Similar to observations in CheXpert, an overlap exists between CXR-FMKD-Direct LP ❹ and CXR-FMKD LP ❺ for *AUC-PR*, specifically for CXR-FMKD-Direct LP (MSE-CS Learned), CXR-FMKD-Direct LP (MAE), and CXR-FMKD-Direct LP (MSE-CS | 0.6-0.4), which fall within the performance range of CXR-FMKD LP ❺.
- **CXR-FMKD LP** ❺:
  - CXR-FMKD LP (MSE-CS Learned) ranks consistently low across all metrics, while CXR-FMKD LP (MSE-CS | 0.7-0.3) sits at the top. CXR-FMKD LP (MSE-CS | 0.9-0.1) leads for *AUC-PR*. Conversely, CXR-FMKD LP (MSE-CS | 0.6-0.4) exhibits the poorest performance for *AUC-PR* and *Youden's J Statistic at 20% FPR*.
- **Original CXR-FM** ❻:
  - Consistent with results from CheXpert, CXR-FM ranks as the least performant compared to the other 46 models featured in the plot.

Additional observations align with those described in the CheXpert section. However, it is worth noting that the MIMIC plot shows slightly more crossings between lines and categories, particularly for the AUC-PR metric, indicating slightly less stable performance trends.

4.1.2. Model Selection

For both the CheXpert and MIMIC datasets, the best-performing models were those developed with fixed weighted combinations of MSE and CS losses during KD, particularly within the CXR-FMKD-Direct FFT category. Consequently, for subsequent experiments and analyses, we focused on this category, selecting the most effective MSE and CS combination along with individual MSE and CS models to enable a thorough comparison. This approach allows us to assess not only the individual contributions of MSE and CS but also their effects when combined through the effective MSE-CS | α-β. Note that choosing the CXR-FMKD-Direct FFT models aligns with the intended use of the CXR-FMKD students created, as these models allow for full fine-tuning, unlike the original teacher model, CXR-FM, which has a 'frozen' backbone (i.e., unavailable) and cannot be fully fine-tuned.

For **CheXpert**, we will therefore focus on CXR-FM, CXR-Model FFT as the benchmark, CXR-FMKD-Direct FFT (CS), CXR-FMKD-Direct FFT (MSE), and CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4). Indeed, the latter two were the top performers for CheXpert.
For **MIMIC**, we will focus on CXR-FM, CXR-Model FFT as the benchmark, CXR-FMKD-Direct FFT (CS), CXR-FMKD-Direct FFT (MSE), and CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1)—the latter stood out as the best-performing model for MIMIC.

This selection streamlines our evaluation for the following sections and allows for a more focused analysis.

4.1.3. CheXpert Performance

Following the model selection described above, we focus on the performance—in terms of *AUC-PR*, *AUC-ROC*, *Maximum Youden's J Statistic*, and *Youden's J Statistic at 20% FPR*—of the CXR-FM, CXR-Model FFT, and selected CXR-FMKD-Direct FFT models (CS, MSE, and MSE-CS | 0.6-0.4) on the CheXpert dataset, as depicted in **Figure 35**. The latter expands on the initial insights from **Figure 33** by not only visualising the average performance across the most significant disease labels (classes 1 to 7) but also detailing the performance for each individual class and the 'Others' category, which includes the remaining seven classes in the disease labels list. Recall that the seven most significant classes evaluated are: Class 1 [Pleural Effusion], Class 2 [No Finding], Class 3 [Cardiomegaly], Class 4 [Pneumothorax], Class 5 [Atelectasis], Class 6 [Consolidation], and Class 7 [Edema].

In line with the detailed discussion in **section 4.1.1**, we observe that CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4) and CXR-FMKD-Direct FFT (MSE) are the top-performing models. They are followed by CXR-FMKD-Direct FFT (CS), CXR-Model FFT, and lastly, CXR-FM, which exhibits the lowest performance.

We will now look into the specific results for each class, reporting the approximate performance ranking for each metric. These rankings will focus primarily on the trends observed for the FFT models (CXR-FMKD-Direct FFT and CXR-Model FFT), mostly followed by CXR-FM:

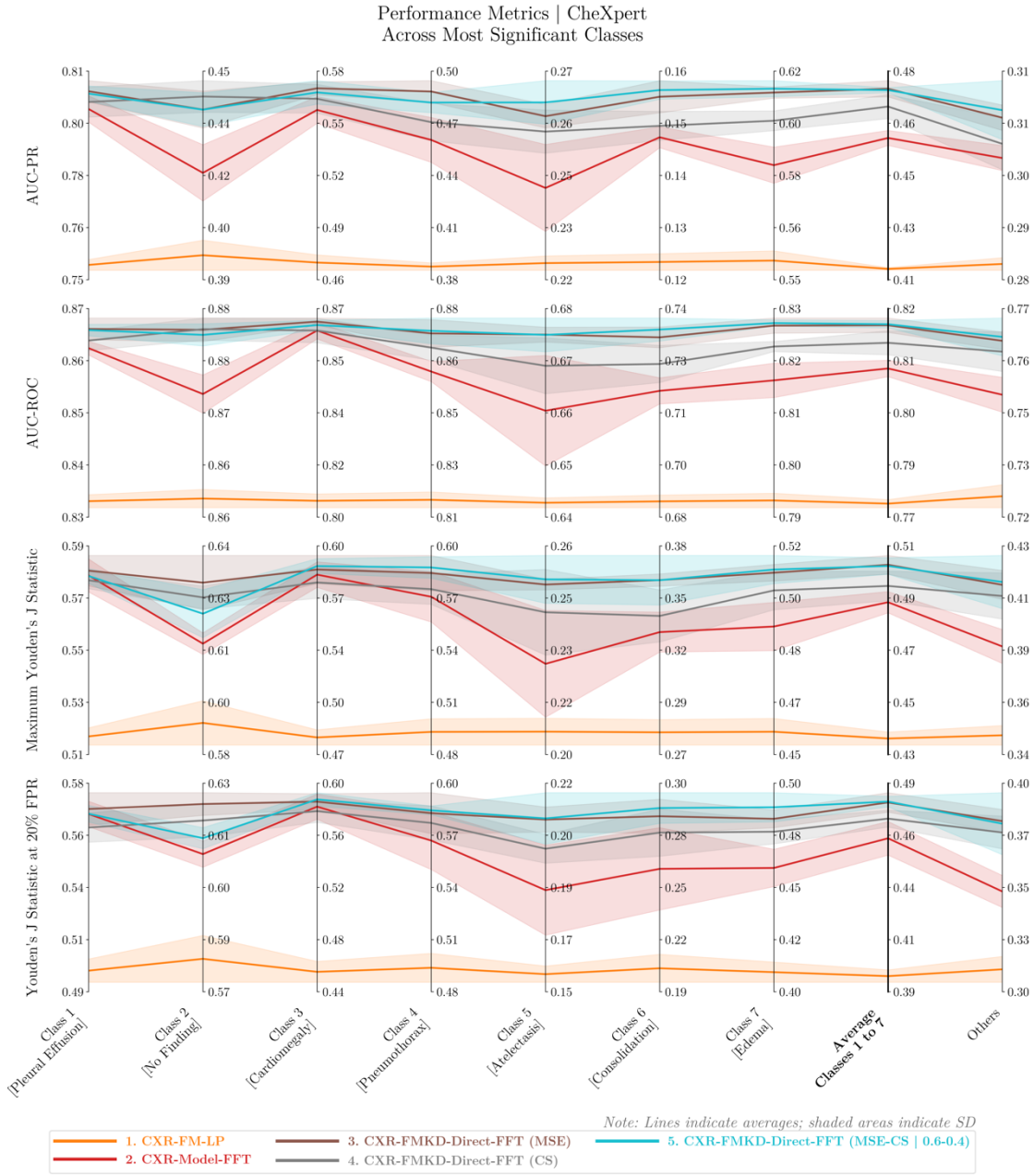

Figure 35. Performance of Selected Models Across Most Significant Classes for CheXpert Dataset. These custom parallel coordinate plots showcase the performance across four metrics: AUC-PR, AUC-ROC, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR for the five selected models tested on the CheXpert dataset. The plots cover the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes in the disease labels list. Each line represents the average results from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4) and CXR-FMKD-Direct FFT (MSE) are the overall top performers across all metrics, followed by CXR-FMKD-Direct FFT (CS), CXR-Model FFT, and lastly, CXR-FM as the least performant.

- For *AUC-PR*, the models exhibit the best performance for Class 1 [Pleural Effusion] and the lowest for Class 6 [Consolidation]. The performance ranking is approximately: Class 1, Class 7, Class 3, Class 4, Class 2, 'Others', Class 5, and finally Class 6.
- For *AUC-ROC*, the highest performance is observed in Class 2 [No Finding], with the lowest in Class 5 [Atelectasis]. The performance ranking is approximately: Class 2, Class 4, Class 1, Class 3, Class 7, 'Others', Class 6, and finally Class 5.

- For *Maximum Youden's J Statistic*, the best performance is also for Class 2 [No Finding], with the poorest performance for Class 5 [Atelectasis]. The ranking is approximately: Class 2, Class 4, Class 3, Class 1, Class 7, 'Others', Class 6, and finally Class 5.
- For *Youden's J Statistic at 20% FPR*, the models perform best again in Class 2 [No Finding] and worst in Class 5 [Atelectasis]. The performance ranking is approximately: Class 2, Class 4, Class 1, Class 3, Class 7, 'Others', Class 6, and finally Class 5.

Significant variations in performance across different classes are evident, such as an *AUC-PR* range from 0.75 to 0.81 for Class 1, contrasting sharply with 0.12 to 0.16 for Class 6; or a *Maximum Youden's J Statistic* range from 0.58 to 0.64 for Class 2, contrasting with 0.20 to 0.26 for Class 5.

Here, the absolute performance trends, indicating where the models achieve higher or lower results per class, are consistent across *AUC-ROC*, *Maximum Youden's J Statistic*, and *Youden's J Statistic at 20% FPR*. These metrics generally show similar patterns of performance across the classes. However, *AUC-PR* deviates slightly; for instance, it exhibits lower performances for Class 2 and the worst scores for Class 6, whereas the other metrics typically show the lowest performance for Class 5.

In terms of relative performance trends, observed across all four plots in **Figure 35**, wider gaps between the performances of the FFT models and CXR-FM are noted in Class 1 and Class 3. Conversely, smaller gaps and more variability in Class 5 suggest that performances are clustered more closely together, with less pronounced improvements from CXR-FM. Further observations reveal different implications of small performance gaps depending on the class. Specifically, the relatively small gap for Class 2 corresponds to uniformly high performance across models, while the narrow gap for Class 5 corresponds to uniformly low performance across all models.

From a more quantitative perspective, detailed results for this experiment for our selected models, along with a relative comparison of model performance to CXR-FM (%Δ w.r.t. CXR-FM) and the CXR-Model FFT baseline (%Δ w.r.t. CXR-Model FFT), are presented in **Table 13** in the *Supplemental Materials*. Focusing on the average for classes 1 to 7, we observe, **on average**, the following improvements:

- For *AUC-PR*: CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4), CXR-FMKD-Direct FFT (MSE), and CXR-FMKD-Direct FFT (CS) are respectively 15.0%, 15.1%, and 13.6% better than CXR-FM; and 3.6%, 3.7%, and 2.4% better than CXR-Model FFT.
- For *AUC-ROC*: They show improvements of 5.3%, 5.2%, and 4.7% over CXR-FM; and 1.3%, 1.2%, and 0.7% over CXR-Model FFT.
- For *Maximum Youden's J Statistic*: They are 15.5%, 15.6%, and 13.7% better than CXR-FM; and 2.9%, 3.0%, and 1.3% better than CXR-Model FFT.
- For *Youden's J Statistic at 20% FPR*: They achieve 19.6%, 19.5%, and 17.7% better performance than CXR-FM; and 3.6%, 3.5%, and 1.9% better than CXR-Model FFT.

Notably, CXR-FMKD-Direct FFT (MSE) and CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4) demonstrate very similar performances. However, the former shows marginally more consistency across metrics where CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4), for example, exhibits the lowest performance among the CXR-FMKD-Direct FFT models for Class 2 [No Finding] across all metrics. Additionally, it should be noted that CXR-Model FFT outperforms CXR-FMKD-Direct FFT (CS) for Class 1 and Class 3 in both *Youden's J Statistic* metrics.

For further information on the detailed results of this performance analysis, please refer to **Figure 52**, **Figure 53**, **Figure 54**, and **Figure 55** in the *Supplemental Material* section. These figures include parallel coordinate plots for the 47 models showed in **Figure 33**, showcasing the performance metrics across the most significant disease classes. Additionally, **Figure 56**, **Figure 57**, **Figure 58**, and **Figure 59** display standard plots for all 49 models, including CXR-Model LP and CXR-FMKD-LP (CS), across all 14 disease classes and their macro average. This helps in observing absolute performance trends. The corresponding parallel coordinate plots for relative performance trends are presented in **Figure 60**, **Figure 61**, **Figure 62**, and **Figure 63**.

Convergence

For completeness, we tracked the evolution and convergence of our selected models' performance during training. Validation loss, along with other performance metrics such as *AUC-PR*, *AUC-ROC*, *Maximum Youden's J Statistic*, and *Youden's J Statistic at 20% FPR*, were evaluated on the validation set at the end of each training epoch. The training spanned 20 epochs in total, with epochs numbered from 1 to 20. Detailed results are provided in **Table 3**, which documents the epochs at which each metric reached its highest (Epoch at Peak) and lowest (Epoch at Lowest) values on the validation set.

As for all data presented in this section, the values for each model in the table reflect the average outcomes from testing five distinct instances of the model, each developed and trained with a different seed. Results are therefore displayed as Avg ± SD, where 'Avg' represents the average value and 'SD' the corresponding standard deviation.

			CheXpert				
			CXR-FM	CXR-Model FFT	CXR-FMKD-Direct FFT		
Metric at Epoch End (Validation Set)	*Class*	*Metric Tracking Action*			MSE	CS	MSE-CS | 0.6-0.4
			Avg ± SD	Avg ± SD	Avg ± SD	Avg ± SD	Avg ± SD
Validation Loss	All Classes	**Epoch at Lowest**	**8.40 ± 2.70**	**13.40 ± 1.82**	**3.00 ± 0.71**	**1.80 ± 0.45**	**3.20 ± 0.84**
		Epoch at Peak	6.80 ± 6.83	14.00 ± 6.60	19.80 ± 0.45	19.40 ± 0.89	19.40 ± 0.89
AUC-PR	Average Classes 1 to 7	Epoch at Lowest	1.00 ± 0.00	1.03 ± 0.06	1.54 ± 1.21	3.06 ± 1.93	1.51 ± 1.15
		Epoch at Peak	**14.17 ± 1.39**	**12.89 ± 1.69**	**4.49 ± 0.33**	**3.60 ± 0.43**	**4.54 ± 0.51**
AUC-ROC	Average Classes 1 to 7	Epoch at Lowest	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.54 ± 1.21	1.00 ± 0.00
		Epoch at Peak	**14.03 ± 1.44**	**12.54 ± 1.69**	**4.69 ± 0.19**	**3.57 ± 0.39**	**4.46 ± 0.51**
Maximum Youden's J Statistic	Average Classes 1 to 7	Epoch at Lowest	1.14 ± 0.32	1.57 ± 1.12	1.00 ± 0.00	2.40 ± 1.36	2.51 ± 1.40
		Epoch at Peak	**12.97 ± 2.09**	**13.40 ± 1.88**	**4.89 ± 0.41**	**3.77 ± 0.41**	**4.69 ± 0.55**
Youden's J Statistic at 20% FPR	Average Classes 1 to 7	Epoch at Lowest	1.00 ± 0.00	1.03 ± 0.06	1.00 ± 0.00	1.86 ± 1.24	1.49 ± 1.09
		Epoch at Peak	**11.86 ± 2.20**	**12.94 ± 2.03**	**4.63 ± 0.16**	**3.71 ± 0.35**	**4.63 ± 0.42**

Table 3. Training Performance Tracking of Selected Models on CheXpert Dataset.
This table reports the evolution and convergence of our models' performance during training on the CheXpert dataset, detailing the epochs at which the highest (Epoch at Peak) and lowest (Epoch at Lowest) values for each metric were recorded on the validation set—out of 20 epochs in total, with epochs numbered from 1 to 20. It includes validation loss, AUC-PR, AUC-ROC, Maximum Youden's J Statistic, and Youden's J Statistic at 20% FPR. The data displayed represent the averages (Avg) with standard deviations (SD) derived from testing five distinct instances of the same model type, each developed and trained with a different seed. Notably, the CXR-FMKD-Direct FFT student models demonstrate faster convergence compared to CXR-FM and CXR-Model FFT.

In model development, our goal is to minimise the loss and maximise other key performance metrics such as those being considered here. For validation loss, which was used for model selection, our CXR-FMKD-Direct FFT student models demonstrated notably early convergence. For instance, on average, CXR-FMKD-Direct FFT (MSE) reached its minimum validation loss at epoch 3.00 ± 0.71, CXR-FMKD-Direct FFT (CS) at epoch 1.80 ± 0.45, and CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4) at epoch 3.20 ± 0.84. These figures are in stark contrast to those for CXR-FM, which reaches its minimum loss at 8.40 ± 2.70, and CXR-Model FFT at 13.40 ± 1.82. Indeed, relative to CXR-Model FFT, our CXR-FMKD-Direct FFT models (MSE, CS, and

MSE-CS | 0.6-0.4) achieve minimum validation loss in epochs that are approximately 78%, 87%, and 76% fewer, respectively. When compared to CXR-FM, these models require 64%, 79%, and 62% fewer epochs. Additionally, CXR-FM reaches the minimum validation loss in 37% fewer epochs than CXR-Model FFT.

The training phases were optimised using the loss metric, so observing the latter is more pertinent for this convergence analysis. However, it is interesting to track the other performance metrics as well. Our student models reached the maximum results much earlier than both CXR-Model FFT and CXR-FM, typically within 3.5-5 epochs. By comparison, CXR-FM generally achieves these maximum values around 12-14 epochs, and CXR-Model FFT around 12.5-13.5 epochs. Here, CXR-FM reaches peak values around 1.4 epochs after CXR-Model FFT for *AUC-PR* and *AUC-ROC*, but approximately 0.8 epochs sooner for *Maximum Youden's J Statistic* and *Youden's J Statistic at 20% FPR.*

Our CXR-FMKD-Direct FFT student models therefore exhibit significantly faster convergence compared to CXR-FM and, more notably, the CXR-Model FFT baseline. Within these student models, the CS variant achieves slightly faster convergence than the MSE and MSE-CS | 0.6-0.4 variants, which show similar convergence epochs.

4.1.4. MIMIC Performance

Similar to **Figure 35** for CheXpert, **Figure 36** displays the performance—in terms of *AUC-ROC*, *AUC-PR*, *Maximum Youden's J Statistic*, and *Youden's J Statistic at 20% FPR*—for the CXR-FM, CXR-Model FFT as the benchmark, CXR-FMKD-Direct FFT (CS), CXR-FMKD-Direct FFT (MSE), and CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1) selected models tested on MIMIC. The figure consists of four parallel coordinate subplots, each representing one of the metrics, and focuses on the most significant classes (1 to 7), their average, and the 'Others' category which includes the remaining seven disease labels.

Aligning with the discussions in **section 4.1.1**, we observe that CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1) is the overall top performer, followed closely by CXR-FMKD-Direct FFT (CS) and CXR-FMKD-Direct FFT (MSE), with CXR-Model FFT next, and CXR-FM ranking the lowest in the performance hierarchy.

Like we did for CheXpert, we will now examine the outcomes for each class and provide an approximate performance ranking for each of the four metrics. These rankings will primarily focus on the trends observed among the FFT models:

- For *AUC-PR*, the best performance is observed for Class 1 [Pleural Effusion], with the lowest for Class 6 [Consolidation]. The approximate performance ranking is: Class 1, Class 2, Class 7, Class 3, Class 5, Class 4, 'Others', and finally Class 6.
- For *AUC-ROC*, the highest performance is also noted for Class 1 [Pleural Effusion], with the lowest for 'Others'. The performance ranking is approximately: Class 1, Class 7, Class 4, Class 2, Class 6, Class 3, Class 5, and finally 'Others'.
- For *Maximum Youden's J Statistic*, the best performance is once again for Class 1 [Pleural Effusion], with the poorest for 'Others'. The ranking is approximately: Class 1, Class 7, Class 4, Class 2, Class 6, Class 5, Class 3, and finally 'Others'.
- For *Youden's J Statistic at 20% FPR*, the models again perform best in Class 1 [Pleural Effusion] and worst in 'Others'. The ranking is approximately: Class 1, Class 7, Class 4, Class 2, Class 6, Class 3, Class 5, and finally 'Others'.

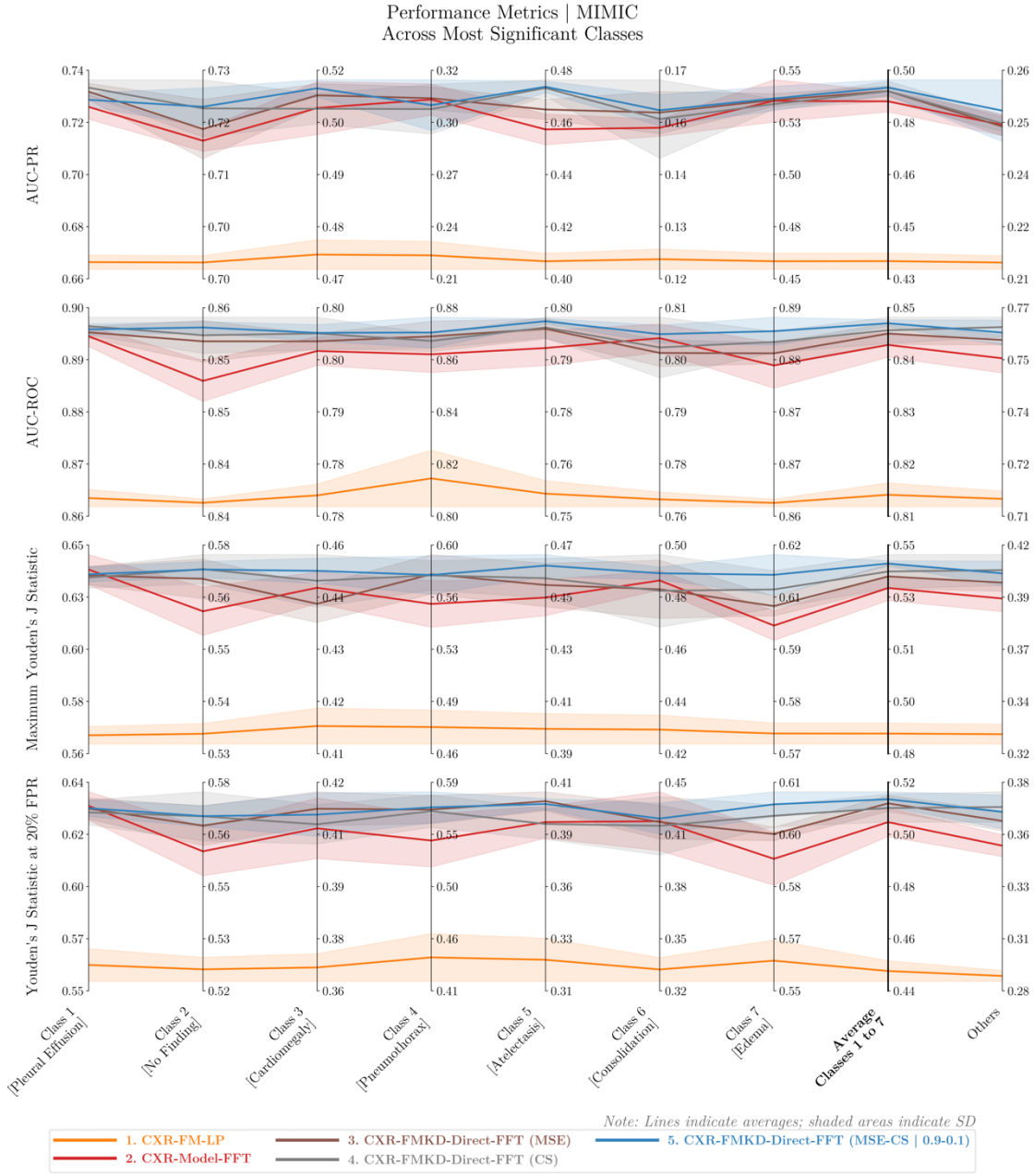

Figure 36. Performance of Selected Models Across Most Significant Classes for MIMIC Dataset. These custom parallel coordinate plots showcase the performance across four metrics: AUC-PR, AUC-ROC, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR for the five selected models tested on the MIMIC dataset. The plots cover the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes in the disease labels list. Each line represents the average results from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1) is the overall top performer across all metrics, followed closely by CXR-FMKD-Direct FFT (CS), CXR-FMKD-Direct FFT (MSE), CXR-Model FFT, and lastly, CXR-FM as the least performant.

Similar to our findings with CheXpert, we observe notable variations in performance across the different classes for MIMIC. For example, the *AUC-PR* range for Class 1 is from 0.66 to 0.74, in contrast with 0.12 to 0.17 for Class 6. The *Youden's J Statistic at 20% FPR* ranges from 0.55 to 0.64 for Class 1, versus 0.28 to 0.38 for 'Others'.

The absolute performance trends also indicate consistency across the *AUC-ROC*, *Maximum Youden's J Statistic*, and *Youden's J Statistic at 20% FPR* metrics, while *AUC-PR*

deviates slightly, particularly underperforming for Class 4 and Class 6 but performing better for Class 3. For all four metrics, the models consistently perform best for Class 1 and rank well for Class 7, but show poorer performance for 'Others'.

Compared to CheXpert, for *AUC-PR*, MIMIC models maintain high rankings for Class 1 and Class 7 and low scores for Class 6 and 'Others', but they perform worse for Class 4 and better for Class 2 and Class 5. For the other three metrics, which share nearly identical performance patterns cross the classes for each dataset, both MIMIC and CheXpert models rank high for Class 1 and Class 4. However, MIMIC models perform better for Class 7 and worse for Class 2, Class 3, and 'Others'.

Overall, models on MIMIC achieve higher absolute scores than on CheXpert. In terms of relative performance trends observed through the parallel coordinate plots, the gaps between FFT models and CXR-FM are more consistent across all classes for the four metrics in MIMIC compared to CheXpert, with narrower gaps among the FFT models themselves. Here, a slightly wider gap between FFT models and CXR-FM is seen in Class 1 for MIMIC, where CXR-Model FFT's performance closely aligns with that of the CXR-FMKD-Direct FFT models—similar to what was observed for CheXpert.

Table 14 in the *Supplemental Materials* presents the detailed numerical results for this experiment for our selected models, including a relative comparison of model performance to CXR-FM (%Δ w.r.t. CXR-FM) and the CXR-Model FFT baseline (%Δ w.r.t. CXR-Model FFT). Focusing on the average for classes 1 to 7, we note the following **average** improvements:

- For *AUC-PR*: CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1), CXR-FMKD-Direct FFT (MSE), and CXR-FMKD-Direct FFT (CS) are respectively 12.8%, 12.5%, and 12.5% better than CXR-FM; and 0.9%, 0.7%, and 0.7% better than CXR-Model FFT.
- For *AUC-ROC*: They show improvements of 4.0%, 3.7%, and 3.8% over CXR-FM; and 0.5%, 0.3%, and 0.3% over CXR-Model FFT.
- For *Maximum Youden's J Statistic*: They are 12.0%, 11.1%, and 11.4% better than CXR-FM; and 1.6%, 0.7%, and 1.1% better than CXR-Model FFT.
- For *Youden's J Statistic at 20% FPR*: They achieve 16.1%, 15.8%, and 15.3% better performance than CXR-FM; and 1.9%, 1.5%, and 1.2% better than CXR-Model FFT.

While the models generally score higher on the metrics in MIMIC compared to CheXpert, the relative improvements over CXR-FM and CXR-Model FFT are smaller. The CXR-FMKD-Direct FFT models perform closely to the CXR-Model FFT, reflecting the narrower gaps between the FFT models as mentioned previously. These FFT models are also slightly closer in performance to CXR-FM but still demonstrate notable improvements.

Here, for MIMIC, CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1) demonstrates the best performance overall. Given the more clustered performance ranges of the FFT models, we also observe more overlap, as depicted by the lines crossing in the parallel coordinate plots. This overlap is particularly evident where the CXR-Model FFT surpasses some of the student models in specific classes. For example, in Class 4, CXR-Model FFT ranks as the second-best for *AUC-PR*, directly behind CXR-FMKD-Direct FFT (MSE). Similarly, in Class 6, it is second-best for the other three metrics (*AUC-ROC, Maximum Youden's J Statistic, and Youden's J Statistic at 20% FPR*), directly behind CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1). In Class 1, it ranks first for both *Youden's J Statistic* metrics.

For further information on the detailed results of this performance analysis, please refer to **Figure 65**, **Figure 66**, **Figure 67**, and **Figure 68** in the *Supplemental Material* section. These

figures include parallel coordinate plots for the 47 models showed in **Figure 34**, showcasing the performance metrics across the most significant disease classes. Additionally, **Figure 69**, **Figure 70**, **Figure 71**, and **Figure 72** display standard plots for all 49 models, including CXR-Model LP and CXR-FMKD-LP (CS), across all 14 disease classes and their macro average. This helps in observing absolute performance trends. The corresponding parallel coordinate plots for relative performance trends are presented in **Figure 73**, **Figure 74**, **Figure 75**, and **Figure 76**.

Convergence

Similar to our approach with CheXpert, we tracked the evolution and convergence of our selected models' performance during training on the MIMIC Dataset. Detailed results are provided in **Table 4**, which documents the epochs at which each metric reached its highest (Epoch at Peak) and lowest (Epoch at Lowest) values on the validation set, across 20 training epochs in total.

			MIMIC				
			CXR-FM	CXR-Model FFT	CXR-FMKD-Direct FFT		
Metric at Epoch End (Validation Set)	*Class*	*Metric Tracking Action*			MSE	CS	MSE-CS | 0.9-0.1
			Avg ± SD	Avg ± SD	Avg ± SD	Avg ± SD	Avg ± SD
Validation Loss	All Classes	**Epoch at Lowest**	**8.60 ± 5.37**	**9.60 ± 1.14**	**4.20 ± 1.30**	**2.60 ± 0.55**	**3.20 ± 1.10**
		Epoch at Peak	7.80 ± 6.91	5.60 ± 7.80	19.40 ± 0.89	19.60 ± 0.89	18.80 ± 1.30
AUC-PR	Average Classes 1 to 7	Epoch at Lowest	1.00 ± 0.00	1.03 ± 0.06	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		Epoch at Peak	**12.60 ± 1.44**	**10.97 ± 0.69**	**4.97 ± 0.62**	**3.49 ± 0.42**	**5.14 ± 0.48**
AUC-ROC	Average Classes 1 to 7	Epoch at Lowest	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		Epoch at Peak	**10.60 ± 1.00**	**11.83 ± 0.66**	**5.29 ± 0.71**	**3.66 ± 0.41**	**4.83 ± 0.51**
Maximum Youden's J Statistic	Average Classes 1 to 7	Epoch at Lowest	1.03 ± 0.06	1.00 ± 0.00	1.00 ± 0.00	2.09 ± 1.49	1.00 ± 0.00
		Epoch at Peak	**12.34 ± 1.39**	**12.34 ± 0.87**	**5.57 ± 0.83**	**3.74 ± 0.43**	**5.20 ± 0.24**
Youden's J Statistic at 20% FPR	Average Classes 1 to 7	Epoch at Lowest	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		Epoch at Peak	**12.94 ± 0.93**	**12.46 ± 1.10**	**5.49 ± 0.54**	**3.63 ± 0.54**	**5.46 ± 0.73**

Table 4. Training Performance Tracking of Selected Models on MIMIC Dataset.
This table reports the evolution and convergence of our models' performance during training on the MIMIC dataset, detailing the epochs at which the highest (Epoch at Peak) and lowest (Epoch at Lowest) values for each metric were recorded on the validation set—out of 20 epochs in total, with epochs numbered from 1 to 20. It includes validation loss, AUC-PR, AUC-ROC, Maximum Youden's J Statistic, and Youden's J Statistic at 20% FPR. The data displayed represent the averages (Avg) with standard deviations (SD) derived from testing five distinct instances of the same model type, each developed and trained with a different seed. Notably, the CXR-FMKD-Direct FFT student models demonstrate faster convergence compared to CXR-FM and CXR-Model FFT.

For the validation loss, our CXR-FMKD-Direct FFT student models also demonstrated early convergence. For instance, on average, CXR-FMKD-Direct FFT (MSE) reached its minimum validation loss at epoch 4.20 ± 1.30, CXR-FMKD-Direct FFT (CS) at epoch 2.60 ± 0.55, and CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1) at epoch 3.20 ± 1.10. In contrast, CXR-FM reaches its minimum loss at 8.60 ± 5.37, and CXR-Model FFT at 9.60 ± 1.14. Relative to CXR-Model FFT, our CXR-FMKD-Direct FFT models (MSE, CS, and MSE-CS | 0.9-0.1) reach minimum validation loss in approximately 56%, 73%, and 67% fewer epochs, respectively. Compared to CXR-FM, these epochs are reduced by 51%, 70%, and 63%. Additionally, although CXR-FM reaches its minimum validation loss in about 10% fewer epochs than CXR-Model FFT, it shows significant variability in convergence time, with an SD of 5.37 epochs, occasionally performing 'slower' than CXR-Model FFT, but consistently 'slower' than the CXR-FMKD-Direct FFT models.

Regarding the other performance metrics, our student models reached their peak results much earlier than both CXR-Model FFT and CXR-FM, typically within 3.5-5.5 epochs. By comparison, CXR-FM generally achieves these maximum values around 10.5-13 epochs, while CXR-Model FFT does so around 11-12.5 epochs. Here, CXR-FM reaches peak values

approximately 1.6 epochs after CXR-Model FFT for *AUC-PR* and 1.2 epochs before for *AUC-ROC*, with similar convergence epochs for *Maximum Youden's J Statistic* and *Youden's J Statistic at 20% FPR.*

Our CXR-FMKD-Direct FFT student models thus demonstrate significantly faster convergence compared to both CXR-FM and the CXR-Model FFT baseline. Echoing findings from CheXpert, the CS variant among these student models shows slightly quicker convergence than the MSE and MSE-CS | 0.9-0.1 variants.

Overall, the improvements in convergence speed for MIMIC are less pronounced compared to CheXpert, aligning with the relatively subtler performance improvements observed for our student models relative to CXR-Model FFT and CXR-FM in MIMIC. In detail, observing the validation loss, the student models—the MSE, CS, and MSE-CS | 0.9-0.1 variants—converge slightly later in MIMIC (averaging 4.20, 2.60, and 3.20 epochs, respectively) than in CheXpert (3.00, 1.80, and 3.20 epochs on average). Conversely, CXR-Model FFT reaches its convergence earlier in MIMIC (9.60 epochs on average) compared to CheXpert (13.40 epochs on average). CXR-FM shows similar convergence times in both datasets, at approximately 8.50 epochs.

4.1.5. Discussion

For both the CheXpert and MIMIC datasets, we observed a clear stratification in model performance, grouped by model type, as illustrated in **Figure 33** for CheXpert and **Figure 34** and MIMIC. As reported previously, this performance stratification reveals the following hierarchy, from best to worst:

1. **CXR-FMKD-Direct FFT** models ❶
2. **CXR-FMKD FFT** models ❷
3. **Benchmark** models ❸ (CXR-Model FFT, ResNet50 FFT, and DenseNet121 FFT)
4. **CXR-FMKD-Direct LP** models ❹
5. **CXR-FMKD LP** models ❺
6. Original **CXR-FM** ❻

This stratification offers valuable insights into the KD and subsequent adaptation processes, highlighting their impact on the performance of the CXR-FMKD student models.

FFT vs. LP Models

Firstly, LP models, including the CXR-FMKD LP-variants (❹ and ❺) and the CXR-FM teacher ❻, underperform relative to the FFT models, which consist of the CXR-FMKD FFT-variants (❶ and ❷) and the benchmark models ❸. This performance disparity aligns with expectations grounded in the fine-tuning approaches employed.

As delineated in **Chapter 2**, LP is characterised as part of the *Targeted Fine-tuning* approach where, in our context, only the final fully-connected layer—that is, *nn.Linear-T* for CXR-FM and the CXR-FMKD non-Direct-variants, and *nn.Linear* for the CXR-FMKD Direct-variants, as described in **Chapter 3**—is adjusted during training. This layer, which maps the features extracted from the model backbone to the 14 disease outputs, has its weights updated to learn task-specific distinctions. However, the backbone itself, thus its input-to-feature mapping capabilities, remains unchanged, preserving the original feature representations derived from the input CXR images. This static feature handling limits the model's capacity to adapt and refine its discriminatory abilities for the disease detection task, as it cannot modify the underlying features that inform the final classification decisions.

In contrast, FFT falls under the *Comprehensive Fine-tuning* approach, where all model weights, including those of the backbone, are updated during training. This holistic update strategy allows the models to optimise and adapt the entire network's feature extraction and mapping processes, not just the last classification layer. By adjusting the non-linear transformations that process input CXR images into discriminative features, FFT models enhance their flexibility and capability to develop robust, task-specific features. This, in turn, enables the final classification layer to work with more informative and relevant feature representations, significantly improving the model's effectiveness in this multilabel classification scenario. This fundamental difference in fine-tuning strategies explains the observed performance disparities, with FFT models outperforming LP models by leveraging their ability to undergo more extensive and deeper adjustments tailored to the task at hand.

Lastly, it is noted that FFT models exhibit greater variability compared to LP models, with large shaded areas around their respective lines indicating a higher SD. This increased variance is expected since FFT involves updating significantly more parameters, including the entire backbone and its non-linear transformations, not just the classifier. This naturally leads to greater performance variability across different training instances.

Direct- vs. non-Direct- CXR-FMKD Variants

Now, within each of the CXR-FMKD LP-variants and FFT-variants, the Direct-variants consistently outperform their non-Direct counterparts—specifically, CXR-FMKD-Direct LP outperforms CXR-FMKD LP, and CXR-FMKD-Direct FFT outperforms CXR-FMKD FFT. This phenomenon reflects practices observed in the SSL literature [200], notably the findings regarding the advantages of removing of the projector, as detailed in **Chapter 3**.

Drawing parallels from Chen et al.'s seminal SimCLR study [114], which pioneered the technique of disregarding the projector in SSL, this approach has been shown to significantly boost top-1 accuracy on ImageNet by about 20% over a 100-epoch training phase. Indeed, it has been demonstrated that projected features tend to underperform in terms of generalisation, and features prior to projection prove more effective. One rationale could be that in SSL, the encoder's output contains the more useful features for downstream tasks such as classification and detection, while the projector's role primarily facilitates the contrastive learning process by creating a space where the contrastive loss can be more effectively minimised.

Interestingly, a similar effect is observed in our experiments. By removing the projector used during the KD process (i.e., *nn.Linar-KD*), we revert to using a larger feature set—specifically, the original 1664 features of the DenseNet169 backbone instead of the compressed 1376 features intended to match CXR-FM's feature size. This adjustment potentially allows the model to retain more detailed representations of the input CXR images, preserving complex patterns and nuances that might be lost in a lower-dimensional space. Consequently, this provides a richer and larger set of inputs to the final classification layer (*nn.Linear*), potentially enhancing the model's ability to discriminate between the classes due to having more parameters to effectively combine these broader feature sets.

Benchmark Models

Among the benchmark models, CXR-Model FFT was selected due to its architecture, DenseNet169, which possesses a feature size (1664 outputs) enabling a 'projection down' to match the 1376 features of CXR-FM for KD. This model outperforms both ResNet50 FFT and DenseNet121 FFT. The latter two serve primarily as reference models, incorporating the popular DenseNet121, widely used in similar disease detection tasks as outlined in [93, 94, 99, 103, 199], and ResNet50, another common architecture also used in KD applications [157]. These models—

CXR-Model FFT, ResNet50 FFT, and DenseNet121 FFT—follow identical development processes but differ in their backbones (DenseNet169, DenseNet121, and ResNet50, respectively) and the linear fully-connected layers that map their features to the 14 disease labels.

The superior performance of CXR-Model FFT can partly be attributed to the higher capacity of DenseNet169, which has approximately 14 million parameters, compared to around 8 million for DenseNet121. Although ResNet50 features around 26 million parameters, its architecture has proven less effective for this disease classification task, lacking the feature-reuse efficiency of DenseNets, which capitalise on reusing feature maps across the network [47]. This confirms that CXR-Model FFT remains a robust benchmark for comparing the performance of our CXR-FMKD student models, which share the same DenseNet169 backbone, demonstrating its competitiveness and relevance. In this context, CXR-Model FFT represents a 'traditional' approach in tackling this disease detection task, where the model, initially pre-trained on ImageNet (i.e., natural images), undergoes comprehensive fine-tuning (FFT) to adapt it for disease classification. This process further ensures it serves as a reliable benchmark against which the CXR-FMKD variants—effectively the 'same' model but enhanced through KD with CXR-FM—are evaluated.

CXR-FMKD FFT-Variants

We base our subsequent interpretations on the data presented in **Figure 33** for CheXpert and **Figure 34** for MIMIC. Notably, our CXR-FMKD FFT-variants (i.e., CXR-FMKD FFT and CXR-FMKD-Direct FFT models) demonstrate superior performance compared to our CXR-Model FFT benchmark for CheXpert, indicating that our student models have effectively leveraged knowledge from CXR-FM. This enhancement suggests that the DenseNet169 backbones, after KD with CXR-FM, have learned useful feature representations from CXRs for the disease detection task. During the subsequent FFT process, these enhanced features have boosted the student models' disease discrimination capabilities right from the onset, improving their training outcomes and performance compared to CXR-Model FFT, which was pre-trained only with natural images.

Particularly, the CXR-FMKD-Direct FFT models provide a direct comparison to CXR-Model FFT. By removing the projector and applying comprehensive fine-tuning (FFT), these models share the same architecture as the benchmark but start with an adaptation advantage through KD. Even without the projector removal, CXR-FMKD FFT models outperform the CXR-Model FFT in CheXpert. However, for MIMIC, the performance advantage is less pronounced, with some overlap between the CXR-Model FFT and CXR-FMKD FFT models depending on the KD loss used. Notably, for *AUC-PR*, CXR-Model FFT slightly outperforms the CXR-FMKD FFT models; however, the CXR-FMKD-Direct FFT models consistently show higher performance compared to the benchmark.

Here, the performance among FFT models in MIMIC is more clustered compared to CheXpert, likely due to the unique characteristics and disease prevalences in the MIMIC dataset, which differs significantly from CheXpert. Indeed, MIMIC primarily originates from an emergency department setting, which may explain the higher incidence of 'No Finding' at 31%, significantly more than CheXpert's 9%, as reported in **Figure 22**. This discrepancy could be attributed to routine comprehensive scans in emergency trauma settings, conducted to confirm the absence of critical thoracic conditions. Such specific characteristics of MIMIC's data might account for the observed narrower clustering of performance, which could indicate a ceiling effect where the models are maximally tuned to the nuances of MIMIC's clinical presentations. This clustering could suggest that the overall higher performance in MIMIC compared to CheXpert may be approaching the limits of performance improvement possible through KD, given the dataset's constraints.

In both datasets, these findings underscore the significance of the direct comparison between CXR-Model FFT and the CXR-FMKD-Direct FFT models. As mentioned earlier, they share the exact same architecture: one only pre-trained on ImageNet, while the others benefit from a form of 'second pre-training' through the KD process with CXR-FM, which is highly relevant and performed through the same disease detection task using CXR inputs. This process allows the student models to assimilate valuable insights from the teacher, enriching their feature set and potentially steering them towards more optimal local minima during training, thus enhancing performance.

Overall, these results validate the effectiveness of KD, which not only allows the reconstructed student models to surpass the CXR-FM teacher but also outperform the traditional CXR-Model FFT benchmark. This advantage accentuates the value of transferring specialised knowledge from the teacher, significantly boosting disease detection capabilities.

CXR-FMKD LP-Variants

While it has been established why the CXR-FMKD LP-variants exhibit lower performance compared to the FFT models, which include the CXR-FMKD FFT-variants and the CXR-Model FFT benchmark—due to the more restrictive LP fine-tuning approach versus the comprehensive FFT—it is notable that these LP-variants still outperform the CXR-FM teacher, which is constrained by the same LP fine-tuning approach. This outcome is counterintuitive given the nature of the KD process, where the student models are essentially attempting to reconstruct the teacher's capabilities.

Specifically, the CXR-FMKD LP models, where the projector was not removed, are expected to generate features from input CXR images that are closely aligned with those produced by the CXR-FM teacher. This alignment is due to the KD process being optimised to minimise the difference between the student's and teacher's features via a KD loss function. Thus, for any given CXR input, the CXR-FMKD LP models should theoretically yield 'imperfect' 1376 features that are a proximate mimicry of the 'ground truth' 1376 features from CXR-FM.

Interestingly, despite these features being 'imperfect', they lead to better performance when processed through the adapted linear classification layer (*nn.Linear-T*)—derived from the LP training—compared to the CXR-FM itself, which utilises what might be considered 'perfect' features. The exact reasons for this enhanced performance are not entirely clear. By 'pre-training' the student through this KD process using a transfer set directly tied to the task at hand, while trying to replicate the teacher's output features, this effectively created a model that generates more useful features to be used by the classification layer for this CXR disease detection. This could potentially be tied to the student model's specific characteristics like its DenseNet169 architecture and pre-initialisation on natural images, which might have maintained directions in the learned features that are specifically favourable for the task and datasets during the KD process, while still trying to align closely with the teacher's outputs. This could also be attributed to the student models possibly exploring different regions of the feature space that, while not exactly replicating the teacher's output features, provide a beneficial diversity in feature representation that enhances overall performance.

When removing the projector to create the CXR-FMKD-Direct LP models, as previously discussed, we further amplify the above-mentioned performance gains.

KD Exploration

From the analysis of **Figure 33** (CheXpert) and **Figure 34** (MIMIC), it is clear that no single KD loss consistently outperforms across all model variants. The varied performance rankings

within the CXR-FMKD categories (LP, FFT, Direct, and non-Direct variants), as discussed in **section 4.1.1**, underscores the nuanced impact of different KD techniques.

Among the various approaches, the models derived from the MSE-CS | α-β fixed weighted combination losses lead the performance metrics within the CXR-FMKD-Direct FFT category, which includes the overall top performers. Generally, MSE-CS combinations maintain high rankings across the other categories as well, indicating their robustness. This success suggests that combining MSE and CS losses can harness their respective strengths: MSE aims to minimise the magnitude differences between the teacher's and student's feature vectors, ensuring the student's outputs closely match the teacher's in scale, while CS focuses on aligning the direction of these vectors, which is crucial for maintaining the relational integrity of features without necessarily preserving their scale. Such complementary nature of CS and MSE could effectively be nudging models towards beneficial optima and enhancing performance.

In contrast, models utilising CS alone often exhibit lower performance, with some exceptions. This inconsistency could stem from CS's sole focus on the directionality of feature vectors, potentially neglecting the importance of their magnitudes, leading to suboptimal feature representations for the specific tasks. Notably, within the CXR-FMKD-Direct LP category, the CS variant excels for CheXpert across all metrics and for *AUC-PR* in MIMIC, yet CXR-FMKD LP (CS) ranks lowest overall, displaying significant variability.

MAE consistently ranks lower within each category. Its uniform treatment of errors may lack the necessary aggressiveness to drive models toward more effective local minima, unlike MSE, which, by amplifying larger discrepancies through its squared error term, guides models towards more discriminative feature learning. Indeed, MSE's consistent efficacy highlights its fundamental utility in minimising direct discrepancies between teacher and student outputs, a simple yet powerful approach popular in KD literature [157].

Lastly, the MSE-CS | Learned variant did not perform as well, with fixed ratios showing better results, though it should be noted that this exploration was not extensively optimised.

Performance Disparities Across Metrics and Classes

The discrepancy in performance trends between *AUC-PR* and the other *AUC-ROC, Maximum Youden's J Statistic, and Youden's J Statistic at 20% FPR* metrics, as discussed in **section 4.1.3** and **section 4.1.4**, prompts a discussion on the distinct implications these metrics have. Here, *AUC-ROC, Maximum Youden's J Statistic, and Youden's J Statistic at 20% FPR* are closely related, as they are all derived from the ROC curve, which plots TPR against FPR. *AUC-ROC* provides a measure of the area under this curve and the Youden's J Statistic metrics are calculated as TPR – FPR, explaining why similar trends are observed for these metrics across **Figure 33**, **Figure 34**, **Figure 35**, and **Figure 36**, as well as in the *Supplemental Material* sections *S.1. Performance Analysis – CheXpert* and *S.2. Performance Analysis – MIMIC*.

We take *AUC-ROC* versus *AUC-PR* for an explanation of the performance disparities. *AUC-ROC* reflect the model's ability to discriminate between classes across various thresholds and is generally less sensitive to class imbalance. This makes it a robust metric for assessing overall model discrimination between positive and negative classes. Conversely, *AUC-PR* focuses on the precision and recall of the positive class, making it more affected in conditions of class imbalance where positive cases are less prevalent. It emphasizes the model's precision, thereby reflecting its ability to identify positive instances correctly amidst a large number of negative samples.

A higher *AUC-ROC* combined with a lower *AUC-PR* can potentially indicate that while the model is generally effective at distinguishing between positive and negative instances, it may yield a higher number of false positives, especially when the positive class is rare, and vice versa.

In MIMIC, where the incidence of 'No Finding' is significantly higher at 31% compared to 9% for CheXpert, our models demonstrate higher *AUC-PR* scores around 0.7 compared to 0.45 in CheXpert for this class (as illustrated in **Figure 36** and **Figure 69** for MIMIC, and **Figure 35** and **Figure 56** for CheXpert). Indeed, the higher prevalence of 'No Finding' in MIMIC may provide more opportunities for the models to correctly predict this class when it is present, contributing to the higher *AUC-PR* scores. This also suggests that *AUC-PR*, which emphasizes precision in predicting positive instances, is particularly relevant for this class due to its imbalanced nature and the high cost associated with false negatives, where potential diseases might be overlooked if 'No Finding' is incorrectly predicted.

The variability in performance rankings across the classes for the various metrics might therefore reflect the specific prevalence of each class within the dataset and the inherent difficulty of detecting certain conditions. For instance, Class 5 [Atelectasis] and Class 6 [Consolidation] typically exhibit lower performance scores across metrics in both MIMIC and CheXpert, suggesting these conditions are generally more challenging to detect. Focusing on CheXpert, we observe narrower relative performance gaps between the FFT models group and CXR-FM for Class 2 [No Finding] and Class 5 [Atelectasis]; however, the significance of these gaps varies by class. The narrow gap in Class 2, associated with high scores, suggests that models generally perform well, showing consistently high performance across the board. Conversely, the small gap observed in Class 5, associated with low scores, indicates that all models struggle with this class, leading to uniformly lower performance.

These observations underline the complexity of model evaluation, where performance can be significantly influenced by the prevalence and detectability of specific classes within the data, revealing disparities across these classes and the metrics used. While these insights are valuable, it is important to note that the primary focus of our study is on the relative improvements among our models, which remain consistent across different classes and metrics, thus allowing for robust interpretations concerning model rankings. A detailed class-based analysis, while insightful, is beyond the scope of this discussion.

Outliers

As noted in **section 4.1.1**, although we tested 49 model types in total, only 47 are displayed in **Figure 33** and **Figure 34**. The CXR-Model LP and CXR-FMKD LP (CS) were omitted due to their significantly lower performance in both CheXpert and MIMIC, which distorted the visual representation of the data (as seen in **Figure 51**, and **Figure 64**). Specifically, CXR-FMKD LP (CS) exhibited the lowest performance among all models tested, with substantial variability indicated by the large shaded area (SD) around its line. For example, the worst instance of CXR-FMKD LP (CS) reached an *AUC-ROC* range of 0.499-0.505 across classes, effectively performing no better than a random classifier. This underperformance suggests that the 1376 features (i.e., keeping the *nn.Linear-KD* projector) generated by the student backbone after KD using CS as the loss are not only challenging but also potentially disadvantageous for the linear classification layer (*nn.Linear-T*), which struggles to use the information they contain to effectively discriminate between the disease labels in the LP setting.

For CXR-Model LP, its notably low performance—still above that of CXR-FMKD LP (CS)—was anticipated and serves as a baseline to confirm that the features produced by CXR-FM are indeed useful for our CXR disease detection task, which is the primary function of such FM. As mentioned earlier, CXR-Model LP employs a DenseNet169 backbone pre-trained on ImageNet, meaning it primarily generates features based on its knowledge of natural images, which substantially differ from CXR data. This 'raw' DenseNet169 backbone therefore provides very rudimentary and not particularly relevant feature representations when applied directly to CXR images. Consequently, these features, when used by the classification layer to produce the

14 prediction outputs, are expected to perform poorly compared to CXR-FM, which underwent a specialised pre-training phase on CXRs. The significant underperformance of CXR-Model LP compared to CXR-FM confirms these expectations, underscoring the relevance and usefulness of CXR-FM for its use in downstream CXR applications.

Convergence

In terms of convergence, our selected CXR-FMKD-Direct FFT models exhibit notably faster convergence, as evidenced by the epoch at which minimum validation loss is achieved, compared to both the CXR-Model FFT baseline and the CXR-FM teacher in CheXpert (**Table 3**) and MIMIC (**Table 4**). CXR-FM also achieves slightly faster convergence than CXR-Model FFT. In general, these convergence improvements are more pronounced in CheXpert than in MIMIC, aligning with relatively greater performance improvements between the models for CheXpert (**Table 13**) than for MIMIC (**Table 14**).

These observations underscore a significant byproduct of the KD process. Our student models not only outperform the teacher and baseline models but also inherit and amplify the teacher FM's strengths. A key premise behind using FMs is their adaptability to downstream tasks with minimal data—crucial in contexts like healthcare where high-quality labelled data is scarce, as highlighted in **Chapter 1** of our report. While this convergence analysis does not directly measure the data volume needed for effective model adaptation, the rapid convergence of our student models within ~1.8-4.2 epochs strongly suggests that they may require less data to achieve optimal performance levels. This rapid achievement of minimum validation loss indicates that the features generated by the student backbone (and so the backbone itself), 'pre-trained' on the knowledge from the CXR-FM teacher during the KD process, are already highly relevant and efficient, guiding the model to performant optima quickly during the FFT phase. Such rapid convergence also suggests reduced computational load through an implied more computationally efficient training of student models.

Moreover, the relatively faster convergence of the teacher model (CXR-FM) compared to CXR-Model FFT, most notable in CheXpert, underscores the utility of its CXR-specific features. However, these features prove more challenging to quickly optimise the model in the LP setting—a scenario that typically requires less data for fine-tuning—compared to the FFT setting. This suggests that while CXR-FM's features are highly relevant for CXR analysis, they do not adapt as swiftly in a constrained fine-tuning environment as they do when all model weights are being adjusted, as is the case with the student models. This also introduces greater variability in CXR-FM's convergence epochs, reflecting the challenges of adapting highly specialised features within a limited tuning framework.

On the other hand, as discussed previously, CXR-Model FFT starts with a feature set derived from natural image pre-training. This baseline requires more epochs to adjust its backbone and classification layers effectively to the CXR disease detection task, highlighting the mismatch between its initial pre-training on natural images and the final CXR-specific application. The slower convergence of CXR-Model FFT compared to the other models further emphasizes the advantage of starting with CXR-tailored features.

Returning to our student models, the CS variant achieves the fastest convergence among them. While this variant may not reach the best performance peaks, it demonstrates how a specific KD adjustment—targeting the directionality of features without necessarily aligning their magnitudes—can lead to quicker, though not optimal, convergence in downstream task adaptation. This is indicative of the CS variant's ability to navigate the feature space quickly but perhaps at the cost of achieving the most discriminative feature set possible.

Conclusion

Our analysis conclusively demonstrates that KD can be used to 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 KD projector and subsequently applying comprehensive fine-tuning (FFT). This approach enables these students not only to surpass the performance of the teacher (CXR-FM), whose backbone remains frozen, but also to exceed that of a traditional benchmark model (CXR-Model FFT) trained from 'scratch' for this specific disease detection task. The superior performance of our CXR-FMKD-Direct FFT student models is further accompanied by faster convergence rates, suggesting an inherited advantage from the FM teacher in swiftly adapting to downstream tasks. These findings underscore the efficacy of KD as a strategy to enhance model performance by leveraging the strengths and knowledge embedded in the original FM.

The consistency of these trends across both the CheXpert and MIMIC datasets, coupled with our methodological approach of averaging results after a fivefold repetition of training with different seeds and subsequent testing for each model variant, reinforces the robustness of our results and interpretations.

4.2. Generalisability Analysis

The goal of the *Generalisability Analysis* is to examine how well our student models perform when tested on a dataset different from the one used in their initial KD and training processes. This analysis effectively assesses the models' ability to maintain performance when applied to OOD data, illustrating how well the distilled CXR-FMKD student models inherit and possibly enhance the robustness and *transfer learning* capabilities of the original CXR-FM teacher.

Notations: For this analysis, models initially trained on the CheXpert dataset underwent subsequent testing on the MIMIC dataset using various levels of fine-tuning: 'Direct Transfer', 'Linear Probing', and 'Full Fine-Tuning'. We will call these the *transfer* models. They are contrasted against corresponding benchmarks which are models sharing the same architecture but exclusively trained and tested on MIMIC. They will be referred to as the *benchmark* models and represent the intended training and development for these models on MIMIC. For further distinction, *benchmark* model names are marked with an asterisk (*).

The *transfer* models examined include the teacher CXR-FM, the baseline CXR-Model FFT, and the selected student models for CheXpert: CXR-FMKD-Direct FFT (MSE), CXR-FMKD-Direct FFT (CS), and CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4). Corresponding *benchmark* models are CXR-FM*, CXR-Model FFT*, and the selected student models for MIMIC: CXR-FMKD-Direct FFT (MSE)*, CXR-FMKD-Direct FFT (CS)*, and CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1)*. The choice of CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4) as a *transfer* model and CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1)* as the corresponding *benchmark* model reflects their status as top performers for CheXpert and MIMIC respectively, chosen for their similar development processes and their demonstrated ability to capture the best representations and discrimination abilities from their respective training phases and datasets.

Like our performance analysis, we focused on the most significant classes (1 to 7), their average, and the 'Others' category, which averages the results for the remaining seven classes.

4.2.1. Direct Transfer

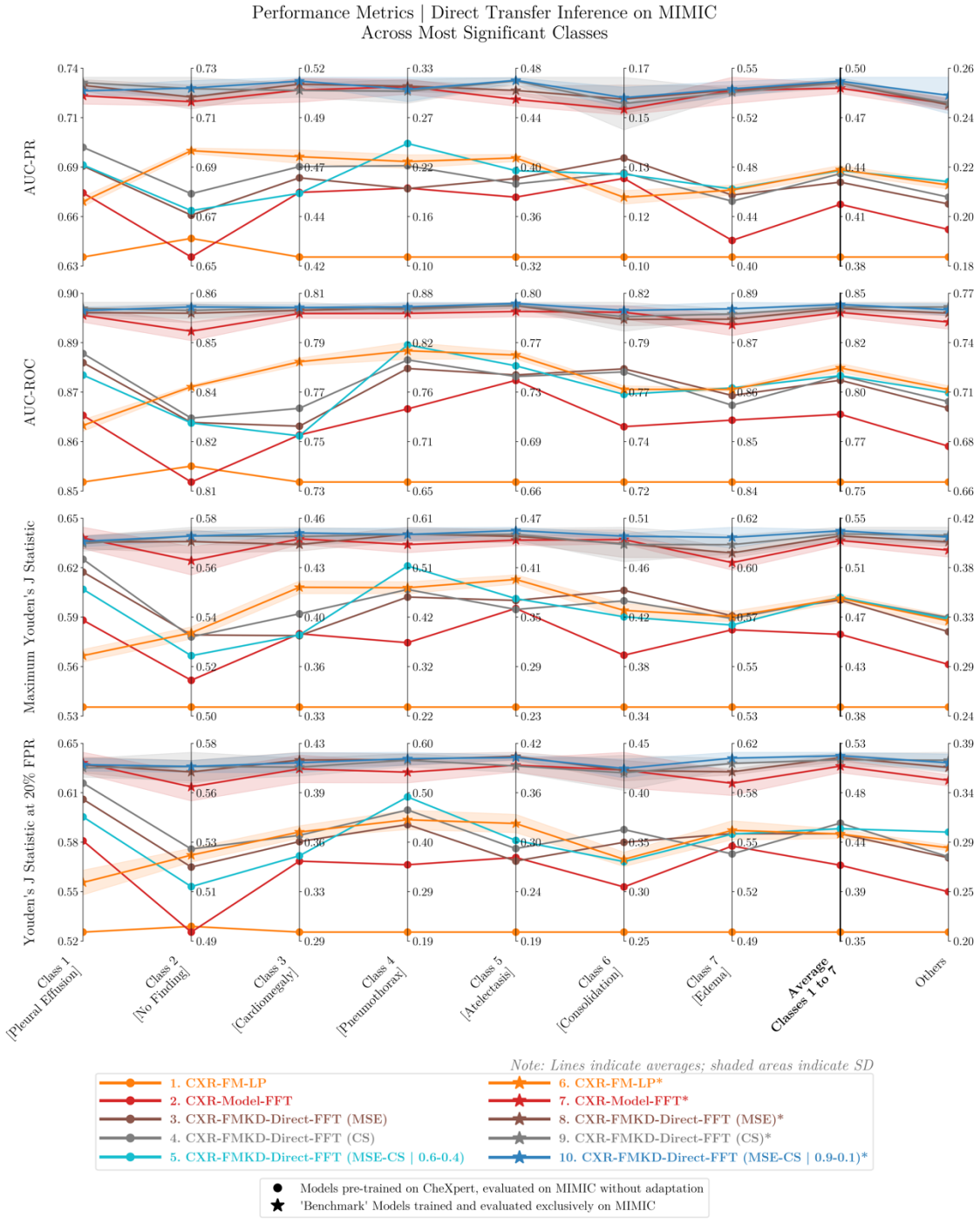

Figure 37. Performance of Selected Transfer Models and Their Benchmarks Across Most Significant Classes After Direct Transfer Inference on MIMIC, Post-CheXpert Pre-training.
These custom parallel coordinate plots display the performance across four metrics: AUC-PR, AUC-ROC, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR for the five selected '*transfer*' models tested on MIMIC without adaptation after pre-training on CheXpert. The models include CXR-FM (teacher), CXR-Model FFT (baseline), and the CXR-FMKD-Direct FFT variants of the three selected student model types (MSE, CS, and MSE-CS | α-β). Each '*transfer*' model, represented by circles, is contrasted against a corresponding *benchmark*, shown with stars, which shares the same architecture but was trained and tested exclusively on MIMIC. The plots cover the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes in the disease labels list. Each line represents the average results from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. For the *transfer* models, the SD is reported as zero because there is no fine-tuning involved; only direct testing on MIMIC occurs.

This section examines how the CheXpert-pre-trained *transfer* models perform on MIMIC without any adaptation, made possible given the shared 14 classes between the datasets. As highlighted in **Figure 37**, the *transfer* models, represented by circles, consistently underperform compared to their *benchmark* counterparts, represented by stars. Indeed, in the figure, each circle (*transfer* model) is plotted below its corresponding star (*benchmark* model) for each class. **Table 5** provides a quantitative comparative analysis focusing on the average for our most significant classes (1 to 7), illustrating negative changes in performance in red for all our *transfer* models compared to the benchmarks. This performance gap is particularly pronounced for CXR-FM and CXR-Model FFT. However, the CXR-FMKD-Direct FFT *transfer* student models manage performances comparable to the CXR-FM* *benchmark* trained exclusively on MIMIC. For instance, looking at the average of classes 1 to 7, the *transfer* CS and MSE-CS | 0.6-0.4 student variants slightly surpass CXR-FM* for *Youden's J Statistic at 20% FPR*. Here, CXR-Model FFT performs lower than the *transfer* student models overall, with the transfer CXR-FM showing the lowest performance with 0.3808 for *AUC-PR*, 0.7535 for *AUC-ROC*, 0.3909 for *Maximum Youden's J Statistic*, and 0.3547 for *Youden's J Statistic at 20% FPR* (**Table 5**).

Direct Transfer Inference on MIMIC

		CXR-FM			CXR-Model FFT			CXR-FMKD-Direct FFT								
								MSE			CS			MSE-CS | α-β		
		Bench-mark (B)	Transfer Model		Bench-mark (B)	Transfer Model		Bench-mark (B)	Transfer Model		Bench-mark (B)	Transfer Model		Bench-mark (B)	Transfer Model	
Metric	***Class***	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)
AUC-PR	Average Classes 1 to 7	0.4359 ± 0.0026	0.3808 ± 0.0000	-12.6%	0.4872 ± 0.0034	0.4140 ± 0.0000	-15.0%	0.4905 ± 0.0017	0.4279 ± 0.0000	-12.7%	0.4905 ± 0.0033	0.4333 ± 0.0000	-11.7%	0.4916 ± 0.0025	0.4353 ± 0.0000	-11.5%
AUC-ROC	Average Classes 1 to 7	0.8113 ± 0.0022	0.7535 ± 0.0000	-7.1%	0.8394 ± 0.0023	0.7878 ± 0.0000	-6.1%	0.8415 ± 0.0020	0.8051 ± 0.0000	-4.3%	0.8421 ± 0.0025	0.8073 ± 0.0000	-4.1%	0.8434 ± 0.0006	0.8074 ± 0.0000	-4.3%
Maximum Youden's J Statistic	Average Classes 1 to 7	0.4841 ± 0.0036	0.3909 ± 0.0000	-19.2%	0.5337 ± 0.0043	0.4533 ± 0.0000	-15.1%	0.5376 ± 0.0041	0.4826 ± 0.0000	-10.2%	0.5393 ± 0.0058	0.4846 ± 0.0000	-10.1%	0.5420 ± 0.0009	0.4851 ± 0.0000	-10.5%
Youden's J Statistic at 20% FPR	Average Classes 1 to 7	0.4448 ± 0.0043	0.3547 ± 0.0000	-20.3%	0.5071 ± 0.0060	0.4161 ± 0.0000	-17.9%	0.5149 ± 0.0033	0.4450 ± 0.0000	-13.6%	0.5129 ± 0.0059	0.4547 ± 0.0000	-11.3%	0.5166 ± 0.0031	0.4497 ± 0.0000	-13.0%

Table 5. Average Performance Comparison of Selected Transfer and Corresponding Benchmark Models After Direct Transfer Inference on MIMIC, Post-CheXpert Pre-training.
This table presents a detailed performance analysis of the five selected '*transfer*' models tested on MIMIC without further adaptation following pre-training on CheXpert. The analysis covers four metrics—AUC-PR, AUC-ROC, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR—focusing on the average for the most significant disease labels (Classes 1 to 7). Results are shown as mean outcomes (Avg) with standard deviations (SD), derived from testing five distinct instances of the same model type, each developed and trained using a different seed. For the *transfer* models, the SD is reported as zero because there is no fine-tuning involved; only direct testing on MIMIC occurs. Performance improvements for each *transfer* model relative to its corresponding *benchmark* (B) are quantified using the formula: (Transfer Model Avg Value − Benchmark Avg Value)/ Benchmark Avg Value × 100% for each metrics.

In terms of the individual classes, we observe similar trends overall (**Figure 37**). Interestingly, for Class 2 [No Finding], the students and CXR-Model FFT *transfer* models see a relative drop in performance across all metrics compared to the teacher models, contrasting with Class 1 [Pleural Effusion], where the opposite trend is observed and they outperform CXR-FM*. The CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4) displays slightly more variable performance across classes compared to the other two student variants. We should note that no SD is observed for our *transfer* models, as no fine-tuning was involved, which introduces variability.

Figure 77 in the *Supplemental Material* extends **Figure 37**'s analysis to all 14 classes, revealing consistent trends with occasional fluctuations. For instance, in Class 9 [Enlarged Cardiomegaly], CXR-FM* shows a notable relative improvement, whereas other *transfer* models underperform, a trend that reverses in Class 11 [Lung Lesion], where *transfer* models outperform CXR-FM* and approach or match the performance of *benchmark* student models.

4.2.2. Linear Probing

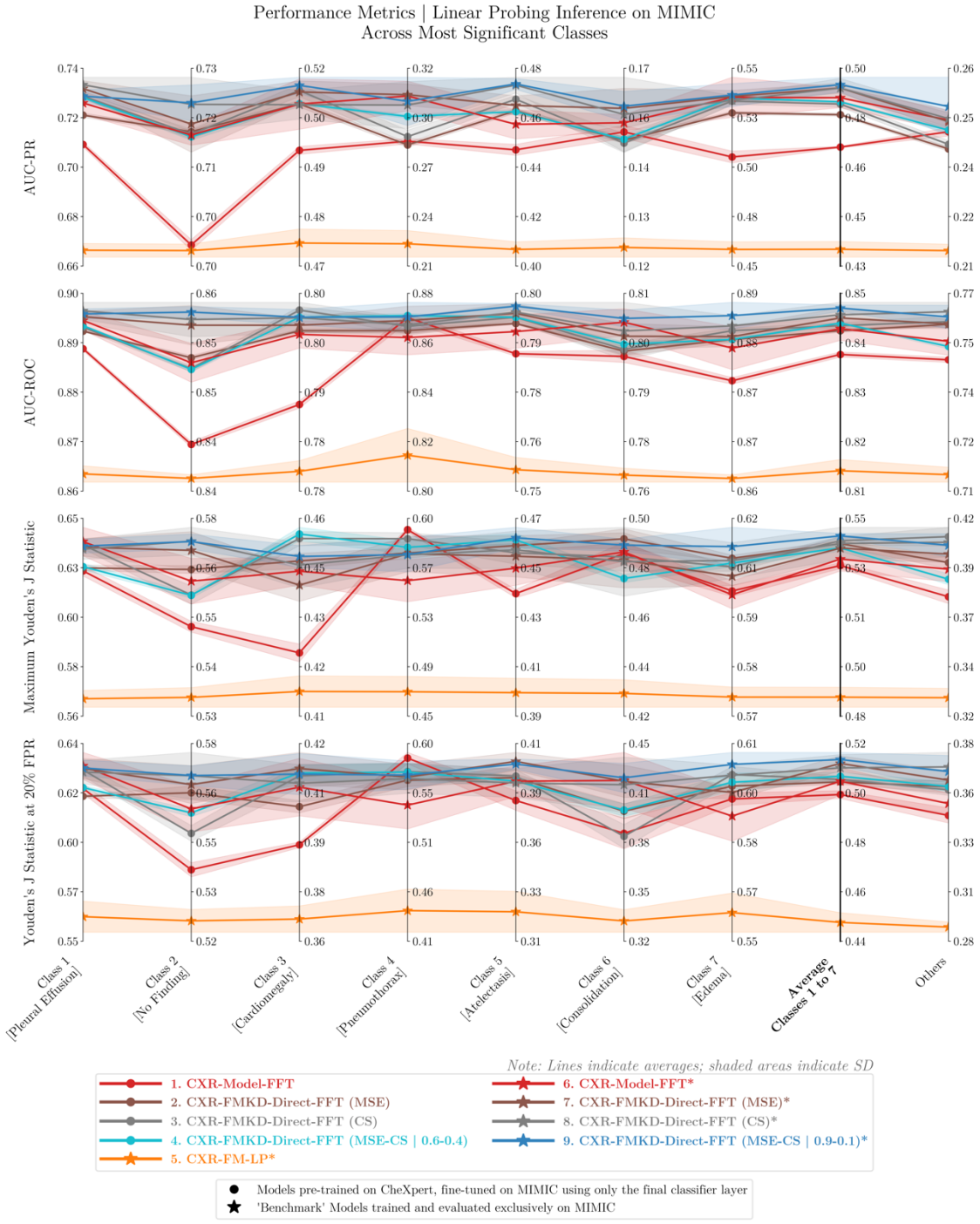

Figure 38. Performance of Selected Transfer Models and Their Benchmarks Across Most Significant Classes After Linear Probing Inference on MIMIC, Post-CheXpert Pre-training.
These custom parallel coordinate plots display the performance across four metrics: AUC-PR, AUC-ROC, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR for the five selected '*transfer*' models that were pre-trained on CheXpert and then fine-tuned on MIMIC using only the final classification layer. The models include CXR-FM (teacher), CXR-Model FFT (baseline), and the CXR-FMKD-Direct FFT variants of the three selected student model types (MSE, CS, and MSE-CS | α-β). Each '*transfer*' model, represented by circles, is contrasted against a corresponding *benchmark*, shown with stars, which shares the same architecture but was trained and tested exclusively on MIMIC. For CXR-FM, both are equivalent. The plots cover the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes in the disease labels list. Each line represents the average results from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation across these tests.

This section evaluates the CheXpert-pre-trained *transfer* models, where their linear classification layer has been replaced and fine-tuned on MIMIC while keeping the backbone frozen. Consequently, for the teacher model CXR-FM, which inherently has a frozen backbone, both its *transfer* and *benchmark* forms are equivalent since the generated feature representations remain unchanged for given CXR inputs and the classification layer's adaptation is the same.

Figure 38 highlights clear improvements for our *transfer* models compared to the 'Direct Transfer' task (**Figure 37**). Across the most significant classes, our CXR-FMKD-Direct FFT *transfer* student models perform comparably to both the CXR-Model FFT* *benchmark* and the CXR-FMKD-Direct FFT* *benchmark* student models. While slightly trailing the *benchmark* students, which are the overall top performers, the *transfer* student models surpass them in some classes, particularly for *Maximum Youden's J Statistic* in Class 3 [Cardiomegaly] with CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4) and CXR-FMKD-Direct FFT (CS), and in Class 6 [Consolidation] with CXR-FMKD-Direct FFT (MSE). The *transfer* CXR-Model FFT lags slightly behind and displays more variability, underperforming in Class 2 [No Finding] and Class 3 [Cardiomegaly], yet achieving some of the highest scores in Class 4 [Pneumothorax]. The teacher model, CXR-FM, consistently exhibits the weakest performance across all classes.

Figure 78 in the *Supplemental Material* extends **Figure 38**'s analysis to all 14 classes. Similar to the 'Direct Transfer' case, the plot reveals consistent trends with occasional fluctuations, such as the relative drop in performance of some models for Class 9 [Enlarged Cardiomegaly].

Linear Probing Inference on MIMIC

		CXR-FM			CXR-Model FFT			CXR-FMKD-Direct FFT								
								MSE			CS			MSE-CS | α-β		
		Bench-mark (B)	Transfer Model		Bench-mark (B)	Transfer Model		Bench-mark (B)	Transfer Model		Bench-mark (B)	Transfer Model		Bench-mark (B)	Transfer Model	
Metric	***Class***	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)
AUC-PR	Average Classes 1 to 7	0.4359 ± 0.0026	0.4359 ± 0.0026	0.0%	0.4872 ± 0.0034	0.4705 ± 0.0003	-3.4%	0.4905 ± 0.0017	0.4814 ± 0.0005	-1.8%	0.4905 ± 0.0033	0.4850 ± 0.0010	-1.1%	0.4916 ± 0.0025	0.4858 ± 0.0003	-1.2%
AUC-ROC	Average Classes 1 to 7	0.8113 ± 0.0022	0.8113 ± 0.0022	0.0%	0.8394 ± 0.0023	0.8343 ± 0.0005	-0.6%	0.8415 ± 0.0020	0.8390 ± 0.0003	-0.3%	0.8421 ± 0.0025	0.8400 ± 0.0006	-0.3%	0.8434 ± 0.0006	0.8405 ± 0.0001	-0.4%
Maximum Youden's J Statistic	Average Classes 1 to 7	0.4841 ± 0.0036	0.4841 ± 0.0036	0.0%	0.5337 ± 0.0043	0.5314 ± 0.0011	-0.4%	0.5376 ± 0.0041	0.5392 ± 0.0011	0.3%	0.5393 ± 0.0058	0.5403 ± 0.0016	0.2%	0.5420 ± 0.0009	0.5377 ± 0.0012	-0.8%
Youden's J Statistic at 20% FPR	Average Classes 1 to 7	0.4448 ± 0.0043	0.4448 ± 0.0043	0.0%	0.5071 ± 0.0060	0.5012 ± 0.0017	-1.2%	0.5149 ± 0.0033	0.5068 ± 0.0010	-1.6%	0.5129 ± 0.0059	0.5079 ± 0.0014	-1.0%	0.5166 ± 0.0031	0.5092 ± 0.0017	-1.4%

Table 6. Average Performance Comparison of Selected Transfer and Corresponding Benchmark Models After Linear Probing Inference on MIMIC, Post-CheXpert Pre-training.
This table presents a detailed performance analysis of the five selected '*transfer*' models that were pre-trained on CheXpert and then fine-tuned on MIMIC using only the classification layer. The analysis covers four metrics—AUC-PR, AUC-ROC, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR—focusing on the average for the most significant disease labels (Classes 1 to 7). Results are shown as mean outcomes (Avg) with standard deviations (SD), derived from testing five distinct instances of the same model type, each developed and trained using a different seed. Performance improvements for each *transfer* model relative to its corresponding *benchmark* (B) are quantified using the formula: (Transfer Model Avg Value − Benchmark Avg Value)/ Benchmark Avg Value × 100% for each metrics. Note that for CXR-FM, *transfer* and *benchmark* models are equivalent due to its frozen backbone constraint.

Focusing on the average for classes 1 to 7 to provide a single overall picture, the CXR-FMKD-Direct FFT* *benchmark* student models remain the top performers across all metrics, closely followed by our *transfer* student models and the *benchmark* CXR-Model FFT*. Here, the *transfer* students perform comparably or slightly better than CXR-Model FFT* across most metrics except *AUC-PR*. Additionally, our *transfer* student models' performance closely aligns with that of the *benchmark* students in *Maximum Youden's J Statistic*. Below these, the *transfer* CXR-Model FFT ranks, separated by a narrow margin in most metrics, but with a more pronounced gap in *AUC-PR*. At the bottom, the teacher model shows a noticeable separation from the rest. Quantifying these improvements for the average of classes 1 to 7, **Table 6** demonstrates that

our *transfer* models are now much closer to their corresponding *benchmarks* than in the 'Direct Transfer' case shown in **Table 5**. While they generally still slightly underperform, the CXR-FMKD-Direct FFT MSE and CS variants do slightly outperform their *benchmarks* in *Maximum Youden's J Statistic*, aligning with our observations from **Figure 38**.

Furthermore, we extended our analysis to include the CXR-FMKD LP, CXR-FMKD-Direct LP, and CXR-FMKD FFT variants of the selected student models, to provide a broader view of how their feature representations perform in this LP inference on MIMIC. **Table 7** builds on **Table 6** by including these variants. While CXR-FMKD FFT reveals similar trends to CXR-FMKD-Direct FFT shown in **Table 6**, the *transfer* LP-variants, especially CXR-FMKD LP, consistently show slight improvements. Notably, CXR-FMKD LP (CS) no longer exhibits the significant underperformance previously discussed. **Table 15** in the *Supplemental Material* combines both **Table 6** and **Table 7**; and **Figure 79**, **Figure 80**, **Figure 81**, **Figure 82**, and **Figure 83** present the corresponding parallel coordinate plots that include these variants.

Linear Probing Inference on MIMIC

		CXR-FM			CXR-Model FFT			CXR-FMKD LP								
								MSE			CS			MSE-CS | α-β		
		Bench-mark (B)	Transfer Model		Bench-mark (B)	Transfer Model		Bench-mark (B)	Transfer Model		Bench-mark (B)	Transfer Model		Bench-mark (B)	Transfer Model	
Metric	***Class***	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)
AUC-PR	Average Classes 1 to 7	0.4359 ± 0.0026	0.4359 ± 0.0026	0.0%	0.4872 ± 0.0034	0.4705 ± 0.0003	-3.4%	0.4457 ± 0.0019	0.4474 ± 0.0003	0.4%	0.2673 ± 0.0671	0.4531 ± 0.0008	69.5%	0.4465 ± 0.0017	0.4518 ± 0.0003	1.2%
AUC-ROC	Average Classes 1 to 7	0.8113 ± 0.0022	0.8113 ± 0.0022	0.0%	0.8394 ± 0.0023	0.8343 ± 0.0005	-0.6%	0.8174 ± 0.0015	0.8206 ± 0.0001	0.4%	0.5948 ± 0.0655	0.8210 ± 0.0007	38.0%	0.8174 ± 0.0014	0.8212 ± 0.0001	0.5%
Maximum Youden's J Statistic	Average Classes 1 to 7	0.4841 ± 0.0036	0.4841 ± 0.0036	0.0%	0.5337 ± 0.0043	0.5314 ± 0.0011	-0.4%	0.4968 ± 0.0025	0.5033 ± 0.0003	1.3%	0.1484 ± 0.1022	0.5042 ± 0.0019	239.8%	0.4949 ± 0.0024	0.5028 ± 0.0004	1.6%
Youden's J Statistic at 20% FPR	Average Classes 1 to 7	0.4448 ± 0.0043	0.4448 ± 0.0043	0.0%	0.5071 ± 0.0060	0.5012 ± 0.0017	-1.2%	0.4616 ± 0.0029	0.4649 ± 0.0007	0.7%	0.1370 ± 0.0977	0.4678 ± 0.0048	241.5%	0.4621 ± 0.0029	0.4673 ± 0.0007	1.1%
								CXR-FMKD-Direct LP								
								MSE			CS			MSE-CS | α-β		
Metric	***Class***															
AUC-PR	Average Classes 1 to 7							0.4479 ± 0.0004	0.4475 ± 0.0002	-0.1%	0.4520 ± 0.0009	0.4528 ± 0.0004	0.2%	0.4477 ± 0.0004	0.4517 ± 0.0003	0.9%
AUC-ROC	Average Classes 1 to 7							0.8203 ± 0.0001	0.8206 ± 0.0001	0.0%	0.8194 ± 0.0006	0.8210 ± 0.0004	0.2%	0.8206 ± 0.0002	0.8212 ± 0.0001	0.1%
Maximum Youden's J Statistic	Average Classes 1 to 7							0.5025 ± 0.0003	0.5037 ± 0.0005	0.2%	0.4978 ± 0.0017	0.5052 ± 0.0009	1.5%	0.5002 ± 0.0003	0.5032 ± 0.0003	0.6%
Youden's J Statistic at 20% FPR	Average Classes 1 to 7							0.4676 ± 0.0007	0.4643 ± 0.0007	-0.7%	0.4662 ± 0.0032	0.4684 ± 0.0032	0.5%	0.4709 ± 0.0012	0.4677 ± 0.0011	-0.7%
								CXR-FMKD FFT								
								MSE			CS			MSE-CS | α-β		
Metric	***Class***															
AUC-PR	Average Classes 1 to 7							0.4857 ± 0.0029	0.4779 ± 0.0001	-1.6%	0.4802 ± 0.0027	0.4685 ± 0.0003	-2.5%	0.4845 ± 0.0030	0.4804 ± 0.0002	-0.9%
AUC-ROC	Average Classes 1 to 7							0.8396 ± 0.0020	0.8371 ± 0.0001	-0.3%	0.8376 ± 0.0013	0.8299 ± 0.0001	-0.9%	0.8403 ± 0.0007	0.8378 ± 0.0001	-0.3%
Maximum Youden's J Statistic	Average Classes 1 to 7							0.5343 ± 0.0041	0.5360 ± 0.0003	0.3%	0.5318 ± 0.0039	0.5196 ± 0.0004	-2.3%	0.5356 ± 0.0008	0.5360 ± 0.0005	0.1%
Youden's J Statistic at 20% FPR	Average Classes 1 to 7							0.5061 ± 0.0036	0.5015 ± 0.0010	-0.9%	0.5045 ± 0.0036	0.4899 ± 0.0010	-2.9%	0.5088 ± 0.0037	0.5009 ± 0.0008	-1.5%

Table 7. Average Performance Comparison of 11 Selected Models and Their Benchmarks After Linear Probing Inference on MIMIC, Post-CheXpert Pre-training.
This table presents a detailed performance analysis of 11 selected '*transfer*' models that were pre-trained on CheXpert and then fine-tuned on MIMIC using only the final classification layer. The models include CXR-FM (teacher), CXR-Model FFT (baseline), and the LP, Direct LP, and FFT variants of each of the three selected student model types (MSE, CS, and MSE-CS | α-β). The analysis covers four metrics—AUC-PR, AUC-ROC, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR—focusing on the average for the most significant disease labels (Classes 1 to 7). Results are shown as mean outcomes (Avg) with standard deviations (SD), derived from testing five distinct instances of the same model type, each developed and trained using a different seed. Performance improvements for each *transfer* model relative to its corresponding *benchmark* (B) are quantified using the formula: (Transfer Model Avg Value − Benchmark Avg Value)/ Benchmark Avg Value × 100% for each metrics. Note that for CXR-FM, the *transfer* and *benchmark* models are equivalent due to its frozen backbone constraint. Additionally, the CXR-FMKD LP (CS)* *benchmark* model shows significant underperformance, which appears to be rectified in the *transfer* setup.

4.2.3. Full Fine-Tuning

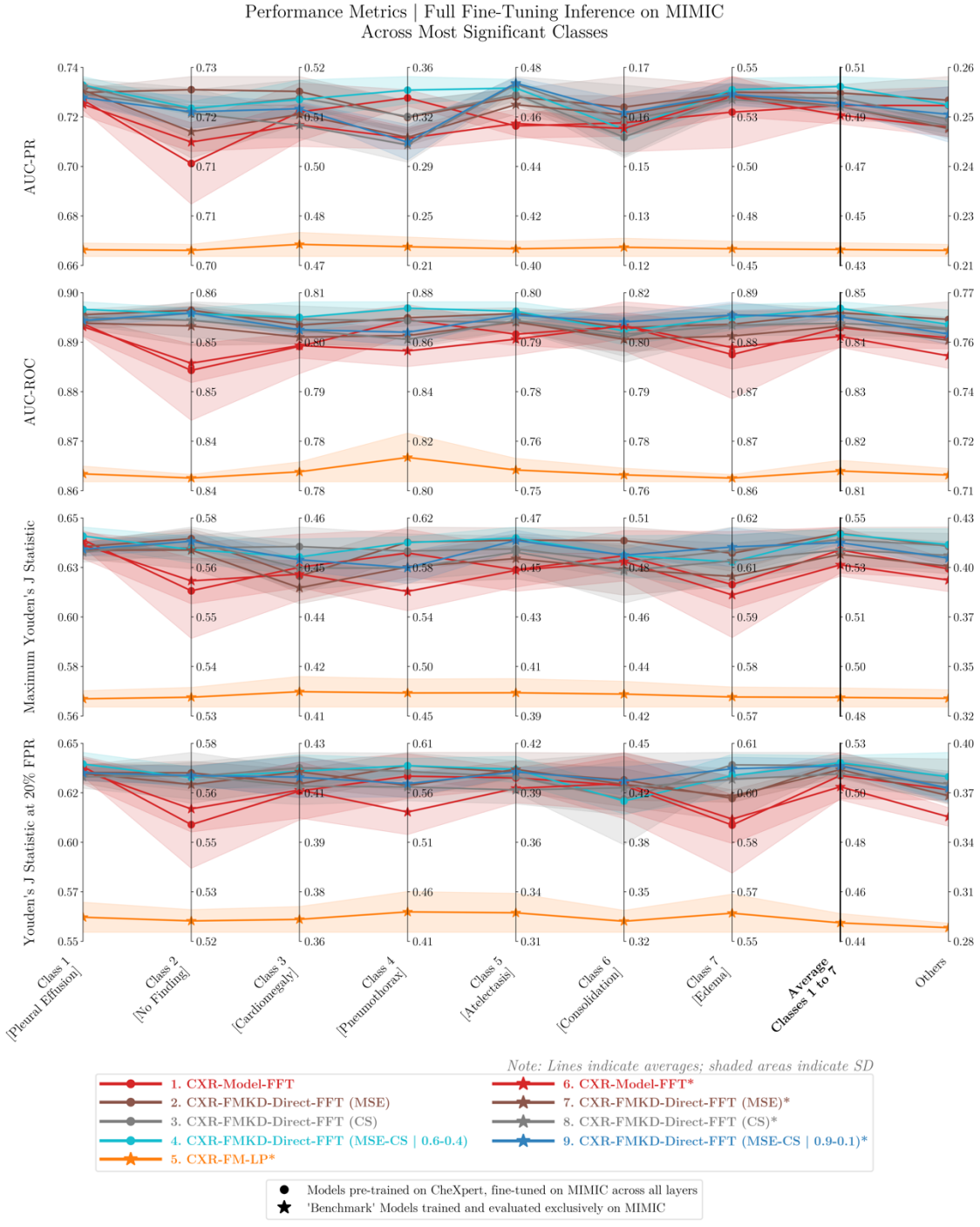

Figure 39. Performance of Selected Transfer Models and Their Benchmarks Across Most Significant Classes After Full Fine-Tuning Inference on MIMIC, Post-CheXpert Pre-training.
These custom parallel coordinate plots display the performance across four metrics: AUC-PR, AUC-ROC, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR for the five selected '*transfer*' models that were pre-trained on CheXpert and then fine-tuned on MIMIC across all layers. The models include CXR-FM (teacher), CXR-Model FFT (baseline), and the CXR-FMKD-Direct FFT variants of the three selected student model types (MSE, CS, and MSE-CS | α-β). Each '*transfer*' model, represented by circles, is contrasted against a corresponding *benchmark*, shown with stars, which shares the same architecture but was trained and tested exclusively on MIMIC. For CXR-FM, both are equivalent. The plots cover the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes in the disease labels list. Each line represents the average results from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation across these tests.

This section evaluates the CheXpert-pre-trained *transfer* models, now fully fine-tuned across all layers, unlike the 'LP' inference where the backbone was frozen. For the teacher model CXR-FM, which inherently has a frozen backbone constraining the fine-tuning process, both its *transfer* and *benchmark* forms are also equivalent as the adaptation of the classification layer remains the same.

Figure 39 demonstrates further improvements in our *transfer* models compared to the 'LP' inference case (**Figure 38**) and, by extension, notably better performance than the 'Direct Transfer' scenario (**Figure 37**). In this 'FFT' context, our CXR-FMKD-Direct FFT *transfer* student models not only match but slightly exceed the CXR-FMKD-Direct FFT* *benchmark* student models across the four metrics overall. For example, in Class 4 [Pneumothorax], our *transfer* models are at the top, whereas the *benchmark* models exhibit lower performance within their range. The *transfer* CXR-Model FFT shows reduced variability and closely aligns with the performance trends of its CXR-Model FFT* *benchmark* counterpart. For instance, there is a slight drop in performance recorded for both models in Class 2 [No Finding] and Class 7 [Edema] relative to the student models. However, CXR-Model FFT slightly outperforms its *benchmark* overall, showcasing a marked improvement in Class 4 [Pneumothorax] for example—a trend also observed in the 'LP' scenario—where it ranks second in *AUC-PR*, just behind CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4). The teacher model, CXR-FM, remains the lowest performer across all classes. Additionally, Larger shaded areas (indicating SD) around the lines representing our *transfer* models are observed in this 'FFT' scenario compared to the 'LP' scenario, which is consistent with our earlier discussion on the variability in performance in the *FFT vs. LP Models* section.

Figure 84 in the *Supplemental Material* extends **Figure 39**'s analysis to all 14 classes. Similar to the 'Direct Transfer' and 'LP' scenarios, the plot reveals consistent trends with occasional fluctuations, such as the relative drop in performance for some models in Class 9 [Enlarged Cardiomegaly] and Class 13 [Fracture].

Full Fine-Tuning Inference on MIMIC

		CXR-FM			CXR-Model FFT			CXR-FMKD-Direct FFT								
								MSE			CS			MSE-CS | α-β		
		Bench-mark (B)	Transfer Model		Bench-mark (B)	Transfer Model		Bench-mark (B)	Transfer Model		Bench-mark (B)	Transfer Model		Bench-mark (B)	Transfer Model	
Metric	***Class***	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)	Avg ± SD	Avg ± SD	%Δ w.r.t. (B)
AUC-PR	Average Classes 1 to 7	0.4359 ± 0.0026	0.4359 ± 0.0026	0.0%	0.4872 ± 0.0034	0.4909 ± 0.0057	**0.8%**	0.4905 ± 0.0017	0.4957 ± 0.0032	**1.1%**	0.4905 ± 0.0033	0.4936 ± 0.0026	**0.6%**	0.4916 ± 0.0025	0.4982 ± 0.0038	**1.3%**
AUC-ROC	Average Classes 1 to 7	0.8113 ± 0.0022	0.8113 ± 0.0022	0.0%	0.8394 ± 0.0023	0.8412 ± 0.0041	**0.2%**	0.8415 ± 0.0020	0.8443 ± 0.0006	**0.3%**	0.8421 ± 0.0025	0.8435 ± 0.0019	**0.2%**	0.8434 ± 0.0006	0.8452 ± 0.0014	**0.2%**
Maximum Youden's J Statistic	Average Classes 1 to 7	0.4841 ± 0.0036	0.4841 ± 0.0036	0.0%	0.5337 ± 0.0043	0.5392 ± 0.0086	**1.0%**	0.5376 ± 0.0041	0.5454 ± 0.0023	**1.5%**	0.5393 ± 0.0058	0.5431 ± 0.0030	**0.7%**	0.5420 ± 0.0009	0.5452 ± 0.0020	**0.6%**
Youden's J Statistic at 20% FPR	Average Classes 1 to 7	0.4448 ± 0.0043	0.4448 ± 0.0043	0.0%	0.5071 ± 0.0060	0.5122 ± 0.0106	**1.0%**	0.5149 ± 0.0033	0.5178 ± 0.0027	**0.6%**	0.5129 ± 0.0059	0.5168 ± 0.0060	**0.8%**	0.5166 ± 0.0031	0.5181 ± 0.0033	**0.3%**

Table 8. Average Performance Comparison of Selected Transfer and Corresponding Benchmark Models After Full Fine-Tuning Inference on MIMIC, Post-CheXpert Pre-training.
This table presents a detailed performance analysis of the five selected '*transfer*' models that were pre-trained on CheXpert and then fine-tuned on MIMIC across all layers. The analysis covers four metrics—AUC-PR, AUC-ROC, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR—focusing on the average for the most significant disease labels (Classes 1 to 7). Results are shown as mean outcomes (Avg) with standard deviations (SD), derived from testing five distinct instances of the same model type, each developed and trained using a different seed. Performance improvements for each *transfer* model relative to its corresponding *benchmark* (B) are quantified using the formula: (Transfer Model Avg Value − Benchmark Avg Value)/ Benchmark Avg Value × 100% for each metrics. Note that for CXR-FM, the *transfer* and *benchmark* models are equivalent due to its frozen backbone constraint.

Focusing on the average for classes 1 to 7 for a concise overview, our CXR-FMKD-Direct FFT *transfer* student models now emerge as the best performers across all metrics, closely followed by the *benchmark* student models. Specifically, CXR-FMKD-Direct FFT (MSE) and FMKD-

Direct FFT (MSE-CS | 0.6-0.4) closely match and are the top contenders, with the latter leading in *AUC-PR* and *AUC-ROC*. CXR-FMKD-Direct FFT (CS) is slightly behind in third place where its performance is matched by CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1)* in *AUC-ROC* and *Youden's J Statistic at 20% FPR*. The *transfer* CXR-Model FFT sits at the lower end of the performance range of these student models, slightly outperforming the *benchmark* CXR-Model FFT*. The teacher model continues to lag significantly behind, echoing trends observed in the other inference scenarios. **Table 8** quantifies these improvements for the average of classes 1 to 7, offering a direct comparison with results from **Table 5** (Direct Transfer) and **Table 6** (LP). We can clearly see the overall improvements with all *transfer* models now showing slight gains over their *benchmarks*, as indicated by positive changes highlighted in green in the table, demonstrating the enhanced performance of the *transfer* models.

4.2.4. Discussion

In the previous section on *Performance Analysis*, we have demonstrated how student models derived from KD, particularly the CXR-FMKD-Direct FFT variants, outperform both the CXR-FM teacher and CXR-Model FFT baseline model within the same dataset used for both KD and subsequent task-specific training. In this section on *Generalisability Analysis*, we aim to verify if the performance improvements observed with these student models on familiar datasets extend to datasets not involved in the KD or task-specific training processes. This approach tests the robustness of the models' learned features when applied to new, unseen data.

It is important to note that the training processes for both the CheXpert and MIMIC datasets were consistent in terms of hyperparameters used, facilitating direct comparisons and interpretations, especially for this section.

Direct Transfer

The 'Direct Transfer' scenario illustrates how the CXR-FMKD-Direct FFT *transfer* student models, which had not been exposed to MIMIC data prior to testing, exhibit notable performance. These models manage to perform on par with, and occasionally surpass, the *benchmark* CXR-FM* teacher model, which was trained directly on MIMIC. This highlights the effectiveness of the feature representations and discrimination capabilities they have acquired from the CheXpert pre-training, further elevated by the knowledge distilled from the CXR-FM teacher during the KD process. This enhancement is evident as these models also outperform the *transfer* CXR-Model FFT, which serves effectively as an equivalent to the CXR-FMKD-Direct FFT students but without the benefit of KD. Overall, the *transfer* models perform reasonably well, actually achieving performance on MIMIC that is not too far from the ranges observed on CheXpert for the four metrics evaluated, without significant degradation.

At the same time, the robust performance of these *transfer* models, matching or outperforming CXR-FM*, reflects poorly on the teacher model's subpar performance. The latter was trained using linear probing—a typical approach for FMs with a frozen backbone—which restricts comprehensive fine-tuning and thus limits its performance on downstream tasks as discussed in **section 4.1.5**. Importantly, this corroborates the need for open access to FMs that allow for flexible backbone adjustments.

It is important to acknowledge the inherent similarities between the co-released CheXpert and MIMIC, such as the shared 14 disease labels enabling this 'Direct Transfer' inference test, which might contribute to the relatively successful performance of *transfer* models. Despite these similarities, significant differences in population characteristics and disease prevalences—such as 25% versus 41% for Class 1 [Pleural Effusion] and 31% versus 9% for Class

2 [No Finding] between MIMIC and CheXpert respectively, as seen in **Figure 22**—introduce distinct challenges. These disparities necessitate robust models that can adapt effectively across varying clinical settings. Interestingly, this could also explain why the *transfer* models perform better for Class 1 [Pleural Effusion], more prevalent in CheXpert, and the *benchmark* models perform better for Class 2 [No Finding], more prevalent in MIMIC, reflecting the inherited characteristics and discrimination capabilities derived from each dataset. This underscores the nuanced impact of training dataset characteristics on model performance across different classes, as also outlined in **section 4.1.5**.

In conclusion, the satisfactory performance of the non-adapted *transfer* models highlights that the knowledge garnered from the CheXpert dataset retains its relevance and discriminatory power when applied to MIMIC, with student models further leveraging the teacher's insights to achieve even better performance.

The forthcoming discussions on the LP and FFT inference scenarios will analyse how the *transfer* models adapt to the new MIMIC task under varying fine-tuning regimes, pertinent to understanding the typical adaptation of an FM to a new task within the context of generalisability being examined. This 'Direct Transfer' case can serve as a baseline to compare how our *transfer* models perform without any fine-tuning against their performance with fine-tuning under LP and FFT scenarios.

Linear Probing

LP can be seen as the fine-tuning method of choice typically when fewer training data are available, as it involves learning only the parameters of the classifier while keeping the input-to-feature mapping of the large backbone intact. In our case, this approach also provides a direct way to evaluate the effectiveness of the features generated by each *transfer* model's backbone. Indeed, by replacing the old classifier and fine-tuning only the new one, we can assess how useful the features being processed—developed through training with CheXpert—are in aiding discrimination for the MIMIC disease detection task.

This rationale underlies our decision to include all four variants of the selected student models to examine the different resulting feature mappings. For instance, *transfer* CXR-FMKD LP retains the exact feature representations learned during the KD process without further task-specific fine-tuning on CheXpert, similar to *transfer* CXR-FMKD-Direct LP, which uses the backbone's input-to-feature mapping learned during KD but omits the projector, resulting in features of size 1664 (from DenseNet169) rather than 1376 (from CXR-FM). Therefore, through the 'LP' inference on MIMIC, these *transfer* LP-variants help evaluate the direct utility of features extracted post-KD, without exposure to the actual disease detection task common to both CheXpert and MIMIC. Conversely, the *transfer* FFT-variants, while also deriving their input-to-feature mapping from KD, enable the test of features that have been further optimised through subsequent fine-tuning tailored specifically to the disease detection task. Note that, by analogy, the *benchmark* LP-variants—which are adapted through LP on MIMIC after their features have been derived from KD using MIMIC as well—also reflect the effectiveness of these features in similar terms to those described above for the *transfer* students. This setup provides a direct comparison point with the *transfer* LP-variants, illustrating differences in feature utility when applied through LP for the MIMIC task based on the dataset (MIMIC or CheXpert) used in the KD process.

From another perspective, the *transfer* LP-variants could be seen as simulating scenarios where the KD transfer set is unlabelled, utilised solely for feature extraction akin to SSL practices, but with high-dimensional 'ground truth' targets being the features from the teacher in response to given CXR inputs. Meanwhile, the *transfer* FFT-variants might represent

situations where the KD transfer set is labelled, allowing the models, post-KD, to be refined further through task-specific training. This is expected to provide an advantage in this 'LP' inference scenario, particularly as CheXpert and MIMIC share the same disease labels.

In terms of results, there is a marked improvement in performance for all *transfer* CXR-FMKD-Direct FFT student models compared with the 'Direct Transfer' scenario. The performance gaps with corresponding *benchmark* models are much smaller, though they still slightly underperform overall. This underperformance is expected, given that the *benchmarks* underwent comprehensive fine-tuning (FFT) on MIMIC, unlike the *transfer* models which only received LP. Yet, the performance of these *transfer* models remains notably close to that of the *benchmarks*. This further highlights the robustness of the features learned by the *transfer* models after their KD and task-specific training on CheXpert, echoing observations from the 'Direct Transfer' case. Notably, in terms of the *Maximum Youden's J Statistic*, the *transfer* CXR-FMKD-Direct FFT student models demonstrate performance overlapping with that of their *benchmark* counterparts. This suggests a high potential for achieving performance on par with *benchmarks* by optimising decision thresholds.

Similar to the 'Direct Transfer' scenario, the *transfer* CXR-FMKD-Direct FFT student models outperform the *transfer* CXR-Model FFT baseline overall, reflecting further performance enhancements provided by the KD process—namely, the knowledge gained from the CXR-FM teacher. This underscores the broad benefits of utilising FMs and KD in this context. Indeed, KD has facilitated the development of more robust features, enhancing generalisability in a medical domain where even incremental improvements in disease detection accuracy are important. The *transfer* CXR-FMKD-Direct FFT students even slightly outperform or match the *benchmark* CXR-Model FFT* across all metrics except *AUC-PR* for the average of classes 1 to 7. This emphasises the effectiveness of these *transfer* student models' features, which are informative enough to enable better discriminative capabilities than the traditional baseline trained exclusively on MIMIC. As expected, our models significantly surpass the teacher CXR-FM in performance, reflecting the previously discussed constraints on its architecture and limited adaptation capabilities.

As mentioned earlier, a detailed analysis of performance variations per class and metric is beyond the scope of this study. We primarily aggregate results to check on the consistency of trends across different classes and metrics, thereby providing a more robust interpretation of the performance of our student models compared to others. Nevertheless, it is interesting to discuss the effects extrapolated from CheXpert through the *transfer* models. For instance, similar to the 'Direct Transfer' scenario, where CheXpert's capabilities were directly transferred to MIMIC, we observe relatively lower performances for Class 2 [No Finding] across metrics for both the *transfer* CXR-Model FFT and the *transfer* CXR-FMKD-Direct FFT student models. Additionally, for Class 4 [Pneumothorax], these *transfer* students exhibit higher performance across most metrics (except for *AUC-PR*) when tested on MIMIC after LP. These variations could be attributed to the retained characteristics from CheXpert that continue to manifest in different inference scenarios, reflecting the specific pre-training data's influence on the models' ability to generalise across OOD tasks.

Other variations in performance, such as the peak in performance for the *transfer* CXR-Model FFT in Class 4 [Pneumothorax], are less straightforward to explain. No clear advantage taken from CheXpert, as demonstrated in the 'Direct Transfer' scenario, is clearly extrapolated to the 'LP' observations for this class. This could be due to useful discrimination encoded in the features from CheXpert, which when utilised in 'LP' on MIMIC, lead to particularly amplified discrimination capabilities. Moreover, the generally lower performances for *AUC-PR* among our

transfer models can be attributed to the imbalances in disease prevalence noted in both CheXpert and MIMIC, and how these factors interplay.

In this 'LP' scenario, we also included the other student model variants for analysis, as previously mentioned. We observe similar ranking trends to those noted in the *Performance Analysis* section, with the *transfer* LP-variants demonstrating lower performances compared to the *transfer* FFT-variants. This echoes, as expected, that features further refined through task-specific fine-tuning on CheXpert post-KD offer a performance advantage for this inference scenario on MIMIC compared to features derived solely through the CheXpert KD process. Generally, this suggests that the more informative these features are for CheXpert, the more useful they also are for MIMIC, aligning with earlier observations about the utility of knowledge acquired from CheXpert for the MIMIC task.

Looking into the FFT-variants separately, similar to the *transfer* CXR-FMKD-Direct FFT, the *transfer* CXR-FMKD FFT models also show close performance gaps with their *benchmarks* while still slightly underperforming. This reflects the effectiveness of the features utilised—whether with or without the projector—after task-specific fine-tuning on CheXpert's disease detection task post-KD. This further confirms that such fine-tuning is particularly effective for adapting to MIMIC, demonstrating the utility of features specifically tailored for this shared task.

On the other hand, the *transfer* LP-variants not only close the gap with their *benchmark* counterparts but also predominantly slightly outperform them, particularly evident in the *transfer* CXR-FMKD LP models. This suggests that the feature mapping learned from the CheXpert KD process may be more effective than that learned from the MIMIC KD process when adapted through LP on MIMIC itself. The more pronounced improvements of the *transfer* models compared to the *benchmark* models for CXR-FMKD LP than for CXR-FMKD-Direct LP can be attributed to these *transfer* LP-variants exhibiting similar, overlapping results in this inference scenario—less distinct than those observed in the *transfer* FFT-variants. Since these *transfer* LP-variants slightly outperform the CXR-FMKD-Direct LP *benchmarks*, which themselves rank higher than the inherently lower-performing CXR-FMKD LP *benchmarks*, the relative boost in performance becomes more noticeable. Notably, the significant underperformance previously observed for CXR-FMKD LP (CS) was rectified when its features were used through LP to fine-tune the classifier. The features generated by KD using CheXpert provided a beneficial pathway to more optimal performance for the model when adapted to MIMIC through LP, further highlighting the intricate relationships in cross-dataset testing and transfer as part of this generalisability analysis.

Lastly, returning to the *transfer* LP-variants that exhibit overlapping performance versus the more distinct results for the *transfer* FFT-variants—with CXR-FMKD-Direct FFT ranking highest—this could be attributed to the *transfer* FFT-variants already possessing very informative features. These features do not require much 'effort' from the classification layer to achieve good discrimination for MIMIC, highlighting more initial advantages. In contrast, the *transfer* LP-variants require more 'effort' in learning discrimination, which in doing so, brings the performance of both *transfer* CXR-FMKD-Direct LP and *transfer* CXR-FMKD LP closer together.

Full Fine-Tuning

The discussion from the 'LP' inference scenario extends naturally into this 'FFT' scenario, illustrating the advantages of adapting our *transfer* models through comprehensive fine-tuning on MIMIC. This approach can be viewed as the optimal strategy, capitalising on the complete

reconstruction of our student models, now fully accessible, in contrast to the limited adjustments possible with the frozen CXR-FM teacher model, which already sits at the bottom of the performance range. Unlike the 'LP' scenario, FFT can also be seen as the method of choice when more training data is available, as it involves adjusting the entire model, including the backbone, which typically requires substantially more data to train effectively. Through FFT, our *transfer* models, already enriched by the KD process and task-specific training on CheXpert, have their entire networks optimised for the MIMIC task. This potentially provides better initial directions for optimisation.

Recalling the proven robustness of the features derived from CheXpert—where the CheXpert KD process yielded more effective features than the MIMIC KD process for the CXR-FMKD LP-variants in 'LP' inference on MIMIC for example—it is reasonable to expect our *transfer* models to now surpass their *benchmark* counterparts. This hypothesis aligns with the previous 'LP' discussion where the CXR-FMKD-Direct FFT *transfer* student models were performing comparably to the *benchmarks*, suggesting they might now outrank them. Indeed, this has been confirmed for 'FFT' where these *transfer* students slightly outperform the *benchmarks*, indicating an upper limit to the performance reached under this configuration. This highlights an important point relating the 'LP' and 'FFT' scenarios: our *transfer* student models already achieved competitive results in 'LP' inference on MIMIC, performing close to those in the 'FFT' scenario, which suggests that the models have captured generalisable, transferrable features. This small improvement in the 'FFT' scenario compared to 'LP' for our *transfer* student models, which are close in performance to the *benchmark* models trained exclusively on MIMIC in both scenarios, contrast with a less desired large improvement that could indicate less useful and missing representations for our models.

On another interesting note, in this 'FFT' inference on MIMIC, the *transfer* models now more closely follow the performance trends of their *benchmark* counterparts across metrics and classes compared to the other inference scenarios. This showcases the impact of FFT, where the entire model's parameters can be overwritten to optimise for the MIMIC detection task more effectively. As previously mentioned, FFT allows for modifications to the input-to-feature mapping and applies more extensive changes, potentially rewriting patterns learned from CheXpert to better optimise for MIMIC. Lastly, similar to the 'LP' scenario, our *transfer* student models outperforming the *transfer* CXR-Model FFT corroborates the added performance improvements in OOD testing provided by the KD process that the students underwent, highlighting their good generalisability characteristics.

Conclusion

'Direct Transfer' serves as a 'zero-shot' performance evaluation, showing that our *transfer* models perform reasonably well on MIMIC, the OOD dataset. Furthermore, it underscores the utility and relevance of the knowledge gained from CheXpert, the in-distribution (ID) dataset, allowing learned features to maintain discrimination power on MIMIC, with *transfer* students even matching the *benchmark* teacher performance.

'LP' isolates the feature representation quality. The ID-pre-trained *transfer* models demonstrated robust results, with the *transfer* students also exhibiting performances close to their *benchmark* counterparts. This indicates that the feature representations learned from the CheXpert dataset are effectively transferable and have maintained their utility in a different dataset context. These models performed similarly to the 'FFT' scenario, suggesting that the features are already generalisable and adaptable.

Moreover, in all cases, the *transfer* CXR-FMKD-Direct FFT students surpass the baseline *transfer* CXR-Model FFT, illustrating the benefits of KD. This advantage is particularly

pronounced in the 'FFT' scenario, where *transfer* models exceed *benchmark* performances, establishing them as the top performers. This not only highlights the effectiveness of the KD process but also underscores the significant potential for enhancing model adaptability and performance through strategic knowledge transfer.

Overall, the results from the 'Direct Transfer', 'LP', and 'FFT' scenarios collectively demonstrate that our student models, by leveraging KD and task-specific training on CheXpert, have successfully captured generalisable and robust features. These features not only enable competitive performance across different settings but also position the distilled student models as superior alternatives in the evaluated disease detection tasks, proving the important value of our approach in this medical context.

4.3. Bias Analysis

This bias analysis section follows similar procedures and output types as those outlined in the paper by Glocker et al. [93], while also introducing a *Novel Bias Score* for **Bias Inspection** to streamline discussions and comparisons across the large number of models we have evaluated.

4.3.1. Bias Inspection

4.3.1.1. CheXpert

Figure 40 presents a feature space analysis of our CXR-FM teacher model using joint scatterplots to examine the marginal distributions for the different subgroups across the first four PCA modes (shown in the first two rows of plots) and the two t-SNE dimensions (shown in the last row of plots). This analysis is based on a balanced random sample of 3000 patients, with 1000 from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), race (White, Asian, and Black), and age (grouped into eight age bins) arranged from left to right in the figure—in the first, second, third, and fourth columns of plots, respectively.

For our bias analysis specifically, we focus primarily on the first three columns relating to disease, sex, and race. Age data, while collected, is less relevant for bias analysis and is omitted in **Figure 41** to simplify the visualisation of marginal distributions. **Figure 40** is presented to provide a foundational view, illustrating that the same sample—represented by dots in the plots—encapsulates various types of information overlaid and is analysed differently across subgroups and dimensions to understand distributional similarities or discrepancies. **Figure 42** further presents the marginal distributions for the CXR-FMKD-Direct FFT (MSE) student model, facilitating a direct comparison with the teacher model showcased in **Figure 41**.

Table 9 and **Table 10** statistically quantify these marginal distribution comparisons in **Figure 41** and **Figure 42** respectively through two-sample Kolmogorov-Smirnov (KS) tests, which generate p-values to test the null hypothesis that distributions for each subgroup pair are identical. These p-values are colour-coded for clarity: green if $p \geq 0.05$ (not significant), yellow if $0.001 \leq p < 0.05$ (significant), and red if $p < 0.001$ (highly significant), enabling easy identification of statistical significance levels across the tables.

Figure 40. Joint Scatterplots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FM Tested on CheXpert.
This figure displays joint scatterplots with corresponding marginal distributions represented at the axes for relevant subgroup comparisons across the first four PCA modes (shown in the first two rows of plots) and the two t-SNE dimensions (shown in the last row of plots) applied to the feature embeddings extracted from CXR-FM tested on CheXpert. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), race (White, Asian, and Black), and age (grouped into eight age bins) arranged from left to right in the figure—in the first, second, third, and fourth columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

For space and clarity, we will not display the corresponding joint plot, marginal plot, and KS tests table for the other CXR-FMKD-Direct FFT student models and the CXR-Model FFT baseline in this section. Instead, the novel bias score calculations provided later in this section will implicitly cover the feature space analysis for each of these models.
All detailed plots and tables for our selected models are available in the *Supplemental Material* section in *S.6. Bias Analysis | Bias Inspection – CheXpert*. The joint plot, marginal plot, and table for the models are as follows:
Figure 85, **Figure 86**, and **Table 16** for CXR-FM;
Figure 87, **Figure 88**, and **Table 17** for CXR-Model FFT;
Figure 89, **Figure 90**, and **Table 18** for CXR-FMKD-Direct FFT (MSE);
Figure 91, **Figure 92**, and **Table 19** for CXR-FMKD-Direct FFT (CS); and
Figure 93, **Figure 94**, and **Table 20** for CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4).

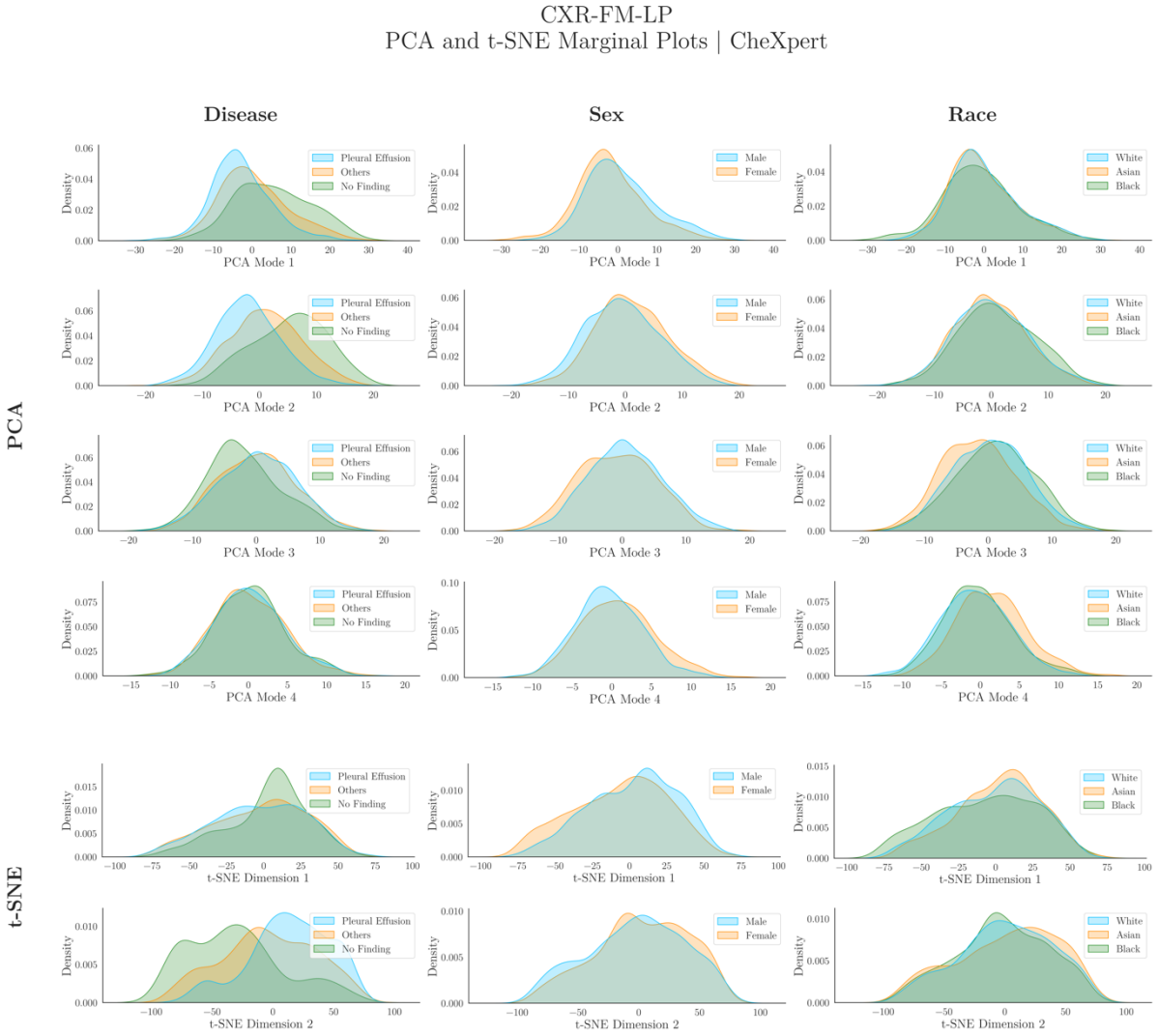

Figure 41. Marginal Plots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FM Tested on CheXpert.
This figure displays the marginal distributions for relevant subgroup comparisons across the first four PCA modes (shown in the first four rows of plots) and the two t-SNE dimensions (shown in the last two rows of plots) applied to the feature embeddings extracted from CXR-FM tested on CheXpert. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), and race (White, Asian, and Black), arranged from left to right in the figure—in the first, second, and third columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

Model (CheXpert)	Mode	Explained Variance	P-Values				
			Pleural Effusion vs No Finding	White vs Asian Patients	White vs Black Patients	Asian vs Black Patients	Male vs Female Patients
CXR-FM	PCA Mode 1	16.71%	1.90E-29	1.00	4.62E-02	7.13E-02	1.55E-11
	PCA Mode 2	8.63%	1.44E-47	1.00	6.64E-02	2.51E-03	2.11E-07
	PCA Mode 3	7.16%	4.09E-08	1.76E-09	7.69E-02	7.89E-16	4.09E-08
	PCA Mode 4	4.21%	1.00	1.25E-13	1.00	4.29E-13	7.18E-08

Table 9. Statistical Comparison of Pairwise Subgroup Marginal Distributions across PCA Modes for CXR-FM Tested on CheXpert.
Two-sample Kolmogorov-Smirnov tests were conducted to compare the pairwise subgroup marginal distributions depicted in **Figure 41** across the first four PCA modes. These tests yielded p-values testing the null hypothesis that the distributions for each subgroup pair are identical. The Benjamini-Yekutieli procedure was then applied to adjust the p-values for multiple testing, with the significance determined at a p-value < 0.05 (95% confidence level). P-values are colour-coded in the table: green if $p \geq 0.05$ (not significant), yellow if $0.001 \leq p < 0.05$ (significant), and red if $p < 0.001$ (highly significant), indicating increasing levels of statistical significance.

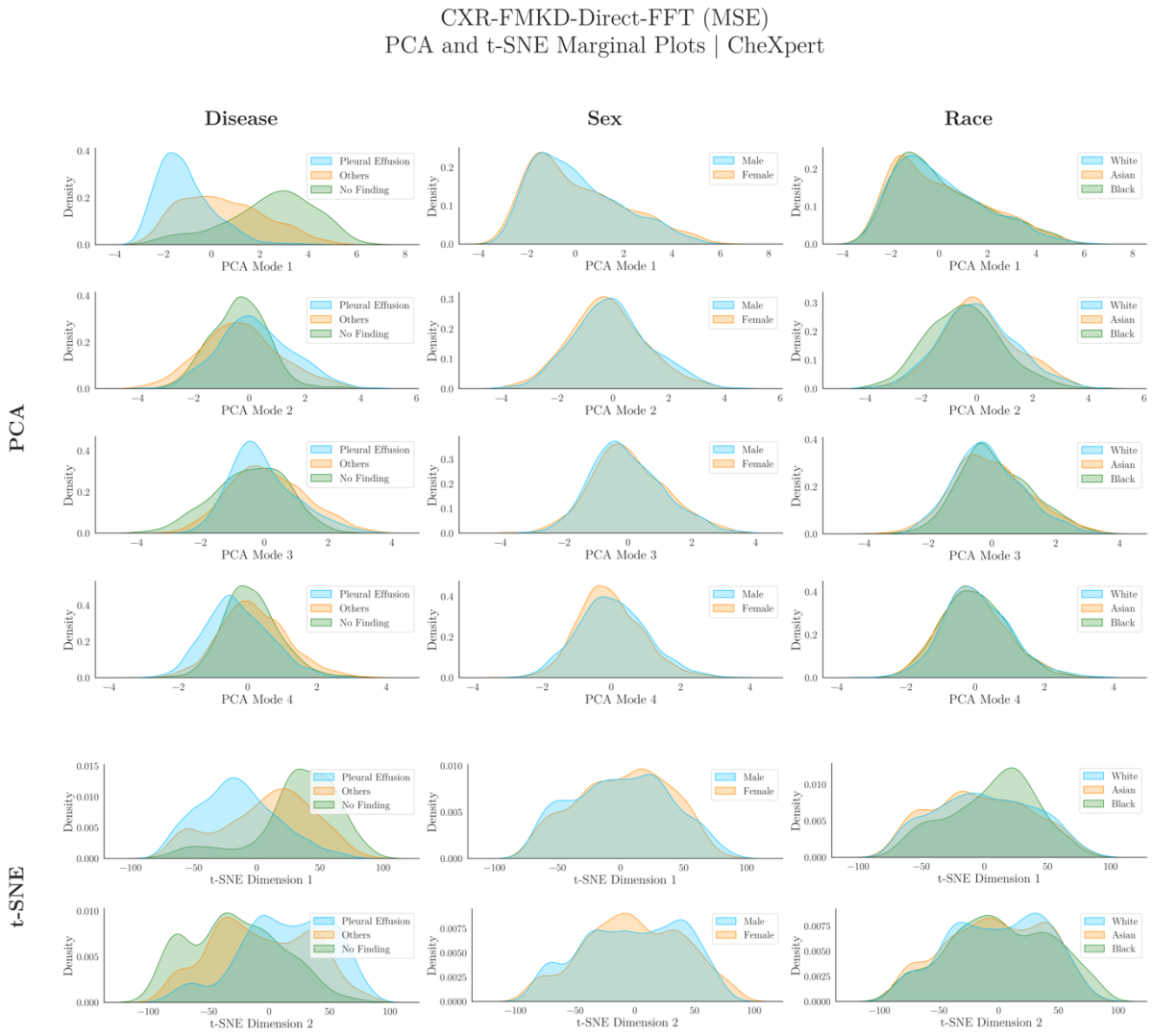

Figure 42. Marginal Plots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FMKD-Direct FFT (MSE) Tested on CheXpert.
This figure displays the marginal distributions for relevant subgroup comparisons across the first four PCA modes (shown in the first four rows of plots) and the two t-SNE dimensions (shown in the last two rows of plots) applied to the feature embeddings extracted from CXR-FMKD-Direct FFT (MSE) tested on CheXpert. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), and race (White, Asian, and Black), arranged from left to right in the figure—in the first, second, and third columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

			P-Values				
Model (CheXpert)	**Mode**	**Explained Variance**	**Pleural Effusion vs No Finding**	**White vs Asian Patients**	**White vs Black Patients**	**Asian vs Black Patients**	**Male vs Female Patients**
CXR-FMKD-Direct FFT (MSE)	PCA Mode 1	19.47%	7.77E-136	9.08E-01	1.00	6.49E-01	1.00
	PCA Mode 2	9.95%	2.35E-09	1.00	2.67E-08	1.27E-10	1.50E-01
	PCA Mode 3	7.16%	1.04E-04	2.18E-01	1.76E-02	3.49E-02	1.00
	PCA Mode 4	4.96%	1.27E-10	1.00	1.00	1.00	1.01E-01

Table 10. Statistical Comparison of Pairwise Subgroup Marginal Distributions across PCA Modes for CXR-FMKD-Direct FFT (MSE) Tested on CheXpert.
Two-sample Kolmogorov-Smirnov tests were conducted to compare the pairwise subgroup marginal distributions depicted in **Figure 42** across the first four PCA modes. These tests yielded p-values testing the null hypothesis that the distributions for each subgroup pair are identical. The Benjamini-Yekutieli procedure was then applied to adjust the p-values for multiple testing, with the significance determined at a p-value < 0.05 (95% confidence level). P-values are colour-coded in the table: green if $p \geq 0.05$ (not significant), yellow if $0.001 \leq p < 0.05$ (significant), and red if $p < 0.001$ (highly significant), indicating increasing levels of statistical significance.

Looking at **Figure 41** and **Figure 42**, we notice more as well as bigger differences in the marginal distributions for the CXR-FM teacher compared to the CXR-FMKD-Direct FFT (MSE) student across the biological sex and race subgroups—i.e., the protected attributes. These discrepancies are particularly notable for the sex subgroup distributions (*Male vs Female*)—shown in the second column—where distinct separations between the distributions are consistently observed across all four PCA modes for CXR-FM, in contrast to the more aligned distributions for CXR-FMKD-Direct FFT (MSE). Similarly, larger shifts in distributions are evident between the race subgroup distributions—third column—in PCA modes 3 and 4 for CXR-FM compared to the student model. In the t-SNE plots, shown in the last two rows, the separation between marginal distributions is less obvious, potentially due to the arbitrary nature of t-SNE's dimension orientations which can obscure these relationships. Still, a slightly larger mismatch between the distributions for the sex subgroups is visible for CXR-FM in the first t-SNE dimension. Notably, for disease labels, the student model exhibits significant separations between 'Pleural Effusion' and 'No Finding', especially in the first PCA mode, which captures most of the variance and is deemed most relevant. This is in contrast to the teacher model, where distributions are closer together, underscoring the student model's enhanced disease discrimination capabilities. This in turn translates into superior performance in the disease detection task for the student model, as seen in the previous *Performance Analysis* section. These differences are also evident in the first t-SNE dimension, further highlighting the student model's advantages.

Let's now examine the corresponding statistical analysis in **Table 9** and **Table 10**. For biological sex, the differences in distributions for *Male vs Female* are 'highly' significant across all four PCA modes for CXR-FM, with p-values less than 0.001 (all indicated in red)—$p = 1.55E^{-11}$, $p = 2.11E^{-07}$, $p = 4.09E^{-08}$, $p = 7.18E^{-08}$. This contrast, these differences are not statistically significant for the CXR-FMKD-Direct FFT (MSE) student, with all p-values shown in green—$p = 1.00$, $p = 1.50E^{-01}$, $p = 1.00$, $p = 1.01E^{-01}$. Regarding race subgroups, the teacher model displays more significant differences compared to the student model. For *White vs Asian*, the teacher model shows highly significant differences in PCA modes 3 and 4, whereas none are observed for the student. For *White vs Black*, the teacher model has one significant difference in the most informative PCA mode 1, while the student exhibits a highly significant result in PCA mode 2 and a significant result in PCA mode 3. For *Asian vs Black*, the teacher model has significant differences in PCA modes 2 to 4, compared to the student model which shows significance in PCA modes 2 and 3 only. The evident larger number of significant p-values observed for CXR-FM compared to CXR-FMKD-Direct FFT (MSE) therefore confirms the visual observations from the marginal plots. Additionally, the notably lower, near-zero p-value for the student in PCA mode 1 at $p = 7.77E^{-136}$—compared to $p = 1.90E^{-29}$ for the teacher—underscores the more pronounced distribution shifts for disease labels. This further points to stronger disease discrimination capabilities for the student model, as discussed earlier.

Recall that the observations described above stem from one balanced random sample of 3000 patients, with 1000 selected from each racial group. To ensure robustness and mitigate sampling variability, the statistical tests were replicated across 5000 bootstrapping-like simulations, each involving a similar balanced stratified sample. This led to the proposed bias quantification approach, which condenses the results from these simulations into single bias scores for each model across subgroups, culminating in an overall averaged bias score. **Table 11** displays the results of this approach for all our selected models, including the teacher CXR-FM, the CXR-FMKD-Direct FFT student variants (MSE, CS, MSE-CS | 0.6-0.4), and the baseline CXR-Model FFT. This table translates the detailed analyses of marginal distributions into comprehensive bias scores ranging from 0 (unbiased) to 150 (highly biased). For a detailed explanation of the bias score development, please refer to the *Novel Bias Score* methodology section.

CheXpert

Model (CheXpert)	Race: Aggregate P-Value Significance	Race: Bias Score	Race: Aggregate P-Value Significance	Race: Combined Bias Score	Sex: Aggregate P-Value Significance	Sex: Combined Bias Score	Overall: Aggregate P-Value Significance	Overall: Combined Bias Score
CXR-FM	*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						
CXR-Model FFT	*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						
CXR-FMKD-Direct FFT (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						
CXR-FMKD-Direct FFT (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						
CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4)	*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						

	P-VALUES RANGE	P-SCORES
FALSE	($p \geq 0.05$)	0
TRUE	($0.001 \leq p < 0.05$)	100
TRUE+	($p < 0.001$)	150

Table 11. Proposed Novel Bias Score Results for Selected Models Tested on CheXpert.
This table displays bias quantification results for our five selected models on CheXpert: teacher CXR-FM, CXR-FMKD-Direct FFT student variants (MSE, CS, MSE-CS | 0.6-0.4), and baseline CXR-Model FFT. Results are derived from 5000 bootstrapping-like simulations of a balanced stratified sample of 3000 patients (1000 from each race), categorised as 'FALSE' ($p \geq 0.05$), 'TRUE' ($0.001 \leq p < 0.05$), and 'TRUE+' ($p < 0.001$), with corresponding p-scores of 0, 100, and 150, yielding bias scores ranging from 0-150. For race, three pairwise subgroup comparisons—*White vs Asian*, *White vs Black*, and *Asian vs Black*—are averaged for race attribute results. The sex attribute includes only *Male vs Female*. These scores are then averaged to determine overall model bias. Colour scales highlight the level of bias across all models, with red indicating higher bias (undesirable) and green indicating lower. For a detailed explanation of the bias score development, please refer to the *Novel Bias Score* methodology section.

First, the results for the teacher CXR-FM and the CXR-FMKD-Direct FFT (MSE) student are consistent with the observations from the marginal distributions of the 3000-patient sample. The interpretation subsection in the *Novel Bias Score* methodology section details how exactly these bias scores are linked to the p-values derived from the marginal distribution comparisons. Higher bias scores for a pairwise subgroup comparison indicate larger separations (mismatch) in marginal distributions across the PCA modes for this pair. Here, the CXR-FM teacher exhibits higher bias scores for both race and sex attributes compared to the CXR-FMKD-Direct FFT (MSE) student, which corresponds to the more pronounced differences in marginal distributions observed for this teacher model.

More generally, the CXR-FM teacher exhibits the highest bias scores compared to the other models. Specifically, CXR-FM presents the highest overall bias with a score of 108.90, significantly more than the others. This is followed by CXR-FMKD-Direct FFT (CS) at 48.82, with CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4) and CXR-Model FFT closely matching at 29.56 and 27.24, respectively. The least biased model is the CXR-FMKD-Direct FFT (MSE) at 16.53, approaching the ideal unbiased score of 0. An advantage of this bias quantification approach is the ability to dissect bias scores by protected attribute. Across race and sex attributes, the bias ranking trends are similar, although CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4) achieves a slightly lower bias score than CXR-Model FFT for the race attribute. Drilling deeper into the race attribute, which consists of comparisons between *White vs Asian*, *White vs Black*, and *Asian vs Black*, we notice variations in trends. For instance, CXR-FM shows the most bias in the *White vs Asian* and *Asian vs Black* comparisons but the least in the *White vs Black*. Conversely, CXR-Model FFT shows the most bias for the *White vs Black* comparison, closely followed by CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4). Interestingly, models generally exhibiting lower biases often present higher biases for this *White vs Black* comparison and vice versa. The other *White vs Asian* and *Asian vs Black* race comparisons generally follow the overall bias score trends.

Lastly, note that the raw data from these 5000 simulations, displayed as frequencies of categorised 'FALSE' ($p \geq 0.05$), 'TRUE' ($0.001 \leq p < 0.05$), and 'TRUE+' ($p < 0.001$), and before the application of the bias score calculation steps, is detailed in **Table 21** in the *Supplemental Material* section. This table also presents the explained variances per PCA mode for each of the five selected models. The combined four PCA modes account for approximately 41.83% of the variance in CXR-Model FFT, 41.55% in CXR-FMKD-Direct FFT (MSE), 37.40% in CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4), 36.72% in CXR-FM, and the least at 25.42% in CXR-FMKD-Direct FFT (CS). Notably, a lower percentage suggests that a model's feature representation captures more information which may not necessarily relate to the disease detection.

4.3.1.2. MIMIC

Before presenting the bias score results, let's first examine the marginal distributions for one balanced random stratified sample of 3000 patients, to visually gauge any initial differences. Similar to the CheXpert section, we inspect the marginal plots for the CXR-FM teacher and the CXR-FMKD-Direct FFT (MSE) student, displayed in **Figure 43** and **Figure 44** respectively.

Unlike the distinct differences seen for CheXpert, the disparities in the marginal distributions between CXR-FM and CXR-FMKD-Direct FFT (MSE) on MIMIC are not as pronounced across the protected characteristics. The trends across the race and sex subgroups appear similar between the two models, with the teacher model occasionally showing fewer shifts in some racial distributions in the PCA modes. However, bias score calculations post-5000 simulations reveal

Figure 43. Marginal Plots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FM Tested on MIMIC.
This figure displays the marginal distributions for relevant subgroup comparisons across the first four PCA modes (shown in the first four rows of plots) and the two t-SNE dimensions (shown in the last two rows of plots) applied to the feature embeddings extracted from CXR-FM tested on MIMIC. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), and race (White, Asian, and Black), arranged from left to right in the figure—in the first, second, and third columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

that the teacher model exhibits relatively larger differences, as indicated by a higher overall bias score. This exemplifies a sampling instance where both models show closer results in PCA feature space analysis, underscoring the significance of conducting multiple simulations for robustness in results. Nevertheless, we do note more marked differences for the teacher in the t-SNE dimensions (shown in the last two rows of plots) for both race and sex subgroups compared to the student model. Additionally, this student model displays a significantly larger separation for *Pleural Effusion vs No Finding* in the most informative PCA mode 1, consistent with previous observations on CheXpert. This supports the student model's superior discrimination capabilities in disease detection, which translate into improved performance—as shown in the previous *Performance Analysis* section. These broader shifts are also prominently visible in the t-SNE dimensions for these disease labels.

Figure 44. Marginal Plots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FMKD-Direct FFT (MSE) Tested on MIMIC.
This figure displays the marginal distributions for relevant subgroup comparisons across the first four PCA modes (shown in the first four rows of plots) and the two t-SNE dimensions (shown in the last two rows of plots) applied to the feature embeddings extracted from CXR-FMKD-Direct FFT (MSE) tested on MIMIC. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), and race (White, Asian, and Black), arranged from left to right in the figure—in the first, second, and third columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

We restricted these initial observations to the CXR-FM and CXR-FMKD-Direct FFT (MSE) models for clarity and conciseness. However, all detailed plots and tables for our selected models are also available in the *Supplemental Material* section in *S.7. Bias Analysis | Bias Inspection – MIMIC*. The joint plot, marginal plot, and KS table for the models, based on one balanced random sample of 3000 patients, are as follows:
Figure 95, **Figure 96**, and **Table 22** for CXR-FM;
Figure 97, **Figure 98**, and **Table 23** for CXR-Model FFT;
Figure 99, **Figure 100**, and **Table 24** for CXR-FMKD-Direct FFT (MSE);
Figure 101, **Figure 102**, and **Table 25** for CXR-FMKD-Direct FFT (CS); and
Figure 103, **Figure 104**, and **Table 26** for CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1).

MIMIC

Model (MIMIC)	Race: Aggregate P-Value Significance	Race: Bias Score	Race: Aggregate P-Value Significance	Race: Combined Bias Score	Sex: Aggregate P-Value Significance	Sex: Combined Bias Score	Overall: Aggregate P-Value Significance	Overall: Combined Bias Score
CXR-FM	*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						
CXR-Model FFT	*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						
CXR-FMKD-Direct FFT (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						
CXR-FMKD-Direct FFT (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						
CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1)	*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						

	P-VALUES RANGE	P-SCORES
FALSE	($p \geq 0.05$)	0
TRUE	($0.001 \leq p < 0.05$)	100
TRUE+	($p < 0.001$)	150

Table 12. Proposed Novel Bias Score Results for Selected Models Tested on MIMIC.
This table displays bias quantification results for our five selected models on MIMIC: teacher CXR-FM, CXR-FMKD-Direct FFT student variants (MSE, CS, MSE-CS | 0.9-0.1), and baseline CXR-Model FFT. Results are derived from 5000 bootstrapping-like simulations of a balanced stratified sample of 3000 patients (1000 from each race), categorised as 'FALSE' ($p \geq 0.05$), 'TRUE' ($0.001 \leq p < 0.05$), and 'TRUE+' ($p < 0.001$), with corresponding p-scores of 0, 100, and 150, yielding bias scores ranging from 0-150. For race, three pairwise subgroup comparisons—*White vs Asian*, *White vs Black*, and *Asian vs Black*—are averaged for race attribute results. The sex attribute includes only *Male vs Female*. These scores are then averaged to determine overall model bias. Colour scales highlight the level of bias across all models, with red indicating higher bias (undesirable) and green indicating lower. For a detailed explanation of the bias score development, please refer to the *Novel Bias Score* methodology section.

Table 12 presents the bias scores for our selected models tested on MIMIC, across the protected attributes and their average. Contrary to findings from the CheXpert dataset, the CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1) model now exhibits the highest bias scores on MIMIC, recording an overall score of 85.35. This is closely followed by the teacher CXR-FM, with a score of 80.47. The CXR-FMKD-Direct FFT (CS) model comes next at 78.70, then CXR-FMKD-Direct FFT (MSE) at 69.22, and finally CXR-Model FFT at 63.16, making it the least biased model.

Across the race and sex attributes, similar bias ranking trends are observed, although there are slight variations among models. For the sex attribute, CXR-FMKD-Direct FFT (MSE) is slightly less biased than CXR-Model FFT, and CXR-FM is slightly more biased than CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1). Examining the pairwise subgroup comparisons within the race attribute also reveals some variations in trends. For instance, consistent with observations from CheXpert, the CXR-FM exhibits some of the highest bias scores in the *White vs Asian* and *Asian vs Black* comparisons, yet the lowest in *White vs Black*. Conversely, CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1) shows the highest bias scores in *White vs Black* and *Asian vs Black* comparisons, but ranks near the bottom in *White vs Asian*, second only to CXR-Model FFT. Notably, this CXR-Model FFT consistently records the lowest bias scores across all comparisons except for *White vs Black*, where it ranks just above CXR-FM. CXR-FMKD-Direct FFT (MSE) also maintains relatively low bias scores across subgroups.

It is important to note that the differences between the bias scores for models tested on MIMIC are much less pronounced than those observed with CheXpert, indicating a narrower range from 63.16 to 85.35 compared to CheXpert's range of 16.53 to 108.90 for the overall bias for instance. This point aligns with the less pronounced discrepancies in marginal distribution differences observed between CXR-FM and CXR-FMKD-Direct FFT (MSE). Additionally, it contextualises the bias scores for the *White vs Asian* comparison among our models—particularly highlighting the contrast between the teacher and the other models. Here, CXR-FM exhibits a relatively significantly higher bias score of 30.11, in contrast to the notably lower scores of the other models: CXR-Model FFT at 0.92, CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1) at 2.60, CXR-FMKD-Direct FFT (MSE) at 5.40, and CXR-FMKD-Direct FFT (CS) at 12.28. Conversely, for the *White vs Black* comparison, the teacher model shows considerably lower bias at 20.46, whereas the other models exhibit higher scores, ranging from 42.33 for CXR-Model FFT to 76.79 for CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1).

The raw data from these 5000 simulations, displayed as frequencies of categorised 'FALSE' ($p \geq 0.05$), 'TRUE' ($0.001 \leq p < 0.05$), and 'TRUE+' ($p < 0.001$), and before applying bias score calculation steps, is detailed in **Table 27** in the *Supplemental Material* section. This table also shows the explained variances per PCA mode for each of the five selected models. The combined four PCA modes account for approximately 64.04% of the variance in CXR-Model FFT, 51.09% in CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1), 43.98% in CXR-FMKD-Direct FFT (MSE), 40.74% in CXR-FM, and the least at 31.38% in CXR-FMKD-Direct FFT (CS). This ranking is similar to that observed with CheXpert, though the (MSE-CS | 0.9-0.1) and (MSE) CXR-FMKD-Direct FFT student variants have switched places. The cumulative explained variances are also higher, suggesting models with feature representations potentially more focused and tailored to the MIMIC disease detection task—where we can also recall the greater absolute performances achieved on this dataset compared to CheXpert.

In the next section, we will examine subgroup performances to uncover potential discrepancies and link them to the varying bias levels exhibited by each model.

4.3.2. Subgroup Performance Analysis

4.3.2.1. CheXpert

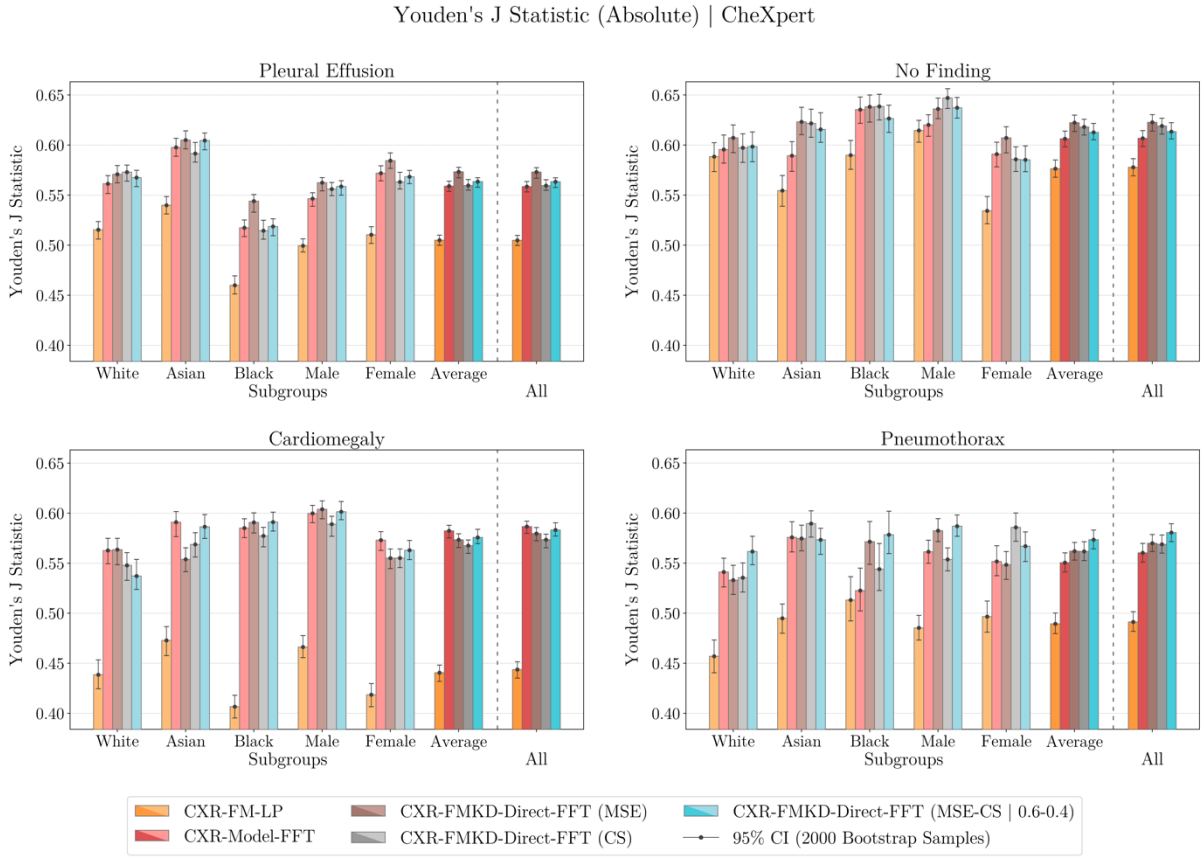

Figure 45. Comparison of Youden's J Statistic Disease Detection Performance Across Patient Subgroups for Selected Models Tested on CheXpert.
This figure illustrates the mean Youden's J Statistic values, depicted by dots, and their corresponding 95% confidence intervals (CIs), shown with whiskers and derived from bootstrapping with 2000 samples, across the relevant patient subgroups (for race and biological sex), their average, and the entire patient sample—denoted by 'All'—for our five selected models developed and tested on CheXpert. These models include the teacher CXR-FM; the CXR-FMKD-Direct FFT variants of the three selected student model types (MSE, CS, and MSE-CS | 0.6-0.4); and the traditional baseline CXR-Model FFT. Note that the Youden's J Statistic values were determined at a fixed (global) decision threshold, optimised to achieve an FPR of 20% across the entire patient sample. The models were assessed for their ability (average absolute classification performance) to detect 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax'. The teacher CXR-FM consistently underperforms compared to the other models, with the student models exhibiting the highest performances overall. Performance disparities can be observed across subgroups, particularly pronounced for the teacher CXR-FM. For instance, CXR-FM shows a notable underperformance in the Black patient subgroup for 'Pleural Effusion' and 'Cardiomegaly', and in the Female patient subgroup for 'No Finding'. Additionally, the teacher model shows a significant drop in performance for 'Cardiomegaly' compared to other disease labels.

Figure 45 compares the absolute disease detection performance of our five selected models tested on CheXpert across relevant patient subgroups, their average, and the entire patient sample—denoted by 'All'—using *Youden's J Statistic*. This metric is calculated across subgroups based on a fixed global decision threshold, optimised to achieve an FPR of 20% across the entire patient sample, with data presented as means with 95% CIs derived from 2000 bootstrap samples for disease labels including 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax'.

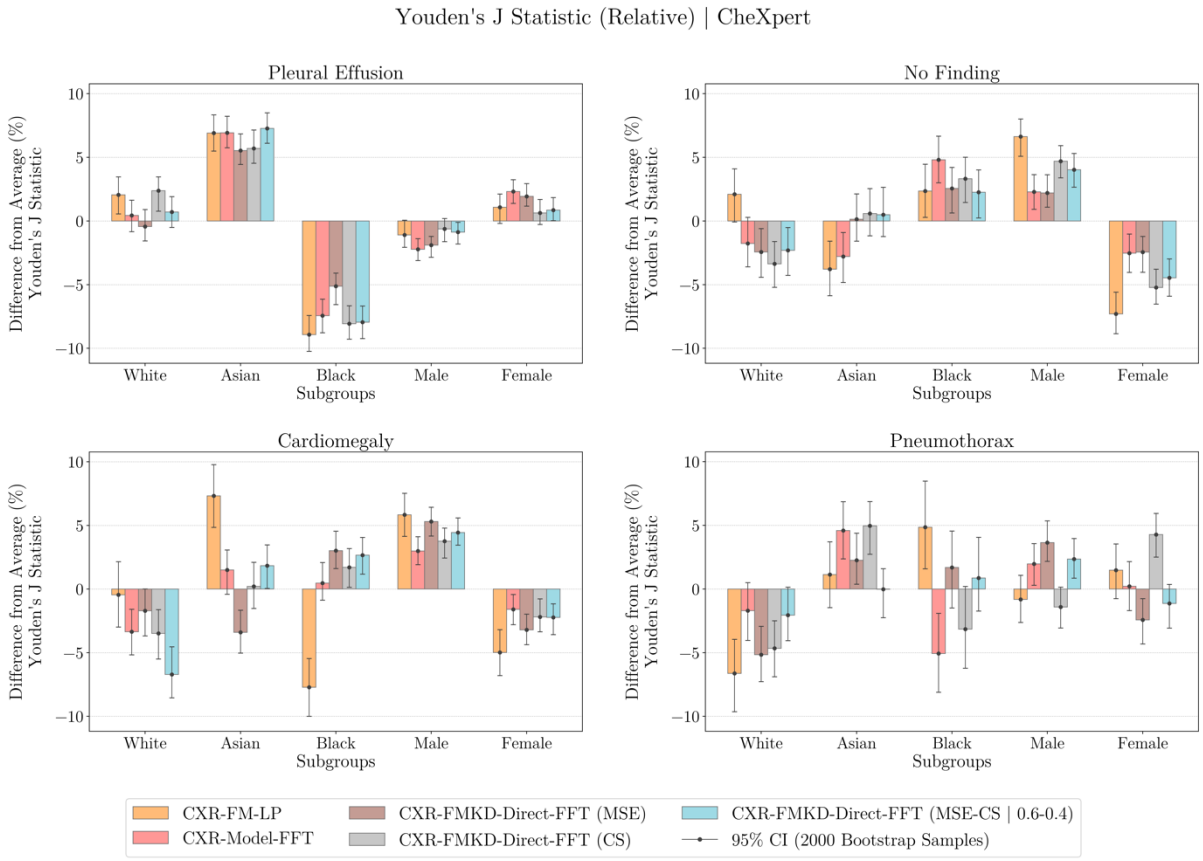

Figure 46. Relative Change in Youden's J Statistic Disease Detection Performance Across Patient Subgroups for Selected Models Tested on CheXpert.
This figure illustrates the mean relative changes in Youden's J Statistic performance, depicted by dots, and their corresponding 95% confidence intervals (CIs), shown with whiskers and derived from bootstrapping with 2000 samples, across the relevant patient subgroups (for race and biological sex) for our five selected models developed and tested on CheXpert. For each model, the relative performance change for each subgroup for a specific disease label was computed by comparing the subgroup's performance (Subgroup Value) with the average performance across all subgroups (Average Value) for that label using the formula: (Subgroup Value – Average Value)/ Average Value × 100%. The models evaluated include the teacher CXR-FM; the CXR-FMKD-Direct FFT variants of the three selected student model types (MSE, CS, and MSE-CS | 0.6-0.4); and the traditional baseline CXR-Model FFT. Note that the Youden's J Statistic values were determined at a fixed (global) decision threshold, optimised to achieve an FPR of 20% across the entire patient sample. The models were assessed for their ability (average relative classification performance) to detect 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax'. Disparities in relative performance can be observed across subgroups, particularly pronounced for the teacher CXR-FM.

Firstly, it is evident that the teacher model, CXR-FM, consistently underperforms compared to the other models—the CXR-FMKD-Direct FFT students and the baseline CXR-Model FFT—across all subgroups and disease labels. This reaffirms findings from our *Performance Analysis* section. Notably, there is a significant drop in performance for the teacher model in detecting 'Cardiomegaly' compared to the other three disease labels. In contrast, the student models generally slightly outperform the CXR-Model FFT baseline across the entire sample 'All', except for 'Cardiomegaly', where the baseline model performs slightly better.

To better examine performance discrepancies across subgroups, **Figure 46** plots the relative changes in *Youden's J Statistic* for each subgroup, comparing these with the corresponding model's average performance achieved across all subgroups. This visualisation clearly highlights the variation in relative performance among the evaluated models, where it is particularly relevant to look at negative changes that indicate underperformance in specific subgroups. The teacher model shows the largest negative performance changes across all disease

labels compared to the other models, with notable decreases in detecting 'Pleural Effusion' (-8.93%) for the Black subgroup, 'No Finding' (-7.30%) for the Female subgroup, 'Cardiomegaly' (-7.71%) also for the Black subgroup, and 'Pneumothorax' (-6.63%) for the White subgroup. Moreover, we observe larger discrepancies in relative performance across the protected characteristics for the teacher compared to the other models. This is evidenced by the more pronounced fluctuations in the histograms, which indicate a greater variation in performance across different subgroups for specific diseases.

Now, let's evaluate how our student models stack up against the CXR-Model FFT baseline. The CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4) and CXR-FMKD-Direct FFT (CS) students show more variation in performance relative to the CXR-Model FFT and the CXR-FMKD-Direct FFT (MSE), which displays the least variability overall. For example, in detecting 'Pleural Effusion' within the Black subgroup, the CS and MSE-CS | 0.6-0.4 variants have more significant decreases in performance, at -8.07% and -7.95% respectively, compared to -7.44% for the baseline and -5.12% for the MSE variant. Similarly, for 'No Finding' in the Female subgroup, the decreases are -5.23% for the CS variant, -4.46% for the MSE-CS | 0.6-0.4 variant, -2.53% for the baseline, and -2.45% for the MSE variant. Overall, this MSE variant demonstrates the least variation in performance, indicating fewer discrepancies across subgroups, followed closely by the baseline. The CS and MSE-CS | 0.6-0.4 variants show greater variation, with the CXR-FM teacher model displaying the most pronounced disparities of all.

Equivalent plots for *AUC-ROC* are available in the *Supplemental Material* section, subsection *S.8. Bias Analysis | Subgroup Performance Analysis – CheXpert*, specifically **Figure 110** (absolute plots) and **Figure 111** (relative plots), which exhibit similar trends. Furthermore, detailed tabular results for this subgroup performance analysis across various metrics (including *AUC-ROC* and *Youden's J Statistic*) for the four disease labels are also provided in this section in **Table 28** ('Pleural Effusion'), **Table 29** ('No Finding'), **Table 30** ('Cardiomegaly'), and **Table 31** ('Pneumothorax'). Lastly, we also plotted *ROC* curves across subgroups for each model, for all four disease labels. These plots demonstrate some TPR/FPR shifts across the patient subgroups, with TPR and FPR determined at a fixed decision threshold optimised to achieve an FPR of 20% across the entire patient sample. These TPR/FPR shifts also indicate performance disparities among subgroups— where higher FPR for 'No Finding' in some patients could lead to increased underdiagnosis for example, which is concerning. These findings further substantiate the need for the bias analysis conducted, as evidenced by persistent disparities that suggest potential bias. The *ROC* curves are shown in this *Supplemental Material* section in **Figure 105** (CXR-FM), **Figure 106** (CXR-Model FFT), **Figure 107** (CXR-FMKD-Direct FFT (MSE)), **Figure 108** (CXR-FMKD-Direct FFT (CS)), and **Figure 109** (CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4)).

4.3.2.2. MIMIC

Figure 47 compares the absolute *Youden's J Statistic* performance of our five selected models, now tested on MIMIC, across the relevant patient subgroups, their average, and the entire patient sample—denoted by 'All'.

Similar to observations for CheXpert, the CXR-FM teacher consistently underperforms compared to the other models—the CXR-FMKD-Direct FFT students and the baseline CXR-Model FFT—across all subgroups and disease labels. This reaffirms findings from our previous

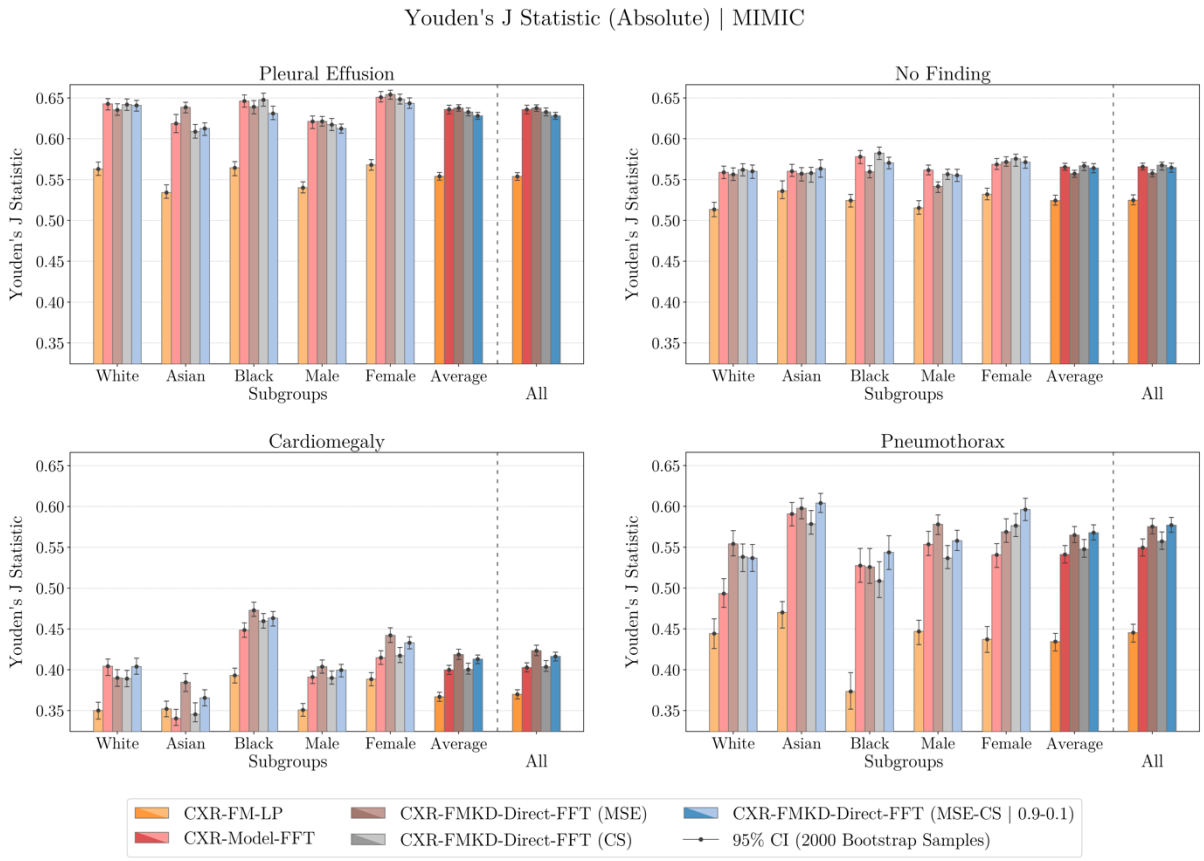

Figure 47. Comparison of Youden's J Statistic Disease Detection Performance Across Patient Subgroups for Selected Models Tested on MIMIC.
This figure illustrates the mean Youden's J Statistic values, depicted by dots, and their corresponding 95% confidence intervals (CIs), shown with whiskers and derived from bootstrapping with 2000 samples, across the relevant patient subgroups (for race and biological sex), their average, and the entire patient sample—denoted by 'All'—for our five selected models developed and tested on MIMIC. These models include the teacher CXR-FM; the CXR-FMKD-Direct FFT variants of the three selected student model types (MSE, CS, and MSE-CS | 0.9-0.1); and the traditional baseline CXR-Model FFT. Note that the Youden's J Statistic values were determined at a fixed (global) decision threshold, optimised to achieve an FPR of 20% across the entire patient sample. The models were assessed for their ability (average absolute classification performance) to detect 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax'. The teacher CXR-FM consistently underperforms compared to the other models, with the student models exhibiting the highest performances overall. Performance disparities can be observed across subgroups and there is also a significant drop in overall performance across all subgroups for our models in detecting 'Cardiomegaly'.

Performance Analysis section. In addition, there is a significant drop in performance for all models in detecting 'Cardiomegaly', which contrasts with 'Pleural Effusion' where models achieve higher performance ranges across all subgroups. Here, the student models generally slightly outperform the CXR-Model FFT baseline across the entire sample.

Figure 48 plots the relative changes in *Youden's J Statistic* of our five selected models across the relevant patient subgroups. Unlike the plots for CheXpert, it is less apparent which model exhibits the largest disparities in relative performance across subgroups. Indeed, for 'Pleural Effusion' and 'No Finding', all models show relatively consistent performance across subgroups, indicated by low relative changes. To provide some distinction, the CXR-FMKD-Direct FFT (MSE) do exhibit slightly less performance variation, maintaining nearly uniform performance across the race subgroups (i.e., White, Asian, and Black patients with relative changes of -0.36%, 0.13%, and 0.22% respectively for 'Pleural Effusion' and -0.18%, 0.01%, and 0.41% respectively for 'No Finding'), while the other models exhibit slightly more variation in these subgroups.

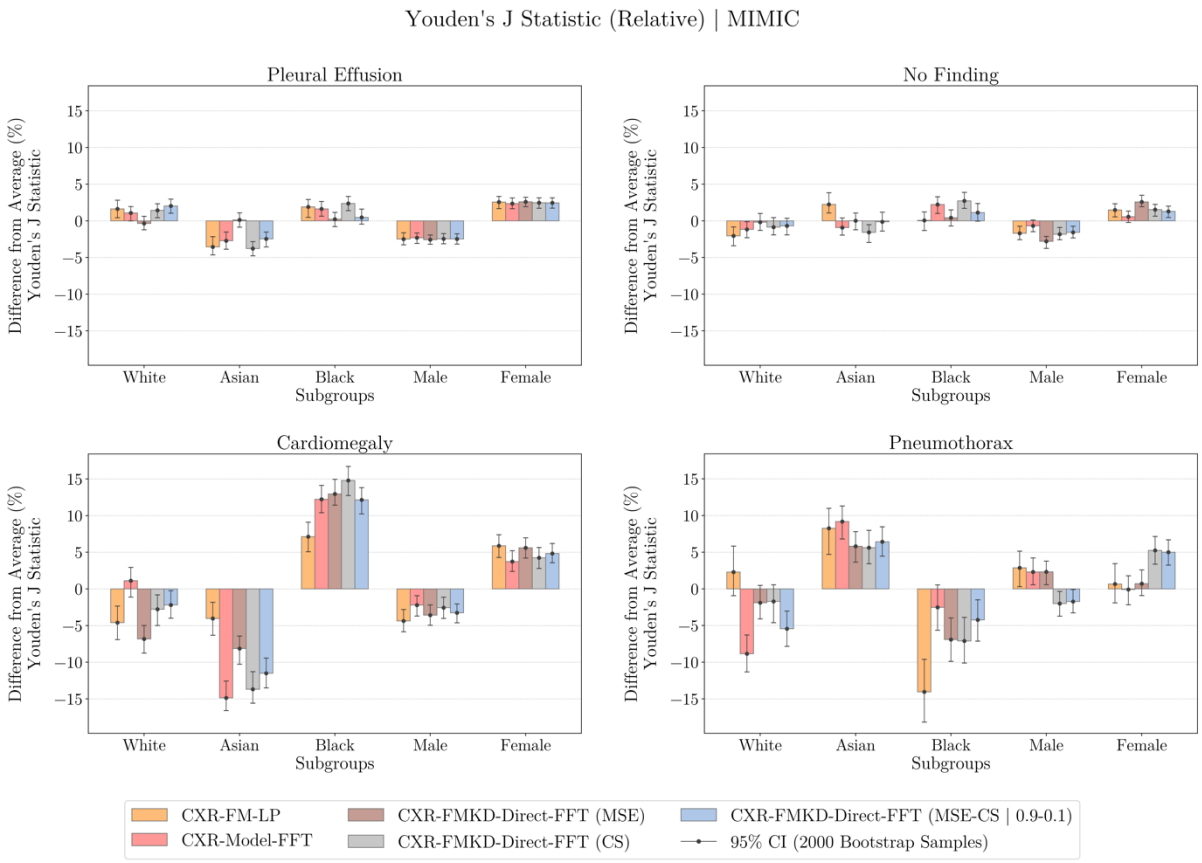

Figure 48. Relative Change in Youden's J Statistic Disease Detection Performance Across Patient Subgroups for Selected Models Tested on MIMIC.
This figure illustrates the mean relative changes in Youden's J Statistic performance, depicted by dots, and their corresponding 95% confidence intervals (CIs), shown with whiskers and derived from bootstrapping with 2000 samples, across the relevant patient subgroups (for race and biological sex) for our five selected models developed and tested on MIMIC. For each model, the relative performance change for each subgroup for a specific disease label was computed by comparing the subgroup's performance (Subgroup Value) with the average performance across all subgroups (Average Value) for that label using the formula: (Subgroup Value – Average Value)/ Average Value × 100%. The models evaluated include the teacher CXR-FM; the CXR-FMKD-Direct FFT variants of the three selected student model types (MSE, CS, and MSE-CS | 0.9-0.1); and the traditional baseline CXR-Model FFT. Note that the Youden's J Statistic values were determined at a fixed (global) decision threshold, optimised to achieve an FPR of 20% across the entire patient sample. The models were assessed for their ability (average relative classification performance) to detect 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax'. Disparities in relative performance can be observed across subgroups for the models being evaluated.

More pronounced variations are noted, however, for 'Cardiomegaly' and 'Pneumothorax' across our models. For 'Cardiomegaly', CXR-Model FFT shows the largest negative change in performance within the Asian patient subgroup, with a decrease of -14.87%. Close behind are the CXR-FMKD-Direct FFT (CS) with -13.70% and the CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1) with -11.50%. The CXR-FMKD-Direct FFT (MSE) exhibits less underperformance at -8.14%, while CXR-FM has the smallest decrease at -4.03%, significantly lower than the others. Note that, overall, the three student models tend to display similar trends and magnitudes in relative performance variation across these labels. Here, CXR-FM now shows the least variation among the race subgroups for 'Cardiomegaly', rather than CXR-FMKD-Direct FFT (MSE) which presents more disparity compared with the 'Pleural Effusion' and 'No Finding' labels. Conversely, for 'Pneumothorax', CXR-FM demonstrates the most variation among race subgroups, particularly with a -14.05% decrease in performance for Black patients—the largest negative change observed for this label. Overall, these observations do not offer straightforward interpretations with the models failing to exhibit consistent trends in performance disparities

and magnitudes across these labels. It is interesting to note, however, that performance variations for sex subgroups remain relatively smaller compared to the broader differences observed across racial subgroups for these four labels.

Equivalent plots for *AUC-ROC* are available in the *Supplemental Material* section, subsection *S.9. Bias Analysis | Subgroup Performance Analysis – MIMIC*, specifically **Figure 117** (absolute plots) and **Figure 118** (relative plots), which display similar trends. Also, detailed tabular results for this subgroup performance analysis across various metrics (including *AUC-ROC* and *Youden's J Statistic*) for the four disease labels are also provided in this section in **Table 32** ('Pleural Effusion'), **Table 33** ('No Finding'), **Table 34** ('Cardiomegaly'), and **Table 35** ('Pneumothorax'). Similar to the analysis for CheXpert, *ROC* curves across subgroups for each model and for all four disease labels were also plotted, showcasing TPR/FPR shifts that indicate performance disparities among subgroups. These *ROC* curves are included in this *Supplemental Material* section in **Figure 112** (CXR-FM), **Figure 113** (CXR-Model FFT), **Figure 114** (CXR-FMKD-Direct FFT (MSE)), **Figure 115** (CXR-FMKD-Direct FFT (CS)), and **Figure 116** (CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1)).

4.3.3. Discussion

CheXpert

To build on the work by Glocker et al. [93], we start our discussion by examining the teacher CXR-FM and comparing it to the CXR-Model FFT baseline in terms of identified biases, as they follow similar development methodologies as those outlined in the study that initially demonstrated bias within the teacher and associated subgroup performance disparities. As anticipated, the teacher exhibited significant bias, scoring 108.90 overall—nearly four times higher than the 27.24 score of the CXR-Model FFT baseline. This substantial bias, particularly reflecting the near-maximal score of 149.24 obtained for the sex attribute, suggests that the teacher model encodes protected characteristics like biological sex and racial identity more strongly than the baseline. Additionally, when assessed in the CheXpert downstream detection task, CXR-FM significantly underperformed relative to the baseline and presented notable subgroup performance disparities such as a -8.93% decrease in detecting 'Pleural Effusion' for the Black subgroup, and a -7.30% decrease for 'No Finding' in the Female subgroup. These outcomes corroborate the findings of Glocker et al. as well as Seyyed-Kalantari et al. [94] which highlighted persistent performance disparities in CXR disease detection across underrepresented patient subgroups.

Here lies the repeatedly mentioned key incentive for using KD to reconstruct the typically inaccessible teacher FMs with a frozen backbone, enabling more comprehensive bias mitigation strategies by gaining full access to the parameters and backbone for fine-tuning. This could help improve performance on downstream tasks and reduce potential subgroup performance disparities. Our selected student models—CXR-FMKD-Direct FFT (MSE), CXR-FMKD-Direct FFT (CS), and CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4)—derived from KD with CXR-FM, have not only shown enhanced performance as noted previously in the *Performance Analysis* section but also significantly lower bias scores than the CXR-FM teacher. Specifically, the CXR-FMKD-Direct FFT (MSE) model stands out as the least biased, with an overall bias score of 16.53—approximately 6.6 times lower than that of the teacher—representing an **85% reduction** in the teacher's bias score. This model performs better than the CXR-Model FFT baseline, which has bias scores close to that of the CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4). Meanwhile, the

CXR-FMKD-Direct FFT (CS) also shows a reduction in bias, with scores roughly half that of the teacher.

It is important to note that the reduced bias in the student models, compared to the teacher, was an unintended but beneficial byproduct of using KD to robustly reconstruct the teacher model. This robust reconstruction was achieved, as evidenced by the positive outcomes in the *Performance Analysis* and *Generalisability Analysis* sections. However, this process also indirectly mitigated bias, resulting in student models with lower bias scores. These less biased models have learned feature representations that do not disproportionately rely on protected characteristics like biological sex or race. Instead, they focus on other, hopefully medically relevant features, ensuring that the model does not take shortcuts in disease prediction. For example, rather than predicting 'Pleural Effusion' based on sex—a dataset artifact where 90% of cases might be female—the models are forced to identify more diagnostic features relevant to the condition. This approach reduces the likelihood of misdiagnosis in underrepresented groups and ensures more equitable health outcomes.

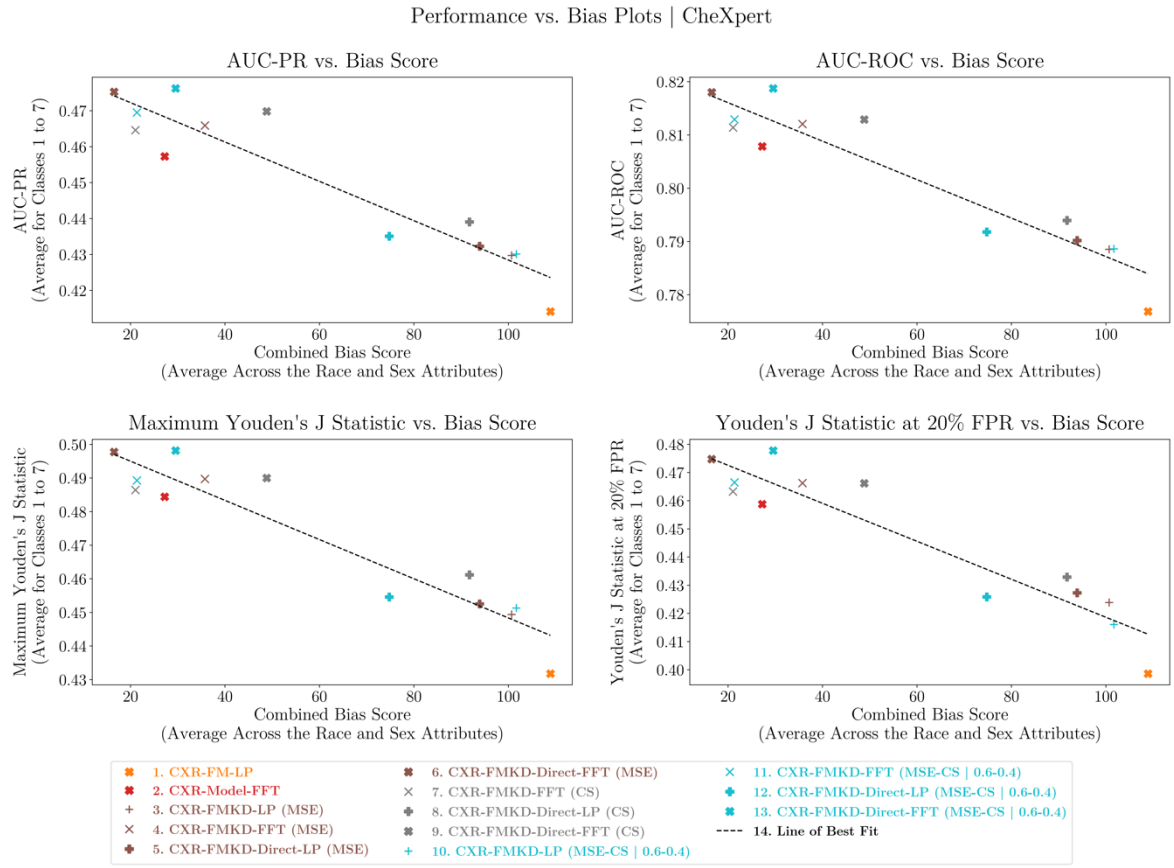

Figure 49. Performance versus Bias Plots for 13 Selected Models Tested on CheXpert.
This figure explores the relationship between performance and bias for the 13 selected models tested on CheXpert. It presents performance data for four metrics—AUC-ROC, AUC-PR, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR—focusing on the average results for the most significant disease labels (Classes 1 to 7). The models evaluated include the teacher CXR-FM, all variants of the three selected student model types (MSE, CS, and MSE-CS | 0.6-0.4) except for CXR-FMKD LP (CS), and the baseline CXR-Model FFT. Performance results for each metric are plotted against the Combined Bias Score, calculated as the average of bias scores across the race and sex attributes for each model, using the bias quantification method proposed in this study. A general negative correlation is observed, where models with higher performance typically exhibit lower bias, as indicated by the lines of best fit for each plot. The CXR-FMKD LP (CS) and baseline CXR-Model LP models are excluded from this analysis due to their significant underperformance, which sets them apart as outliers.

One should understand that bias, as discussed in this study, is inherently linked to a model's disease detection mechanisms through the feature representations it develops, which directly influence performance. Bias evaluation focused on ensuring models do not exploit undesirable patterns in the feature space that might favour certain subgroups unduly, ultimately impacting discrimination capabilities. Therefore, understanding the relationship between performance and bias is important. In **Figure 49**, we plot overall model performance—defined by the average results for the most significant disease labels (classes 1 to 7) across four metrics (*AUC-ROC, AUC-PR, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR*)—against overall bias scores, which average the sex and race bias scores. This analysis excludes the CXR-FMKD LP (CS) and baseline CXR-Model LP due to their outlier status from significant underperformance. A general negative correlation is observed, indicating that models with higher performance typically exhibit lower bias, and vice versa. This suggests that, in general, higher performance on CheXpert is associated with a reduced reliance on protected characteristics for disease prediction, as evidenced by lower bias scores, and instead, a greater focus on more diagnostically relevant features. The extension of **Figure 49** by including the outliers is presented in **Figure 119** in the *Supplemental Material* section, *S.10. Performance vs. Bias Analysis – CheXpert*, still showing a negative correlation trend.

Here, recall the distinct performance of the CS variant, which exhibited higher bias compared to the other two students, the MSE and MSE-CS | 0.6-0.4 variants. The nature of KD employing CS, which fundamentally focuses on the directionality of feature vectors rather than their magnitudes—as MSE does—may lead to suboptimal feature representations for specific tasks. This focus on directionality without considering magnitude could inadvertently encourage the model to rely on shortcuts or superficial features linked to protected characteristics, thereby leading to higher bias scores considering the trend in In **Figure 49**.

In terms of subgroup performance analysis, our student models demonstrate reduced disparities compared to the teacher model. As noted earlier, there is a correlation between higher bias scores and larger disparities in subgroup performance. For example, the approximate ranking from models showing lower to higher subgroup performance disparities—CXR-FMKD-Direct FFT (MSE), CXR-Model FFT, CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4), CXR-FMKD-Direct FFT (CS), and lastly CXR-FM—matches their ranking from lowest to highest overall bias score. This might be explained by the models' reduced reliance on protected characteristics for making predictions, leading to more equitable outcomes across subgroups.

Differentiating overall performance from subgroup disparities is also crucial. A model could lead in overall performance yet show significant subgroup disparities. Additionally, this study highlights that models exhibit varied biases across racial comparisons—*White vs Asian*, *White vs Black*, and *Asian vs Black*. For instance, CXR-FM showed the most bias in the *White vs Asian* and *Asian vs Black* comparisons but less in the *White vs Black* comparison, a pattern inversely observed in the CXR-Model FFT. These observations suggest that models may react differently to specific subgroup dynamics, influenced by the training data's nature, the features models emphasize, or the subgroups' statistical properties. This underscores the complexity of bias within AI models, where reducing bias in one area can unintentionally exacerbate it in another. Effective bias mitigation requires tailored strategies that account for the unique characteristics of each subgroup, highlighting that a one-size-fits-all approach is often not adequate.

Considering this new bias analysis, the MSE variant—occupying the top left corner of the plots in **Figure 49**—emerges as the preferred student model, despite the MSE-CS | 0.6-0.4 variant

achieving technically the best overall performance. Indeed, the MSE model shows the lowest subgroup performance disparities and bias scores, while still maintaining one of the top overall performances, comparable to the MSE-CS | 0.6-0.4 variant. In contrast, the CXR-FM teacher occupies the least desirable position at the bottom right corner of the plots, exhibiting the highest bias score and the lowest performances overall.

MIMIC

In contrast to the bias inspection on CheXpert, CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1) now exhibits the highest overall bias score on MIMIC, surpassing the CXR-FM teacher, which comes in second. The CXR-Model FFT baseline and CXR-FMKD-Direct FFT (MSE) continue to display the lowest bias scores, with the baseline now recording the smallest bias. However, as noted earlier, the range of these overall bias scores, from 63.16 to 85.35, is much narrower than observed in CheXpert. Additionally, for these scores, over 50% of the aggregated p-values from the 5000 simulations indicate significance (TRUE or TRUE+), revealing models that are considered more on the biased side, with a tendency towards bias—especially when compared to

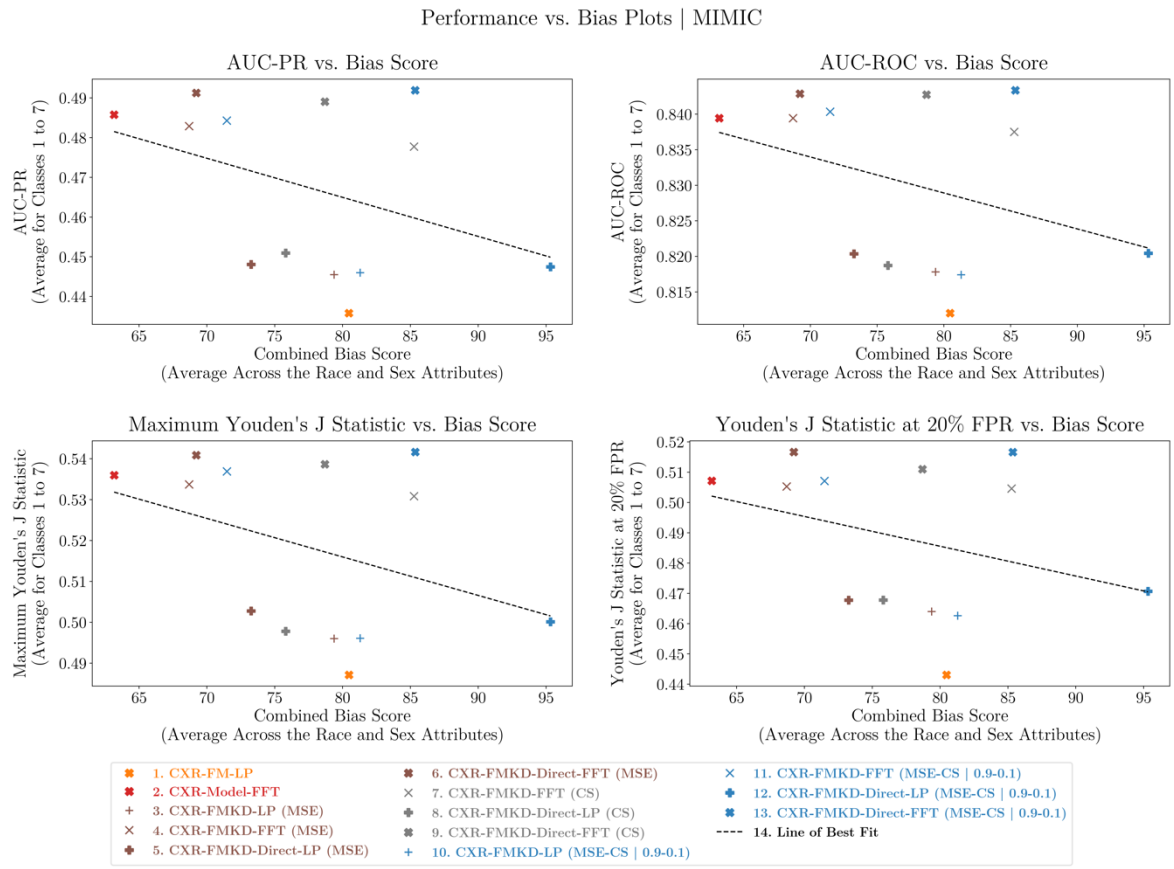

Figure 50. Performance versus Bias Plots for 13 Selected Models Tested on MIMIC.
This figure explores the relationship between performance and bias for the 13 selected models tested on MIMIC. It presents performance data for four metrics—AUC-ROC, AUC-PR, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR—focusing on the average results for the most significant disease labels (Classes 1 to 7). The models evaluated include the teacher CXR-FM, all variants of the three selected student model types (MSE, CS, and MSE-CS | 0.9-0.1) except for CXR-FMKD LP (CS), and the baseline CXR-Model FFT. Performance results for each metric are plotted against the Combined Bias Score, calculated as the average of bias scores across the race and sex attributes for each model, using the bias quantification method proposed in this study. A general negative correlation is observed, where models with higher performance typically exhibit lower bias, as indicated by the lines of best fit for each plot. The CXR-FMKD LP (CS) and baseline CXR-Model LP models are excluded from this analysis due to their significant underperformance, which sets them apart as outliers.

their generally lower bias levels in CheXpert. In terms of subgroup performance analysis on MIMIC, there are no clear trends in performance disparities among the models, with mixed results across different disease labels, such as the varying disparities trends in 'Pneumothorax' and 'Cardiomegaly', making it challenging to rank them conclusively.

Overall, these results suggest that models developed for MIMIC are more homogenous in terms of bias and their feature space encoding. **Figure 50** illustrates overall model performance—defined by the average results for the most significant disease labels (classes 1 to 7) across four metrics (*AUC-ROC, AUC-PR, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR*)—plotted against overall bias scores. While a general negative correlation is observed, it is less pronounced than in CheXpert, with data points more dispersed across the plot. An extension of **Figure 50**, including the CXR-FMKD LP (CS) and CXR-Model LP outliers, is shown in **Figure 120** in the *Supplemental Material* section, *S.11. Performance vs. Bias Analysis – MIMIC*, still showing a negative correlation trend. The student models are spread across the bias score range, with the MSE, CS, and MSE-CS | 0.9-0.1 CXR-FMKD-Direct FFT student variants showing similar top performance results, but the MSE-CS | 0.9-0.1 variant being more biased to the right, the CS variant toward the middle, and the MSE variant toward the left. It should be noted that these student models tend to display similar trends and magnitudes in relative performance variation across disease labels, which suggests that no one student model significantly outperforms the others in terms of equitable performance across subgroups.

Importantly, this demonstrates that for the MIMIC dataset, top-level performance can be achieved differently across varying levels of bias. Despite the relatively minor differences in bias scores, this suggests an important design choice: prioritising models with lower bias at comparable performance levels. Thus, the CXR-FMKD-Direct FFT (MSE), exhibiting the lowest bias, emerges as the model of choice following this bias analysis as well. Furthermore, the relatively close clustering of high bias scores across all models—reflected also in close performance ranges—indicate that all models developed for MIMIC notably use protected characteristics to some advantage in disease prediction. This is further evidenced by the highest performance achieved by CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1), albeit at a higher bias cost compared to other student models. This contrasts with findings from CheXpert, where reducing reliance on protected attributes was deemed preferable, leading to better discrimination capabilities. Here, the majority of the overall bias derives from the sex attribute, while bias scores from the race attribute are comparatively lower. Yet, the subgroup performance analysis reveals that disparities for biological sex were smaller than those observed for race, hinting at potential benefits linked to leveraging protected characteristics on MIMIC.

This analysis underscores the complexity of bias in AI models and the necessity of integrating bias inspection and subgroup performance analysis into the traditional model development workflow, especially in healthcare settings. It emphasizes the importance of considering each dataset individually, illustrating that even datasets as similar as MIMIC and CheXpert, which share the same 14 disease labels, can require distinct approaches to bias mitigation and performance optimisation. These differences are influenced by each dataset's unique characteristics, subgroup prevalences, and their interplay, highlighting that there are no one-size-fits-all solutions in selecting diagnostic models and effectively addressing biases.

Conclusion

The application of KD in this study was not initially intended as a bias mitigation strategy; however, it resulted in the creation of student models—CXR-FMKD-Direct FFT variants—that demonstrated a significant reduction in bias compared to their teacher model, CXR-FM, for the

CheXpert dataset. Specifically, the KD process achieved up to an **85% reduction** in the teacher's bias score and notably diminished the subgroup performance disparities that were previously observed.

From our analysis, it was evident that models with lower bias not only showed fewer subgroup performance disparities but also higher overall performance. This suggests that superior performance on CheXpert is generally linked to a reduced reliance on protected characteristics for disease prediction, favoring a greater emphasis on more diagnostically relevant features instead.

The scenario was markedly different for the MIMIC dataset, where all models exhibited closely clustered and relatively high bias scores. This uniformity did not allow for clear trends in subgroup performance disparities among the models. Observations of performance versus bias indicated that top-level performance on MIMIC is achieved differently across varying levels of bias. This suggests a general key model selection criterion: when models demonstrate similar overall performance and comparable trends in subgroup performance disparities, preference should be given to the model with the lower bias score. Interestingly, all models developed for MIMIC appear to benefit from using protected characteristics to achieve better performance results, contrasting with findings from CheXpert where reducing reliance on protected attributes was deemed more useful.

Across both datasets, of all the models evaluated, the CXR-FMKD-Direct FFT (MSE) student variant emerged as the most balanced, offering the best combination of high performance and low bias. It surpassed the baseline CXR-Model FFT in performance and, in terms of bias, matched it closely for MIMIC and achieved the lowest scores for CheXpert.

These divergent findings between CheXpert and MIMIC underline the importance of dataset-specific considerations in model selection and bias mitigation. The distinct results highlight the need for integrating bias inspection and subgroup performance analysis into the traditional model development workflow, especially in healthcare applications where the equitable treatment of all subgroups is crucial for clinical deployment.

Chapter 5

Conclusion and Future Work

This paper successfully demonstrates the potential of Knowledge Distillation (KD) as a strategy to robustly reconstruct Foundation Models (FMs), distilling their knowledge into student models that inherit strengths of their teachers. These models are not only performant on downstream tasks but also enhance transparency, tunability, and the capacity for bias mitigation—qualities often restricted by the transparency issues that plague many FMs. Focusing on chest radiography (CXR) and Google's proprietary CXR-FM, which Glocker et al. [93] identified for biases and performance disparities across patient subgroups, we developed distilled, robust CXR-FMKD student models.

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.

Generalisability Analysis

Our generalisability analysis assessed the adaptability of the '*transfer*' models initially trained on CheXpert and subsequently tested on MIMIC, using the shared 14 disease labels. We evaluated various testing strategies—Direct Transfer without adaptation, Linear Probing (LP) by fine-tuning only the final classifier layer, and Full Fine-Tuning (FFT) across all layers—to compare the performance of the *transfer* models with corresponding '*benchmark*' models that share the same architecture but were exclusively trained and tested on MIMIC. In the forthcoming analyses, the term 'students' specifically refers to the selected top-performing CXR-FMKD-Direct FFT student variants.

'Direct Transfer' acted as a 'zero-shot' performance evaluation, demonstrating that our *transfer* models perform reasonably well on MIMIC, the out-of-distribution (OOD) dataset. This scenario highlighted the relevance of the knowledge acquired from CheXpert, the in-distribution (ID) dataset, enabling the learned features to retain discrimination power on MIMIC, with some *transfer* students even matching the performance of the *benchmark* teacher.

'LP' served to directly assess the robustness and transferability of features generated by each model's backbone. The *transfer* models showed robust results, with performances closely trailing their *benchmark* counterparts, indicating effective transferability of the features developed on CheXpert to an OOD context.

In the 'FFT' scenario, the *transfer* students not only matched but slightly outperformed their *benchmark* counterparts, becoming the top performers. This marginal performance improvement from 'LP' to 'FFT' suggests that the features were already well-generalised and adaptable, reflecting the high quality of feature representation fostered by KD and task-specific training on CheXpert.

The results from the 'Direct Transfer', 'LP', and 'FFT' scenarios collectively illustrate that our student models, by leveraging KD and specific training on CheXpert, successfully captured robust and generalisable features. These features not only ensure competitive performance across different datasets but also establish the distilled student models as superior alternatives in the evaluated disease detection tasks. This efficacy, particularly evident when using 'FFT' on OOD data, underscores the significant potential of strategic knowledge transfer to enhance model adaptability and performance in medical contexts where even small gains are important.

Bias Analysis

We developed a novel bias scoring method to systematically evaluate biases within our models, focusing on protected characteristics such as racial identity and biological sex.

While not originally intended as a bias mitigation technique, our use of KD led to the formation of student models—the CXR-FMKD-Direct FFT variants—that exhibited a considerable decrease in bias compared to their teacher, the CXR-FM, for the CheXpert dataset. Notably, the KD process facilitated an up to **85%** reduction in the teacher model's bias score, reducing subgroup performance disparities as well.

Our findings for CheXpert indicate that models with lower bias levels not only exhibited fewer subgroup disparities but also enhanced overall performance. Notably, since a higher bias score indicates a stronger encoding of protected characteristics in the model, this correlation suggests that superior overall performance on CheXpert typically involves minimising reliance on these characteristics for disease prediction, thereby prioritising more clinically relevant diagnostic features.

Conversely, the MIMIC dataset presented a different scenario, with all models showing tightly grouped, relatively elevated bias scores. This close clustering prevented the emergence of distinct trends in subgroup performance disparities. Observations of performance versus overall bias score indicated that top-level performance on MIMIC was achieved differently across varying levels of bias. Consequently, a general key model selection criterion emerged: among models with similar overall performance and subgroup disparity trends, those with lower bias should be favoured. Notably, all models seemed to leverage protected characteristics beneficially for enhanced performance outcomes on MIMIC, a stark contrast to the CheXpert findings where reduced reliance on such attributes was associated with better performance.

Among all assessed models, the CXR-FMKD-Direct FFT (MSE) student variant stood out as the most balanced, achieving the best combination of high performance and low bias—outperforming the baseline CXR-Model FFT and achieving the lowest bias scores on CheXpert while closely matching it on MIMIC. Conversely, the student variant developed using the fixed weighted combinations of MSE and CS, although achieving the best performance overall, presented higher bias scores than the MSE student variant.

Implications

The contrasting results from the **Bias Analysis** between the CheXpert and MIMIC datasets further underscore the need for dataset-specific considerations in model selection and bias mitigation, indicating that a one-size-fits-all approach is often inadequate. These distinct findings emphasize the importance of incorporating bias inspection and subgroup performance analysis into the conventional model development workflow, especially in healthcare applications, to ensure fair treatment across all patient subgroups, which is crucial for clinical deployment. Here, the notable contrast in how leveraging protected characteristics for disease detection affects outcomes differently in CheXpert compared to MIMIC draws attention to the broader context of fairness in ML. While there has been significant progress in making ML fairer, especially in scenarios where the causes of observed disparities are identifiable and the use of protected characteristics could be detrimental, this is not universally applicable. In some contexts, strategically using such protected attributes is actually essential for achieving fairness [204, 205].

The significant performance improvements of the student models over the teacher underscore the effectiveness of FFT strategies over LP in fine-tuning for downstream tasks. This highlights the critical role that comprehensive model tuning plays in enhancing the capabilities of FMs, further advocating for the availability of open, fully tuneable FMs. Indeed, a pivotal point of this study is to emphasize the importance of open access to FMs, which is crucial not only for enhanced performance but also for effective bias mitigation, particularly in healthcare applications. Recently, in alignment with the themes of this research, the weights for CXR-FM were made public [97], setting a significant precedent in a competitive AI landscape where sharing insights from proprietary models is not typically incentivised. This shift underscores the ongoing need for open access to medical FMs and transparency in their development to foster innovation and enhance safety across healthcare applications.

To formalise the KD exploration undertaken in this study for reconstructing closed FMs similar to the CXR-FM, with a frozen backbone and feature embeddings as the only output, we recommend the following initial strategies:

- Begin with MSE as the KD loss, which is simpler than MSE-CS combinations and widely used in KD literature [157]. Our study confirms its effectiveness in achieving good performance and mitigating bias.
- Employ a learning rate scheduler, like the *'OneCycle LR'* [201], which has been effective in enhancing the convergence of the KD student model training.
- After completing the KD process, remove any projectors used to match teacher and student feature embeddings, which were optimised during KD, echoing similar practices in SSL [114, 200].
- Where applicable, apply FFT to the resulting student models to adapt them for downstream tasks. This approach led to the development of our CXR-FMKD-Direct FFT student variants, which emerged as the top performers.

Lastly, this study introduced a novel bias quantification score, designed to streamline the comparison of bias across multiple models. This new metric provides a practical, single score approach that simplifies the traditionally complex process of analysing bias through feature space examinations, which typically involves multiple separate results for comparing marginal distributions.

Limitations and Future Work

Some limitations for our study and corresponding future work are detailed below.

KD Process: This thesis presents a comprehensive exploration of KD for reconstructing the CXR-FM teacher model. However, while our findings contribute to important insights, several areas could benefit from further investigation:

1. Future research could explore alternative KD loss functions beyond those currently used or refine the MSE-CS combinations with dynamic weighting. Particularly, expanding on the preliminary exploration of MSE-CS with learned weights could potentially optimise student model performance.
2. Investigating architectures similar to CXR-FM (an EfficientNet-L2), such as those within the EfficientNet family, or exploring fundamentally different architectures like the popular Vision Transformers [131], could provide valuable comparisons and insights into the adaptability and efficacy of KD across diverse model frameworks.
3. Integrating more advanced KD methods, such as relation-based KD which examines relationships between inputs and outputs, could offer deeper insights into knowledge transfer and enhance model robustness.
4. The transfer set used in this study was quite large, being identical to the task-specific training set. Future studies could investigate the effects of varying the size and composition of the transfer set on the effectiveness of KD and, ultimately, the performance of the student.

Generalisability Analysis: While our generalisability analysis demonstrated reasonable performance of transfer models on MIMIC after trained solely on CheXpert, the similarity between these two datasets may not sufficiently challenge the models' adaptability. Indeed, both datasets are aligned closely in terms of disease labels and imaging, which might not accurately reflect the models' performance across genuinely OOD scenarios. Further research could involve:

1. Assessing the generalisability of the student models on datasets significantly different from CheXpert, featuring varied labels and diagnostic tasks.
2. Evaluating the models' performance when pre-trained on MIMIC and subsequently tested on CheXpert, to compare with the reverse training scenario observed in this study.

3. Investigating the volume (minimal amount) of data required for the student models to perform effectively in new environments, highlighting the practical transfer learning advantages of typical FMs in reducing label dependency—crucial in healthcare settings where data can be scarce and expensive to annotate.

Similarly to the extensive experiments performed in the original study on the CXR-FM teacher model [89], our future work could replicate such evaluations for our student models. For example, training on a small dataset of around 45 chest radiographs to measure diagnostic performance on tuberculosis, similar to past benchmarks, could further demonstrate the generalisability and robustness of our KD-enhanced models.

Bias Analysis: In this study, the novel bias score was developed using a robust method that involved 5000 simulations, akin to statistical bootstrapping. This approach was chosen to enhance the reliability of our findings by mitigating variations in p-value outputs. While this method effectively strengthens our confidence in the relative performance comparisons among models, further consultation with statisticians is recommended to validate the statistical guarantees of this technique. Moreover, our study confronts the complex nature of fairness in ML, recognising that the strategic use of protected attributes can sometimes be necessary to achieve fairness [204, 205], as previously noted. This challenges our conventional definition of bias, suggesting that out bias score calculations and interpretations must be contextually grounded and sensitive to the specific fairness objectives of each application.

Subgroup Performance Analysis: The current analysis is limited to four specific disease labels: 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax', chosen due to their relatively high performance ranges. However, this may not fully capture the comprehensive landscape of model performance across the different subgroups. This limitation suggests a need for a more extensive examination that includes a broader range of disease labels. Future research could aim to develop a more nuanced methodology that quantifies subgroup performance disparities across all disease labels associated with the task.

Bibliography

1. NHS (2022) Cancer. https://www.nhs.uk/conditions/cancer/#. Accessed 20 Jun 2024
2. Cancer Research UK (2015) Survival three times higher when cancer is diagnosed early. In: Cancer News. https://news.cancerresearchuk.org/2015/08/10/survival-three-times-higher-when-cancer-is-diagnosed-early/#. Accessed 20 Jun 2024
3. Ehman RL, Hendee WR, Welch MJ, Dunnick NR, Bresolin LB, Arenson RL, Baum S, Hricak H, Thrall JH (2007) Blueprint for imaging in biomedical research. Radiology 244:12–27
4. Hillman BJ (2006) Introduction to the special issue on medical imaging in oncology. Journal of Clinical Oncology 24:3223–3224
5. Atri M (2006) New technologies and directed agents for applications of cancer imaging. Journal of Clinical Oncology 24:3299–3308
6. Lee KS, Jeong YJ, Han J, Kim BT, Kim H, Kwon OJ (2004) T1 non-small cell lung cancer: imaging and histopathologic findings and their prognostic implications. Radiographics. https://doi.org/10.1148/RG.246045018
7. Hussain S, Mubeen I, Ullah N, Shah SSUD, Khan BA, Zahoor M, Ullah R, Khan FA, Sultan MA (2022) Modern Diagnostic Imaging Technique Applications and Risk Factors in the Medical Field: A Review. Biomed Res Int 2022:5164970
8. Fass L (2008) Imaging and cancer: A review. Mol Oncol 2:115
9. Islam SKMS, Nasim MA Al, Hossain I, Ullah MA, Gupta KD, Bhuiyan MMH (2023) Introduction of Medical Imaging Modalities. In: Zheng B, Andrei S, Sarker MK, Gupta KD (eds) Data Driven Approaches on Medical Imaging. Springer, Cham, pp 1–25
10. Sternbach G, Varon J (1993) Wilhelm Konrad Roentgen: A new kind of rays. J Emerg Med 11:743–745
11. Riesz PB (1995) The life of Wilhelm Conrad Roentgen. AJR Am J Roentgenol 165:1533–1537
12. Glasser O (1995) W. C. Roentgen and the discovery of the Roentgen rays. AJR Am J Roentgenol 165:1033–1040
13. Röntgen WC (1896) On a new kind of rays. Science (1979) 3:227–231
14. Knutsson F (1969) Röntgen and the Nobel Prize with notes from his correspondence with Svante Arrhenius. Acta Radiol Diagn (Stockh) 8:449–460
15. Roobottom CA, Mitchell G, Morgan-Hughes G (2010) Radiation-reduction strategies in cardiac computed tomographic angiography. Clin Radiol 65:859–867
16. Darby MJ, Barron DA, Hyland RE (2011) Oxford Handbook of Medical Imaging. In: Darby MJ, Barron DA, Hyland RE (eds) Oxford Medical Handbooks. Oxford University Press, pp 1–21
17. Goertz L, Al-Sewaidi Y, Habib M, Zopfs D, Reichardt B, Ranft A, Kabbasch C (2024) State-of-the-art mobile head CT scanner delivers nearly the same image quality as a conventional stationary CT scanner. Sci Rep 14:6393
18. Dudink J, Jeanne Steggerda S, Horsch S, et al (2020) State-of-the-art neonatal cerebral ultrasound: technique and reporting. Pediatr Res 87:3–12
19. Wattjes MP, Rovira À, Miller D, et al (2015) MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis—establishing disease prognosis and monitoring patients. Nat Rev Neurol 11:597–606

20. O'Brien JT, Firbank MJ, Davison C, Barnett N, Bamford C, Donaldson C, Olsen K, Herholz K, Williams D, Lloyd J (2014) 18F-FDG PET and Perfusion SPECT in the Diagnosis of Alzheimer and Lewy Body Dementias. Journal of Nuclear Medicine 55:1959–1965
21. Ghoncheh M, Pournamdar Z, Salehiniya H (2016) Incidence and Mortality and Epidemiology of Breast Cancer in the World. Asian Pac J Cancer Prev 17:43–46
22. Azamjah N, Soltan-Zadeh Y, Zayeri F (2019) Global Trend of Breast Cancer Mortality Rate: A 25-Year Study. Asian Pac J Cancer Prev 20:2015–2020
23. Forouzanfar MH, Foreman KJ, Delossantos AM, Lozano R, Lopez AD, Murray CJL, Naghavi M (2011) Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis. Lancet 378:1461–1484
24. World Health Organization (2024) Breast cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer. Accessed 16 Jun 2024
25. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 71:209–249
26. Marmot M, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M (2012) The benefits and harms of breast cancer screening: An independent review. The Lancet 380:1778–1786
27. Tabár L, Yen AMF, Wu WYY, Chen SLS, Chiu SYH, Fann JCY, Ku MMS, Smith RA, Duffy SW, Chen THH (2015) Insights from the breast cancer screening trials: how screening affects the natural history of breast cancer and implications for evaluating service screening programs. Breast J 21:13–20
28. Nicosia L, Gnocchi G, Gorini I, et al (2023) History of Mammography: Analysis of Breast Imaging Diagnostic Achievements over the Last Century. Healthcare (Basel) 11:1596
29. Nyström L, Wall S, Rutqvist LE, et al (1993) Breast cancer screening with mammography: overview of Swedish randomised trials. The Lancet 341:973–978
30. National Breast Cancer Foundation Inc (2024) Breast Cancer Facts & Stats. https://www.nationalbreastcancer.org/breast-cancer-facts/. Accessed 16 Jun 2024
31. World Health Organization (2023) Lung cancer. https://www.who.int/news-room/fact-sheets/detail/lung-cancer. Accessed 20 Jun 2024
32. Nooreldeen R, Bach H (2021) Current and Future Development in Lung Cancer Diagnosis. Int J Mol Sci. https://doi.org/10.3390/IJMS22168661
33. Amicizia D, Piazza MF, Marchini F, et al (2023) Systematic Review of Lung Cancer Screening: Advancements and Strategies for Implementation. Healthcare (Basel). https://doi.org/10.3390/HEALTHCARE11142085/S1
34. Kim J, Kim KH (2020) Role of chest radiographs in early lung cancer detection. Transl Lung Cancer Res 9:522
35. Panunzio A, Sartori P (2020) Lung Cancer and Radiological Imaging. Curr Radiopharm 13:238
36. World Health Organization (2021) Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed 16 Jun 2024
37. Mc Namara K, Alzubaidi H, Jackson JK (2019) Cardiovascular disease as a leading cause of death: how are pharmacists getting involved? Integr Pharm Res Pract 8:1–11
38. Wang YR, Yang K, Wen Y, et al (2024) Screening and diagnosis of cardiovascular disease using artificial intelligence-enabled cardiac magnetic resonance imaging. Nat Med 30:1471–1480

39. Pontone G, Rossi A, Guglielmo M, et al (2022) Clinical applications of cardiac computed tomography: a consensus paper of the European Association of Cardiovascular Imaging—part I. Eur Heart J Cardiovasc Imaging 23:314
40. Narula J, Chandrashekhar Y, Ahmadi A, et al (2021) SCCT 2021 Expert Consensus Document on Coronary Computed Tomographic Angiography: A Report of the Society of Cardiovascular Computed Tomography. J Cardiovasc Comput Tomogr 15:217
41. Tseng WYI, Su MYM, Tseng YHE (2016) Introduction to Cardiovascular Magnetic Resonance: Technical Principles and Clinical Applications. Acta Cardiol Sin 32:144
42. NHS England (2023) Diagnostic Imaging Dataset Statistical Release.
43. Lecun Y, Bengio Y, Hinton G (2015) Deep Learning. Nature 521:436–444
44. Najjar R (2023) Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging. Diagnostics (Basel) 13:2760
45. Pinto-Coelho L (2023) How Artificial Intelligence Is Shaping Medical Imaging Technology: A Survey of Innovations and Applications. Bioengineering (Basel) 10:1435
46. Hu J, Shen L, Sun G (2018) Squeeze-and-Excitation Networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, Salt Lake City, UT, USA, pp 7132–7141
47. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Institute of Electrical and Electronics Engineers Inc., Honolulu, HI, USA, pp 2261–2269
48. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, Las Vegas, NV, USA, pp 770–778
49. Wang X, Shrivastava A, Gupta A (2017) A-Fast-RCNN: Hard Positive Generation via Adversary for Object Detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Institute of Electrical and Electronics Engineers Inc., Honolulu, HI, USA, pp 3039–3048
50. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, Columbus, OH, USA, pp 580–587
51. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2018) DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans Pattern Anal Mach Intell 40:834–848
52. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid Scene Parsing Network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Institute of Electrical and Electronics Engineers Inc., Honolulu, HI, USA, pp 6230–6239
53. Kumar Y, Koul A, Singla R, Ijaz MF (2022) Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Humaniz Comput 14:8486
54. Plested J, Gedeon T (2022) Deep transfer learning for image classification: a survey.
55. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL (2018) Artificial intelligence in radiology. Nat Rev Cancer 18:510
56. Ghaffar Nia N, Kaplanoglu E, Nasab A (2023) Evaluation of artificial intelligence techniques in disease diagnosis and prediction. Discover Artificial Intelligence 3:5
57. Jabeen K, Khan MA, Balili J, Alhaisoni M, Almujally NA, Alrashidi H, Tariq U, Cha JH (2023) BC2NetRF: Breast Cancer Classification from Mammogram Images Using Enhanced Deep Learning Features and Equilibrium-Jaya Controlled Regula Falsi-Based Features Selection. Diagnostics (Basel) 13:1238

58. Yu X, Pang W, Xu Q, Liang M (2020) Mammographic image classification with deep fusion learning. Sci Rep 10:14361
59. Behrendt F, Bengs M, Bhattacharya D, Krüger J, Opfer R, Schlaefer A (2023) A systematic approach to deep learning-based nodule detection in chest radiographs. Sci Rep 13:10120
60. Schultheiss M, Schober SA, Lodde M, Bodden J, Aichele J, Müller-Leisse C, Renger B, Pfeiffer F, Pfeiffer D (2020) A robust convolutional neural network for lung nodule detection in the presence of foreign bodies. Sci Rep 10:12987
61. Ma J, He Y, Li F, Han L, You C, Wang B (2024) Segment anything in medical images. Nature Communications 15:654
62. van der Graaf JW, van Hooff ML, Buckens CFM, Rutten M, van Susante JLC, Kroeze RJ, de Kleuver M, van Ginneken B, Lessmann N (2024) Lumbar spine segmentation in MR images: a dataset and a public benchmark. Sci Data 11:264
63. Singh A, Salehi SSM, Gholipour A (2020) Deep Predictive Motion Tracking in Magnetic Resonance Imaging: Application to Fetal Imaging. IEEE Trans Med Imaging 39:3534
64. Islam KT, Wijewickrema S, O'Leary S (2021) A deep learning based framework for the registration of three dimensional multi-modal medical images of the head. Sci Rep 11:1860
65. Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X (2020) Deep learning in medical image registration: a review. Phys Med Biol 65:20TR01
66. Khalifa M, Albadawy M (2024) AI in diagnostic imaging: Revolutionising accuracy and efficiency. Computer Methods and Programs in Biomedicine Update 5:100146
67. Zhou SK, Rueckert D, Fichtinger G (2019) Handbook of Medical Image Computing and Computer Assisted Intervention, 1st ed. Elsevier and MICCAI Society Book Series
68. Bhatt N, Bhatt N, Prajapati P, Sorathiya V, Alshathri S, El-Shafai W (2024) A Data-Centric Approach to improve performance of deep learning models. Sci Rep 14:22329
69. Leming MJ, Bron EE, Bruffaerts R, Ou Y, Iglesias JE, Gollub RL, Im H (2023) Challenges of implementing computer-aided diagnostic models for neuroimages in a clinical setting. NPJ Digit Med 6:129
70. Varoquaux G, Cheplygina V (2022) Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit Med. https://doi.org/10.1038/s41746-022-00592-y
71. Cohen JP, Hashir M, Brooks R, Bertrand H (2020) On the limits of cross-domain generalization in automated X-ray prediction. In: Proc Mach Learn Res. ML Research Press, pp 136–149
72. Castro DC, Walker I, Glocker B (2020) Causality matters in medical imaging. Nat Commun 11:1–10
73. Su Z, Guo J, Yang X, Wang Q, Coenen F, Huang K (2024) Navigating Distribution Shifts in Medical Image Analysis: A Survey.
74. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, Kohane IS, Saria S (2021) The Clinician and Dataset Shift in Artificial Intelligence. N Engl J Med 385:283–286
75. Bommasani R, Hudson DA, Adeli E, et al (2021) On the Opportunities and Risks of Foundation Models.
76. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, Rajpurkar P (2023) Foundation models for generalist medical artificial intelligence. Nature 616:259–265
77. Willemink MJ, Roth HR, Sandfort V (2022) Toward Foundational Deep Learning Models for Medical Imaging in the New Era of Transformer Networks. Radiol Artif Intell. https://doi.org/10.1148/RYAI.210284

78. Azad B, Azad R, Eskandari S, Bozorgpour A, Kazerouni A, Rekik I, Merhof D (2023) Foundational Models in Medical Imaging: A Comprehensive Survey and Future Vision.
79. Hosna A, Merry E, Gyalmo J, Alom Z, Aung Z, Azim MA (2022) Transfer learning: a friendly introduction. J Big Data. https://doi.org/10.1186/S40537-022-00652-W/FIGURES/6
80. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is All you Need. Advances in Neural Information Processing Systems 30 (NIPS 2017) 30:
81. Ghesu FC, Georgescu B, Mansoor A, et al (2022) Self-supervised Learning from 100 Million Medical Images.
82. Azizi S, Culp L, Freyberg J, et al (2022) Robust and Efficient Medical Imaging with Self-Supervision.
83. United Nations (2008) United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR), 2008 Report on sources and effects of ionizing radiation.
84. Broder J (2011) Imaging the Chest: The Chest Radiograph. Diagnostic Imaging for the Emergency Physician 296
85. Jones CM, Buchlak QD, Oakden-Rayner L, Milne M, Seah J, Esmaili N, Hachey B (2021) Chest radiographs and machine learning – Past, present and future. J Med Imaging Radiat Oncol 65:544
86. Çallı E, Sogancioglu E, van Ginneken B, van Leeuwen KG, Murphy K (2021) Deep learning for chest X-ray analysis: A survey. Med Image Anal 72:102125
87. Raoof S, Feigin D, Sung A, Raoof S, Irugulpati L, Rosenow EC (2012) Interpretation of plain chest roentgenogram. Chest 141:545–558
88. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, Maschinot A, Liu C, Krishnan D (2021) Supervised Contrastive Learning.
89. Sellergren AB, Chen C, Nabulsi Z, et al (2022) Simplified Transfer Learning for Chest Radiography Models Using Less Data. Radiology 305:454–465
90. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM (2017) ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Institute of Electrical and Electronics Engineers Inc., Honolulu, HI, USA, pp 3462–3471
91. Wójcik MA (2022) Foundation Models in Healthcare: Opportunities, Biases and Regulatory Prospects in Europe. In: Electronic Government and the Information Systems Perspective. Springer International Publishing, pp 32–46
92. Wiggins WF, Tejani AS (2022) On the Opportunities and Risks of Foundation Models for Natural Language Processing in Radiology. Radiol Artif Intell 4:e220119
93. Glocker B, Jones C, Roschewitz M, Winzeck S (2023) Risk of Bias in Chest Radiography Deep Learning Foundation Models. Radiol Artif Intell 5:230060
94. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M (2021) Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 27:2176–2182
95. Adleberg J, Wardeh A, Doo FX, Marinelli B, Cook TS, Mendelson DS, Kagen A (2022) Predicting Patient Demographics From Chest Radiographs With Deep Learning. Journal of the American College of Radiology 19:1151–1161
96. Gichoya JW, Banerjee I, Bhimireddy AR, et al (2022) AI recognition of patient race in medical imaging: a modelling study. Lancet Digit Health 4:e406–e414
97. Thelin T, Kirmizibayrak C (2024) Helping everyone build AI for healthcare applications with open foundation models. In: Google Research. https://research.google/blog/helping-

everyone-build-ai-for-healthcare-applications-with-open-foundation-models/. Accessed 1 Jan 2025
98. Nazer LH, Zatarah R, Waldrip S, et al (2023) Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digital Health 2:e0000278
99. Glocker B, Jones C, Bernhardt M, Winzeck S (2023) Algorithmic encoding of protected characteristics in chest X-ray disease detection models. EBioMedicine 89:104467
100. Marcinkevičs R, Ozkan E, Vogt JE (2022) Debiasing Deep Chest X-Ray Classifiers using Intra- and Post-processing Methods. In: Lipton Z, Ranganath R, Sendak M, Sjoding M, Yeung S (eds) Proc Mach Learn Res. ML Research Press, pp 504–536
101. Oakden-Rayner L, Dunnmon J, Carneiro G, Re C (2020) Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. In: Proceedings of the 2020 ACM Conference on Health, Inference, and Learning (ACM CHIL 2020). Association for Computing Machinery, Inc, pp 151–159
102. DeGrave AJ, Janizek JD, Lee SI (2021) AI for radiographic COVID-19 detection selects shortcuts over signal. Nat Mach Intell 3:610–619
103. 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
104. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng C ying, Mark RG, Horng S (2019) MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 6:317
105. Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019). Association for Computational Linguistics (ACL), Minneapolis, Minnesota, pp 4171–4186
106. Brown TB, Mann B, Ryder N, et al (2020) Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems 33 (NeurIPS 2020)
107. Radford A, Kim JW, Hallacy C, et al (2021) Learning Transferable Visual Models From Natural Language Supervision. In: Proc Mach Learn Res. ML Research Press, pp 8748–8763
108. Zhang S, Metaxas D (2024) On the challenges and perspectives of foundation models for medical image analysis. Med Image Anal 91:102996
109. Zhou Y, Chia MA, Wagner SK, et al (2023) A foundation model for generalizable disease detection from retinal images. Nature 2023 622:7981 622:156–163
110. Chen Z, Varma M, Xu J, et al (2024) A Vision-Language Foundation Model to Enhance Efficiency of Chest X-ray Interpretation.
111. Miles K (2020) Radiomics for personalised medicine: the long road ahead. British Journal of Cancer 2020 122:7 122:929–930
112. Kirillov A, Mintun E, Ravi N, et al (2023) Segment Anything. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Institute of Electrical and Electronics Engineers Inc., pp 3992–4003
113. Awais M, Naseer M, Khan S, Anwer RM, Cholakkal H, Shah M, Yang M-H, Khan FS (2023) Foundational Models Defining a New Era in Vision: A Survey and Outlook.
114. Chen T, Kornblith S, Norouzi M, Hinton G (2020) A Simple Framework for Contrastive Learning of Visual Representations. In: 37th International Conference on Machine Learning, ICML 2020. International Machine Learning Society (IMLS), pp 1575–1585
115. Grill JB, Strub F, Altché F, et al (2020) Bootstrap your own latent: A new approach to self-supervised Learning. Adv Neural Inf Process Syst 2020-December:

116. He K, Chen X, Xie S, Li Y, Dollar P, Girshick R (2022) Masked Autoencoders Are Scalable Vision Learners. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022-June:15979–15988
117. Caron M, Touvron H, Misra I, Jegou H, Mairal J, Bojanowski P, Joulin A (2021) Emerging Properties in Self-Supervised Vision Transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Institute of Electrical and Electronics Engineers Inc., pp 9630–9640
118. Bengio Y, Courville A, Vincent P (2012) Representation Learning: A Review and New Perspectives. IEEE Trans Pattern Anal Mach Intell 35:1798–1828
119. Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22:1345–1359
120. Howard J, Ruder S (2018) Universal Language Model Fine-tuning for Text Classification. In: ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers). Association for Computational Linguistics (ACL), pp 328–339
121. Han Z, Gao C, Liu J, Zhang J, Zhang SQ (2024) Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey.
122. Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, Han X, Chen YW, Wu J (2020) UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. Institute of Electrical and Electronics Engineers Inc., pp 1055–1059
123. Jiang Y, Shen Y (2024) M^{4}oE: A Foundation Model for Medical Multimodal Image Segmentation with Mixture of Experts.
124. Ghojogh B, Crowley M (2019) The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial.
125. Santos CFG dos, Papa JP (2022) Avoiding Overfitting: A Survey on Regularization Methods for Convolutional Neural Networks. ACM Comput Surv. https://doi.org/10.1145/3510413
126. Arjun Singh M, Pandey N, Shirgaonkar A, Manoj P, Aski V (2024) A Study of Optimizations for Fine-tuning Large Language Models.
127. Kachris C (2024) A Survey on Hardware Accelerators for Large Language Models.
128. Hu E, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2021) LoRA: Low-Rank Adaptation of Large Language Models.
129. Snoek J, Larochelle H, Adams RP (2012) Practical Bayesian Optimization of Machine Learning Algorithms. Adv Neural Inf Process Syst 4:2951–2959
130. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O (2016) 3D U-net: Learning dense volumetric segmentation from sparse annotation. In: Ourselin S, Joskowicz L, Sabuncu M, Unal G, Wells W (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Springer International Publishing, Cham, Athens, Greece, pp 424–432
131. Dosovitskiy A, Beyer L, Kolesnikov A, et al (2021) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021 - 9th International Conference on Learning Representations
132. Li J, Chen J, Tang Y, Wang C, Landman BA, Zhou SK (2023) Transforming medical imaging with Transformers? A comparative review of key properties, current progresses, and future perspectives. Med Image Anal 85:102762
133. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86:2324

134. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25 (NIPS 2012). https://doi.org/10.1145/3065386
135. Russakovsky O, Deng J, Su H, et al (2015) ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 115:211–252
136. Deng J, Dong W, Socher R, Li L-J, Kai Li, Li Fei-Fei (2009) ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Institute of Electrical and Electronics Engineers (IEEE), Miami, FL, USA, pp 248–255
137. Simonyan K, Zisserman A (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations, ICLR 2015
138. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going Deeper with Convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, Boston, MA, USA, pp 1–9
139. Lo S-CB, Chan HP, Lin J-S, Li H, Freedman MT, Mun SK (1995) Artificial convolution neural network for medical image pattern recognition. Neural Networks 8:1201–1214
140. Sahiner B, Chan HP, Petrick N, Wei D, Helvie MA, Adler DD, Goodsitt MM (1996) Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images. IEEE Trans Med Imaging 15:598–610
141. Lo S-CB, Lou S-LA, Lin J-S, Freedman MT, Chien M V., Mun SK (1995) Artificial convolution neural network techniques and applications for lung nodule detection. IEEE Trans Med Imaging 14:711–718
142. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, Munich, Germany, pp 234–241
143. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH (2021) nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 2020 18:2 18:203–211
144. Zhou B, Chen X, Zhou SK, Duncan JS, Liu C (2022) DuDoDR-Net: Dual-domain data consistent recurrent network for simultaneous sparse view and metal artifact reduction in computed tomography. Med Image Anal. https://doi.org/10.1016/J.MEDIA.2021.102289
145. Siddique N, Paheding S, Elkin CP, Devabhaktuni V (2021) U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications. IEEE Access 9:82031–82057
146. Oktay O, Schlemper J, Folgoc L Le, et al (2018) Attention U-Net: Learning Where to Look for the Pancreas.
147. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y (2021) TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation.
148. Longpre S, Mahari R, Obeng-Marnu N, Brannon W, South T, Gero K, Pentland S, Kabbara J (2024) Data Authenticity, Consent, & Provenance for AI are all broken: what will it take to fix them? Proceedings of the 41 st International Conference on Machine Learning, PMLR 235
149. Bommasani R, Klyman K, Longpre S, Kapoor S, Maslej N, Xiong B, Zhang D, Liang P (2023) The Foundation Model Transparency Index.
150. Amann J, Blasimme A, Vayena E, Frey D, Madai VI (2020) Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inform Decis Mak 20:1–9

151. Pagano TP, Loureiro RB, Lisboa FVN, et al (2022) Bias and unfairness in machine learning models: a systematic literature review.
152. Le Quy T, Roy A, Iosifidis V, Zhang W, Ntoutsi E (2021) A survey on datasets for fairness-aware machine learning. Wiley Interdiscip Rev Data Min Knowl Discov. https://doi.org/10.1002/widm.1452
153. Kemker R, McClure M, Abitino A, Hayes TL, Kanan C (2018) Measuring Catastrophic Forgetting in Neural Networks. In: 32nd AAAI Conference on Artificial Intelligence. AAAI press, pp 3390–3398
154. Kirkpatrick J, Pascanu R, Rabinowitz N, et al (2016) Overcoming catastrophic forgetting in neural networks. In: Proc Natl Acad Sci U S A. National Academy of Sciences, pp 3521–3526
155. Le H, Tran T, Venkatesh S (2018) Dual control memory augmented neural networks for treatment recommendations. In: Advances in Knowledge Discovery and Data Mining. PAKDD. Springer, Cham, pp 273–284
156. Cheng Y, Wang D, Zhou P, Zhang T (2018) Model Compression and Acceleration for Deep Neural Networks: The Principles, Progress, and Challenges. IEEE Signal Process Mag 35:126–136
157. Gou J, Yu B, Maybank SJ, Tao D (2020) Knowledge Distillation: A Survey. Int J Comput Vis 129:1789–1819
158. Bucila C, Caruana R, Niculescu-Mizil A (2006) Model Compression. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06). ACM, New York, NY, USA, pp 535–541
159. Hinton G, Vinyals O, Dean J (2015) Distilling the Knowledge in a Neural Network.
160. Müller R, Kornblith S, Hinton G (2019) When Does Label Smoothing Help? Adv Neural Inf Process Syst 32:
161. Ding Q, Wu S, Sun H, Guo J, Xia S-T (2019) Adaptive Regularization of Labels.
162. Zhang F, Zhu X, Ye M (2018) Fast Human Pose Estimation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, pp 3512–3521
163. Chen G, Choi W, Yu X, Han T, Chandraker M (2017) Learning Efficient Object Detection Models with Knowledge Distillation. 31st Conference on Neural Information Processing Systems (NIPS 2017)
164. Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, Bengio Y (2014) FitNets: Hints for Thin Deep Nets. ICLR 2015 - 3rd International Conference on Learning Representations
165. Wang X, Fu T, Liao S, Wang S, Lei Z, Mei T (2020) Exclusivity-Consistency Regularized Knowledge Distillation for Face Recognition. In: Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science(). Springer Science and Business Media Deutschland GmbH, pp 325–342
166. Xu K, Rui L, Li Y, Gu L (2020) Feature Normalized Knowledge Distillation for Image Classification. In: Vedaldi A, Bischof H, Brox T, Frahm J (eds) Computer Vision – ECCV 2020. Lecture Notes in Computer Science. Springer Science, Cham, pp 664–680
167. Zagoruyko S, Komodakis N (2016) Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings
168. Huang Z, Wang N (2017) Like What You Like: Knowledge Distill via Neuron Selectivity Transfer.
169. Heo B, Lee M, Yun S, Choi JY (2018) Knowledge Transfer via Distillation of Activation Boundaries Formed by Hidden Neurons. In: 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence

Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019. AAAI Press, pp 3779–3787

170. Kim J, Park SU, Kwak N (2018) Paraphrasing Complex Network: Network Compression via Factor Transfer. Adv Neural Inf Process Syst 2018-December:2760–2769
171. Passalis N, Tefas A (2018) Learning Deep Representations with Probabilistic Knowledge Transfer. In: Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science(). Springer Verlag, pp 283–299
172. Chen D, Mei JP, Zhang Y, Wang C, Wang Z, Feng Y, Chen C (2020) Cross-Layer Distillation with Semantic Calibration. In: 35th AAAI Conference on Artificial Intelligence, AAAI 2021. Association for the Advancement of Artificial Intelligence, pp 7028–7036
173. Park W, Kim D, Lu Y, Cho M (2019) Relational Knowledge Distillation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, pp 3962–3971
174. Yim J, Joo D, Bae J, Kim J (2017) A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. Institute of Electrical and Electronics Engineers Inc., pp 7130–7138
175. Zhang C, Peng Y (2018) Better and faster: Knowledge transfer from multiple self-supervised learning tasks via graph distillation for video classification. In: IJCAI International Joint Conference on Artificial Intelligence. International Joint Conferences on Artificial Intelligence, pp 1135–1141
176. Lee SH, Kim DH, Song BC (2018) Self-supervised Knowledge Distillation Using Singular Value Decomposition. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11210 LNCS:339–354
177. Passalis N, Tzelepi M, Tefas A (2020) Heterogeneous Knowledge Distillation Using Information Flow Modeling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, pp 2336–2345
178. Lee S, Song BC (2019) Graph-based Knowledge Distillation by Multi-head Attention Network. 30th British Machine Vision Conference 2019, BMVC 2019
179. Mirzadeh SI, Farajtabar M, Li A, Levine N, Matsukawa A, Ghasemzadeh H (2019) Improved Knowledge Distillation via Teacher Assistant. In: AAAI 2020 - 34th AAAI Conference on Artificial Intelligence. AAAI press, pp 5191–5198
180. Zhang Y, Xiang T, Hospedales TM, Lu H (2017) Deep Mutual Learning. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, pp 4320–4328
181. Hou Y, Ma Z, Liu C, Loy CC (2019) Learning lightweight lane detection CNNS by self attention distillation. In: Proceedings of the IEEE International Conference on Computer Vision. Institute of Electrical and Electronics Engineers Inc., pp 1013–1021
182. Schaudt D, von Schwerin R, Hafner A, Riedel P, Späte C, Reichert M, Hinteregger A, Beer M, Kloth C (2023) Leveraging human expert image annotations to improve pneumonia differentiation through human knowledge distillation. Sci Rep 13:1–13
183. Ying M, Wang Y, Yang K, Wang H, Liu X (2024) A deep learning knowledge distillation framework using knee MRI and arthroscopy data for meniscus tear detection. Front Bioeng Biotechnol. https://doi.org/10.3389/FBIOE.2023.1326706
184. Hsu Y-C, Smith J, Shen Y, Kira Z, Jin H (2022) A Closer Look at Knowledge Distillation with Features, Logits, and Gradients.
185. de Rijk P, Schneider L, Cordts M, Gavrila DM (2022) Structural Knowledge Distillation for Object Detection. Adv Neural Inf Process Syst 35:

186. Park DY, Cha MH, Jeong C, Kim DS, Han B (2021) Learning Student-Friendly Teacher Networks for Knowledge Distillation. In: Adv Neural Inf Process Syst. Neural information processing systems foundation, pp 13292–13303
187. Huang T, You S, Wang F, Qian C, Xu C (2022) Knowledge Distillation from A Stronger Teacher. Adv Neural Inf Process Syst 35:
188. Malihi L, Heidemann G (2024) Matching the Ideal Pruning Method with Knowledge Distillation for Optimal Compression. Applied System Innovation 2024, Vol 7, Page 56 7:56
189. Sarfraz F, Arani E, Zonooz B (2020) Knowledge Distillation Beyond Model Compression. In: Proceedings - International Conference on Pattern Recognition. Institute of Electrical and Electronics Engineers Inc., pp 6181–6188
190. Micaelli P, Storkey A (2019) Zero-shot Knowledge Transfer via Adversarial Belief Matching. Adv Neural Inf Process Syst 32:
191. Nayak GK, Reddy Mopuri K, Chakraborty A (2020) Effectiveness of Arbitrary Transfer Sets for Data-free Knowledge Distillation. In: Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021. Institute of Electrical and Electronics Engineers Inc., pp 1429–1437
192. Johnson AEW, Pollard TJ, Greenbaum NR, Lungren MP, Deng C, Peng Y, Lu Z, Mark RG, Berkowitz SJ, Horng S (2019) MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs.
193. Hansell DM, Bankier AA, MacMahon H, McLoud TC, Müller NL, Remy J (2008) Fleischner Society: glossary of terms for thoracic imaging. Radiology 246:697–722
194. Garbin C, Rajpurkar P, Irvin J, Lungren MP, Marques O (2021) Structured dataset documentation: a datasheet for CheXpert.
195. Wick M, panda swetasudha, Tristan J-B (2019) Unlocking Fairness: a Trade-off Revisited. Adv Neural Inf Process Syst 32:
196. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla N V., Herrera F (2012) A unifying view on dataset shift in classification. Pattern Recognit 45:521–530
197. Xie Q, Luong MT, Hovy E, Le Q V. (2020) Self-training with noisy student improves imagenet classification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, Seattle, WA, USA, pp 10684–10695
198. Wang X, Fu T, Liao S, Wang S, Lei Z, Mei T (2020) Exclusivity-Consistency Regularized Knowledge Distillation for Face Recognition. In: Vedaldi A, Bischof H, Brox T, Frahm J (eds) Computer Vision – ECCV 2020. Lecture Notes in Computer Science. Springer, Cham, pp 325–342
199. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M (2021) CheXclusion: Fairness gaps in deep chest X-ray classifiers. Pacific Symposium on Biocomputing 26:232–243
200. Balestriero R, Ibrahim M, Sobal V, et al (2023) A Cookbook of Self-Supervised Learning.
201. Smith LN, Topin N (2017) Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. https://doi.org/10.1117/12.2520589
202. Jollife IT, Cadima J (2016) Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci 374:20150202
203. Van Der Maaten L, Hinton G (2008) Visualizing Data using t-SNE. Journal of Machine Learning Research 9:2579–2605
204. Ustun B, Liu Y, Parkes DC (2019) Fairness without Harm: Decoupled Classifiers with Preference Guarantees. In: Proceedings of the 36th International Conference on Machine Learning. PMLR, pp 6373–6382
205. Wang H, Hsu H, DIaz M, Calmon FP (2021) To Split or not to Split: The Impact of Disparate Treatment in Classification. IEEE Trans Inf Theory 67:6733–6757

Supplemental Material

S.1. Performance Analysis – CheXpert

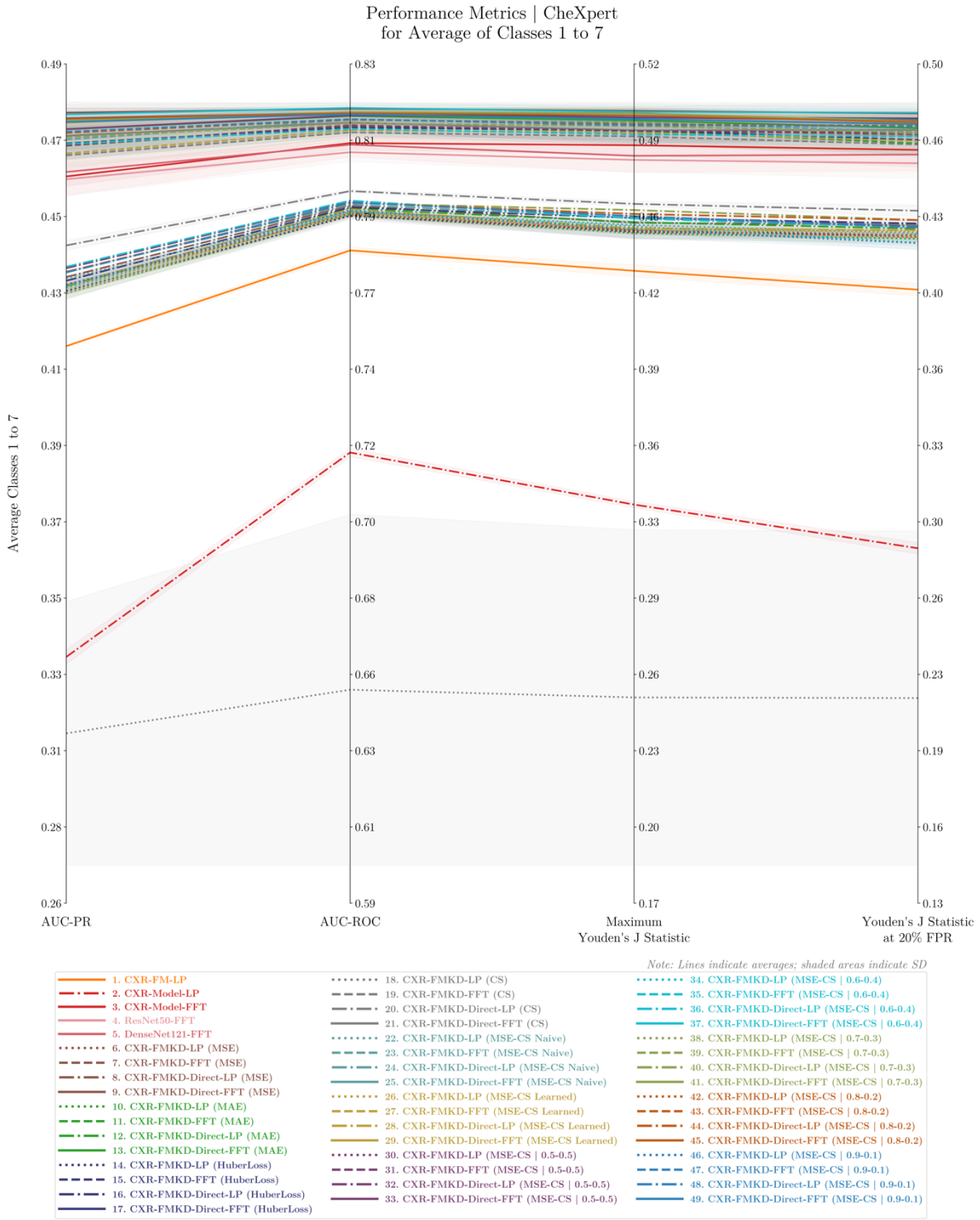

Figure 51. Comparative Analysis of Performance Metrics Across 49 Models for CheXpert Dataset. This custom parallel coordinate plot visualises the performance metrics—AUC-ROC, AUC-PR, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR—for all 49 models tested on the CheXpert dataset, focusing on the average results for the most significant disease labels (Classes 1 to 7). Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. Notably, except for CXR-Model LP and CXR-FMKD LP (CS) which significantly underperformed compared to the rest, we observe a stratification in model performance, grouped by type from top (best performers) to bottom (least performers): CXR-FMKD-Direct FFT models, CXR-FMKD FFT models, Benchmark Models (i.e., CXR-Model FFT, ResNet50 FFT, DenseNet121 FFT), CXR-FMKD-Direct LP models, CXR-FMKD LP models, and lastly the original CXR-FM.

CheXpert

Metric (Test Set)	Class	CXR-FM (1)			CXR-Model FFT (2)			CXR-FMKD-Direct FFT								
								MSE			CS			MSE-CS | 0.6-0.4		
		Avg ± SD (± %SD)	%Δ w.r.t. (1)	%Δ w.r.t. (2)	Avg ± SD (± %SD)	%Δ w.r.t. (1)	%Δ w.r.t. (2)	Avg ± SD (± %SD)	%Δ w.r.t. (1)	%Δ w.r.t. (2)	Avg ± SD (± %SD)	%Δ w.r.t. (1)	%Δ w.r.t. (2)	Avg ± SD (± %SD)	%Δ w.r.t. (1)	%Δ w.r.t. (2)
AUC-PR	Class 1	0.7523 ± 0.0016 (± 0.22%)	0.0%	-6.0%	0.8000 ± 0.0040 (± 0.50%)	6.3%	0.0%	0.8054 ± 0.0033 (± 0.40%)	7.1%	0.7%	0.8022 ± 0.0046 (± 0.58%)	6.6%	0.3%	0.8047 ± 0.0024 (± 0.30%)	7.0%	0.6%
	Class 2	0.3947 ± 0.0048 (± 1.21%)	0.0%	-6.2%	0.4209 ± 0.0090 (± 2.13%)	6.6%	0.0%	0.4410 ± 0.0059 (± 1.33%)	11.7%	4.8%	0.4452 ± 0.0050 (± 1.13%)	12.8%	5.8%	0.4410 ± 0.0054 (± 1.22%)	11.7%	4.8%
	Class 3	0.4688 ± 0.0046 (± 0.97%)	0.0%	-16.2%	0.5595 ± 0.0079 (± 1.42%)	19.3%	0.0%	0.5723 ± 0.0038 (± 0.67%)	22.1%	2.3%	0.5661 ± 0.0085 (± 1.51%)	20.8%	1.2%	0.5699 ± 0.0070 (± 1.23%)	21.6%	1.9%
	Class 4	0.3846 ± 0.0022 (± 0.57%)	0.0%	-16.4%	0.4601 ± 0.0134 (± 2.91%)	19.6%	0.0%	0.4890 ± 0.0065 (± 1.33%)	27.2%	6.3%	0.4707 ± 0.0111 (± 2.37%)	22.4%	2.3%	0.4824 ± 0.0061 (± 1.26%)	25.4%	4.9%
	Class 5	0.2216 ± 0.0020 (± 0.88%)	0.0%	-8.5%	0.2421 ± 0.0119 (± 4.90%)	9.3%	0.0%	0.2617 ± 0.0032 (± 1.22%)	18.1%	8.1%	0.2575 ± 0.0059 (± 2.29%)	16.2%	6.3%	0.2655 ± 0.0059 (± 2.24%)	19.8%	9.6%
	Class 6	0.1220 ± 0.0017 (± 1.40%)	0.0%	-17.4%	0.1477 ± 0.0022 (± 1.46%)	21.1%	0.0%	0.1560 ± 0.0033 (± 2.12%)	27.9%	5.7%	0.1500 ± 0.0030 (± 1.99%)	23.0%	1.6%	0.1574 ± 0.0019 (± 1.20%)	29.0%	6.6%
	Class 7	0.5532 ± 0.0033 (± 0.59%)	0.0%	-5.5%	0.5853 ± 0.0061 (± 1.04%)	5.8%	0.0%	0.6098 ± 0.0019 (± 0.31%)	10.2%	4.2%	0.6003 ± 0.0033 (± 0.56%)	8.5%	2.6%	0.6111 ± 0.0027 (± 0.44%)	10.5%	4.4%
	Average Classes 1 to 7	**0.4139 ± 0.0005 (± 0.11%)**	**0.0%**	**-9.9%**	**0.4594 ± 0.0026 (± 0.56%)**	**11.0%**	**0.0%**	**0.4765 ± 0.0028 (± 0.59%)**	**15.1%**	**3.7%**	**0.4703 ± 0.0042 (± 0.89%)**	**13.6%**	**2.4%**	**0.4760 ± 0.0009 (± 0.19%)**	**15.0%**	**3.6%**
	Others	0.2811 ± 0.0011 (± 0.39%)	0.0%	-6.1%	0.2994 ± 0.0022 (± 0.72%)	6.5%	0.0%	0.3064 ± 0.0022 (± 0.72%)	9.0%	2.3%	0.3019 ± 0.0045 (± 1.49%)	7.4%	0.8%	0.3077 ± 0.0051 (± 1.66%)	9.5%	2.8%
AUC-ROC	Class 1	0.8332 ± 0.0013 (± 0.16%)	0.0%	-3.6%	0.8647 ± 0.0014 (± 0.16%)	3.8%	0.0%	0.8687 ± 0.0023 (± 0.26%)	4.3%	0.5%	0.8663 ± 0.0021 (± 0.24%)	4.0%	0.2%	0.8684 ± 0.0012 (± 0.14%)	4.2%	0.4%
	Class 2	0.8581 ± 0.0012 (± 0.14%)	0.0%	-1.6%	0.8717 ± 0.0025 (± 0.29%)	1.6%	0.0%	0.8801 ± 0.0015 (± 0.17%)	2.6%	1.0%	0.8802 ± 0.0014 (± 0.16%)	2.6%	1.0%	0.8794 ± 0.0014 (± 0.16%)	2.5%	0.9%
	Class 3	0.8077 ± 0.0023 (± 0.28%)	0.0%	-6.5%	0.8640 ± 0.0028 (± 0.32%)	7.0%	0.0%	0.8669 ± 0.0012 (± 0.14%)	7.3%	0.3%	0.8640 ± 0.0040 (± 0.47%)	7.0%	0.0%	0.8658 ± 0.0016 (± 0.18%)	7.2%	0.2%
	Class 4	0.8198 ± 0.0024 (± 0.29%)	0.0%	-4.5%	0.8587 ± 0.0031 (± 0.36%)	4.7%	0.0%	0.8703 ± 0.0042 (± 0.48%)	6.2%	1.3%	0.8660 ± 0.0039 (± 0.44%)	5.6%	0.8%	0.8710 ± 0.0039 (± 0.45%)	6.3%	1.4%
	Class 5	0.6439 ± 0.0010 (± 0.15%)	0.0%	-2.8%	0.6624 ± 0.0111 (± 1.68%)	2.9%	0.0%	0.6776 ± 0.0015 (± 0.22%)	5.2%	2.3%	0.6714 ± 0.0056 (± 0.84%)	4.3%	1.4%	0.6776 ± 0.0034 (± 0.50%)	5.2%	2.3%
	Class 6	0.6837 ± 0.0020 (± 0.29%)	0.0%	-4.8%	0.7183 ± 0.0041 (± 0.57%)	5.1%	0.0%	0.7351 ± 0.0032 (± 0.44%)	7.5%	2.3%	0.7267 ± 0.0058 (± 0.80%)	6.3%	1.2%	0.7375 ± 0.0036 (± 0.49%)	7.9%	2.7%
	Class 7	0.7958 ± 0.0013 (± 0.17%)	0.0%	-2.7%	0.8183 ± 0.0032 (± 0.39%)	2.8%	0.0%	0.8285 ± 0.0014 (± 0.17%)	4.1%	1.3%	0.8246 ± 0.0009 (± 0.11%)	3.6%	0.8%	0.8290 ± 0.0006 (± 0.07%)	4.2%	1.3%
	Average Classes 1 to 7	**0.7774 ± 0.0009 (± 0.12%)**	**0.0%**	**-3.8%**	**0.8083 ± 0.0019 (± 0.23%)**	**4.0%**	**0.0%**	**0.8182 ± 0.0016 (± 0.20%)**	**5.2%**	**1.2%**	**0.8142 ± 0.0027 (± 0.34%)**	**4.7%**	**0.7%**	**0.8184 ± 0.0008 (± 0.10%)**	**5.3%**	**1.3%**
	Others	0.7262 ± 0.0030 (± 0.41%)	0.0%	-3.5%	0.7525 ± 0.0045 (± 0.60%)	3.6%	0.0%	0.7665 ± 0.0023 (± 0.30%)	5.6%	1.9%	0.7637 ± 0.0051 (± 0.67%)	5.2%	1.5%	0.7676 ± 0.0049 (± 0.64%)	5.7%	2.0%
Maximum Youden's J Statistic	Class 1	0.5169 ± 0.0033 (± 0.64%)	0.0%	-10.6%	0.5782 ± 0.0063 (± 1.09%)	11.9%	0.0%	0.5801 ± 0.0058 (± 1.00%)	12.2%	0.3%	0.5764 ± 0.0032 (± 0.56%)	11.5%	-0.3%	0.5781 ± 0.0028 (± 0.49%)	11.8%	0.0%
	Class 2	0.5892 ± 0.0064 (± 1.08%)	0.0%	-3.7%	0.6120 ± 0.0031 (± 0.51%)	3.9%	0.0%	0.6295 ± 0.0078 (± 1.24%)	6.8%	2.9%	0.6253 ± 0.0033 (± 0.53%)	6.1%	2.2%	0.6206 ± 0.0069 (± 1.11%)	5.3%	1.4%
	Class 3	0.4792 ± 0.0050 (± 1.03%)	0.0%	-18.2%	0.5855 ± 0.0083 (± 1.42%)	22.2%	0.0%	0.5888 ± 0.0092 (± 1.57%)	22.9%	0.6%	0.5804 ± 0.0104 (± 1.79%)	21.1%	-0.9%	0.5911 ± 0.0048 (± 0.82%)	23.4%	1.0%
	Class 4	0.4959 ± 0.0075 (± 1.50%)	0.0%	-13.4%	0.5725 ± 0.0143 (± 2.51%)	15.5%	0.0%	0.5860 ± 0.0100 (± 1.71%)	18.2%	2.4%	0.5768 ± 0.0086 (± 1.50%)	16.3%	0.8%	0.5892 ± 0.0061 (± 1.04%)	18.8%	2.9%
	Class 5	0.2091 ± 0.0037 (± 1.75%)	0.0%	-8.1%	0.2276 ± 0.0146 (± 6.41%)	8.9%	0.0%	0.2493 ± 0.0014 (± 0.58%)	19.2%	9.5%	0.2417 ± 0.0117 (± 4.85%)	15.6%	6.2%	0.2507 ± 0.0065 (± 2.61%)	19.9%	10.1%
	Class 6	0.2770 ± 0.0068 (± 2.45%)	0.0%	-16.2%	0.3305 ± 0.0107 (± 3.24%)	19.3%	0.0%	0.3583 ± 0.0033 (± 0.92%)	29.4%	8.4%	0.3392 ± 0.0139 (± 4.11%)	22.5%	2.6%	0.3582 ± 0.0133 (± 3.70%)	29.4%	8.4%
	Class 7	0.4607 ± 0.0040 (± 0.86%)	0.0%	-6.4%	0.4921 ± 0.0072 (± 1.46%)	6.8%	0.0%	0.5081 ± 0.0025 (± 0.49%)	10.3%	3.3%	0.5028 ± 0.0058 (± 1.15%)	9.1%	2.2%	0.5090 ± 0.0042 (± 0.83%)	10.5%	3.4%
	Average Classes 1 to 7	**0.4325 ± 0.0025 (± 0.58%)**	**0.0%**	**-10.9%**	**0.4855 ± 0.0041 (± 0.85%)**	**12.2%**	**0.0%**	**0.5000 ± 0.0036 (± 0.73%)**	**15.6%**	**3.0%**	**0.4918 ± 0.0058 (± 1.17%)**	**13.7%**	**1.3%**	**0.4995 ± 0.0032 (± 0.65%)**	**15.5%**	**2.9%**
	Others	0.3491 ± 0.0043 (± 1.22%)	0.0%	-10.0%	0.3881 ± 0.0074 (± 1.92%)	11.2%	0.0%	0.4152 ± 0.0063 (± 1.53%)	18.9%	7.0%	0.4103 ± 0.0101 (± 2.47%)	17.5%	5.7%	0.4164 ± 0.0117 (± 2.80%)	19.3%	7.3%
Youden's J Statistic at 20% FPR	Class 1	0.4995 ± 0.0050 (± 1.01%)	0.0%	-12.0%	0.5675 ± 0.0055 (± 0.97%)	13.6%	0.0%	0.5697 ± 0.0071 (± 1.24%)	14.0%	0.4%	0.5616 ± 0.0064 (± 1.13%)	12.4%	-1.0%	0.5676 ± 0.0036 (± 0.64%)	13.6%	0.0%
	Class 2	0.5815 ± 0.0063 (± 1.09%)	0.0%	-4.6%	0.6098 ± 0.0036 (± 0.58%)	4.9%	0.0%	0.6234 ± 0.0031 (± 0.49%)	7.2%	2.2%	0.6189 ± 0.0045 (± 0.72%)	6.4%	1.5%	0.6141 ± 0.0029 (± 0.47%)	5.6%	0.7%
	Class 3	0.4519 ± 0.0081 (± 1.79%)	0.0%	-22.1%	0.5800 ± 0.0098 (± 1.69%)	28.3%	0.0%	0.5838 ± 0.0062 (± 1.07%)	29.2%	0.7%	0.5765 ± 0.0087 (± 1.51%)	27.6%	-0.6%	0.5856 ± 0.0052 (± 0.89%)	29.6%	1.0%
	Class 4	0.4899 ± 0.0084 (± 1.71%)	0.0%	-13.1%	0.5636 ± 0.0168 (± 2.99%)	15.1%	0.0%	0.5795 ± 0.0119 (± 2.05%)	18.3%	2.8%	0.5738 ± 0.0099 (± 1.72%)	17.1%	1.8%	0.5811 ± 0.0026 (± 0.44%)	18.6%	3.1%
	Class 5	0.1564 ± 0.0028 (± 1.80%)	0.0%	-15.7%	0.1855 ± 0.0156 (± 8.39%)	18.6%	0.0%	0.2099 ± 0.0043 (± 2.06%)	34.2%	13.2%	0.1998 ± 0.0050 (± 2.50%)	27.7%	7.7%	0.2103 ± 0.0089 (± 4.23%)	34.4%	13.4%
	Class 6	0.2049 ± 0.0074 (± 3.60%)	0.0%	-20.4%	0.2574 ± 0.0217 (± 8.44%)	25.6%	0.0%	0.2852 ± 0.0090 (± 3.14%)	39.2%	10.8%	0.2766 ± 0.0125 (± 4.52%)	35.0%	7.4%	0.2895 ± 0.0081 (± 2.80%)	41.3%	12.5%
	Class 7	0.4055 ± 0.0051 (± 1.27%)	0.0%	-11.5%	0.4585 ± 0.0095 (± 2.07%)	13.1%	0.0%	0.4834 ± 0.0045 (± 0.94%)	19.2%	5.4%	0.4769 ± 0.0062 (± 1.30%)	17.6%	4.0%	0.4893 ± 0.0074 (± 1.52%)	20.7%	6.7%
	Average Classes 1 to 7	**0.3985 ± 0.0028 (± 0.69%)**	**0.0%**	**-13.4%**	**0.4603 ± 0.0075 (± 1.64%)**	**15.5%**	**0.0%**	**0.4764 ± 0.0044 (± 0.92%)**	**19.5%**	**3.5%**	**0.4691 ± 0.0038 (± 0.81%)**	**17.7%**	**1.9%**	**0.4768 ± 0.0024 (± 0.50%)**	**19.6%**	**3.6%**
	Others	0.3114 ± 0.0061 (± 1.97%)	0.0%	-10.6%	0.3484 ± 0.0076 (± 2.18%)	11.9%	0.0%	0.3817 ± 0.0055 (± 1.44%)	22.6%	9.6%	0.3763 ± 0.0078 (± 2.07%)	20.9%	8.0%	0.3805 ± 0.0147 (± 3.87%)	22.2%	9.2%

Table 13. Absolute and Relative Performance of Selected Models Across Most Significant Classes for the CheXpert Dataset.
This table provides detailed performance results across four metrics—AUC-PR, AUC-ROC, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR—for the five selected models tested on the CheXpert dataset. It includes data for the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes in the disease labels list. Results are shown as average outcomes (Avg) with standard deviations (SD and %SD = SD/Avg × 100%), derived from testing five distinct instances of the same model type, each developed and trained using a different seed. Performance improvements are quantified relative to CXR-FM and the CXR-Model FFT baseline, calculated using the formula (Model Avg Value − Baseline Avg Value)/Baseline Avg Value × 100%, where the baseline is either CXR-FM (1) or CXR-Model FFT (2). This analysis highlights the enhancements achieved by the CXR-FMKD-Direct FFT student models in terms of AUC-PR, AUC-ROC, Maximum Youden's J Statistic, and Youden's J Statistic at 20% FPR, underscoring their relative and absolute performance gains.

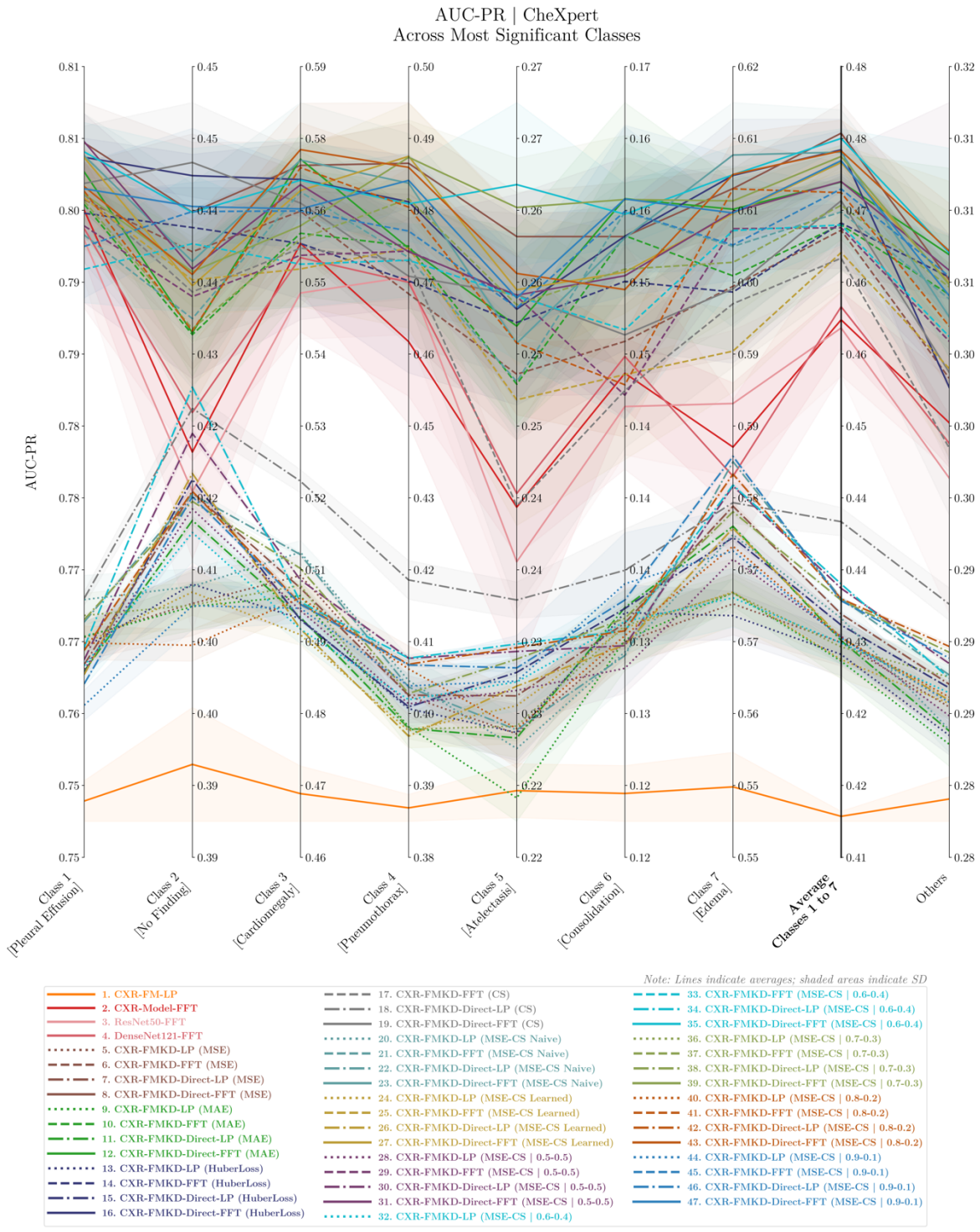

Figure 52. Parallel Coordinate Plot of AUC-PR Performance Across the Most Significant Classes for 47 Models Tested on the CheXpert Dataset.

This custom parallel coordinate plot visualises the AUC-PR metric across 47 models tested on the CheXpert dataset. It displays performance results for the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes in the disease labels list. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. We observe an overall stratification in model performance, grouped by type from top (best performers) to bottom (least performers): CXR-FMKD-Direct FFT models, CXR-FMKD FFT models, Benchmark Models (i.e., CXR-Model FFT, ResNet50 FFT, DenseNet121 FFT), CXR-FMKD-Direct LP models, CXR-FMKD LP models, and lastly the original CXR-FM. Generally, the models exhibit the best performance for Class 1 [Pleural Effusion] and the poorest for Class 6 [Consolidation].

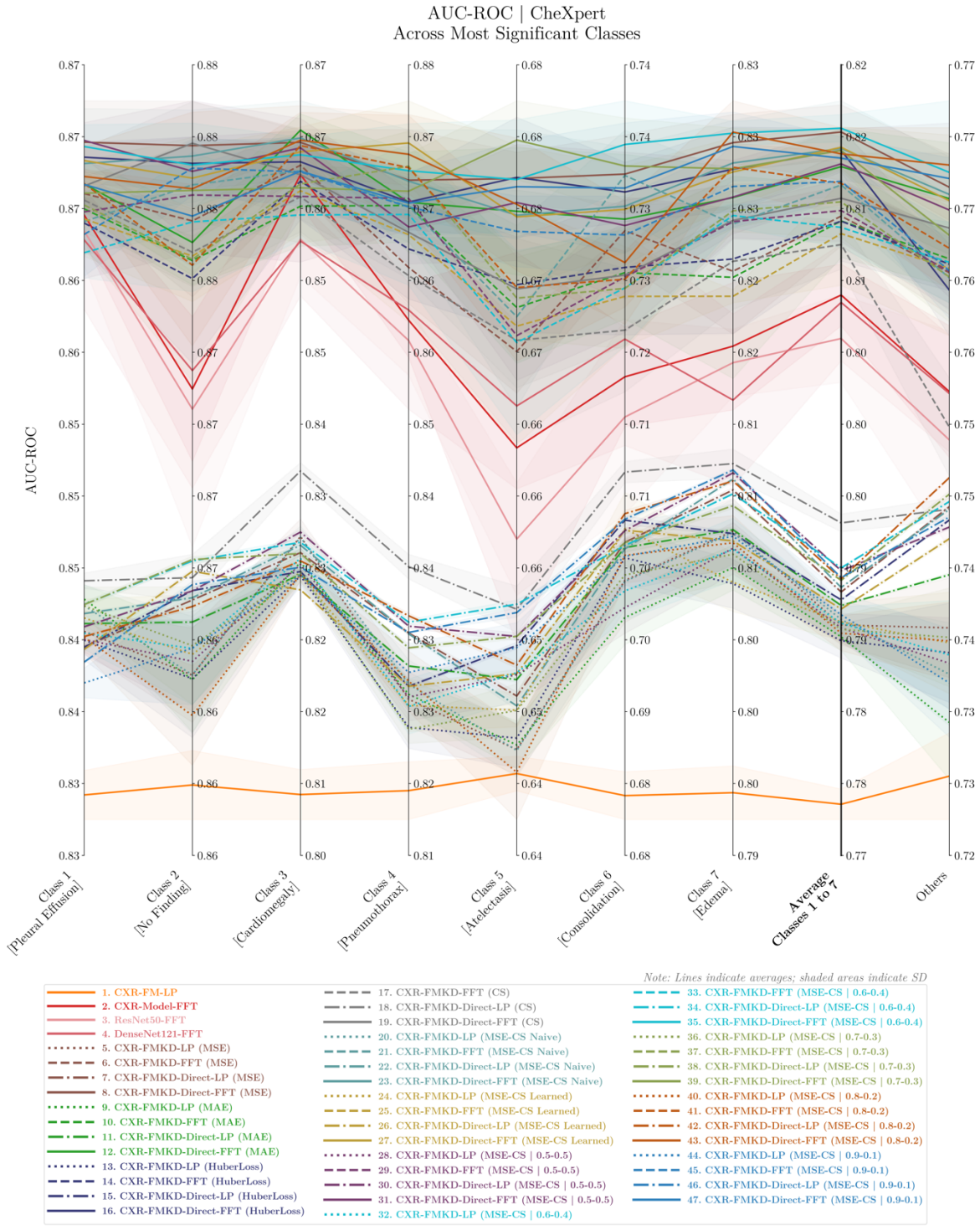

Figure 53. Parallel Coordinate Plot of AUC-ROC Performance Across the Most Significant Classes for 47 Models Tested on the CheXpert Dataset.

This custom parallel coordinate plot visualises the AUC-ROC metric across 47 models tested on the CheXpert dataset. It displays performance results for the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes in the disease labels list. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. We observe an overall stratification in model performance, grouped by type from top (best performers) to bottom (least performers): CXR-FMKD-Direct FFT models, CXR-FMKD FFT models, Benchmark Models (i.e., CXR-Model FFT, ResNet50 FFT, DenseNet121 FFT), CXR-FMKD-Direct LP models, CXR-FMKD LP models, and lastly the original CXR-FM. Generally, the models exhibit the best performance for Class 1 [Pleural Effusion], Class 2 [No Finding], Class 3 [Cardiomegaly], and Class 4 [Pneumothorax]; and the poorest for Class 5 [Atelectasis].

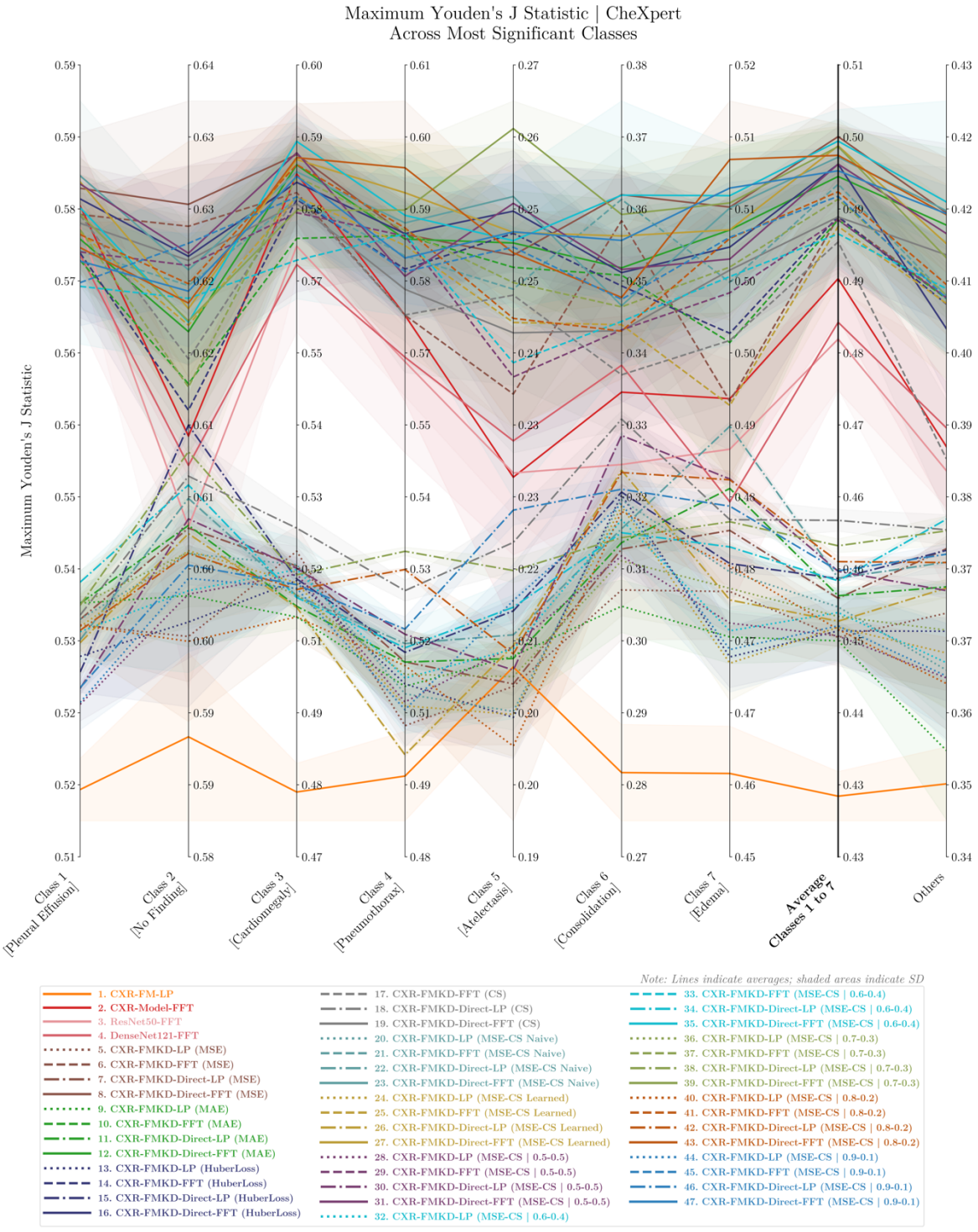

Figure 54. Parallel Coordinate Plot of Maximum Youden's J Statistic Performance Across the Most Significant Classes for 47 Models Tested on the CheXpert Dataset.
This custom parallel coordinate plot visualises the Maximum Youden's J Statistic metric across 47 models tested on the CheXpert dataset. It displays performance results for the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes in the disease labels list. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. We observe an overall stratification in model performance, grouped by type from top (best performers) to bottom (least performers): CXR-FMKD-Direct FFT models, CXR-FMKD FFT models, Benchmark Models (i.e., CXR-Model FFT, ResNet50 FFT, DenseNet121 FFT), CXR-FMKD-Direct LP models, CXR-FMKD LP models, and lastly the original CXR-FM. Generally, the models exhibit the best performance for Class 2 [No Finding] and the poorest for Class 5 [Atelectasis].

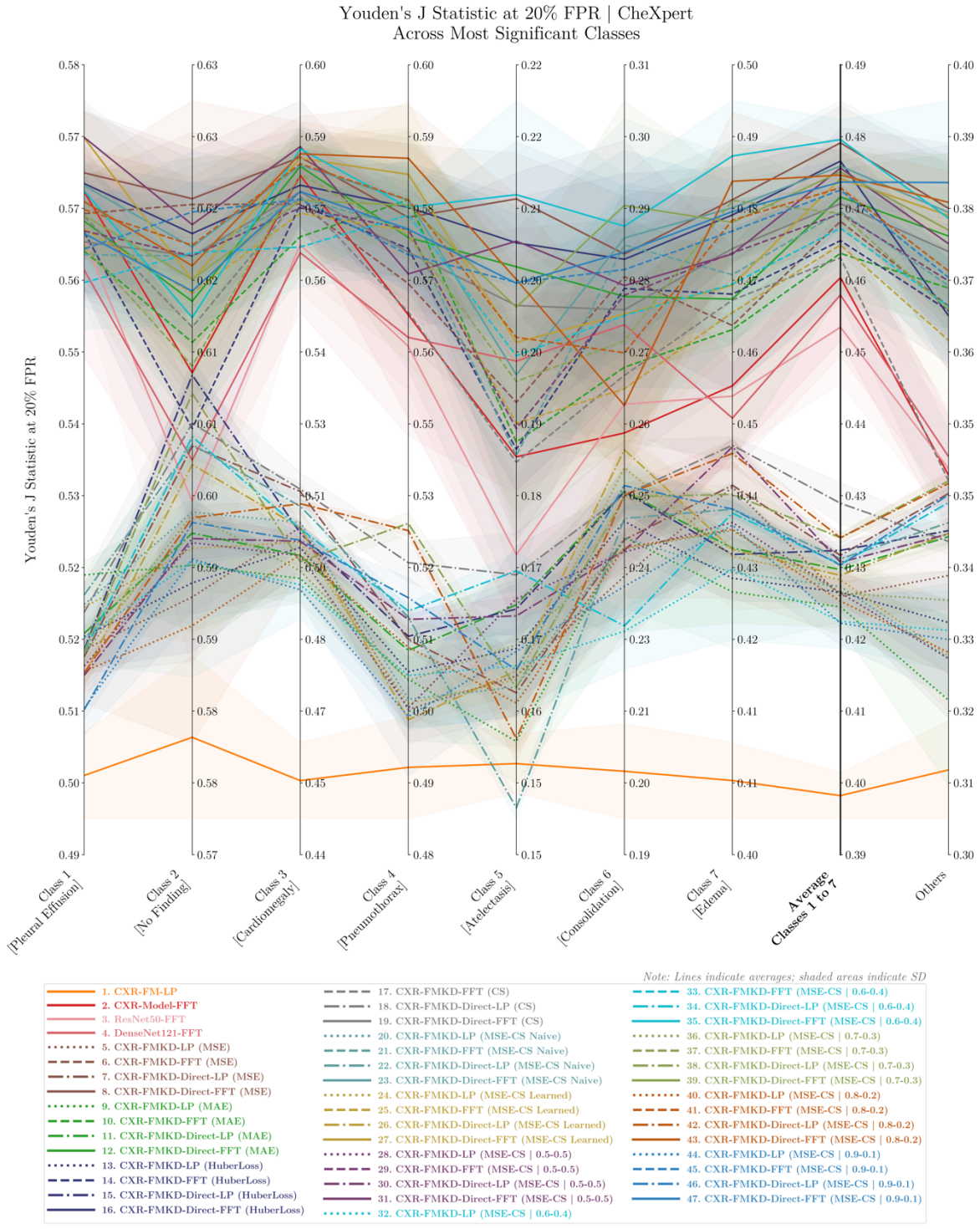

Figure 55. Parallel Coordinate Plot of Youden's J Statistic at 20% FPR Performance Across the Most Significant Classes for 47 Models Tested on the CheXpert Dataset.

This custom parallel coordinate plot visualises the Youden's J Statistic at 20% FPR metric across 47 models tested on the CheXpert dataset. It displays performance results for the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes in the disease labels list. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. We observe an overall stratification in model performance, grouped by type from top (best performers) to bottom (least performers): CXR-FMKD-Direct FFT models, CXR-FMKD FFT models, Benchmark Models (i.e., CXR-Model FFT, ResNet50 FFT, DenseNet121 FFT), CXR-FMKD-Direct LP models, CXR-FMKD LP models, and lastly the original CXR-FM. Generally, the models exhibit the best performance for Class 2 [No Finding] and the poorest for Class 5 [Atelectasis].

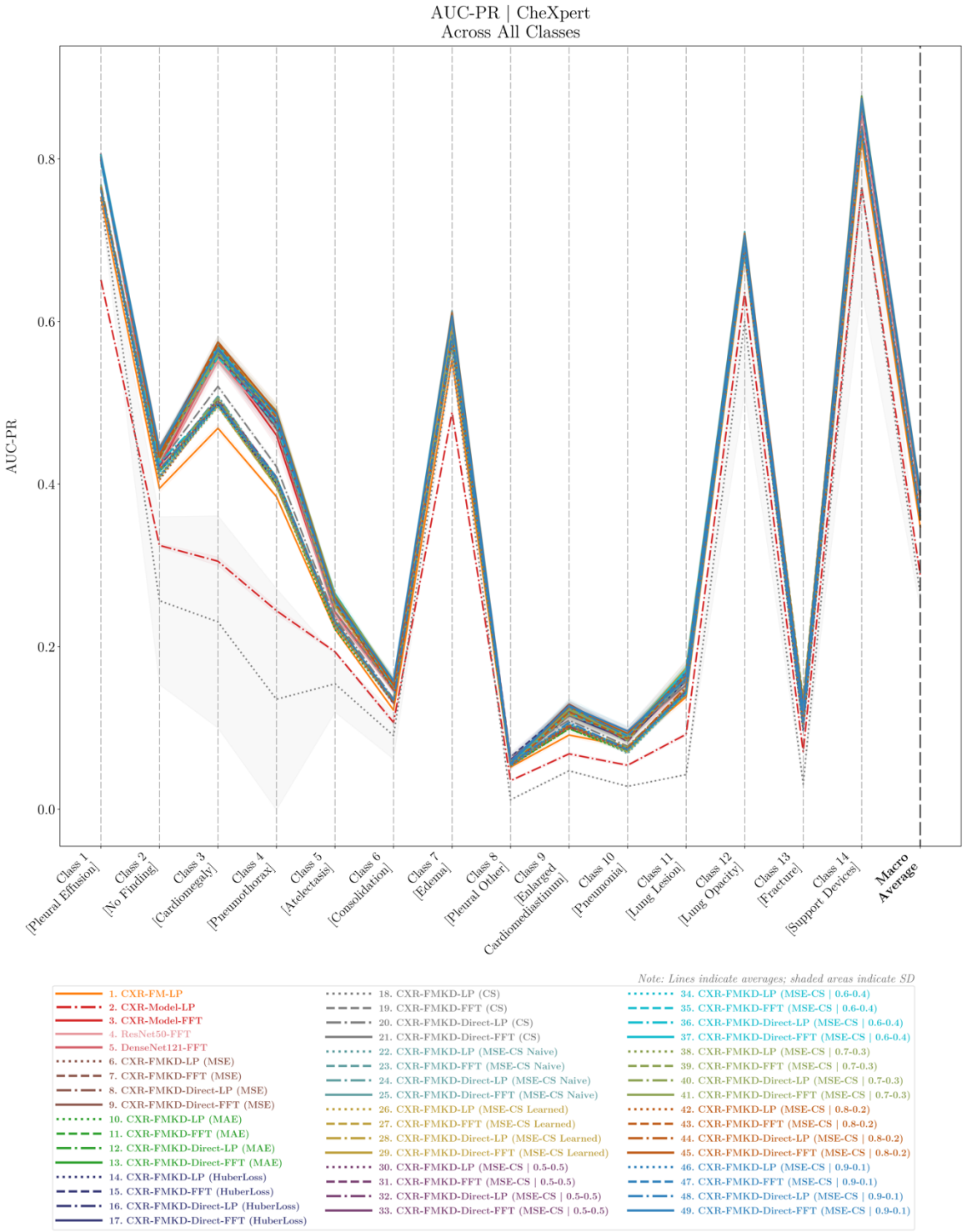

Figure 56. AUC-PR Performance Across All 14 Classes for 49 Models Tested on the CheXpert Dataset. This plot visualises the AUC-PR metric across 49 models tested on the CheXpert dataset. It displays performance results for each of the 14 disease classes as well as their macro-average. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. Generally, the models exhibit the best performance for Class 14 [Support Devices] and Class 1 [Pleural Effusion]; and the poorest for Class 8 [Pleural Others], Class 9 [Enlarged Cardiomediastinum], and Class 10 [Pneumonia].

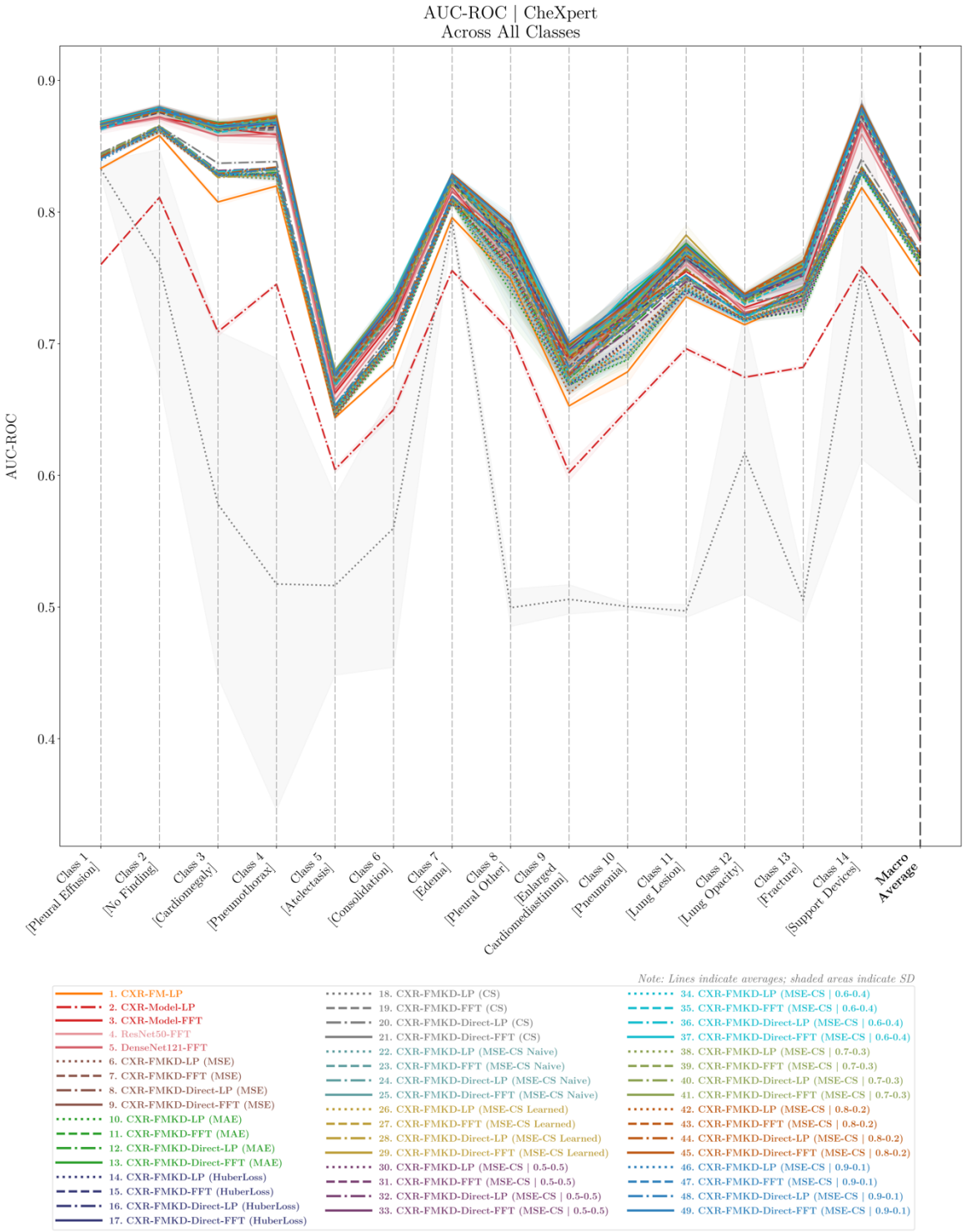

Figure 57. AUC-ROC Performance Across All 14 Classes for 49 Models Tested on the CheXpert Dataset.

This plot visualises the AUC-ROC metric across 49 models tested on the CheXpert dataset. It displays performance results for each of the 14 disease classes as well as their macro-average. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. Generally, the models exhibit the best performance for Class 14 [Support Devices], Class 1 [Pleural Effusion], Class 2 [No Finding], Class 3 [Cardiomegaly], and Class 4 [Pneumothorax]; and the poorest for Class 5 [Atelectasis] and Class 9 [Enlarged Cardiomediastinum].

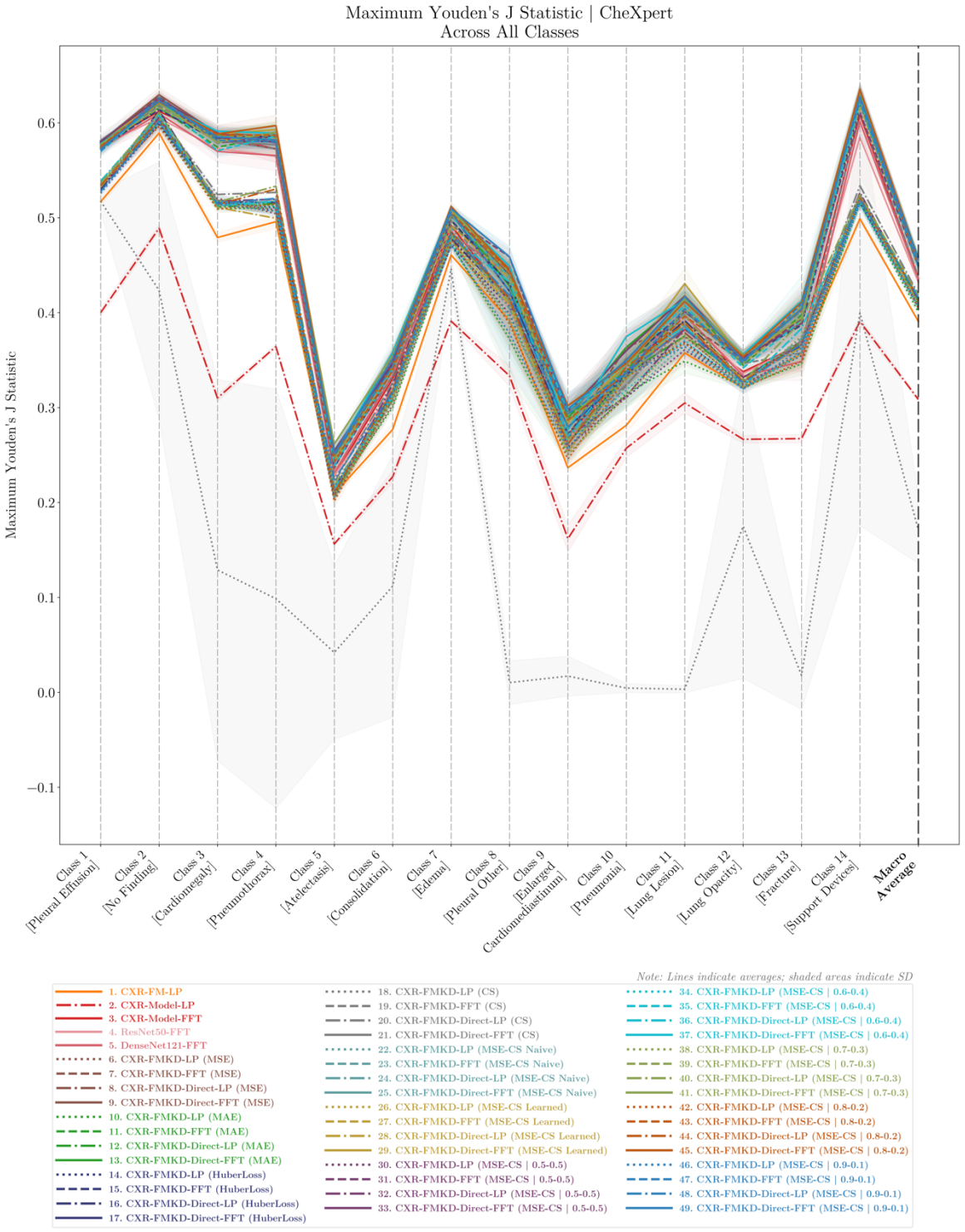

Figure 58. Maximum Youden's J Statistic Performance Across All 14 Classes for 49 Models Tested on the CheXpert Dataset.

This plot visualises the Maximum Youden's J Statistic metric across 49 models tested on the CheXpert dataset. It displays performance results for each of the 14 disease classes as well as their macro-average. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. Generally, the models exhibit the best performance for Class 14 [Support Devices] and Class 2 [No Finding]; and the poorest for Class 5 [Atelectasis] and Class 9 [Enlarged Cardiomediastinum].

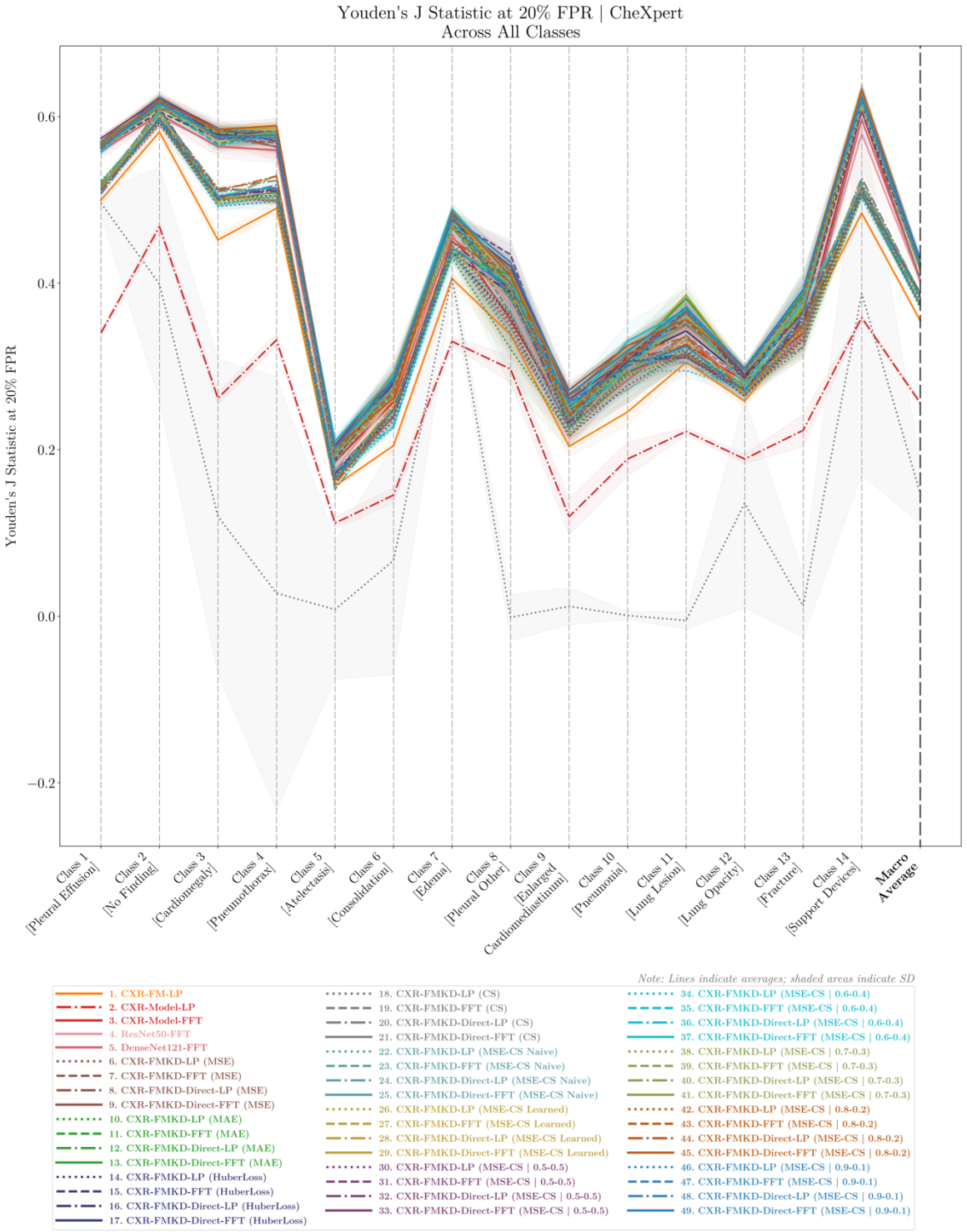

Figure 59. Youden's J Statistic at 20% FPR Performance Across All 14 Classes for 49 Models Tested on the CheXpert Dataset.

This plot visualises the Youden's J Statistic at 20% FPR metric across 49 models tested on the CheXpert dataset. It displays performance results for each of the 14 disease classes as well as their macro-average. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. Generally, the models exhibit the best performance for Class 14 [Support Devices] and Class 2 [No Finding]; and the poorest for Class 5 [Atelectasis] and Class 9 [Enlarged Cardiomediastinum].

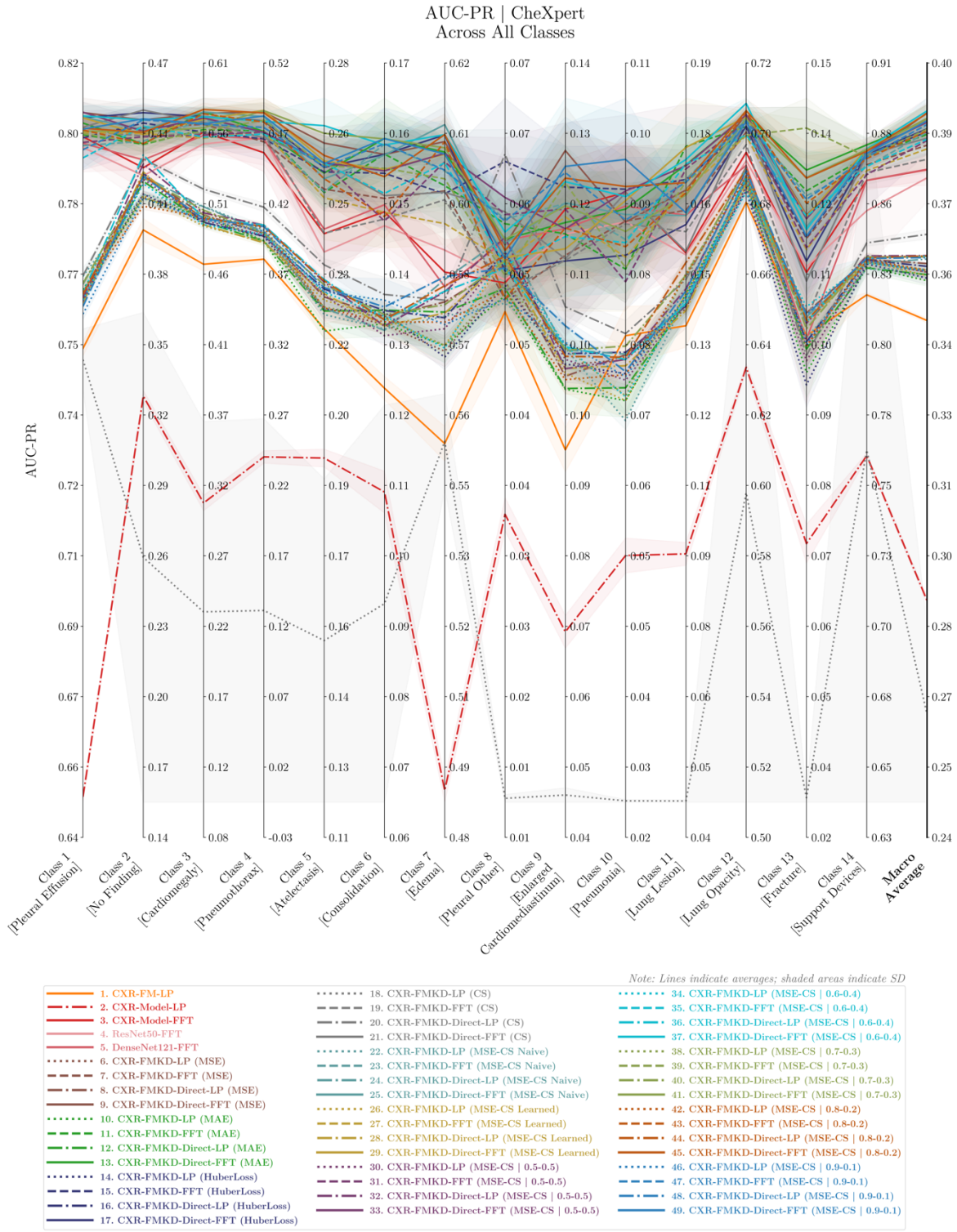

Figure 60. Parallel Coordinate Plot of AUC-PR Performance Across All 14 Classes for 49 Models Tested on the CheXpert Dataset.

This custom parallel coordinate plot visualises the AUC-PR metric across 49 models tested on the CheXpert dataset. It displays performance results for each of the 14 disease classes as well as their macro-average. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. Generally, the models exhibit the best performance for Class 14 [Support Devices] and Class 1 [Pleural Effusion]; and the poorest for Class 8 [Pleural Others], Class 9 [Enlarged Cardiomediastinum], and Class 10 [Pneumonia].

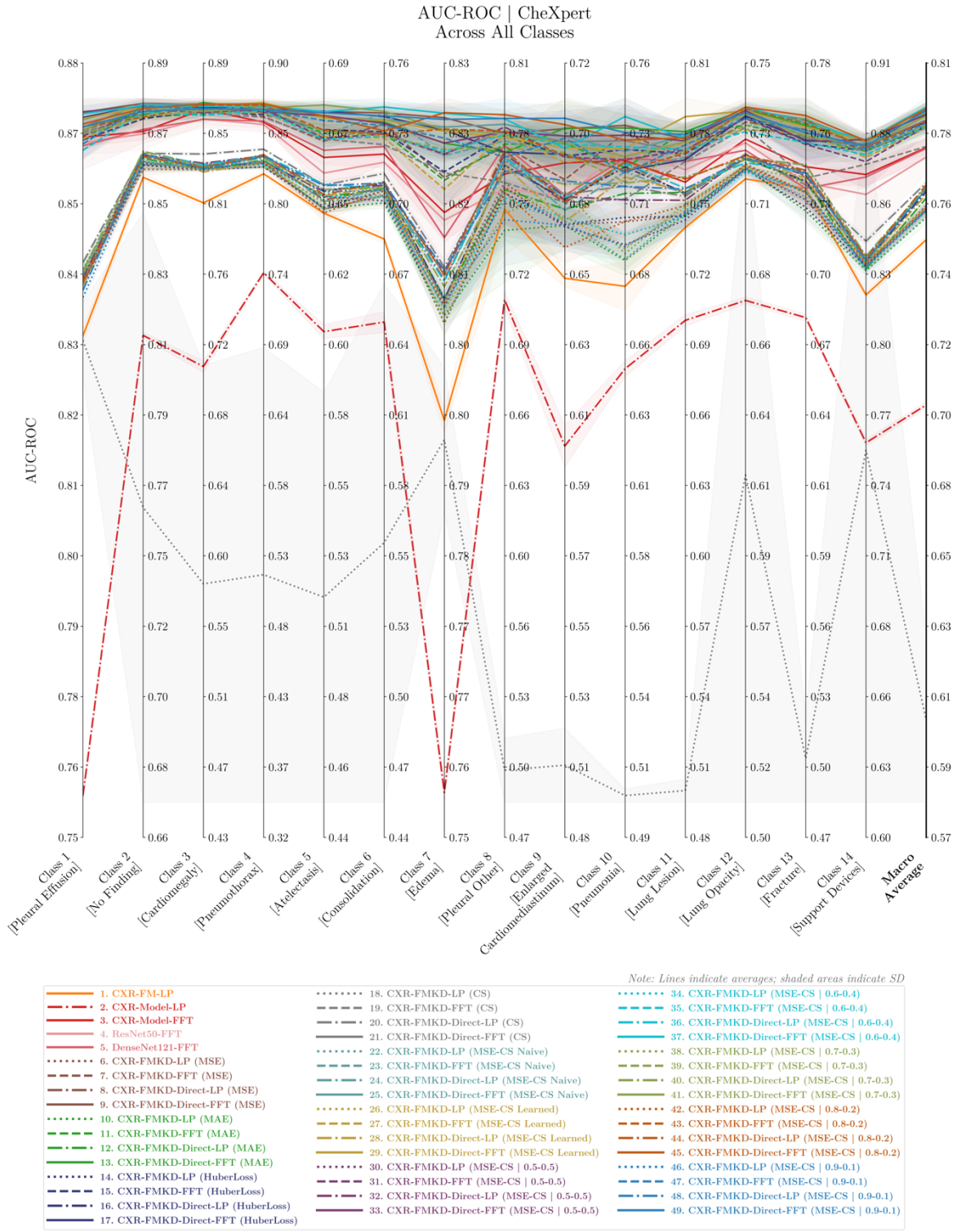

Figure 61. Parallel Coordinate Plot of AUC-ROC Performance Across All 14 Classes for 49 Models Tested on the CheXpert Dataset.

This custom parallel coordinate plot visualises the AUC-ROC metric across 49 models tested on the CheXpert dataset. It displays performance results for each of the 14 disease classes as well as their macro-average. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. Generally, the models exhibit the best performance for Class 14 [Support Devices], Class 1 [Pleural Effusion], Class 2 [No Finding], Class 3 [Cardiomegaly], and Class 4 [Pneumothorax]; and the poorest for Class 5 [Atelectasis] and Class 9 [Enlarged Cardiomediastinum].

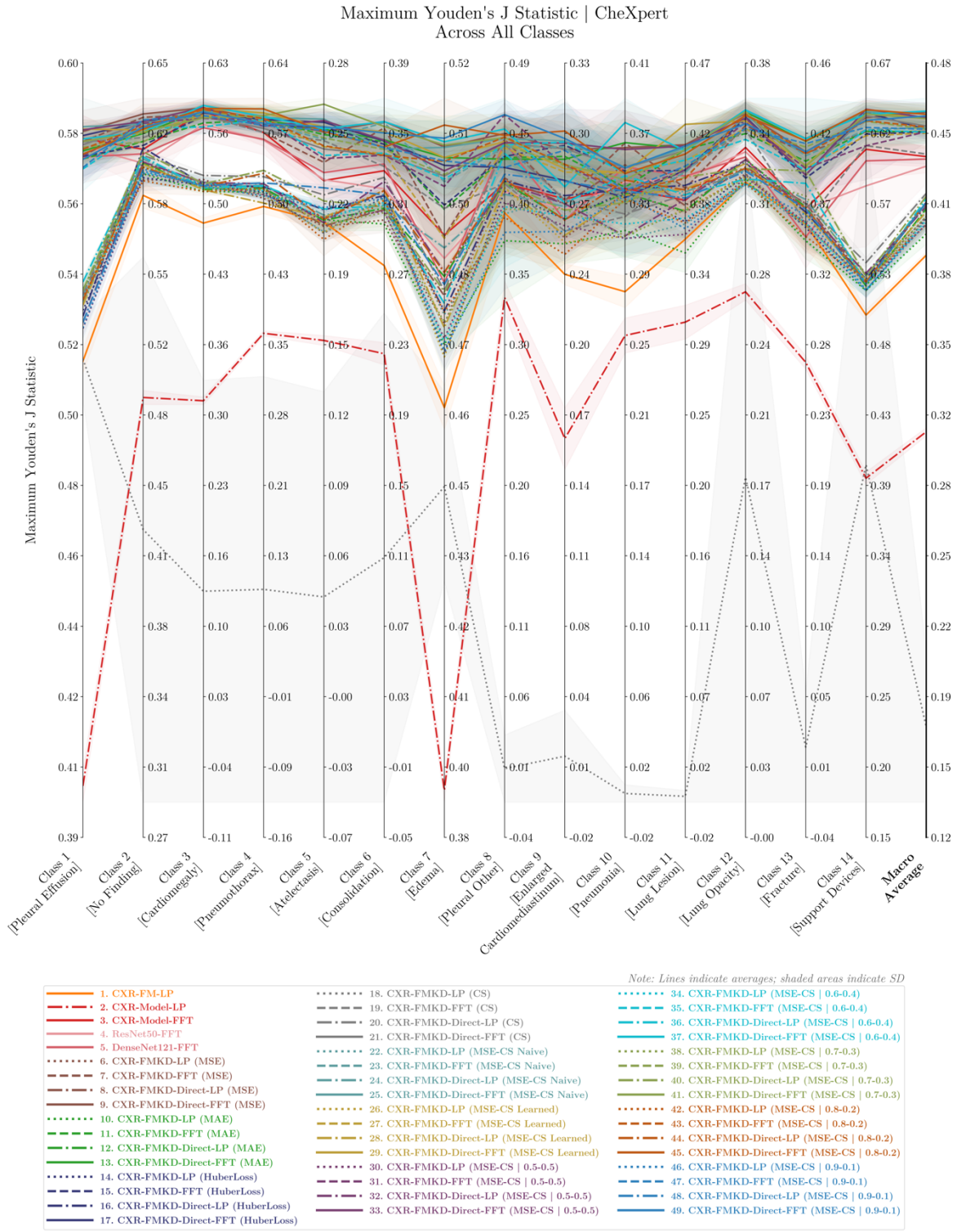

Figure 62. Parallel Coordinate Plot of Maximum Youden's J Statistic Performance Across All 14 Classes for 49 Models Tested on the CheXpert Dataset.

This custom parallel coordinate plot visualises the Maximum Youden's J Statistic metric across 49 models tested on the CheXpert dataset. It displays performance results for each of the 14 disease classes as well as their macro-average. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. Generally, the models exhibit the best performance for Class 14 [Support Devices], Class 2 [No Finding], Class 3 [Cardiomegaly], and Class 4 [Pneumothorax]; and the poorest for Class 5 [Atelectasis] and Class 9 [Enlarged Cardiomediastinum].

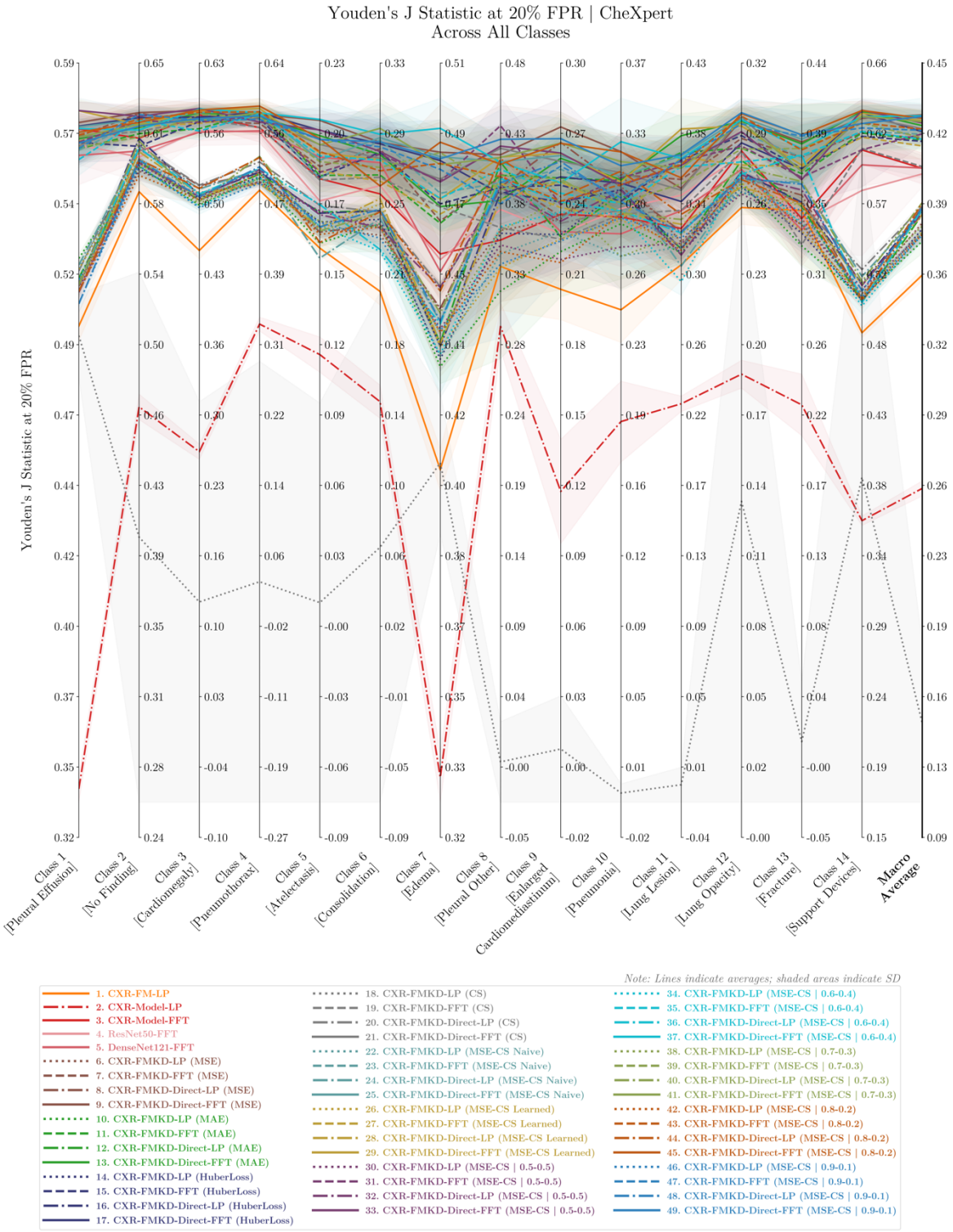

Figure 63. Parallel Coordinate Plot of Youden's J Statistic at 20% FPR Performance Across All 14 Classes for 49 Models Tested on the CheXpert Dataset.

This custom parallel coordinate plot visualises the Youden's J Statistic at 20% FPR metric across 49 models tested on the CheXpert dataset. It displays performance results for each of the 14 disease classes as well as their macro-average. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. Generally, the models exhibit the best performance for Class 14 [Support Devices], Class 2 [No Finding], Class 3 [Cardiomegaly], and Class 4 [Pneumothorax]; and the poorest for Class 5 [Atelectasis] and Class 9 [Enlarged Cardiomediastinum].

S.2. Performance Analysis – MIMIC

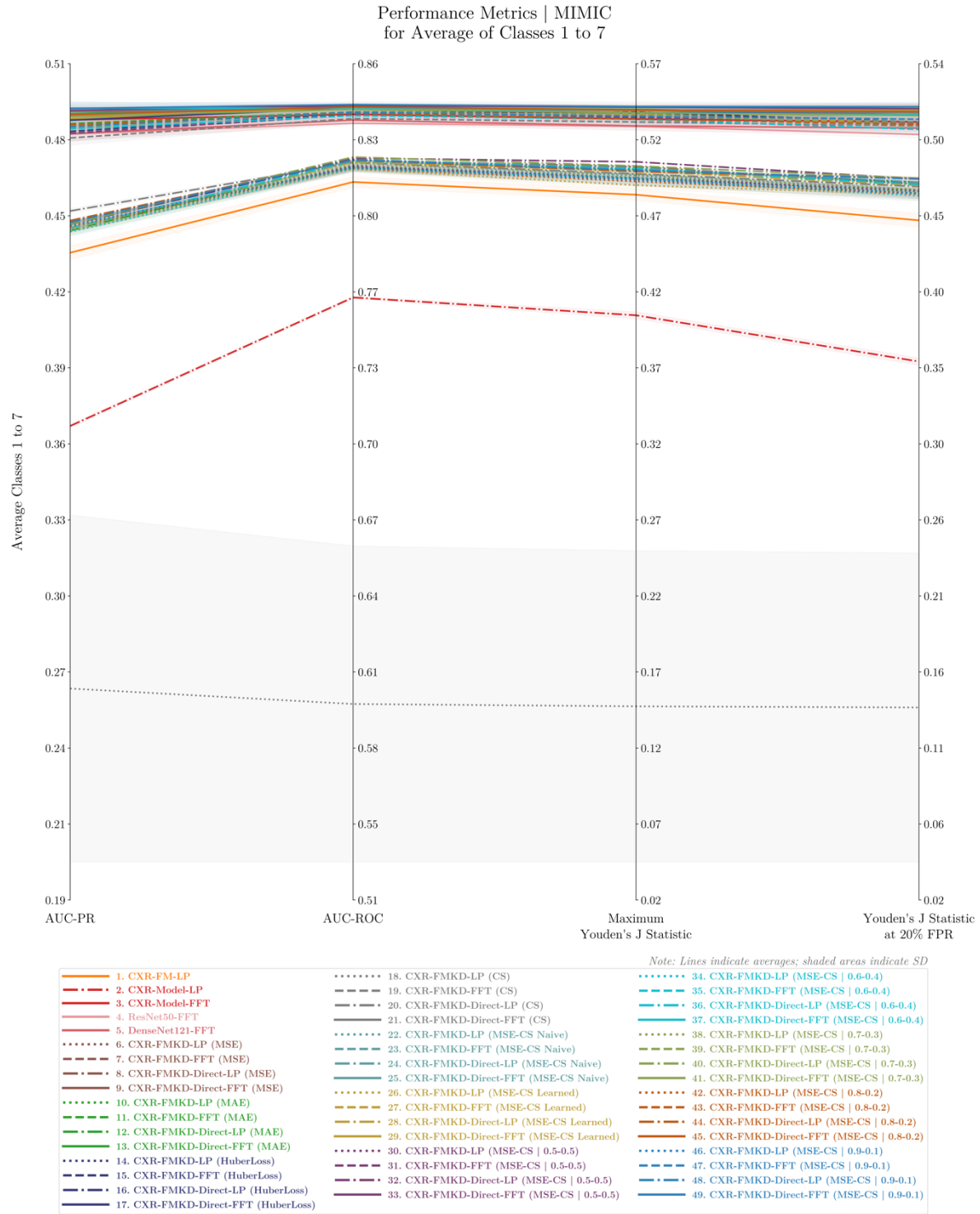

Figure 64. Comparative Analysis of Performance Metrics Across 49 Models for MIMIC Dataset. This custom parallel coordinate plot visualises the performance metrics—AUC-ROC, AUC-PR, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR—for all 49 models tested on the MIMIC dataset, focusing on the average results for the most significant disease labels (Classes 1 to 7). Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. Notably, except for CXR-Model LP and CXR-FMKD LP (CS) which significantly underperformed compared to the rest, we observe a stratification in model performance, grouped by type from top (best performers) to bottom (least performers): CXR-FMKD-Direct FFT models, CXR-FMKD FFT models, Benchmark Models (i.e., CXR-Model FFT, ResNet50 FFT, DenseNet121 FFT), CXR-FMKD-Direct LP models, CXR-FMKD LP models, and lastly the original CXR-FM.

MIMIC																	
Metric (Test Set)	Class	CXR-FM (1)			CXR-Model FFT (2)			CXR-FMKD-Direct FFT									
								MSE			CS			MSE-CS | 0.9-0.1			
		Avg ± SD (± %SD)	%Δ w.r.t. (1)	%Δ w.r.t. (2)	Avg ± SD (± %SD)	%Δ w.r.t. (1)	%Δ w.r.t. (2)	Avg ± SD (± %SD)	%Δ w.r.t. (1)	%Δ w.r.t. (2)	Avg ± SD (± %SD)	%Δ w.r.t. (1)	%Δ w.r.t. (2)	Avg ± SD (± %SD)	%Δ w.r.t. (1)	%Δ w.r.t. (2)
AUC-PR	Class 1	0.6665 ± 0.0028 (± 0.42%)	0.0%	-8.3%	0.7267 ± 0.0048 (± 0.66%)	9.0%	0.0%	0.7326 ± 0.0032 (± 0.44%)	9.9%	0.8%	0.7341 ± 0.0031 (± 0.42%)	10.1%	1.0%	0.7294 ± 0.0027 (± 0.37%)	9.4%	0.4%
	Class 2	0.6989 ± 0.0012 (± 0.17%)	0.0%	-2.9%	0.7196 ± 0.0018 (± 0.25%)	3.0%	0.0%	0.7216 ± 0.0051 (± 0.70%)	3.2%	0.3%	0.7251 ± 0.0049 (± 0.67%)	3.7%	0.8%	0.7253 ± 0.0032 (± 0.44%)	3.8%	0.8%
	Class 3	0.4729 ± 0.0035 (± 0.75%)	0.0%	-6.9%	0.5079 ± 0.0063 (± 1.25%)	7.4%	0.0%	0.5109 ± 0.0028 (± 0.54%)	8.0%	0.6%	0.5077 ± 0.0036 (± 0.71%)	7.4%	0.0%	0.5126 ± 0.0020 (± 0.40%)	8.4%	0.9%
	Class 4	0.2217 ± 0.0077 (± 3.49%)	0.0%	-28.0%	0.3079 ± 0.0087 (± 2.82%)	38.9%	0.0%	0.3086 ± 0.0063 (± 2.05%)	39.2%	0.2%	0.3026 ± 0.0137 (± 4.54%)	36.5%	-1.7%	0.3048 ± 0.0141 (± 4.62%)	37.5%	-1.0%
	Class 5	0.4086 ± 0.0029 (± 0.71%)	0.0%	-10.5%	0.4564 ± 0.0056 (± 1.22%)	11.7%	0.0%	0.4636 ± 0.0036 (± 0.79%)	13.4%	1.6%	0.4713 ± 0.0030 (± 0.64%)	15.3%	3.3%	0.4718 ± 0.0021 (± 0.44%)	15.5%	3.4%
	Class 6	0.1231 ± 0.0025 (± 2.04%)	0.0%	-20.9%	0.1556 ± 0.0021 (± 1.34%)	26.4%	0.0%	0.1593 ± 0.0052 (± 3.28%)	29.5%	2.4%	0.1577 ± 0.0097 (± 6.17%)	28.2%	1.4%	0.1599 ± 0.0039 (± 2.41%)	29.9%	2.8%
	Class 7	0.4593 ± 0.0039 (± 0.84%)	0.0%	-14.4%	0.5363 ± 0.0102 (± 1.89%)	16.8%	0.0%	0.5367 ± 0.0031 (± 0.57%)	16.9%	0.1%	0.5348 ± 0.0024 (± 0.45%)	16.4%	-0.3%	0.5375 ± 0.0058 (± 1.08%)	17.0%	0.2%
	Average Classes 1 to 7	**0.4359 ± 0.0026 (± 0.60%)**	**0.0%**	**-10.5%**	**0.4872 ± 0.0034 (± 0.70%)**	**11.8%**	**0.0%**	**0.4905 ± 0.0017 (± 0.35%)**	**12.5%**	**0.7%**	**0.4905 ± 0.0033 (± 0.66%)**	**12.5%**	**0.7%**	**0.4916 ± 0.0025 (± 0.51%)**	**12.8%**	**0.9%**
	Others	0.2163 ± 0.0016 (± 0.74%)	0.0%	-13.1%	0.2489 ± 0.0025 (± 1.01%)	15.1%	0.0%	0.2486 ± 0.0022 (± 0.90%)	14.9%	-0.2%	0.2493 ± 0.0020 (± 0.79%)	15.3%	0.1%	0.2523 ± 0.0073 (± 2.90%)	16.7%	1.4%
AUC-ROC	Class 1	0.8635 ± 0.0016 (± 0.19%)	0.0%	-3.4%	0.8941 ± 0.0019 (± 0.22%)	3.5%	0.0%	0.8949 ± 0.0011 (± 0.13%)	3.6%	0.1%	0.8961 ± 0.0017 (± 0.19%)	3.8%	0.2%	0.8954 ± 0.0010 (± 0.12%)	3.7%	0.1%
	Class 2	0.8365 ± 0.0004 (± 0.05%)	0.0%	-1.5%	0.8496 ± 0.0022 (± 0.26%)	1.6%	0.0%	0.8538 ± 0.0022 (± 0.26%)	2.1%	0.5%	0.8545 ± 0.0020 (± 0.23%)	2.2%	0.6%	0.8553 ± 0.0006 (± 0.07%)	2.3%	0.7%
	Class 3	0.7782 ± 0.0015 (± 0.20%)	0.0%	-2.4%	0.7976 ± 0.0019 (± 0.24%)	2.5%	0.0%	0.7989 ± 0.0010 (± 0.13%)	2.7%	0.2%	0.7999 ± 0.0022 (± 0.27%)	2.8%	0.3%	0.8000 ± 0.0011 (± 0.14%)	2.8%	0.3%
	Class 4	0.8140 ± 0.0101 (± 1.25%)	0.0%	-5.2%	0.8585 ± 0.0065 (± 0.76%)	5.5%	0.0%	0.8649 ± 0.0054 (± 0.63%)	6.2%	0.7%	0.8633 ± 0.0027 (± 0.31%)	6.1%	0.6%	0.8663 ± 0.0055 (± 0.64%)	6.4%	0.9%
	Class 5	0.7574 ± 0.0029 (± 0.38%)	0.0%	-4.0%	0.7893 ± 0.0039 (± 0.49%)	4.2%	0.0%	0.7935 ± 0.0022 (± 0.28%)	4.8%	0.5%	0.7939 ± 0.0023 (± 0.29%)	4.8%	0.6%	0.7952 ± 0.0006 (± 0.08%)	5.0%	0.7%
	Class 6	0.7674 ± 0.0018 (± 0.24%)	0.0%	-4.9%	0.8069 ± 0.0035 (± 0.43%)	5.2%	0.0%	0.8033 ± 0.0033 (± 0.41%)	4.7%	-0.5%	0.8046 ± 0.0074 (± 0.93%)	4.9%	-0.3%	0.8079 ± 0.0022 (± 0.27%)	5.3%	0.1%
	Class 7	0.8624 ± 0.0005 (± 0.06%)	0.0%	-2.0%	0.8797 ± 0.0029 (± 0.32%)	2.0%	0.0%	0.8812 ± 0.0013 (± 0.15%)	2.2%	0.2%	0.8826 ± 0.0013 (± 0.15%)	2.3%	0.3%	0.8840 ± 0.0018 (± 0.20%)	2.5%	0.5%
	Average Classes 1 to 7	**0.8113 ± 0.0022 (± 0.28%)**	**0.0%**	**-3.3%**	**0.8394 ± 0.0023 (± 0.27%)**	**3.5%**	**0.0%**	**0.8415 ± 0.0020 (± 0.24%)**	**3.7%**	**0.3%**	**0.8421 ± 0.0025 (± 0.29%)**	**3.8%**	**0.3%**	**0.8434 ± 0.0006 (± 0.08%)**	**4.0%**	**0.5%**
	Others	0.7136 ± 0.0022 (± 0.30%)	0.0%	-5.1%	0.7523 ± 0.0040 (± 0.54%)	5.4%	0.0%	0.7573 ± 0.0014 (± 0.18%)	6.1%	0.7%	0.7609 ± 0.0027 (± 0.36%)	6.6%	1.1%	0.7593 ± 0.0034 (± 0.45%)	6.4%	0.9%
Maximum Youden's J Statistic	Class 1	0.5689 ± 0.0036 (± 0.63%)	0.0%	-10.6%	0.6366 ± 0.0062 (± 0.97%)	11.9%	0.0%	0.6342 ± 0.0038 (± 0.60%)	11.5%	-0.4%	0.6333 ± 0.0043 (± 0.69%)	11.3%	-0.5%	0.6346 ± 0.0030 (± 0.48%)	11.6%	-0.3%
	Class 2	0.5324 ± 0.0024 (± 0.46%)	0.0%	-5.1%	0.5613 ± 0.0057 (± 1.01%)	5.4%	0.0%	0.5689 ± 0.0047 (± 0.82%)	6.8%	1.4%	0.5712 ± 0.0034 (± 0.60%)	7.3%	1.8%	0.5711 ± 0.0019 (± 0.33%)	7.3%	1.8%
	Class 3	0.4157 ± 0.0041 (± 0.98%)	0.0%	-7.0%	0.4469 ± 0.0036 (± 0.80%)	7.5%	0.0%	0.4433 ± 0.0042 (± 0.95%)	6.6%	-0.8%	0.4485 ± 0.0059 (± 1.32%)	7.9%	0.4%	0.4507 ± 0.0028 (± 0.63%)	8.4%	0.9%
	Class 4	0.4733 ± 0.0116 (± 2.45%)	0.0%	-15.0%	0.5570 ± 0.0159 (± 2.85%)	17.7%	0.0%	0.5772 ± 0.0135 (± 2.33%)	22.0%	3.6%	0.5765 ± 0.0100 (± 1.74%)	21.8%	3.5%	0.5767 ± 0.0130 (± 2.26%)	21.9%	3.5%
	Class 5	0.3978 ± 0.0056 (± 1.41%)	0.0%	-10.8%	0.4462 ± 0.0066 (± 1.49%)	12.2%	0.0%	0.4508 ± 0.0080 (± 1.78%)	13.3%	1.0%	0.4533 ± 0.0073 (± 1.61%)	14.0%	1.6%	0.4580 ± 0.0041 (± 0.89%)	15.1%	2.6%
	Class 6	0.4264 ± 0.0059 (± 1.39%)	0.0%	-12.5%	0.4875 ± 0.0039 (± 0.81%)	14.3%	0.0%	0.4839 ± 0.0119 (± 2.46%)	13.5%	-0.7%	0.4832 ± 0.0149 (± 3.08%)	13.3%	-0.9%	0.4904 ± 0.0026 (± 0.53%)	15.0%	0.6%
	Class 7	0.5743 ± 0.0026 (± 0.44%)	0.0%	-4.3%	0.6003 ± 0.0036 (± 0.60%)	4.5%	0.0%	0.6050 ± 0.0024 (± 0.39%)	5.3%	0.8%	0.6090 ± 0.0045 (± 0.74%)	6.0%	1.4%	0.6125 ± 0.0049 (± 0.80%)	6.7%	2.0%
	Average Classes 1 to 7	**0.4841 ± 0.0036 (± 0.74%)**	**0.0%**	**-9.3%**	**0.5337 ± 0.0043 (± 0.80%)**	**10.2%**	**0.0%**	**0.5376 ± 0.0041 (± 0.77%)**	**11.1%**	**0.7%**	**0.5393 ± 0.0058 (± 1.08%)**	**11.4%**	**1.1%**	**0.5420 ± 0.0009 (± 0.17%)**	**12.0%**	**1.6%**
	Others	0.3282 ± 0.0047 (± 1.44%)	0.0%	-16.4%	0.3927 ± 0.0063 (± 1.59%)	19.6%	0.0%	0.4003 ± 0.0041 (± 1.03%)	21.9%	1.9%	0.4062 ± 0.0074 (± 1.83%)	23.8%	3.4%	0.4047 ± 0.0058 (± 1.43%)	23.3%	3.1%
Youden's J Statistic at 20% FPR	Class 1	0.5593 ± 0.0076 (± 1.37%)	0.0%	-11.7%	0.6334 ± 0.0066 (± 1.04%)	13.3%	0.0%	0.6322 ± 0.0049 (± 0.78%)	13.0%	-0.2%	0.6304 ± 0.0058 (± 0.92%)	12.7%	-0.5%	0.6324 ± 0.0039 (± 0.62%)	13.1%	-0.2%
	Class 2	0.5268 ± 0.0031 (± 0.60%)	0.0%	-5.5%	0.5578 ± 0.0064 (± 1.14%)	5.9%	0.0%	0.5645 ± 0.0052 (± 0.93%)	7.1%	1.2%	0.5669 ± 0.0064 (± 1.13%)	7.6%	1.6%	0.5670 ± 0.0027 (± 0.48%)	7.6%	1.7%
	Class 3	0.3675 ± 0.0043 (± 1.17%)	0.0%	-10.5%	0.4105 ± 0.0094 (± 2.28%)	11.7%	0.0%	0.4165 ± 0.0053 (± 1.26%)	13.3%	1.5%	0.4117 ± 0.0061 (± 1.47%)	12.0%	0.3%	0.4147 ± 0.0067 (± 1.62%)	12.8%	1.0%
	Class 4	0.4428 ± 0.0205 (± 4.63%)	0.0%	-18.5%	0.5430 ± 0.0226 (± 4.15%)	22.6%	0.0%	0.5695 ± 0.0121 (± 2.13%)	28.6%	4.9%	0.5683 ± 0.0140 (± 2.46%)	28.3%	4.7%	0.5715 ± 0.0134 (± 2.34%)	29.1%	5.2%
	Class 5	0.3241 ± 0.0106 (± 3.28%)	0.0%	-17.3%	0.3919 ± 0.0079 (± 2.01%)	20.9%	0.0%	0.4022 ± 0.0047 (± 1.16%)	24.1%	2.6%	0.3910 ± 0.0075 (± 1.93%)	20.6%	-0.2%	0.4008 ± 0.0027 (± 0.66%)	23.7%	2.3%
	Class 6	0.3317 ± 0.0073 (± 2.19%)	0.0%	-21.5%	0.4223 ± 0.0180 (± 4.26%)	27.3%	0.0%	0.4217 ± 0.0104 (± 2.46%)	27.2%	-0.1%	0.4195 ± 0.0178 (± 4.25%)	26.5%	-0.7%	0.4239 ± 0.0094 (± 2.23%)	27.8%	0.4%
	Class 7	0.5616 ± 0.0059 (± 1.05%)	0.0%	-4.9%	0.5905 ± 0.0075 (± 1.27%)	5.2%	0.0%	0.5976 ± 0.0019 (± 0.32%)	6.4%	1.2%	0.6027 ± 0.0031 (± 0.51%)	7.3%	2.1%	0.6060 ± 0.0036 (± 0.59%)	7.9%	2.6%
	Average Classes 1 to 7	**0.4448 ± 0.0043 (± 0.97%)**	**0.0%**	**-12.3%**	**0.5071 ± 0.0060 (± 1.19%)**	**14.0%**	**0.0%**	**0.5149 ± 0.0033 (± 0.65%)**	**15.8%**	**1.5%**	**0.5129 ± 0.0059 (± 1.15%)**	**15.3%**	**1.2%**	**0.5166 ± 0.0031 (± 0.61%)**	**16.1%**	**1.9%**
	Others	0.2884 ± 0.0027 (± 0.94%)	0.0%	-18.4%	0.3532 ± 0.0054 (± 1.54%)	22.5%	0.0%	0.3655 ± 0.0049 (± 1.34%)	26.8%	3.5%	0.3725 ± 0.0076 (± 2.04%)	29.2%	5.5%	0.3700 ± 0.0084 (± 2.27%)	28.3%	4.8%

Table 14. Absolute and Relative Performance of Selected Models Across Most Significant Classes for the MIMIC Dataset.
This table provides detailed performance results across four metrics—AUC-PR, AUC-ROC, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR—for the five selected models tested on the MIMIC dataset. It includes data for the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes in the disease labels list. Results are shown as average outcomes (Avg) with standard deviations (SD and %SD = SD/Avg × 100%), derived from testing five distinct instances of the same model type, each developed and trained using a different seed. Performance improvements are quantified relative to CXR-FM and the CXR-Model FFT baseline, calculated using the formula (Model Avg Value − Baseline Avg Value)/Baseline Avg Value × 100%, where the baseline is either CXR-FM (1) or CXR-Model FFT (2). This analysis highlights the enhancements achieved by the CXR-FMKD-Direct FFT student models in terms of AUC-PR, AUC-ROC, Maximum Youden's J Statistic, and Youden's J Statistic at 20% FPR, underscoring their relative and absolute performance gains.

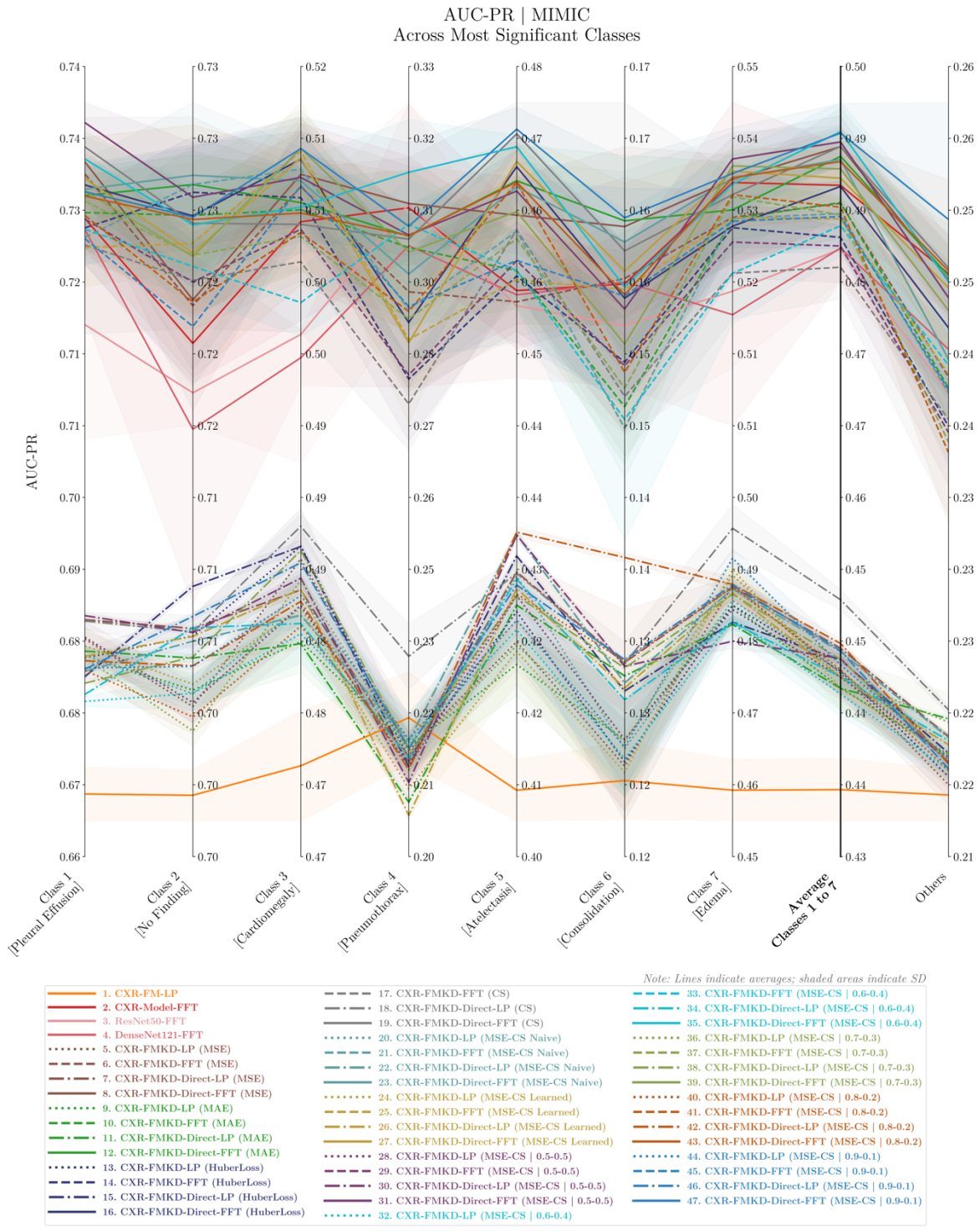

Figure 65. Parallel Coordinate Plot of AUC-PR Performance Across the Most Significant Classes for 47 Models Tested on the MIMIC Dataset.

This custom parallel coordinate plot visualises the AUC-PR metric across 47 models tested on the MIMIC dataset. It displays performance results for the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes in the disease labels list. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. We observe an overall stratification in model performance, grouped by type from top (best performers) to bottom (least performers): CXR-FMKD-Direct FFT models, CXR-FMKD FFT models, Benchmark Models (i.e., CXR-Model FFT, ResNet50 FFT, DenseNet121 FFT), CXR-FMKD-Direct LP models, CXR-FMKD LP models, and lastly the original CXR-FM. Generally, the models exhibit the best performance for Class 1 [Pleural Effusion] and Class 2 [No Finding]; and the poorest for Class 6 [Consolidation].

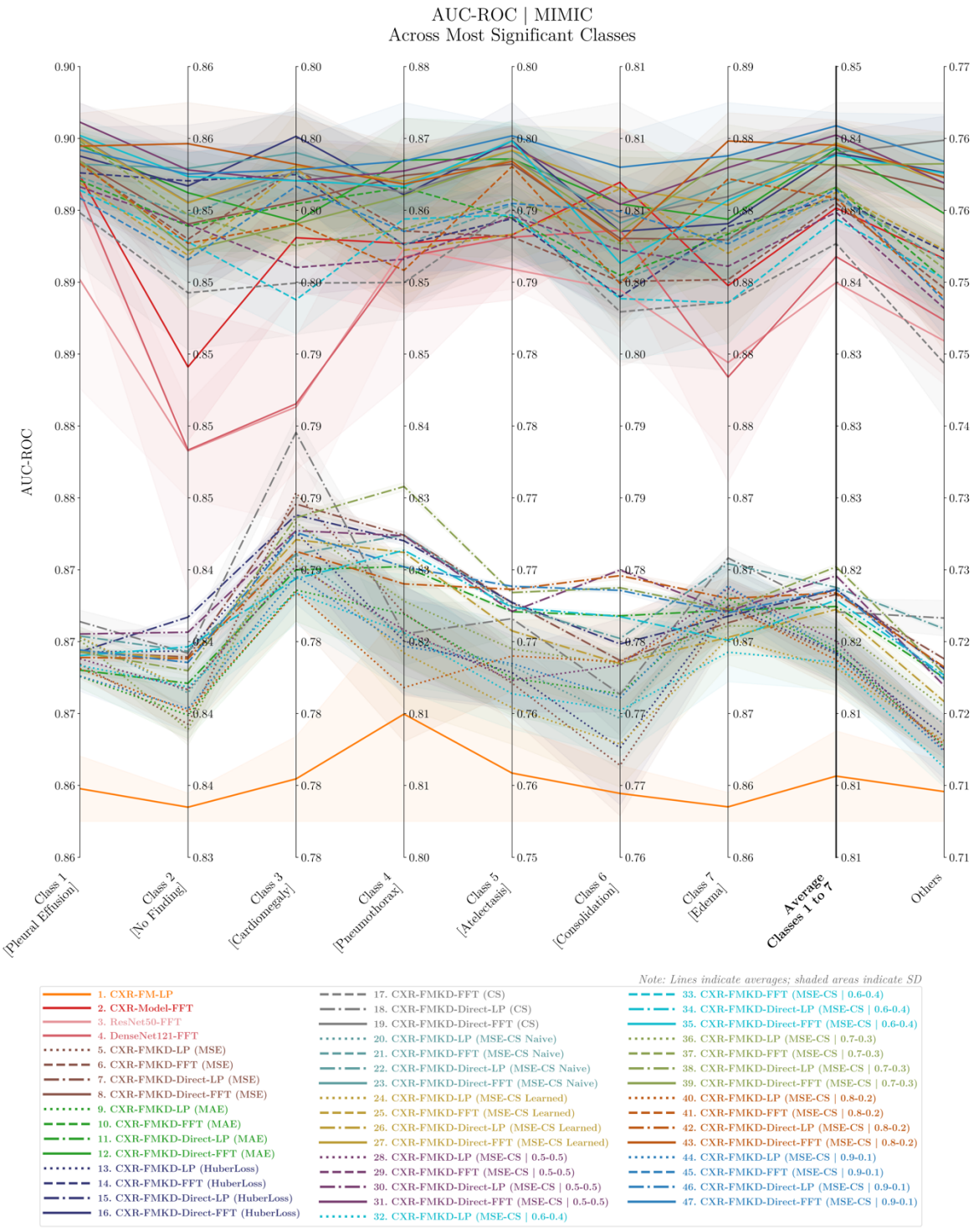

Figure 66. Parallel Coordinate Plot of AUC-ROC Performance Across the Most Significant Classes for 47 Models Tested on the MIMIC Dataset.

This custom parallel coordinate plot visualises the AUC-ROC metric across 47 models tested on the MIMIC dataset. It displays performance results for the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes in the disease labels list. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. We observe an overall stratification in model performance, grouped by type from top (best performers) to bottom (least performers): CXR-FMKD-Direct FFT models, CXR-FMKD FFT models, Benchmark Models (i.e., CXR-Model FFT, ResNet50 FFT, DenseNet121 FFT), CXR-FMKD-Direct LP models, CXR-FMKD LP models, and lastly the original CXR-FM. Generally, the models exhibit the best performance for Class 1 [Pleural Effusion], Class 4 [Pneumothorax], and Class 7 [Edema]; and the poorest for 'Others'.

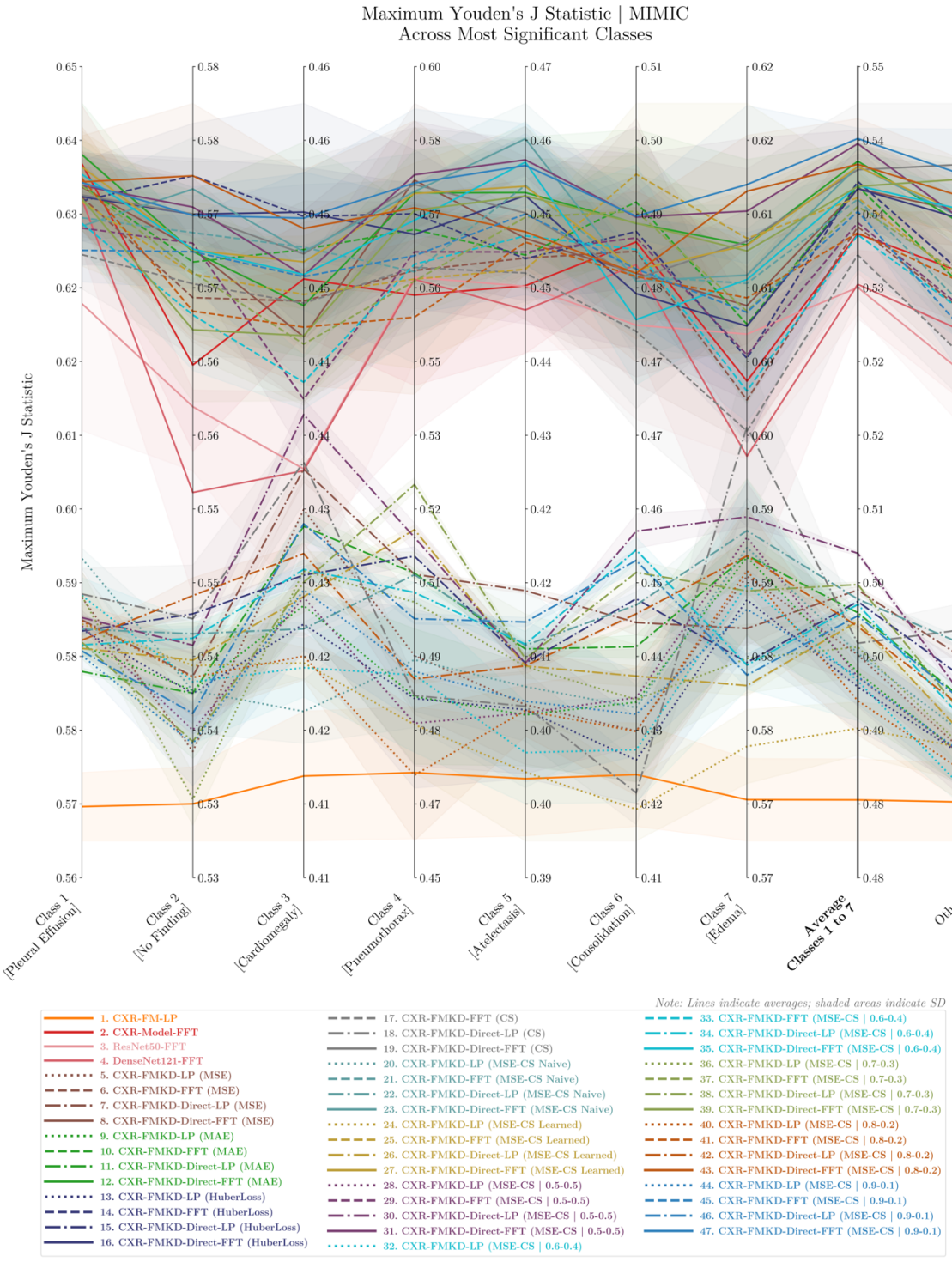

Figure 67. Parallel Coordinate Plot of Maximum Youden's J Statistic Performance Across the Most Significant Classes for 47 Models Tested on the MIMIC Dataset.

This custom parallel coordinate plot visualises the Maximum Youden's J Statistic metric across 47 models tested on the MIMIC dataset. It displays performance results for the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes in the disease labels list. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. We observe an overall stratification in model performance, grouped by type from top (best performers) to bottom (least performers): CXR-FMKD-Direct FFT models, CXR-FMKD FFT models, Benchmark Models (i.e., CXR-Model FFT, ResNet50 FFT, DenseNet121 FFT), CXR-FMKD-Direct LP models, CXR-FMKD LP models, and lastly the original CXR-FM. Generally, the models exhibit the best performance for Class 1 [Pleural Effusion] and the poorest for 'Others'.

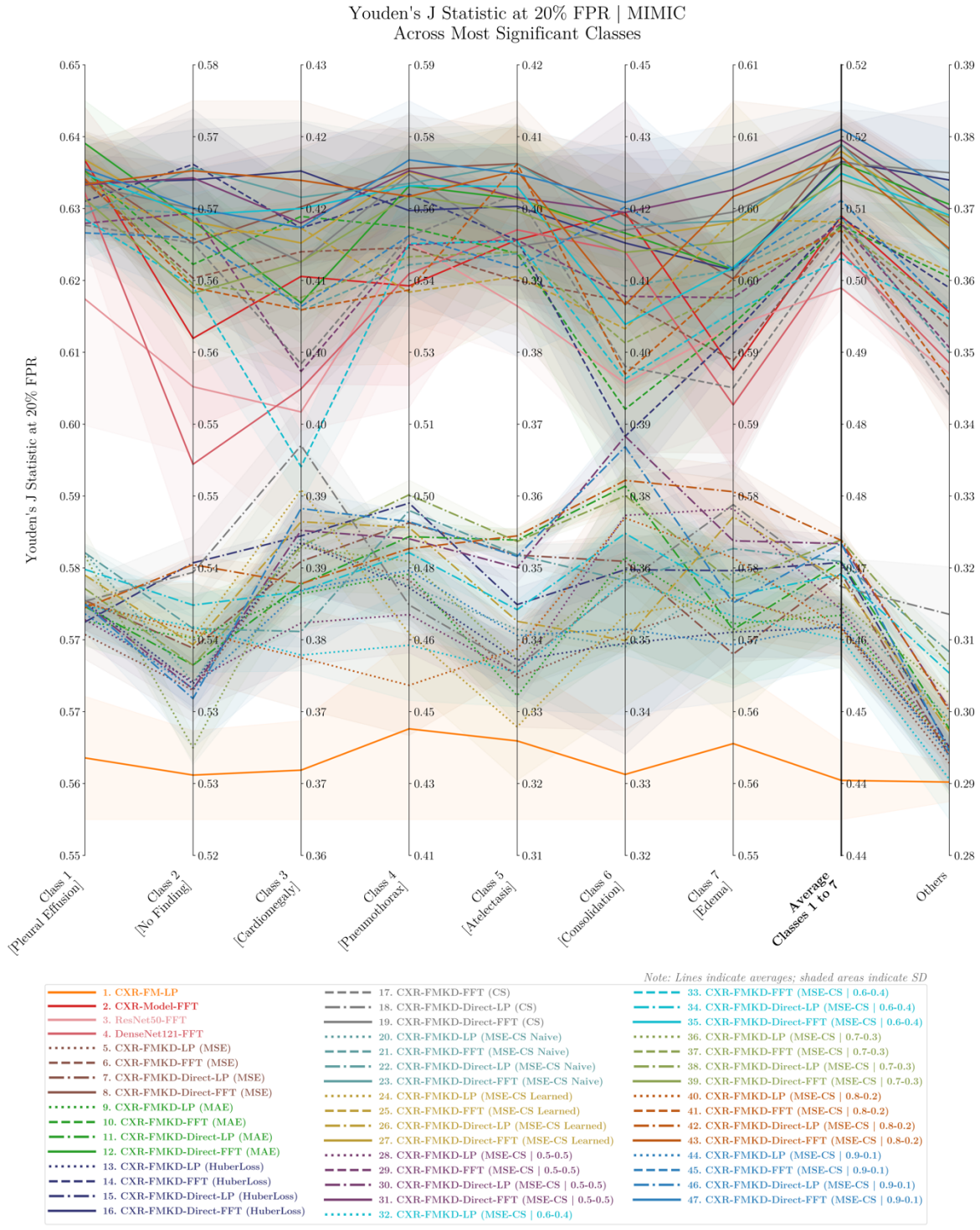

Figure 68. Parallel Coordinate Plot of Youden's J Statistic at 20% FPR Performance Across the Most Significant Classes for 47 Models Tested on the MIMIC Dataset.

This custom parallel coordinate plot visualises the Youden's J Statistic at 20% FPR metric across 47 models tested on the MIMIC dataset. It displays performance results for the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes in the disease labels list. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. We observe an overall stratification in model performance, grouped by type from top (best performers) to bottom (least performers): CXR-FMKD-Direct FFT models, CXR-FMKD FFT models, Benchmark Models (i.e., CXR-Model FFT, ResNet50 FFT, DenseNet121 FFT), CXR-FMKD-Direct LP models, CXR-FMKD LP models, and lastly the original CXR-FM. Generally, the models exhibit the best performance for Class 1 [Pleural Effusion] and the poorest for 'Others'.

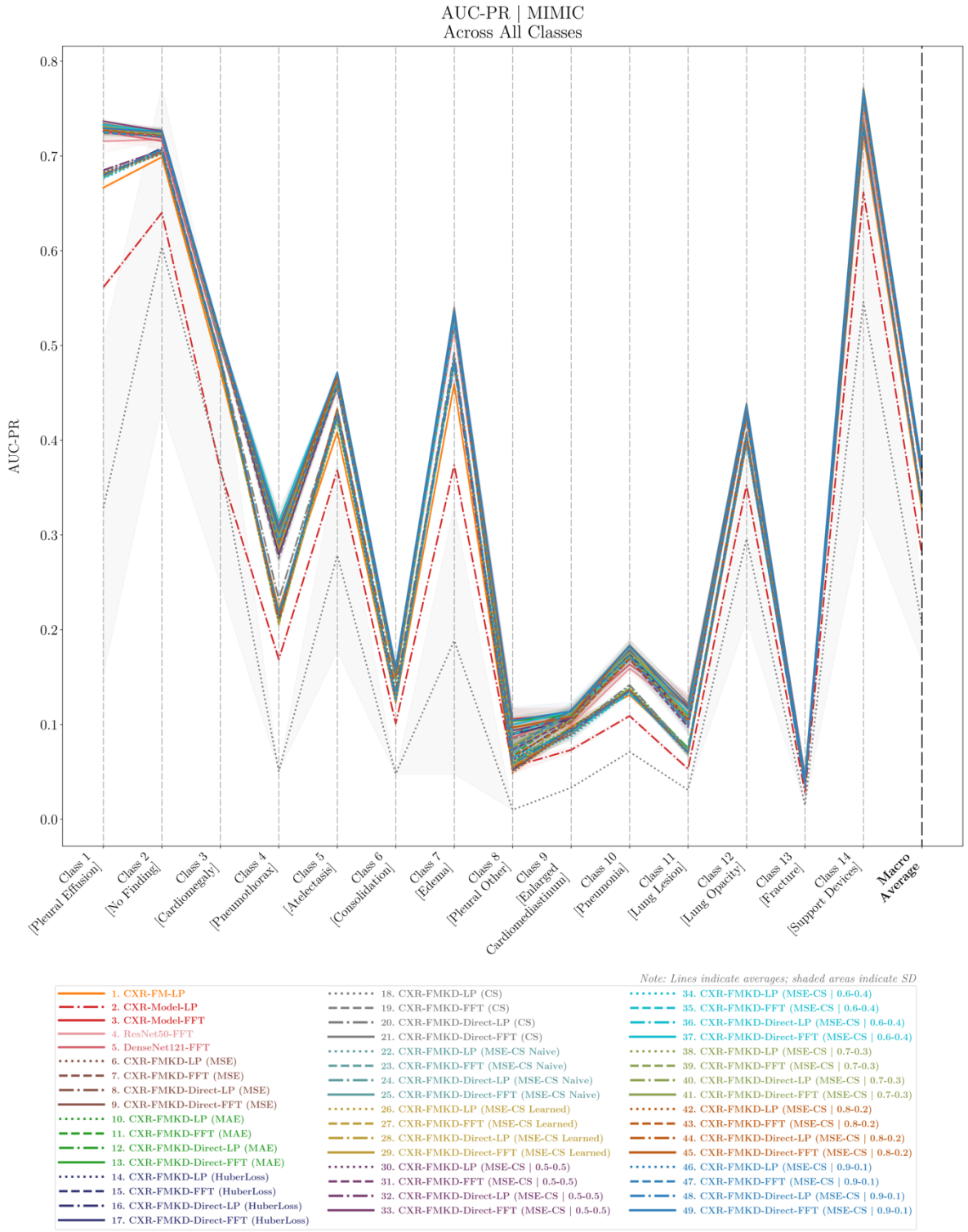

Figure 69. AUC-PR Performance Across All 14 Classes for 49 Models Tested on the MIMIC Dataset. This plot visualises the AUC-PR metric across 49 models tested on the MIMIC dataset. It displays performance results for each of the 14 disease classes as well as their macro-average. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. Generally, the models exhibit the best performance for Class 14 [Support Devices], Class 1 [Pleural Effusion], and Class 2 [No Finding]; and the poorest for Class 13 [Fracture].

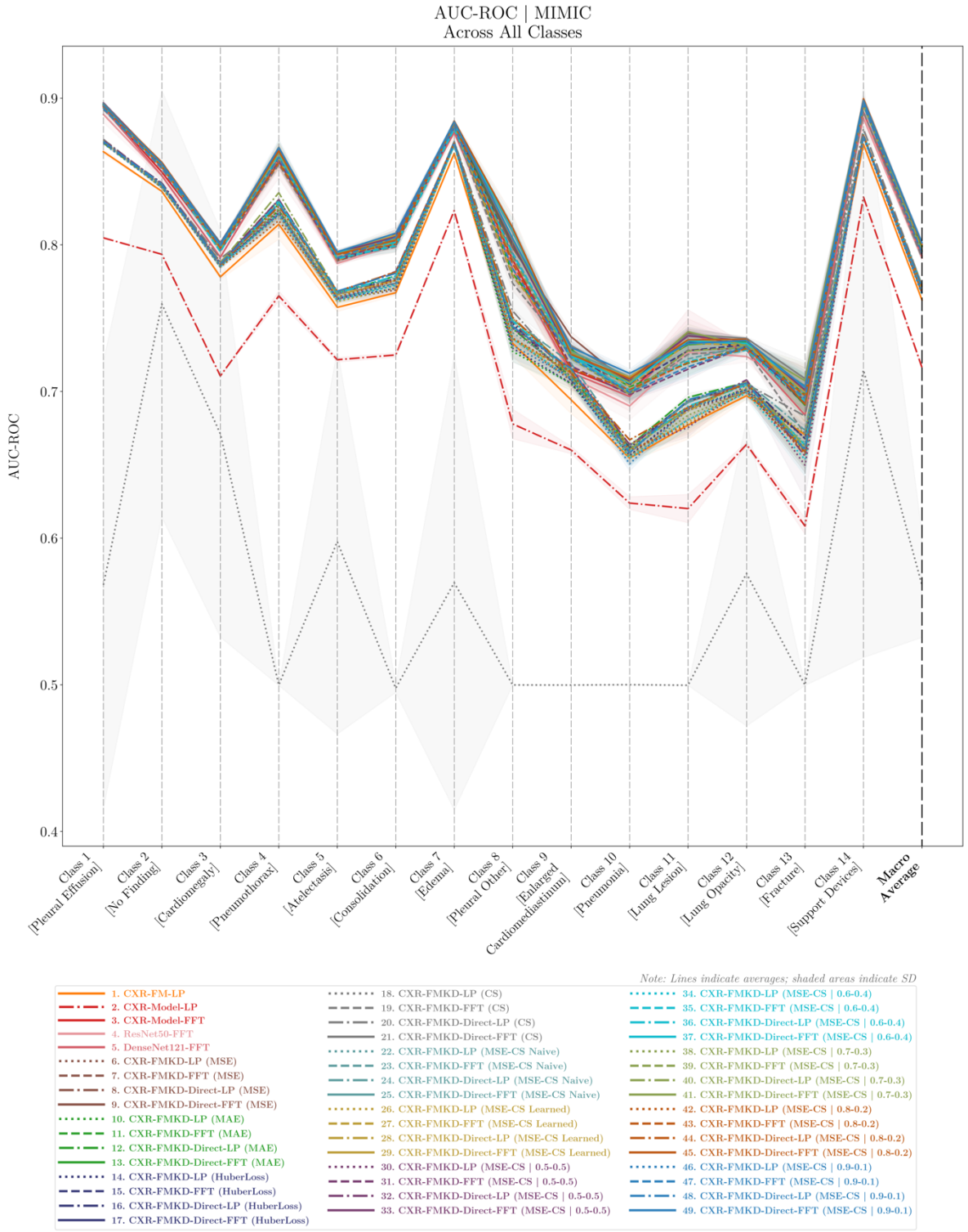

Figure 70. AUC-ROC Performance Across All 14 Classes for 49 Models Tested on the MIMIC Dataset. This plot visualises the AUC-ROC metric across 49 models tested on the MIMIC dataset. It displays performance results for each of the 14 disease classes as well as their macro-average. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. Generally, the models exhibit the best performance for Class 14 [Support Devices], Class 1 [Pleural Effusion], and Class 7 [Edema]; and the poorest for Class 10 [Pneumonia] and Class 13 [Fracture].

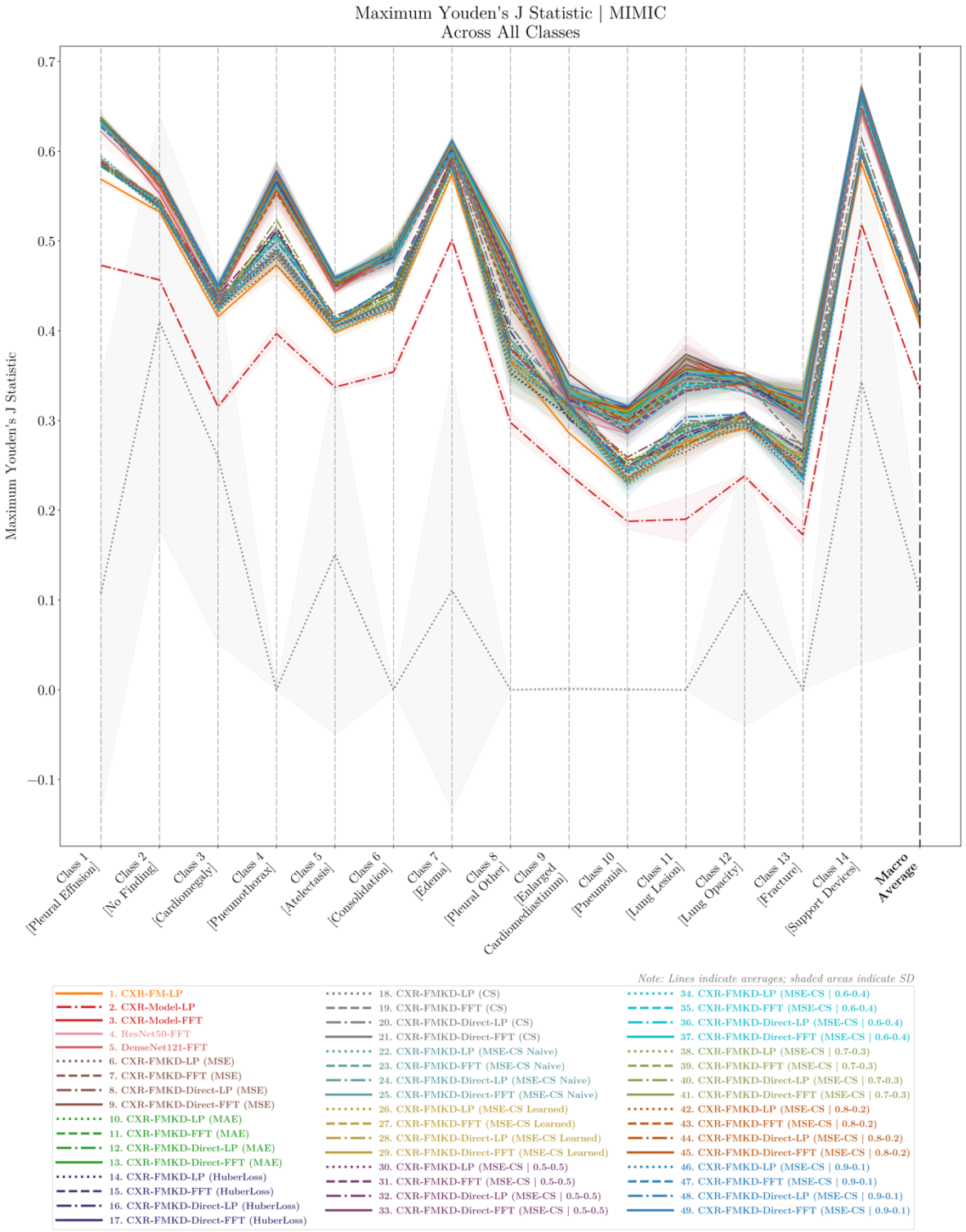

Figure 71. Maximum Youden's J Statistic Performance Across All 14 Classes for 49 Models Tested on the MIMIC Dataset.

This plot visualises the Maximum Youden's J Statistic metric across 49 models tested on the MIMIC dataset. It displays performance results for each of the 14 disease classes as well as their macro-average. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. Generally, the models exhibit the best performance for Class 14 [Support Devices], Class 1 [Pleural Effusion], and Class 7 [Edema]; and the poorest for Class 10 [Pneumonia] and Class 13 [Fracture].

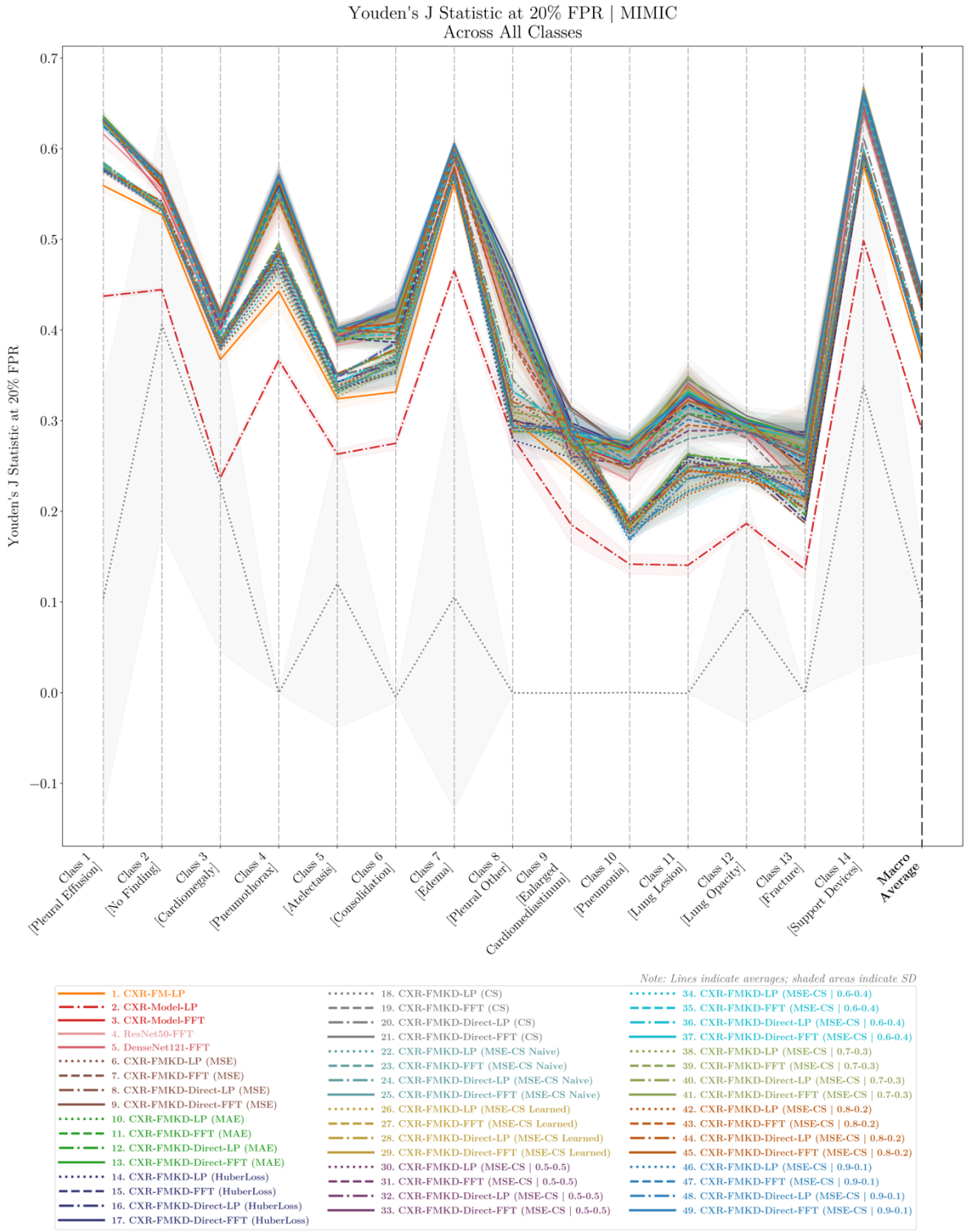

Figure 72. Youden's J Statistic at 20% FPR Performance Across All 14 Classes for 49 Models Tested on the MIMIC Dataset.

This plot visualises the Youden's J Statistic at 20% FPR metric across 49 models tested on the MIMIC dataset. It displays performance results for each of the 14 disease classes as well as their macro-average. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. Generally, the models exhibit the best performance for Class 14 [Support Devices], Class 1 [Pleural Effusion], and Class 7 [Edema]; and the poorest for Class 10 [Pneumonia] and Class 13 [Fracture].

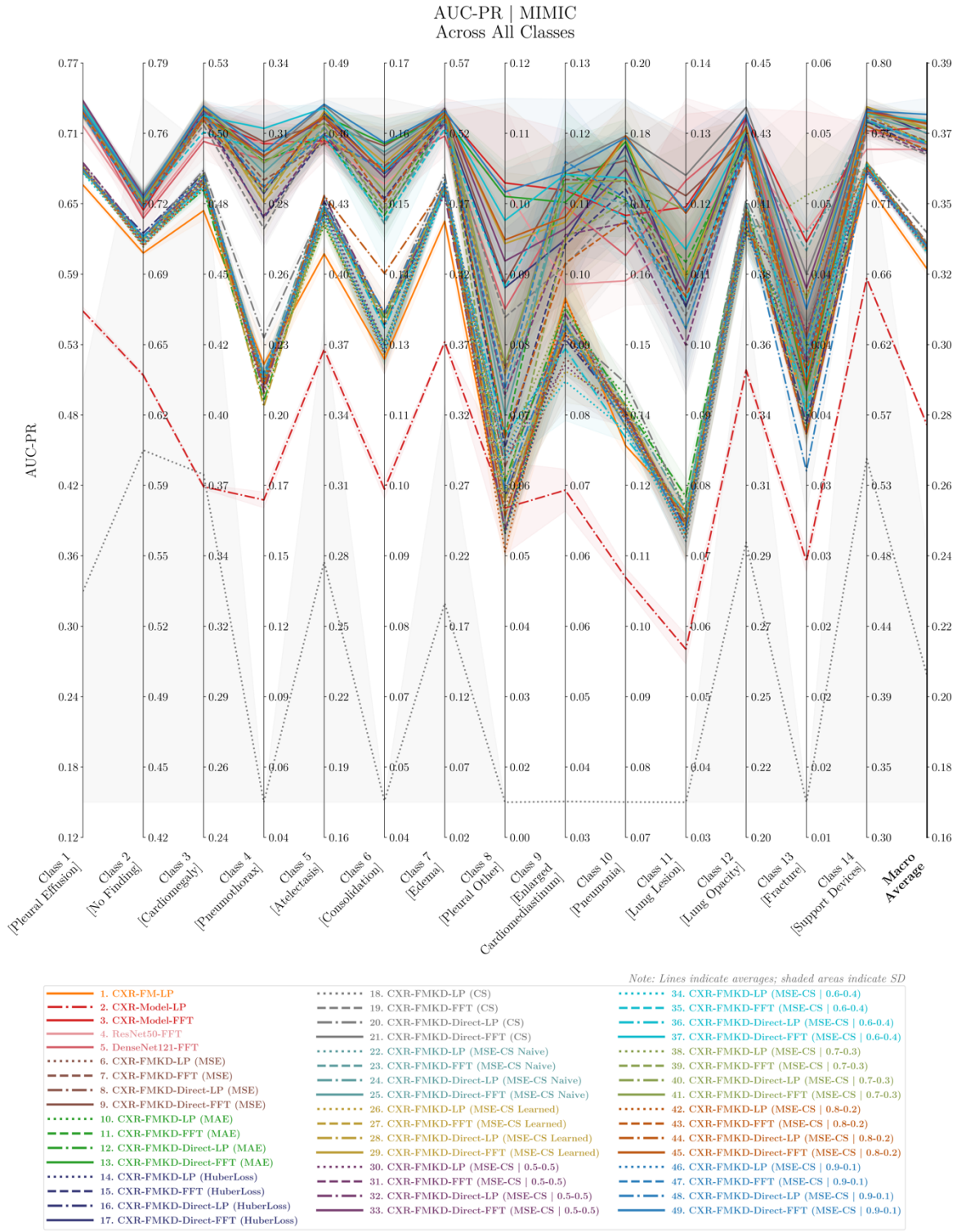

Figure 73. Parallel Coordinate Plot of AUC-PR Performance Across All 14 Classes for 49 Models Tested on the MIMIC Dataset.

This custom parallel coordinate plot visualises the AUC-PR metric across 49 models tested on the MIMIC dataset. It displays performance results for each of the 14 disease classes as well as their macro-average. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. Generally, the models exhibit the best performance for Class 14 [Support Devices], Class 1 [Pleural Effusion], and Class 2 [No Finding]; and the poorest for Class 13 [Fracture].

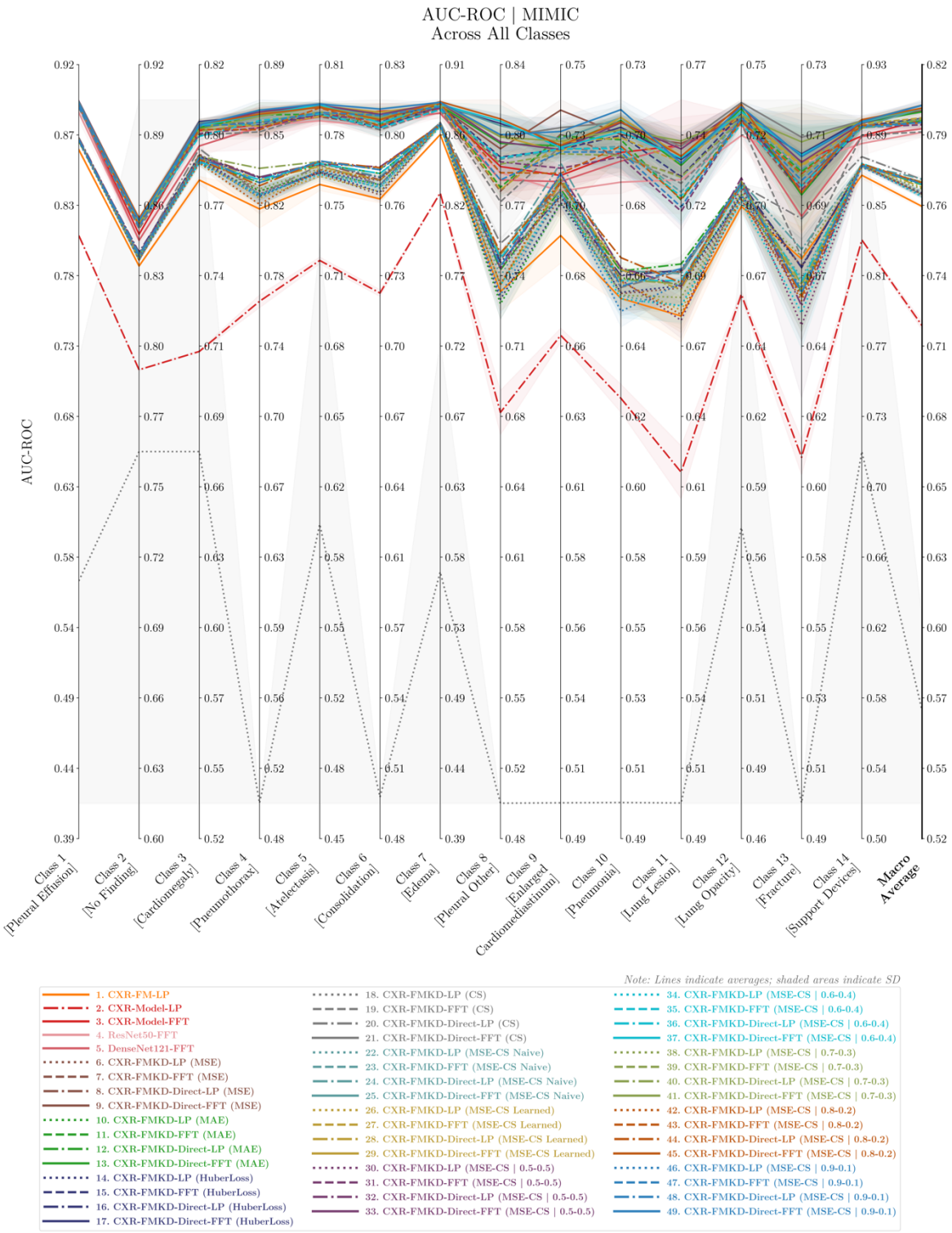

Figure 74. Parallel Coordinate Plot of AUC-ROC Performance Across All 14 Classes for 49 Models Tested on the MIMIC Dataset.

This custom parallel coordinate plot visualises the AUC-ROC metric across 49 models tested on the MIMIC dataset. It displays performance results for each of the 14 disease classes as well as their macro-average. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. Generally, the models exhibit the best performance for Class 14 [Support Devices], Class 1 [Pleural Effusion], and Class 7 [Edema]; and the poorest for Class 10 [Pneumonia] and Class 13 [Fracture].

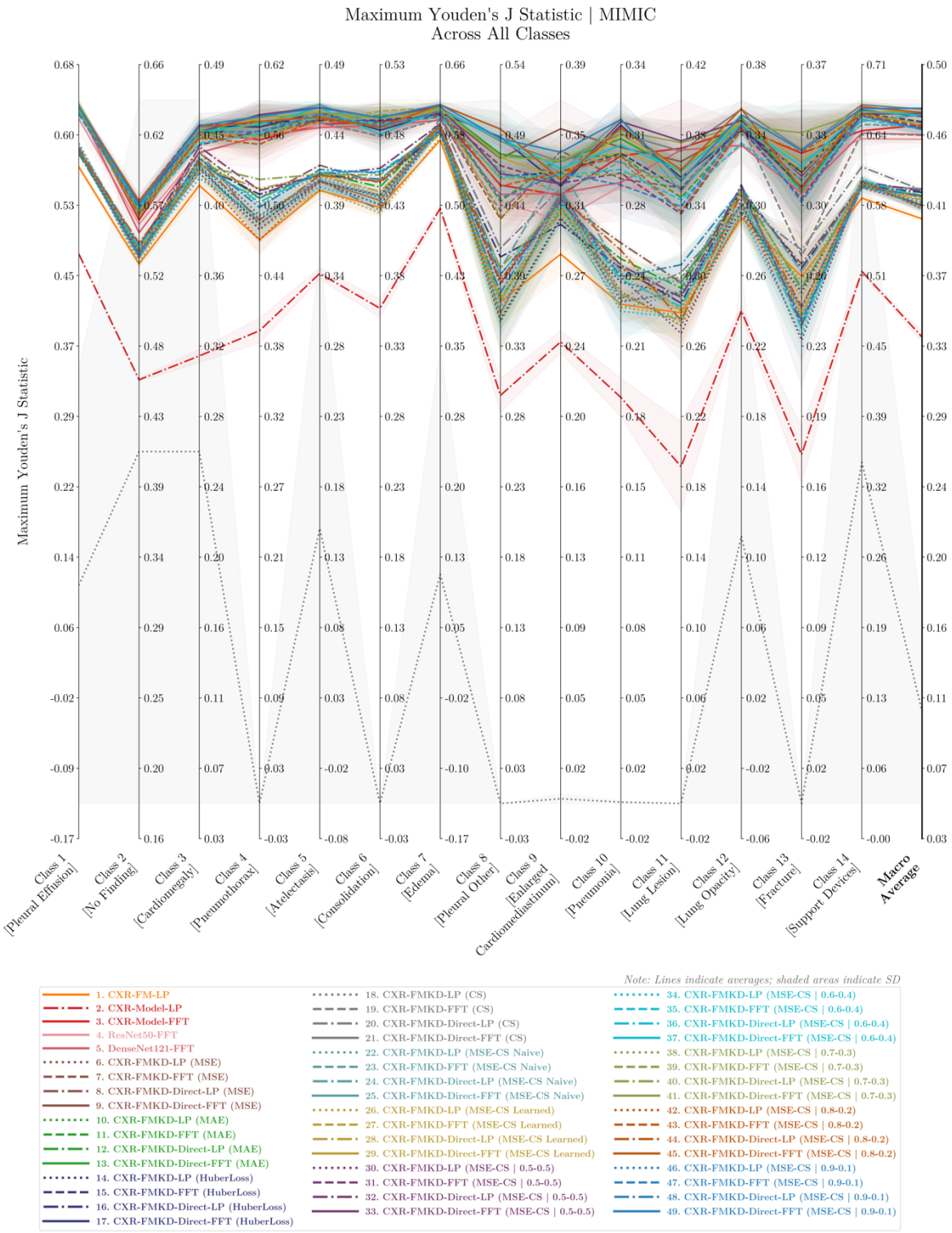

Figure 75. Parallel Coordinate Plot of Maximum Youden's J Statistic Performance Across All 14 Classes for 49 Models Tested on the MIMIC Dataset.

This custom parallel coordinate plot visualises the Maximum Youden's J Statistic metric across 49 models tested on the MIMIC dataset. It displays performance results for each of the 14 disease classes as well as their macro-average. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. Generally, the models exhibit the best performance for Class 14 [Support Devices], Class 1 [Pleural Effusion], and Class 7 [Edema]; and the poorest for Class 10 [Pneumonia] and Class 13 [Fracture].

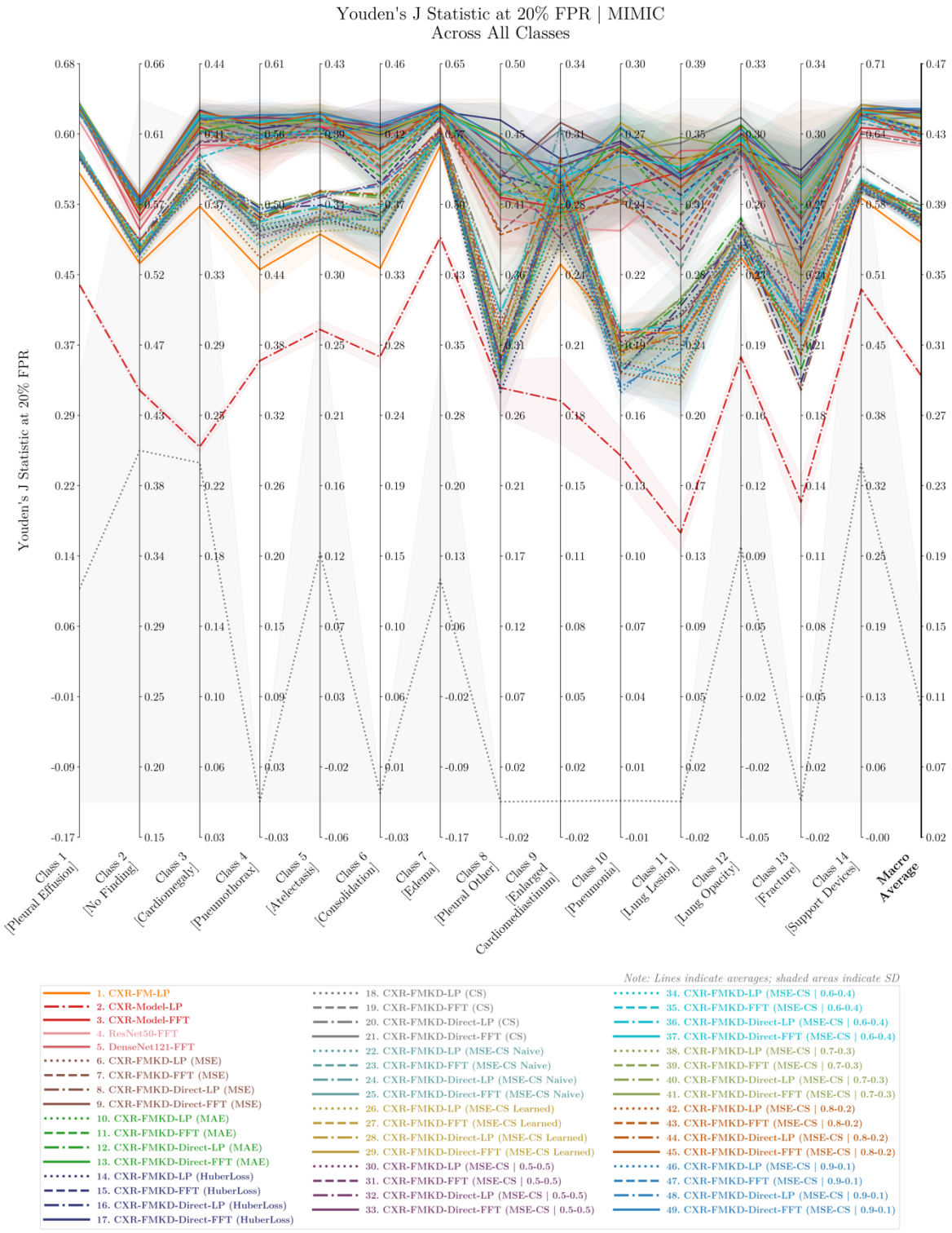

Figure 76. Parallel Coordinate Plot of Youden's J Statistic at 20% FPR Performance Across All 14 Classes for 49 Models Tested on the MIMIC Dataset.

This custom parallel coordinate plot visualises the Youden's J Statistic at 20% FPR metric across 49 models tested on the MIMIC dataset. It displays performance results for each of the 14 disease classes as well as their macro-average. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests. Each CXR-FMKD model line is colour-coded according to the KD loss applied during its development. Additional visual markers denote the CXR-FMKD model types, with LP-variants models featuring dots in their lines, while FFT-variants models do not: dotted lines for LP models, dash-dotted for Direct LP models, dashed for FFT models, and solid for Direct FFT models. Generally, the models exhibit the best performance for Class 14 [Support Devices], Class 1 [Pleural Effusion], and Class 7 [Edema]; and the poorest for Class 10 [Pneumonia] and Class 13 [Fracture].

S.3. Generalisability Analysis – Direct Transfer (DT)

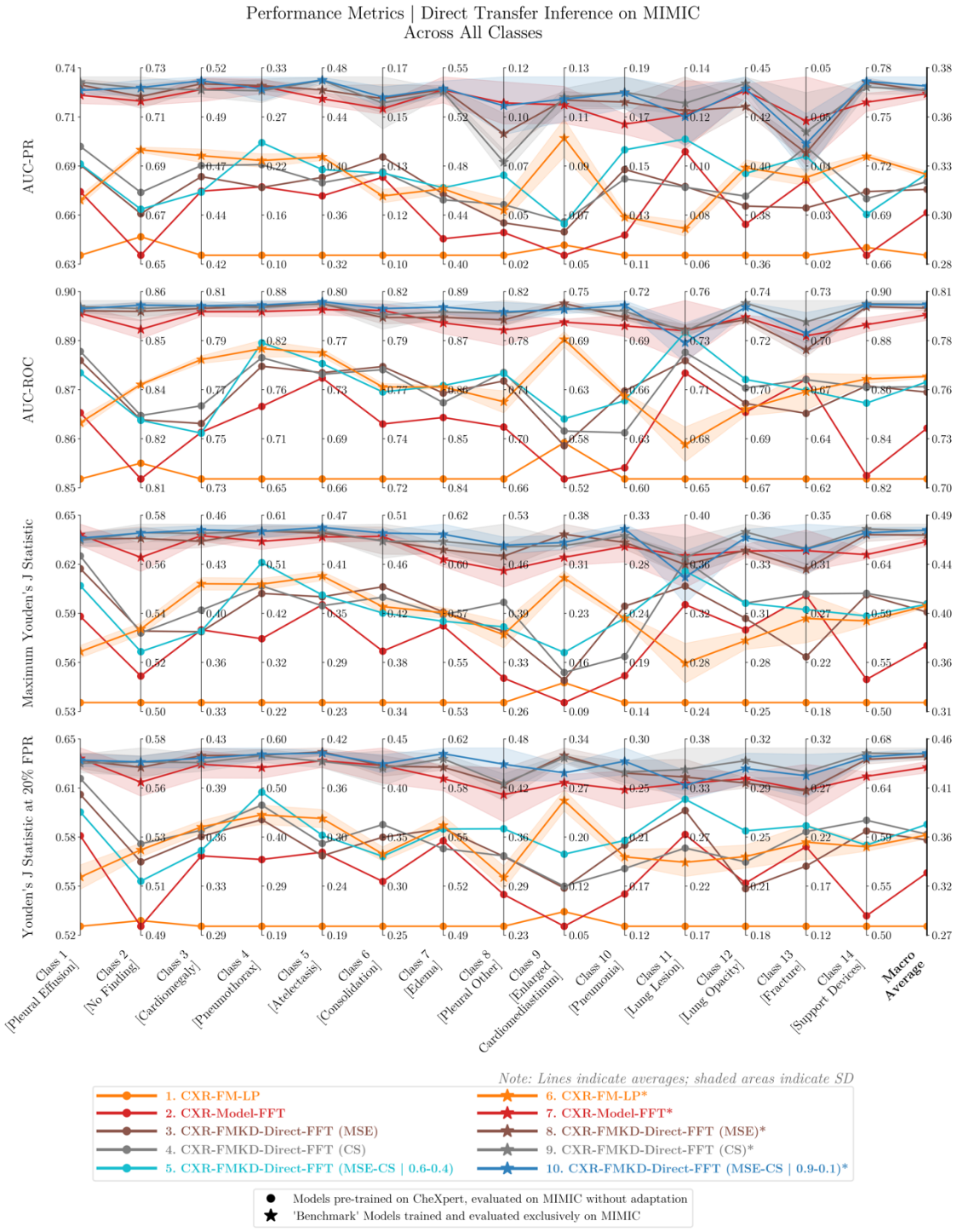

Figure 77. Performance of Selected Transfer Models and Their Benchmarks Across All 14 Classes After Direct Transfer Inference on MIMIC, Post-CheXpert Pre-training.
These custom parallel coordinate plots for our Direct Transfer Inference analysis display the performance across four metrics: AUC-PR, AUC-ROC, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR for the five selected '*transfer*' models tested on MIMIC without adaptation after pre-training on CheXpert. The models include CXR-FM (teacher), CXR-Model FFT (baseline), and the CXR-FMKD-Direct FFT variants of the three selected student model types (MSE, CS, and MSE-CS | α-β). Each *transfer* model, represented by circles, is contrasted against a corresponding *benchmark*, shown with stars, which shares the same architecture but was trained and tested exclusively on MIMIC. The plots display results for each of the 14 disease classes as well as their macro-average. Each line represents the average results from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests.

S.4. Generalisability Analysis – Linear Probing (LP)

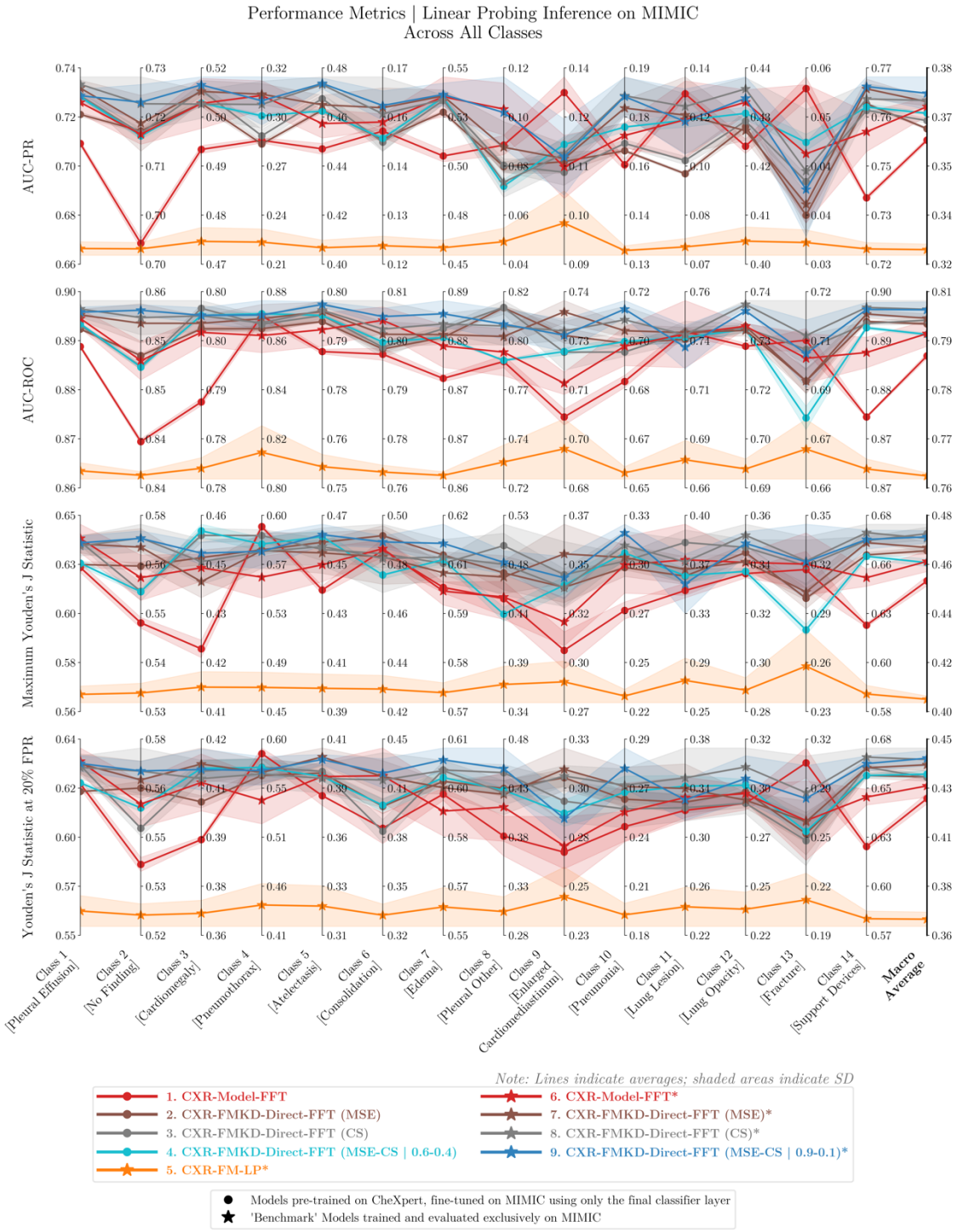

Figure 78. Performance of Selected Transfer Models and Their Benchmarks Across All 14 Classes After Linear Probing Inference on MIMIC, Post-CheXpert Pre-training.
These custom parallel coordinate plots for our Linear Probing Inference analysis display the performance across four metrics: AUC-PR, AUC-ROC, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR for the five selected '*transfer*' models that were pre-trained on CheXpert and then fine-tuned on MIMIC using only the final classification layer. The models include CXR-FM (teacher), CXR-Model FFT (baseline), and the CXR-FMKD-Direct FFT variants of the three selected student model types (MSE, CS, and MSE-CS | α-β). Each *transfer* model, represented by circles, is contrasted against a corresponding *benchmark*, shown with stars, which shares the same architecture but was trained and tested exclusively on MIMIC. For CXR-FM (i.e., CXR-FM-LP), both are equivalent. The plots display results for each of the 14 disease classes as well as their macro-average. Each line represents the average results from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests.

Linear Probing Inference on MIMIC

Metric	Class	CXR-FM Benchmark (B) Avg ± SD	CXR-FM Transfer Model Avg ± SD	CXR-FM Transfer Model %Δ w.r.t. (B)	CXR-Model FFT Benchmark (B) Avg ± SD	CXR-Model FFT Transfer Model Avg ± SD	CXR-Model FFT Transfer Model %Δ w.r.t. (B)	CXR-FMKD LP MSE Benchmark (B) Avg ± SD	CXR-FMKD LP MSE Transfer Model Avg ± SD	CXR-FMKD LP MSE Transfer Model %Δ w.r.t. (B)	CXR-FMKD LP CS Benchmark (B) Avg ± SD	CXR-FMKD LP CS Transfer Model Avg ± SD	CXR-FMKD LP CS Transfer Model %Δ w.r.t. (B)	CXR-FMKD LP MSE-CS | α-β Benchmark (B) Avg ± SD	CXR-FMKD LP MSE-CS | α-β Transfer Model Avg ± SD	CXR-FMKD LP MSE-CS | α-β Transfer Model %Δ w.r.t. (B)
AUC-PR	Average Classes 1 to 7	0.4359 ± 0.0026	0.4359 ± 0.0026	0.0%	0.4872 ± 0.0034	0.4705 ± 0.0003	-3.4%	0.4457 ± 0.0019	0.4474 ± 0.0003	0.4%	0.2673 ± 0.0671	0.4531 ± 0.0008	69.5%	0.4465 ± 0.0017	0.4518 ± 0.0003	1.2%
AUC-ROC	Average Classes 1 to 7	0.8113 ± 0.0022	0.8113 ± 0.0022	0.0%	0.8394 ± 0.0023	0.8343 ± 0.0005	-0.6%	0.8174 ± 0.0015	0.8206 ± 0.0001	0.4%	0.5948 ± 0.0655	0.8210 ± 0.0007	38.0%	0.8174 ± 0.0014	0.8212 ± 0.0001	0.5%
Maximum Youden's J Statistic	Average Classes 1 to 7	0.4841 ± 0.0036	0.4841 ± 0.0036	0.0%	0.5337 ± 0.0043	0.5314 ± 0.0011	-0.4%	0.4968 ± 0.0025	0.5033 ± 0.0003	1.3%	0.1484 ± 0.1022	0.5042 ± 0.0019	239.8%	0.4949 ± 0.0024	0.5028 ± 0.0004	1.6%
Youden's J Statistic at 20% FPR	Average Classes 1 to 7	0.4448 ± 0.0043	0.4448 ± 0.0043	0.0%	0.5071 ± 0.0060	0.5012 ± 0.0017	-1.2%	0.4616 ± 0.0029	0.4649 ± 0.0007	0.7%	0.1370 ± 0.0977	0.4678 ± 0.0048	241.5%	0.4621 ± 0.0029	0.4673 ± 0.0007	1.1%

Metric	Class	CXR-FMKD-Direct LP MSE Avg ± SD	CXR-FMKD-Direct LP MSE %Δ w.r.t. (B)	CXR-FMKD-Direct LP CS Avg ± SD	CXR-FMKD-Direct LP CS %Δ w.r.t. (B)	CXR-FMKD-Direct LP MSE-CS | α-β Avg ± SD	CXR-FMKD-Direct LP MSE-CS | α-β %Δ w.r.t. (B)
AUC-PR	Average Classes 1 to 7	0.4479 ± 0.0004	-0.1%	0.4520 ± 0.0009	0.2%	0.4477 ± 0.0004	0.9%
AUC-ROC	Average Classes 1 to 7	0.8203 ± 0.0001	0.0%	0.8194 ± 0.0006	0.2%	0.8206 ± 0.0002	0.1%
Maximum Youden's J Statistic	Average Classes 1 to 7	0.5025 ± 0.0003	0.2%	0.4978 ± 0.0017	1.5%	0.5002 ± 0.0003	0.6%
Youden's J Statistic at 20% FPR	Average Classes 1 to 7	0.4676 ± 0.0007	-0.7%	0.4662 ± 0.0032	0.5%	0.4709 ± 0.0012	-0.7%

Metric	Class	CXR-FMKD FFT MSE Avg ± SD	CXR-FMKD FFT MSE %Δ w.r.t. (B)	CXR-FMKD FFT CS Avg ± SD	CXR-FMKD FFT CS %Δ w.r.t. (B)	CXR-FMKD FFT MSE-CS | α-β Avg ± SD	CXR-FMKD FFT MSE-CS | α-β %Δ w.r.t. (B)
AUC-PR	Average Classes 1 to 7	0.4857 ± 0.0029	-1.6%	0.4802 ± 0.0027	-2.5%	0.4845 ± 0.0030	-0.9%
AUC-ROC	Average Classes 1 to 7	0.8396 ± 0.0020	-0.3%	0.8376 ± 0.0013	-0.9%	0.8403 ± 0.0007	-0.3%
Maximum Youden's J Statistic	Average Classes 1 to 7	0.5343 ± 0.0041	0.3%	0.5318 ± 0.0039	-2.3%	0.5356 ± 0.0008	0.1%
Youden's J Statistic at 20% FPR	Average Classes 1 to 7	0.5061 ± 0.0036	-0.9%	0.5045 ± 0.0036	-2.9%	0.5088 ± 0.0037	-1.5%

Metric	Class	CXR-FMKD-Direct FFT MSE Avg ± SD	CXR-FMKD-Direct FFT MSE %Δ w.r.t. (B)	CXR-FMKD-Direct FFT CS Avg ± SD	CXR-FMKD-Direct FFT CS %Δ w.r.t. (B)	CXR-FMKD-Direct FFT MSE-CS | α-β Avg ± SD	CXR-FMKD-Direct FFT MSE-CS | α-β %Δ w.r.t. (B)
AUC-PR	Average Classes 1 to 7	0.4905 ± 0.0017	-1.8%	0.4905 ± 0.0033	-1.1%	0.4916 ± 0.0025	-1.2%
AUC-ROC	Average Classes 1 to 7	0.8415 ± 0.0020	-0.3%	0.8421 ± 0.0025	-0.3%	0.8434 ± 0.0006	-0.4%
Maximum Youden's J Statistic	Average Classes 1 to 7	0.5376 ± 0.0041	0.3%	0.5393 ± 0.0058	0.2%	0.5420 ± 0.0009	-0.8%
Youden's J Statistic at 20% FPR	Average Classes 1 to 7	0.5149 ± 0.0033	-1.6%	0.5129 ± 0.0059	-1.0%	0.5166 ± 0.0031	-1.4%

Table 15. Average Performance Comparison of 14 Selected Models and Their Benchmarks After Linear Probing Inference on MIMIC, Post-CheXpert Pre-training.

This table presents a detailed performance analysis of 14 selected '*transfer*' models that were pre-trained on CheXpert and then fine-tuned on MIMIC using only the final classification layer. The models include CXR-FM (teacher), CXR-Model FFT (baseline), and all four variants of each of the three selected student model types (MSE, CS, and MSE-CS | α-β). The analysis covers four metrics—AUC-PR, AUC-ROC, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR—focusing on the average for the most significant disease labels (Classes 1 to 7). Results are shown as mean outcomes (Avg) with standard deviations (SD), derived from testing five distinct instances of the same model type, each developed and trained using a different seed. Performance improvements for each *transfer* model relative to its corresponding *benchmark* (B) are quantified using the formula: (Transfer Model Avg Value − Benchmark Avg Value)/ Benchmark Avg Value × 100% for each metrics. Note that for CXR-FM, the *transfer* and *benchmark* models are equivalent due to its frozen backbone constraint. Additionally, the CXR-FMKD LP (CS)* *benchmark* model shows significant underperformance, which appears to be rectified in the *transfer* model setup.

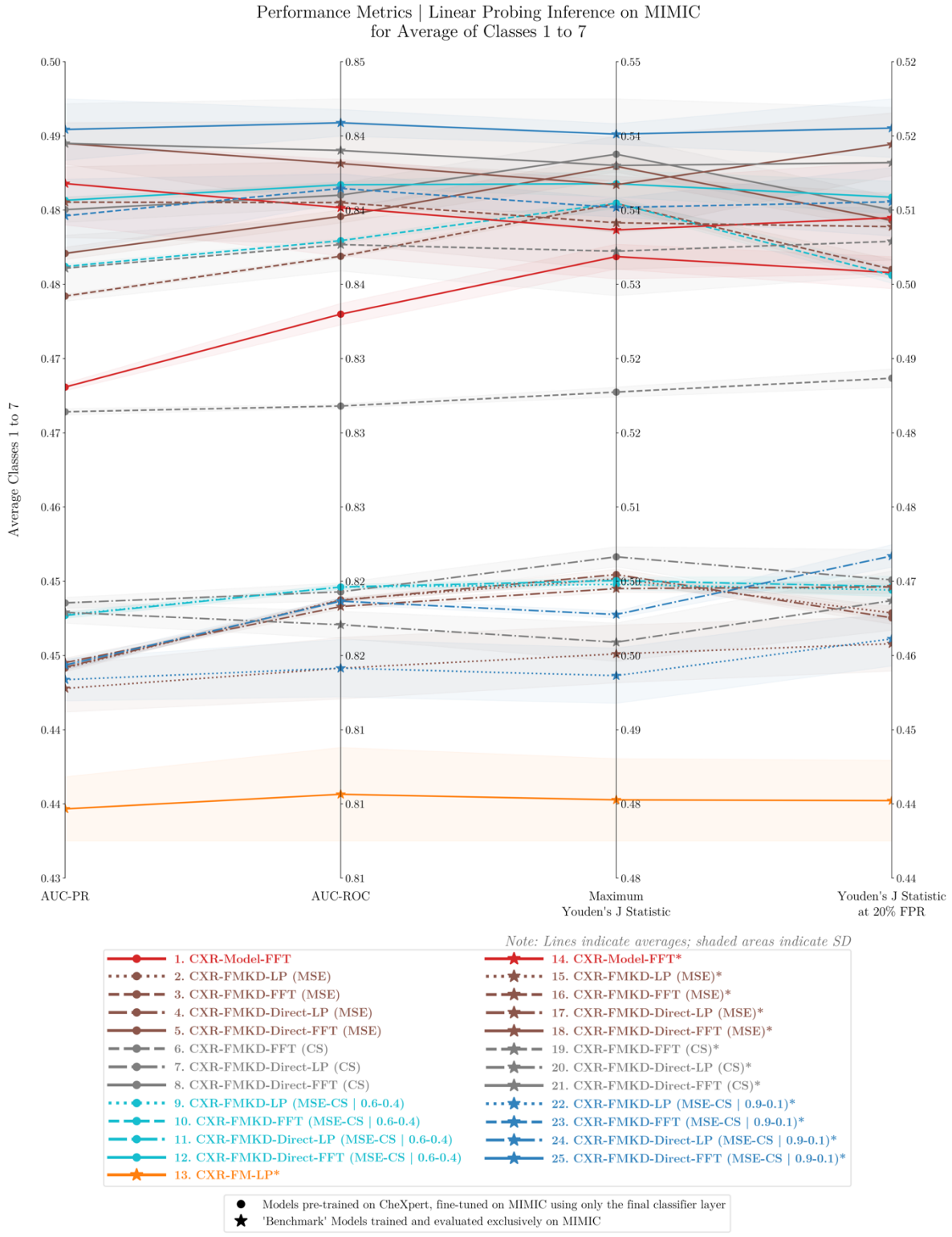

Figure 79. Comparative Analysis of Performance Metrics for 13 Models and Their Benchmarks After Linear Probing Inference on MIMIC, Post-CheXpert Pre-training.

This custom parallel coordinate plot for our Linear Probing Inference analysis visualises the performance metrics—AUC-ROC, AUC-PR, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR—for 13 models that were pre-trained on CheXpert and then fine-tuned on MIMIC using only the final classification layer. The models include CXR-FM (teacher), CXR-Model FFT (baseline), and the variants of the three selected student model types (MSE, CS, and MSE-CS | α-β). CXR-FMKD LP (CS) and CXR-Model LP were omitted due to significant underperformance which distorted the visual representation of the data. Each model, represented by circles, is contrasted against a corresponding benchmark, shown with stars, which shares the same architecture but was trained and tested exclusively on MIMIC. For CXR-FM, both are equivalent. The plot focuses on the average results for the most significant disease labels (Classes 1 to 7). Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests.

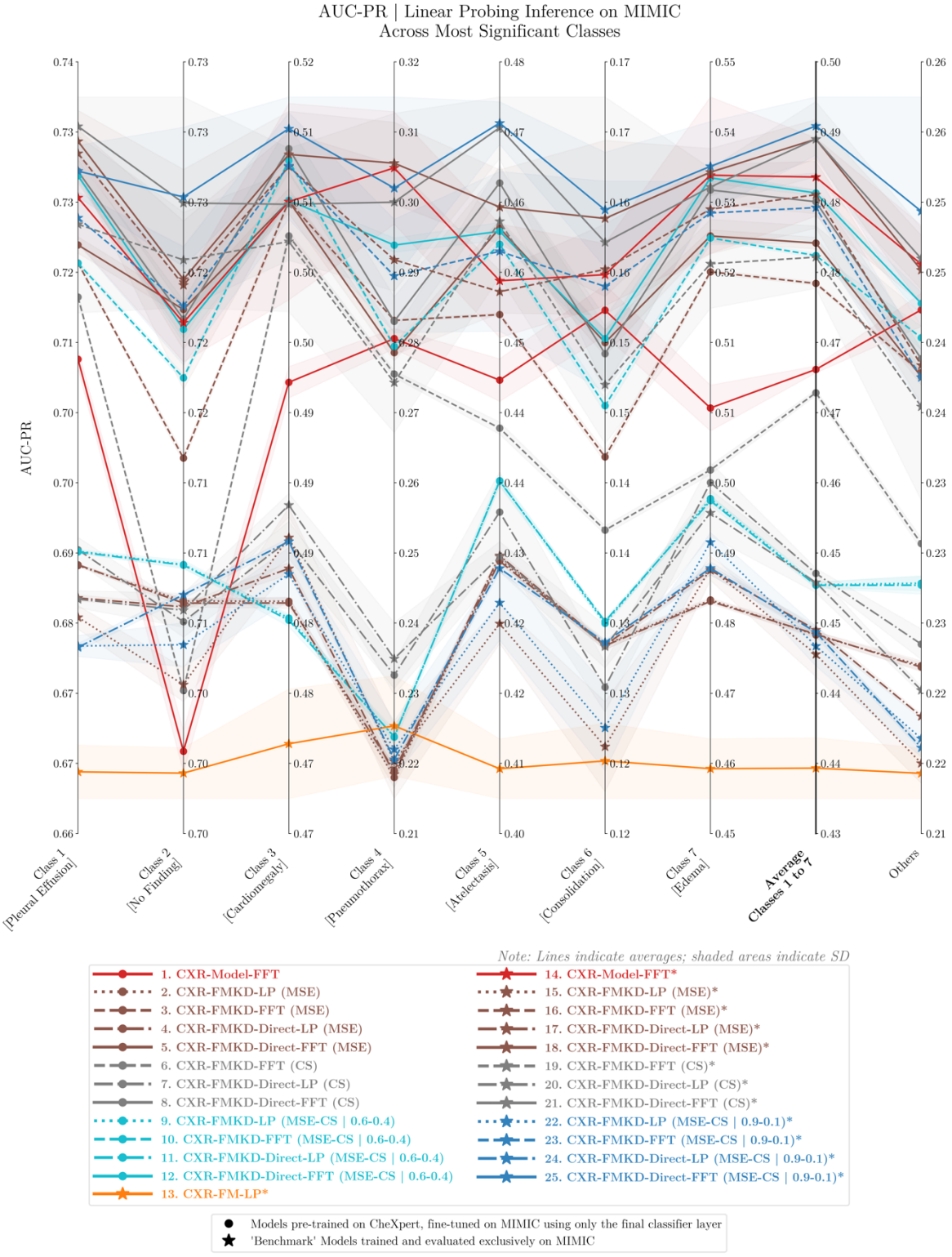

Figure 80. Parallel Coordinate Plot of AUC-PR Performance Across the Most Significant Classes for 13 Models and Their Benchmarks After Linear Probing Inference on MIMIC, Post-CheXpert Pre-training.

This custom parallel coordinate plot visualises the AUC-PR metric for 13 models that were pre-trained on CheXpert and then fine-tuned on MIMIC using only the final classification layer. The models include CXR-FM (teacher), CXR-Model FFT (baseline), and the variants of the three selected student model types (MSE, CS, and MSE-CS | α-β). CXR-FMKD LP (CS) and CXR-Model LP were omitted due to significant underperformance which distorted the visual representation of the data. Each model, represented by circles, is contrasted against a corresponding benchmark, shown with stars, which shares the same architecture but was trained and tested exclusively on MIMIC. For CXR-FM, both are equivalent. The plot displays performance results for the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes in the disease labels list. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests.

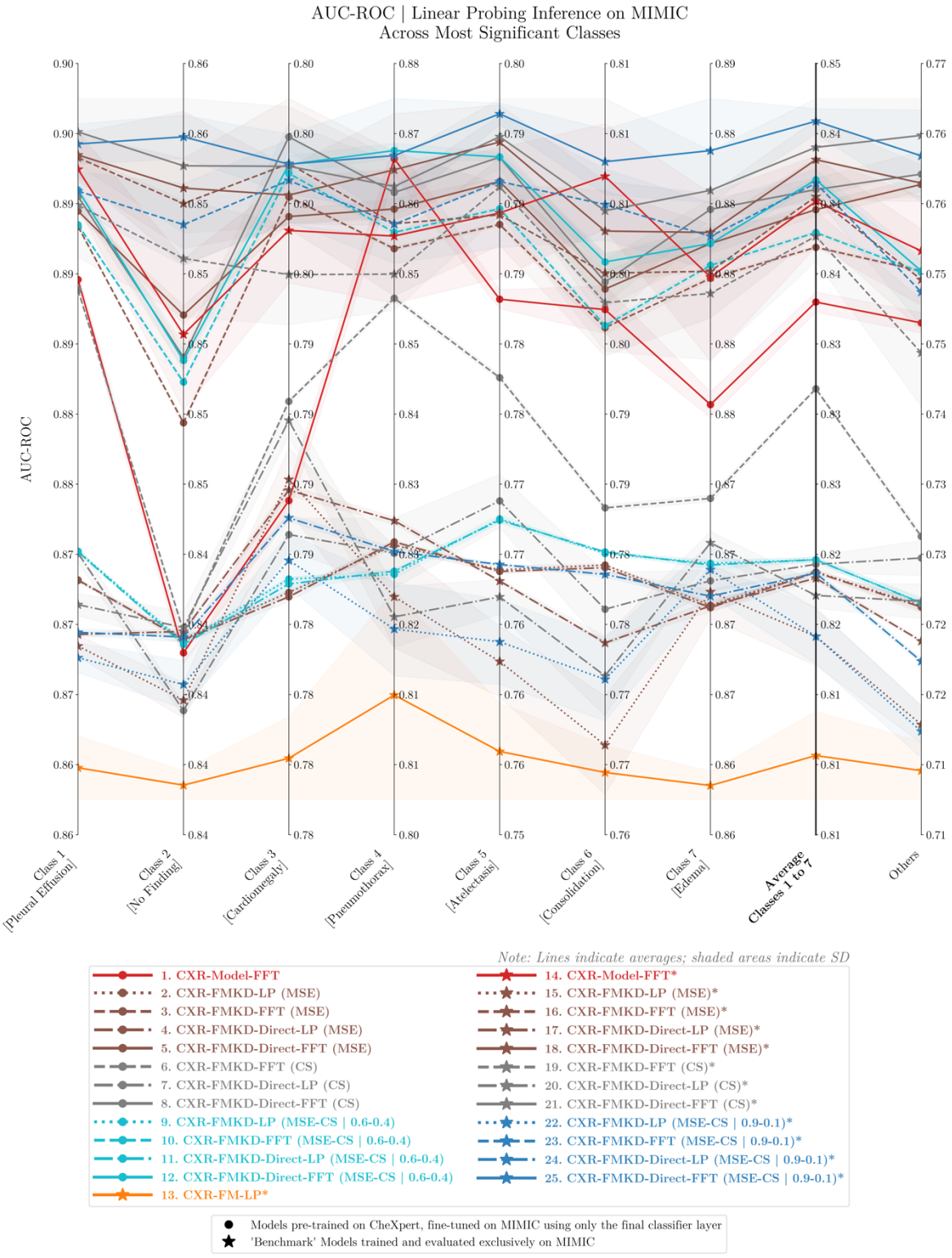

Figure 81. Parallel Coordinate Plot of AUC-ROC Performance Across the Most Significant Classes for 13 Models and Their Benchmarks After Linear Probing Inference on MIMIC, Post-CheXpert Pre-training.

This custom parallel coordinate plot visualises the AUC-ROC metric for 13 models that were pre-trained on CheXpert and then fine-tuned on MIMIC using only the final classification layer. The models include CXR-FM (teacher), CXR-Model FFT (baseline), and the variants of the three selected student model types (MSE, CS, and MSE-CS | α-β). CXR-FMKD LP (CS) and CXR-Model LP were omitted due to significant underperformance which distorted the visual representation of the data. Each model, represented by circles, is contrasted against a corresponding benchmark, shown with stars, which shares the same architecture but was trained and tested exclusively on MIMIC. For CXR-FM, both are equivalent. The plot displays performance results for the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes in the disease labels list. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests.

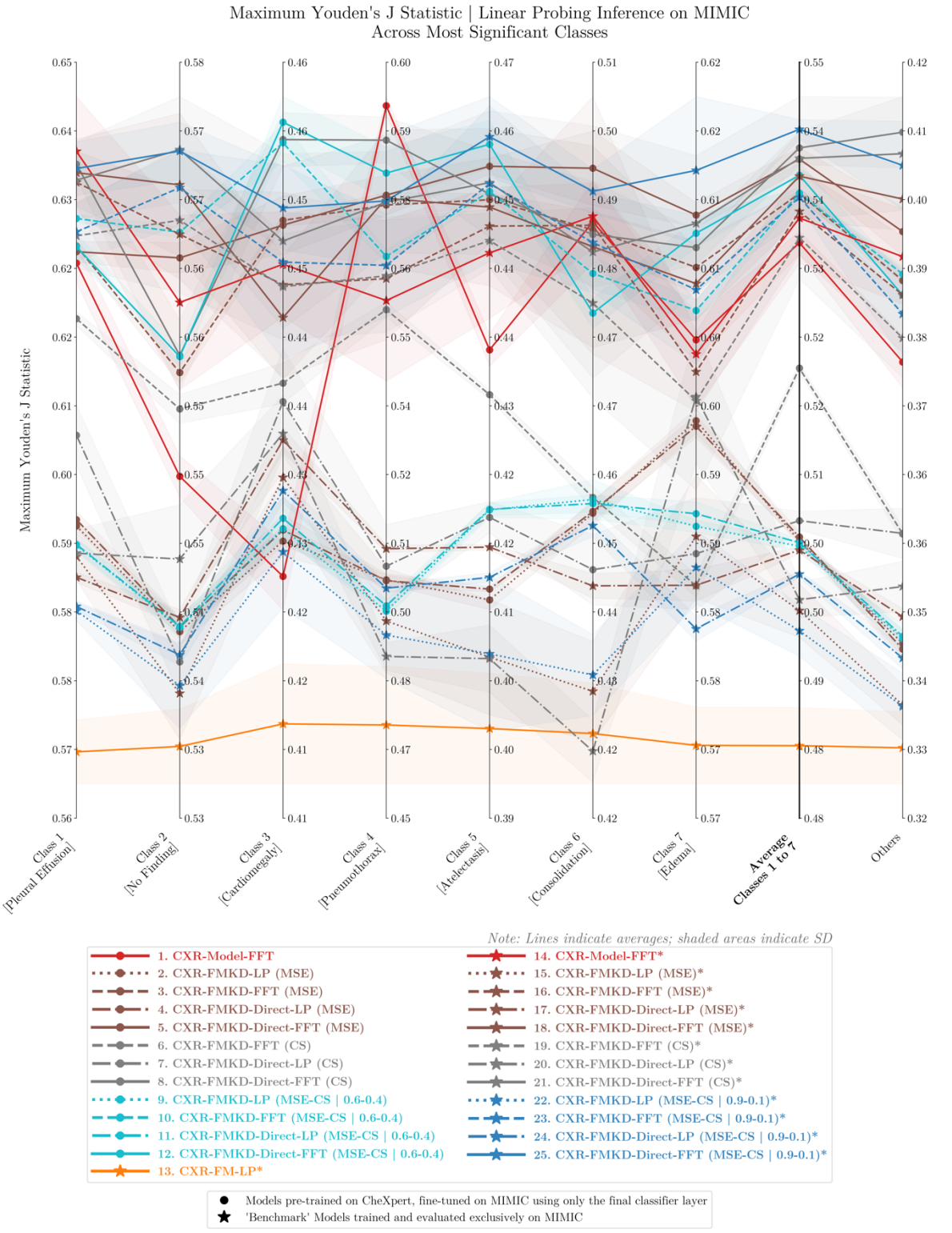

Figure 82. Parallel Coordinate Plot of Maximum Youden's J Statistic Performance Across the Most Significant Classes for 13 Models and Their Benchmarks After Linear Probing Inference on MIMIC, Post-CheXpert Pre-training.

This custom parallel coordinate plot visualises the Maximum Youden's J Statistic metric for 13 models that were pre-trained on CheXpert and then fine-tuned on MIMIC using only the final classification layer. The models include CXR-FM (teacher), CXR-Model FFT (baseline), and the variants of the three selected student model types (MSE, CS, and MSE-CS | α-β). CXR-FMKD LP (CS) and CXR-Model LP were omitted due to significant underperformance which distorted the visual representation of the data. Each model, represented by circles, is contrasted against a corresponding benchmark, shown with stars, which shares the same architecture but was trained and tested exclusively on MIMIC. For CXR-FM, both are equivalent. The plot displays performance results for the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests.

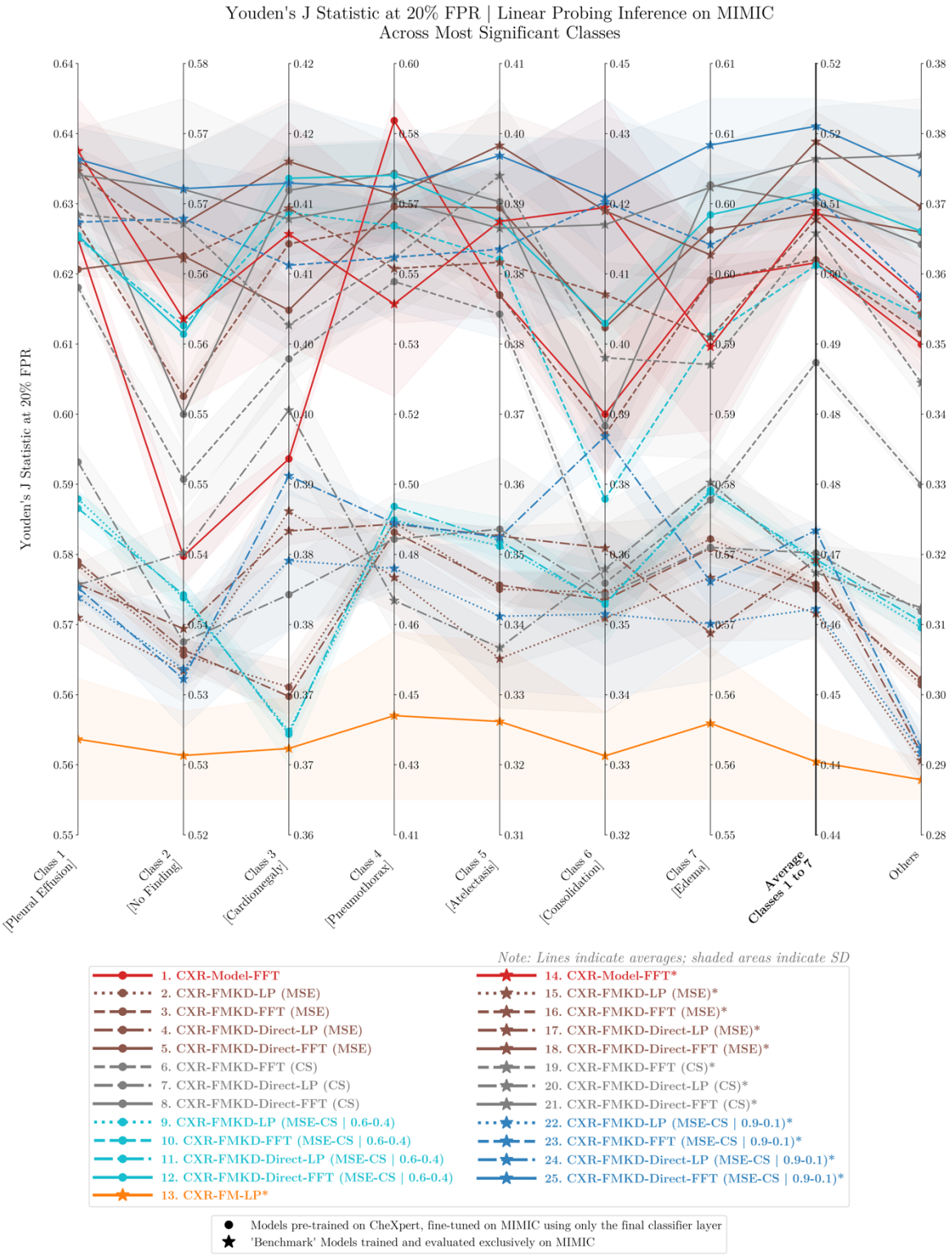

Figure 83. Parallel Coordinate Plot of Youden's J Statistic at 20% FPR Performance Across the Most Significant Classes for 13 Models and Their Benchmarks After Linear Probing Inference on MIMIC, Post-CheXpert Pre-training.

This custom parallel coordinate plot visualises the Youden's J Statistic at 20% FPR metric for 13 models that were pre-trained on CheXpert and then fine-tuned on MIMIC using only the final classification layer. The models include CXR-FM (teacher), CXR-Model FFT (baseline), and the variants of the three selected student model types (MSE, CS, and MSE-CS | α-β). CXR-FMKD LP (CS) and CXR-Model LP were omitted due to significant underperformance which distorted the visual representation of the data. Each model, represented by circles, is contrasted against a corresponding benchmark, shown with stars, which shares the same architecture but was trained and tested exclusively on MIMIC. For CXR-FM, both are equivalent. The plot displays performance results for the most significant disease labels (Classes 1 to 7), their average, and the 'Others' category which encompasses the remaining seven classes. Each line represents the average outcomes from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests.

S.5. Generalisability Analysis – Full Fine-Tuning (FFT)

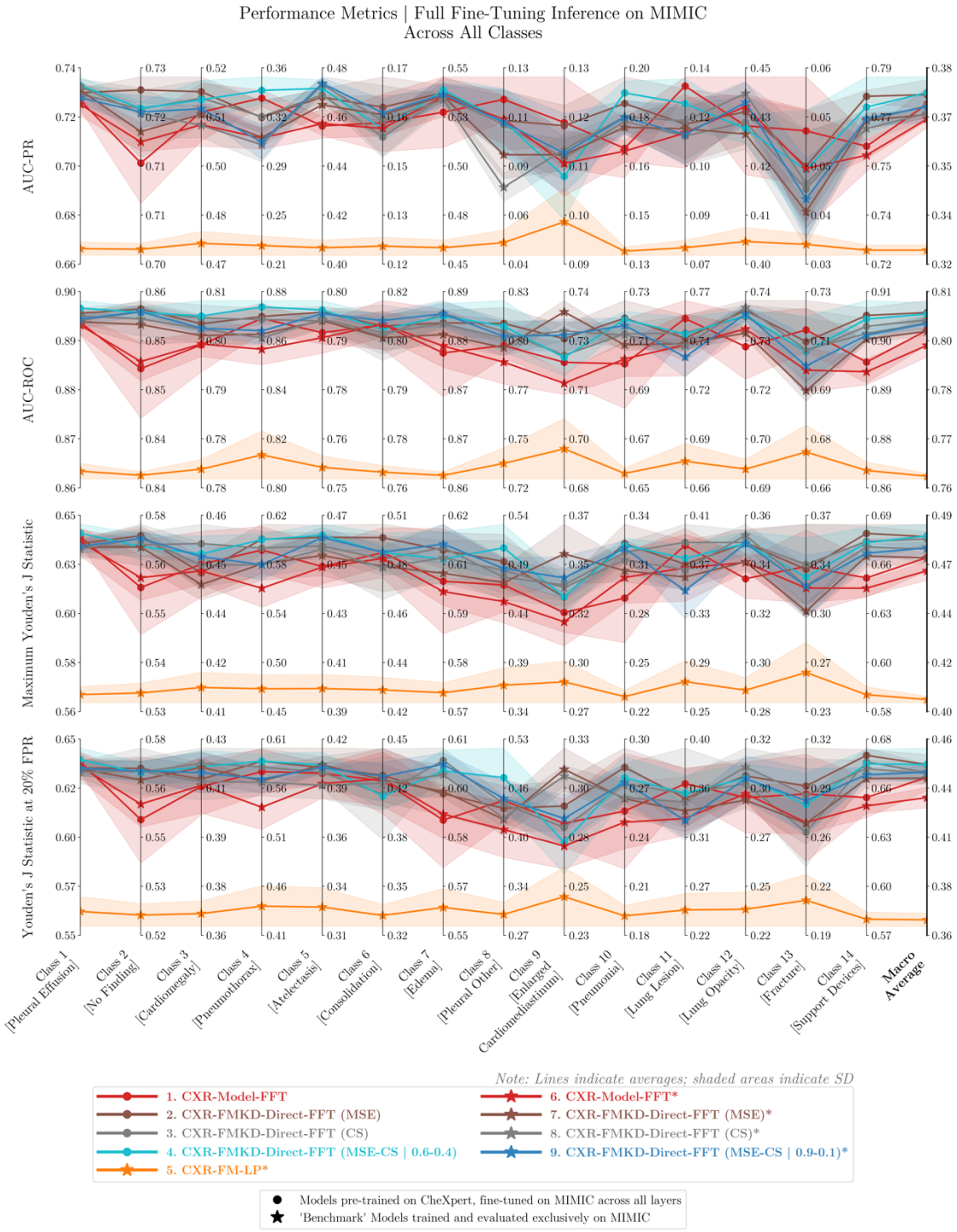

Figure 84. Performance of Selected Transfer Models and Their Benchmarks Across All 14 Classes After Full Fine-Tuning Inference on MIMIC, Post-CheXpert Pre-training.
These custom parallel coordinate plots for our Full Fine-Tuning Inference analysis display the performance across four metrics: AUC-PR, AUC-ROC, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR for the five selected '*transfer*' models that were pre-trained on CheXpert and then fine-tuned on MIMIC across all layers. The models include CXR-FM (teacher), CXR-Model FFT (baseline), and the CXR-FMKD-Direct FFT variants of the three selected student model types (MSE, CS, and MSE-CS | α-β). Each *transfer* model, represented by circles, is contrasted against a corresponding *benchmark*, shown with stars, which shares the same architecture but was trained and tested exclusively on MIMIC. For CXR-FM (i.e., CXR-FM-LP), both are equivalent. The plots display results for each of the 14 disease classes as well as their macro-average. Each line represents the average results from testing five distinct instances of the same model type, each developed and trained using a different seed, with the shaded areas indicating the standard deviation (SD) across these tests.

S.6. Bias Analysis | Bias Inspection – CheXpert

Figure 85. Joint Scatterplots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FM Tested on CheXpert [Repeated for Appendix].
This figure displays joint scatterplots with corresponding marginal distributions represented at the axes for relevant subgroup comparisons across the first four PCA modes (shown in the first two rows of plots) and the two t-SNE dimensions (shown in the last row of plots) applied to the feature embeddings extracted from CXR-FM tested on CheXpert. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), race (White, Asian, and Black), and age (grouped into eight age bins) arranged from left to right in the figure—in the first, second, third, and fourth columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

Model (CheXpert)	Mode	Explained Variance	P-Values				
			Pleural Effusion vs No Finding	White vs Asian Patients	White vs Black Patients	Asian vs Black Patients	Male vs Female Patients
CXR-FM	PCA Mode 1	16.71%	1.90E-29	1.00	4.62E-02	7.13E-02	1.55E-11
	PCA Mode 2	8.63%	1.44E-47	1.00	6.64E-02	2.51E-03	2.11E-07
	PCA Mode 3	7.16%	4.09E-08	1.76E-09	7.69E-02	7.89E-16	4.09E-08
	PCA Mode 4	4.21%	1.00	1.25E-13	1.00	4.29E-13	7.18E-08

Table 16. Statistical Comparison of Pairwise Subgroup Marginal Distributions across PCA Modes for CXR-FM Tested on CheXpert [Repeated for Appendix].
Two-sample Kolmogorov-Smirnov tests were conducted to compare all relevant pairwise subgroup marginal distributions for disease detection, biological sex, and race across the first four PCA modes, as indicated in the last five columns. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. These tests yielded p-values testing the null hypothesis that the distributions for each subgroup pair are identical. The Benjamini-Yekutieli procedure was then applied to adjust the p-values for multiple testing, with the significance determined at a p-value $<$ 0.05 (95% confidence level). P-values are colour-coded in the table: green if $p \geq 0.05$ (not significant), yellow if $0.001 \leq p < 0.05$ (significant), and red if $p < 0.001$ (highly significant), indicating increasing levels of statistical significance.

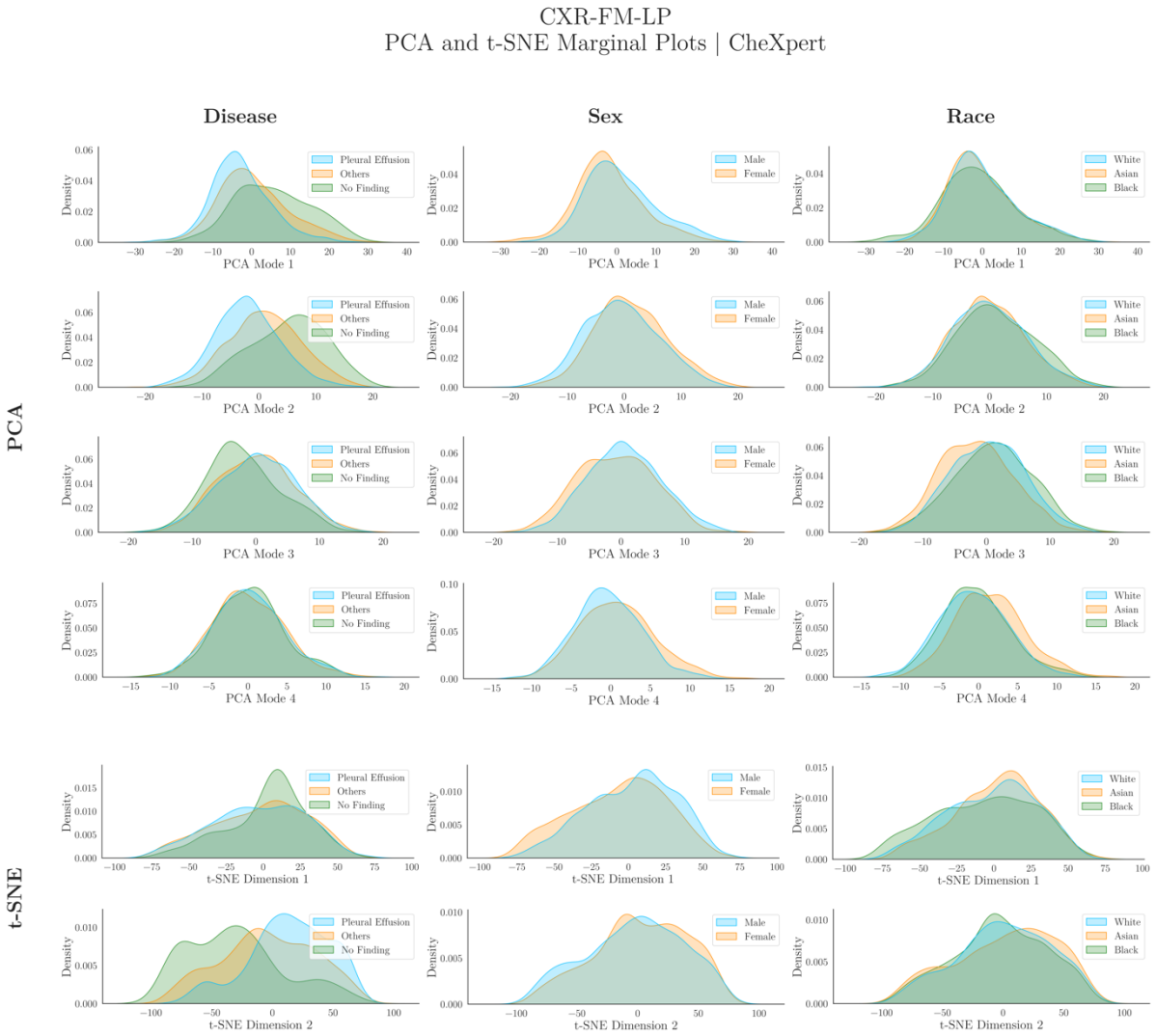

Figure 86. Marginal Plots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FM Tested on CheXpert [Repeated for Appendix].
This figure displays the marginal distributions for relevant subgroup comparisons across the first four PCA modes (shown in the first four rows of plots) and the two t-SNE dimensions (shown in the last two rows of plots) applied to the feature embeddings extracted from CXR-FM tested on CheXpert. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), and race (White, Asian, and Black), arranged from left to right in the figure—in the first, second, and third columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

Figure 87. Joint Scatterplots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-Model FFT Tested on CheXpert.
This figure displays joint scatterplots with corresponding marginal distributions represented at the axes for relevant subgroup comparisons across the first four PCA modes (shown in the first two rows of plots) and the two t-SNE dimensions (shown in the last row of plots) applied to the feature embeddings extracted from CXR-Model FFT tested on CheXpert. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), race (White, Asian, and Black), and age (grouped into eight age bins) arranged from left to right in the figure—in the first, second, third, and fourth columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

Model (CheXpert)	Mode	Explained Variance	P-Values				
			Pleural Effusion vs No Finding	White vs Asian Patients	White vs Black Patients	Asian vs Black Patients	Male vs Female Patients
CXR-Model FFT	PCA Mode 1	18.83%	1.16E-131	1.00	1.00	1.00	9.97E-01
	PCA Mode 2	9.31%	7.96E-13	1.00	1.00	5.18E-02	3.78E-01
	PCA Mode 3	7.45%	1.43E-05	1.00	7.32E-06	1.62E-06	3.91E-01
	PCA Mode 4	6.23%	1.00	1.10E-02	4.23E-05	4.24E-02	4.07E-03

Table 17. Statistical Comparison of Pairwise Subgroup Marginal Distributions across PCA Modes for CXR-Model FFT Tested on CheXpert.
Two-sample Kolmogorov-Smirnov tests were conducted to compare all relevant pairwise subgroup marginal distributions for disease detection, biological sex, and race across the first four PCA modes, as indicated in the last five columns. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. These tests yielded p-values testing the null hypothesis that the distributions for each subgroup pair are identical. The Benjamini-Yekutieli procedure was then applied to adjust the p-values for multiple testing, with the significance determined at a p-value $<$ 0.05 (95% confidence level). P-values are colour-coded in the table: green if $p \geq 0.05$ (not significant), yellow if $0.001 \leq p < 0.05$ (significant), and red if $p < 0.001$ (highly significant), indicating increasing levels of statistical significance.

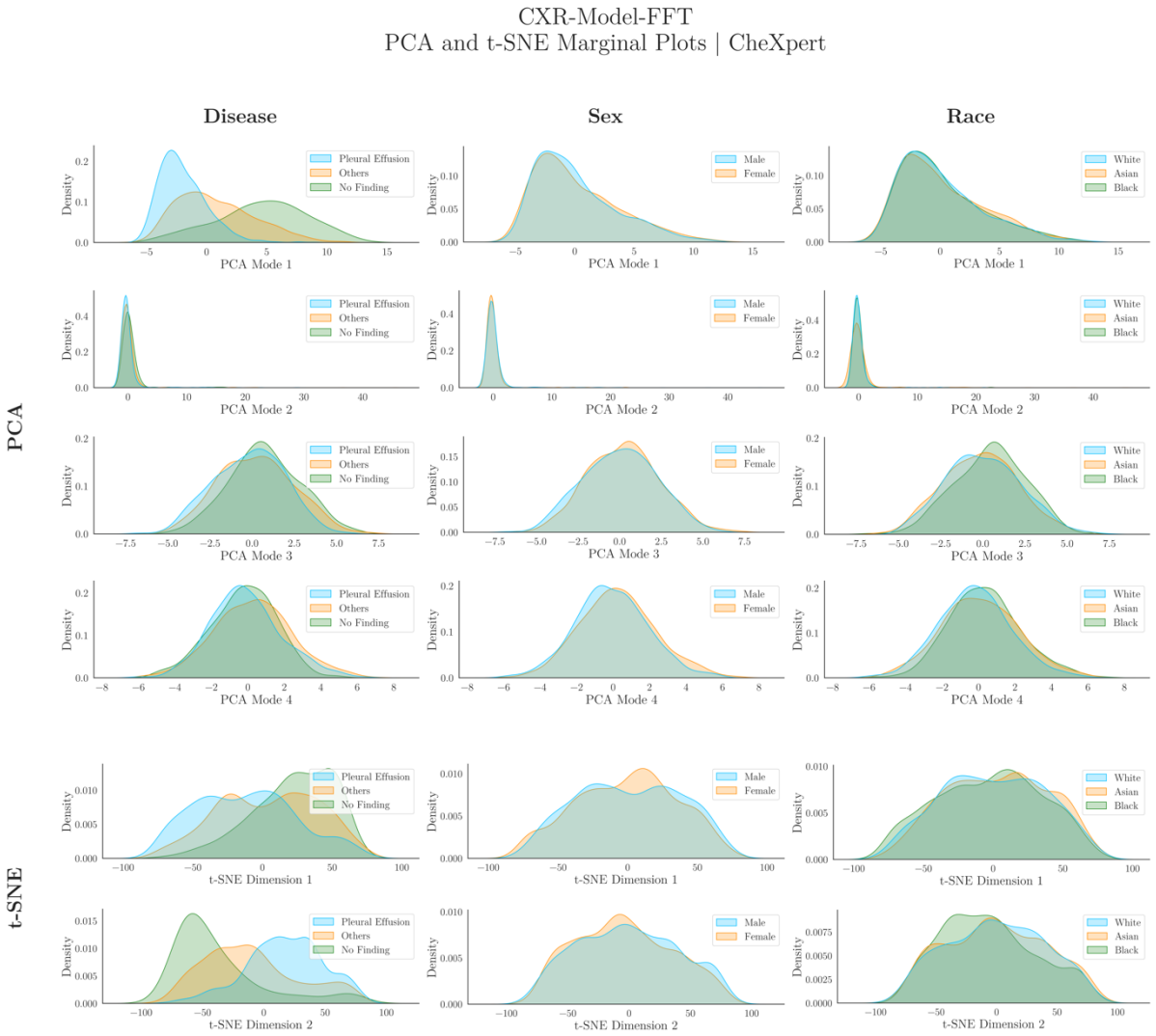

Figure 88. Marginal Plots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-Model-FFT Tested on CheXpert.
This figure displays the marginal distributions for relevant subgroup comparisons across the first four PCA modes (shown in the first four rows of plots) and the two t-SNE dimensions (shown in the last two rows of plots) applied to the feature embeddings extracted from CXR-Model-FFT tested on CheXpert. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), and race (White, Asian, and Black), arranged from left to right in the figure—in the first, second, and third columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

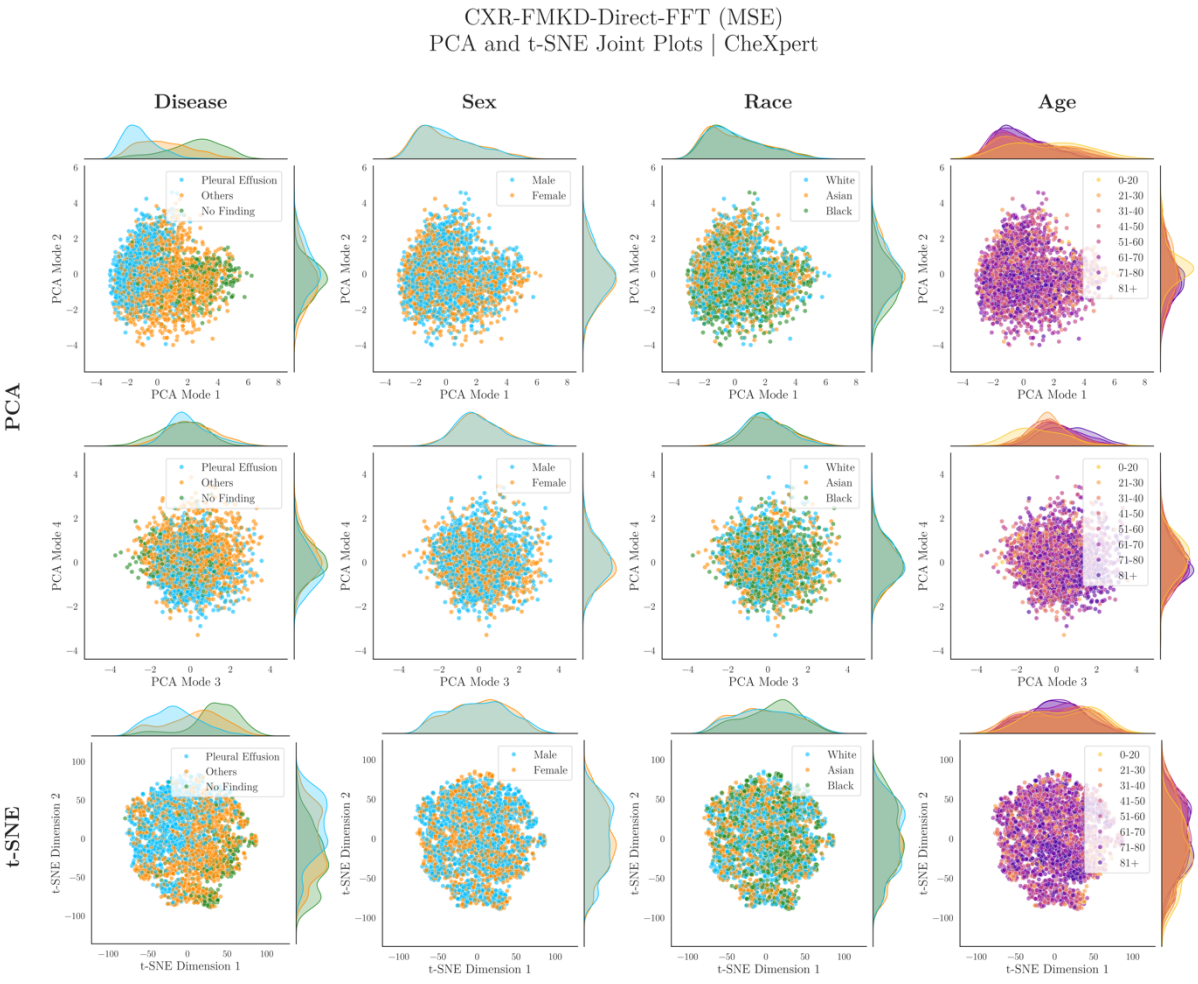

Figure 89. Joint Scatterplots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FMKD-Direct FFT (MSE) Tested on CheXpert.
This figure displays joint scatterplots with corresponding marginal distributions represented at the axes for relevant subgroup comparisons across the first four PCA modes (shown in the first two rows of plots) and the two t-SNE dimensions (shown in the last row of plots) applied to the feature embeddings extracted from CXR-FMKD-Direct FFT (MSE) tested on CheXpert. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), race (White, Asian, and Black), and age (grouped into eight age bins) arranged from left to right in the figure—in the first, second, third, and fourth columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

			P-Values				
Model (CheXpert)	**Mode**	**Explained Variance**	**Pleural Effusion vs No Finding**	**White vs Asian Patients**	**White vs Black Patients**	**Asian vs Black Patients**	**Male vs Female Patients**
CXR-FMKD-Direct FFT (MSE)	PCA Mode 1	19.47%	7.77E-136	9.08E-01	1.00	6.49E-01	1.00
	PCA Mode 2	9.95%	2.35E-09	1.00	2.67E-08	1.27E-10	1.50E-01
	PCA Mode 3	7.16%	1.04E-04	2.18E-01	1.76E-02	3.49E-02	1.00
	PCA Mode 4	4.96%	1.27E-10	1.00	1.00	1.00	1.01E-01

Table 18. Statistical Comparison of Pairwise Subgroup Marginal Distributions across PCA Modes for CXR-FMKD-Direct FFT (MSE) Tested on CheXpert [Repeated for Appendix].
Two-sample Kolmogorov-Smirnov tests were conducted to compare all relevant pairwise subgroup marginal distributions for disease detection, biological sex, and race across the first four PCA modes, as indicated in the last five columns. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. These tests yielded p-values testing the null hypothesis that the distributions for each subgroup pair are identical. The Benjamini-Yekutieli procedure was then applied to adjust the p-values for multiple testing, with the significance determined at a p-value $<$ 0.05 (95% confidence level). P-values are colour-coded in the table: green if $p \geq 0.05$ (not significant), yellow if $0.001 \leq p < 0.05$ (significant), and red if $p < 0.001$ (highly significant), indicating increasing levels of statistical significance.

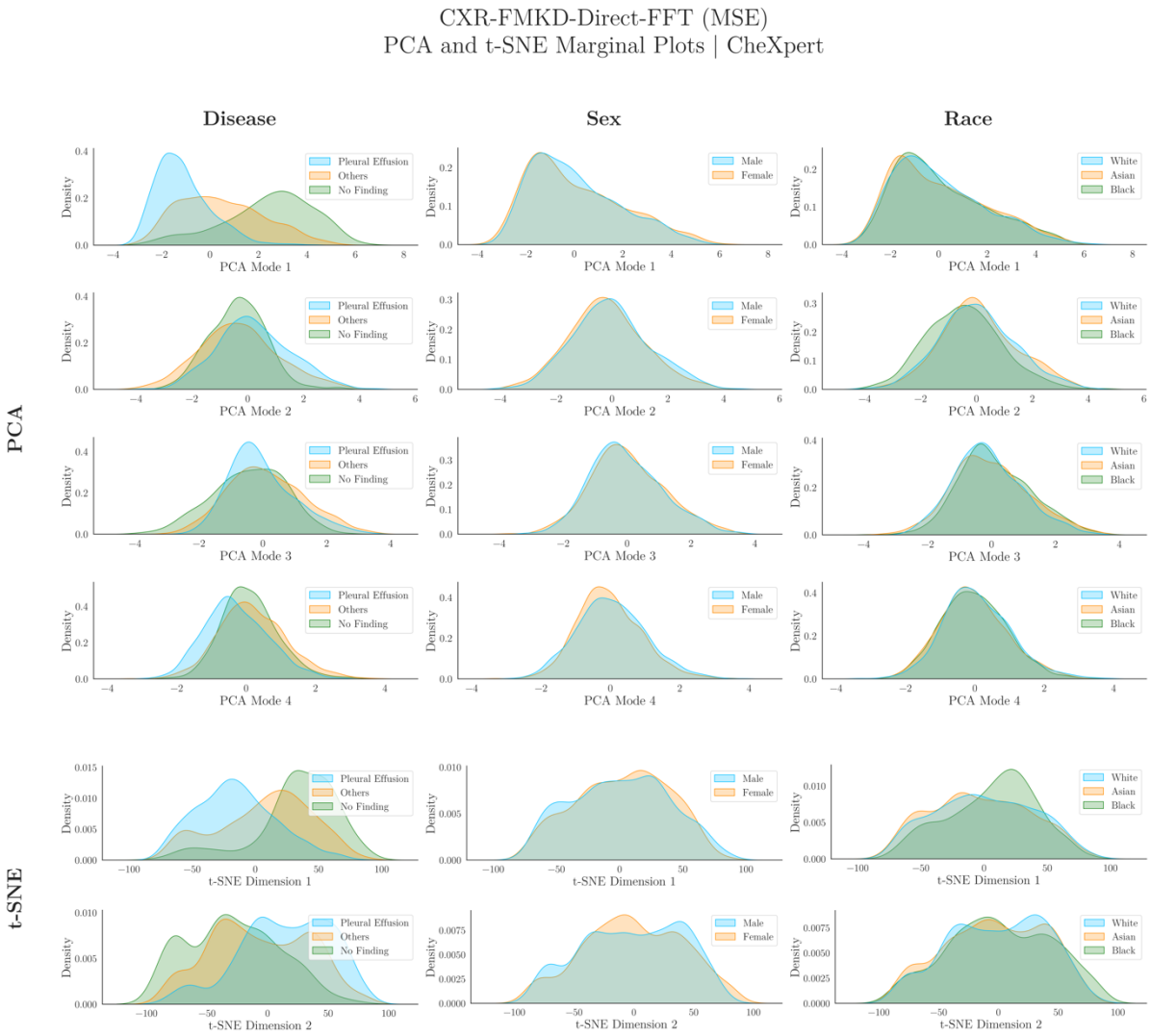

Figure 90. Marginal Plots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FMKD-Direct FFT (MSE) Tested on CheXpert [Repeated for Appendix]. This figure displays the marginal distributions for relevant subgroup comparisons across the first four PCA modes (shown in the first four rows of plots) and the two t-SNE dimensions (shown in the last two rows of plots) applied to the feature embeddings extracted from CXR-FMKD-Direct FFT (MSE) tested on CheXpert. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), and race (White, Asian, and Black), arranged from left to right in the figure—in the first, second, and third columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

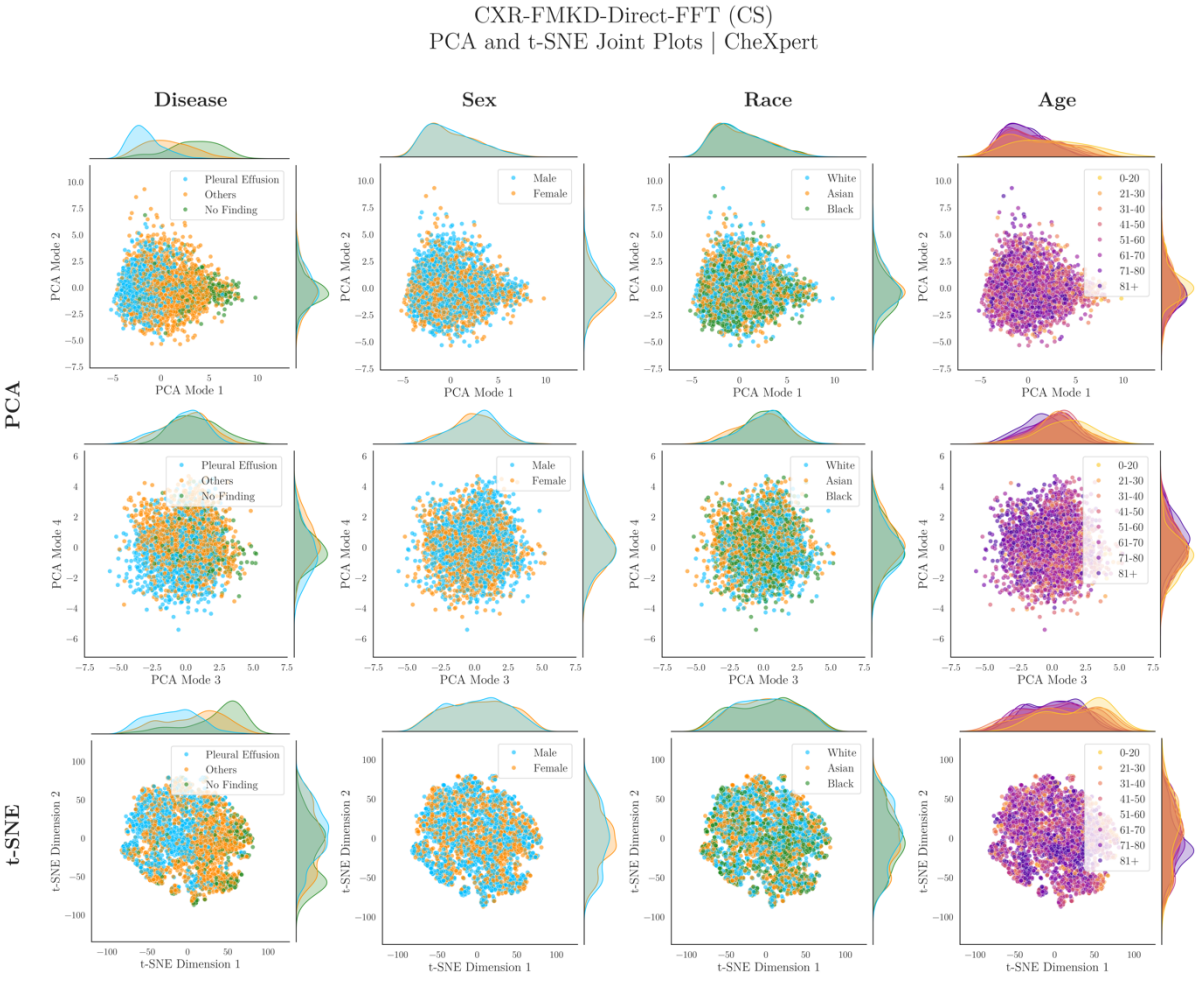

Figure 91. Joint Scatterplots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FMKD-Direct FFT (CS) Tested on CheXpert.
This figure displays joint scatterplots with corresponding marginal distributions represented at the axes for relevant subgroup comparisons across the first four PCA modes (shown in the first two rows of plots) and the two t-SNE dimensions (shown in the last row of plots) applied to the feature embeddings extracted from CXR-FMKD-Direct FFT (CS) tested on CheXpert. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), race (White, Asian, and Black), and age (grouped into eight age bins) arranged from left to right in the figure—in the first, second, third, and fourth columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

Model (CheXpert)	Mode	Explained Variance	P-Values				
			Pleural Effusion vs No Finding	White vs Asian Patients	White vs Black Patients	Asian vs Black Patients	Male vs Female Patients
CXR-FMKD-Direct FFT (CS)	PCA Mode 1	11.86%	9.64E-132	1.00	1.00	1.00	1.00
	PCA Mode 2	5.18%	2.94E-04	8.98E-01	2.44E-07	3.18E-07	6.93E-07
	PCA Mode 3	4.59%	1.20E-09	4.31E-04	9.46E-02	5.16E-03	2.81E-04
	PCA Mode 4	3.80%	1.93E-05	1.77E-06	1.50E-01	3.29E-03	2.18E-01

Table 19. Statistical Comparison of Pairwise Subgroup Marginal Distributions across PCA Modes for CXR-FMKD-Direct FFT (CS) Tested on CheXpert.
Two-sample Kolmogorov-Smirnov tests were conducted to compare all relevant pairwise subgroup marginal distributions for disease detection, biological sex, and race across the first four PCA modes, as indicated in the last five columns. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. These tests yielded p-values testing the null hypothesis that the distributions for each subgroup pair are identical. The Benjamini-Yekutieli procedure was then applied to adjust the p-values for multiple testing, with the significance determined at a p-value $<$ 0.05 (95% confidence level). P-values are colour-coded in the table: green if $p \geq 0.05$ (not significant), yellow if $0.001 \leq p < 0.05$ (significant), and red if $p < 0.001$ (highly significant), indicating increasing levels of statistical significance.

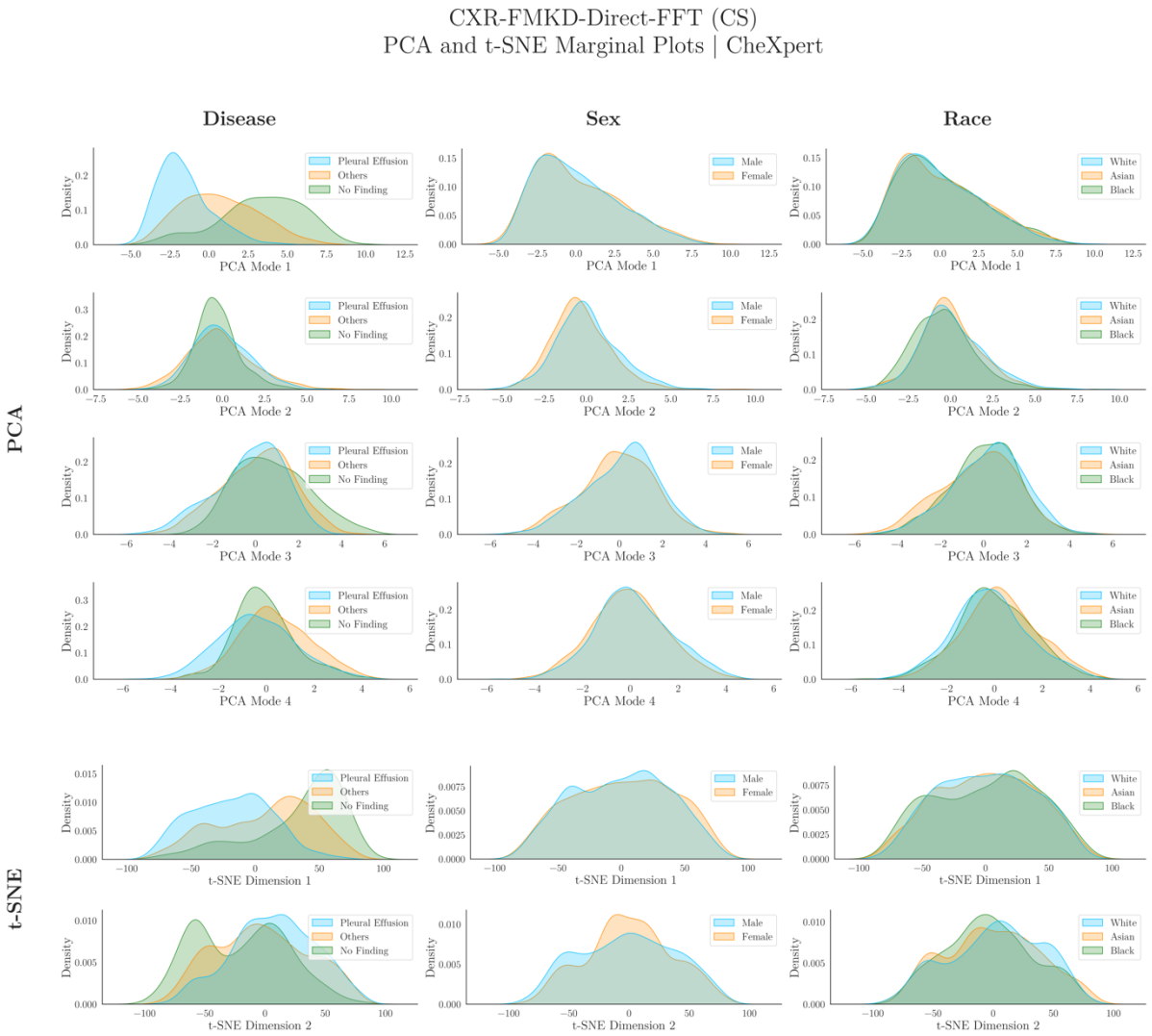

Figure 92. Marginal Plots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FMKD-Direct FFT (CS) Tested on CheXpert.
This figure displays the marginal distributions for relevant subgroup comparisons across the first four PCA modes (shown in the first four rows of plots) and the two t-SNE dimensions (shown in the last two rows of plots) applied to the feature embeddings extracted from CXR-FMKD-Direct FFT (CS) tested on CheXpert. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), and race (White, Asian, and Black), arranged from left to right in the figure—in the first, second, and third columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

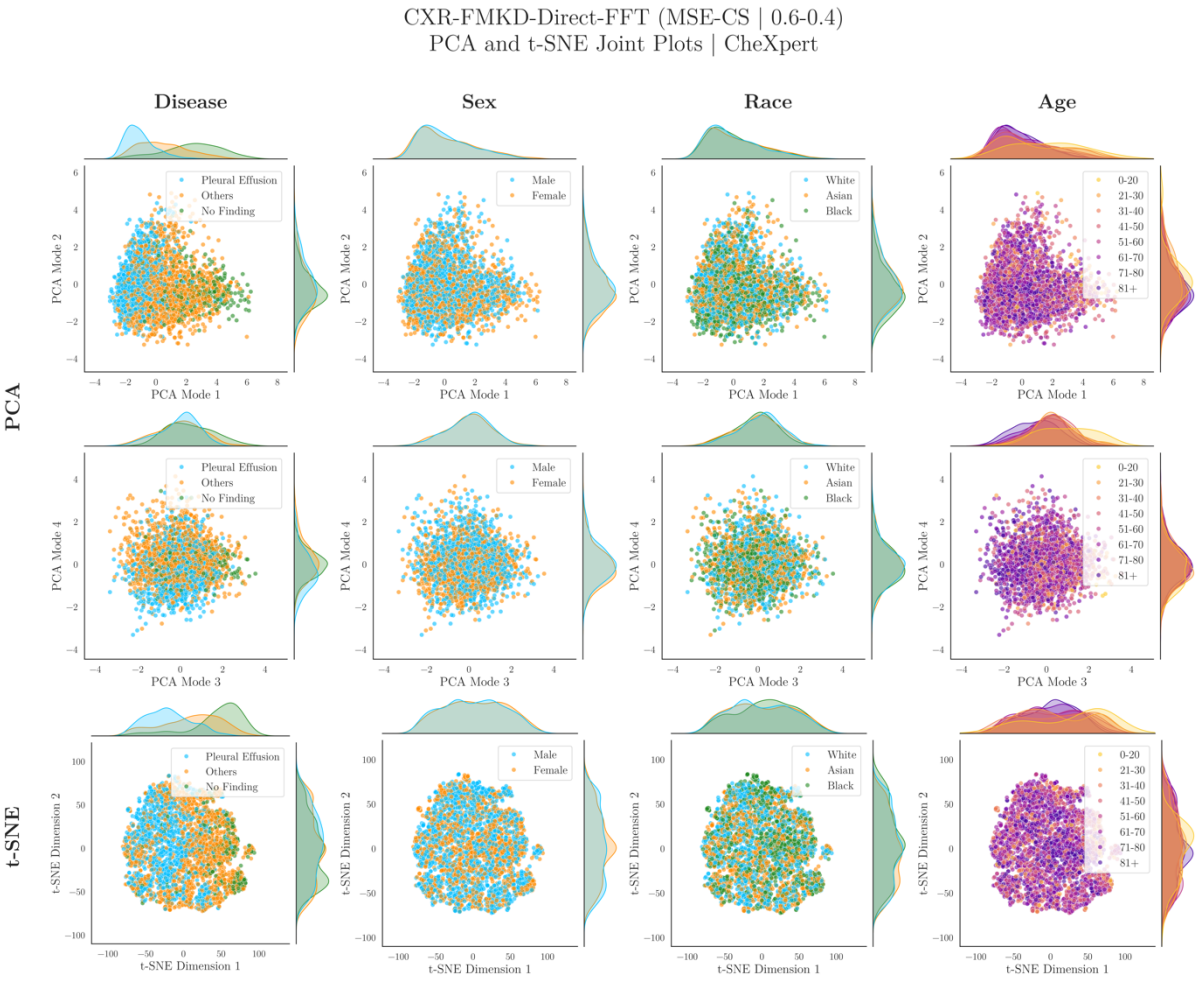

Figure 93. Joint Scatterplots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4) Tested on CheXpert.
This figure displays joint scatterplots with corresponding marginal distributions represented at the axes for relevant subgroup comparisons across the first four PCA modes (shown in the first two rows of plots) and the two t-SNE dimensions (shown in the last row of plots) applied to the feature embeddings extracted from CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4) tested on CheXpert. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), race (White, Asian, and Black), and age (grouped into eight age bins) arranged from left to right in the figure—in the first, second, third, and fourth columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

Model (CheXpert)	Mode	Explained Variance	P-Values				
			Pleural Effusion vs No Finding	White vs Asian Patients	White vs Black Patients	Asian vs Black Patients	Male vs Female Patients
CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4)	PCA Mode 1	17.20%	3.42E-141	7.87E-01	7.87E-01	7.87E-01	5.77E-01
	PCA Mode 2	9.01%	1.09E-06	1.00	4.38E-07	4.61E-11	2.57E-04
	PCA Mode 3	6.15%	8.68E-10	1.46E-02	2.28E-05	5.82E-01	1.00
	PCA Mode 4	5.04%	4.36E-13	1.00	1.00	1.00	8.22E-04

Table 20. Statistical Comparison of Pairwise Subgroup Marginal Distributions across PCA Modes for CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4) Tested on CheXpert.
Two-sample Kolmogorov-Smirnov tests were conducted to compare all relevant pairwise subgroup marginal distributions for disease detection, biological sex, and race across the first four PCA modes, as indicated in the last five columns. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. These tests yielded p-values testing the null hypothesis that the distributions for each subgroup pair are identical. The Benjamini-Yekutieli procedure was then applied to adjust the p-values for multiple testing, with the significance determined at a p-value $<$ 0.05 (95% confidence level). P-values are colour-coded in the table: green if $p \geq 0.05$ (not significant), yellow if $0.001 \leq p < 0.05$ (significant), and red if $p < 0.001$ (highly significant), indicating increasing levels of statistical significance.

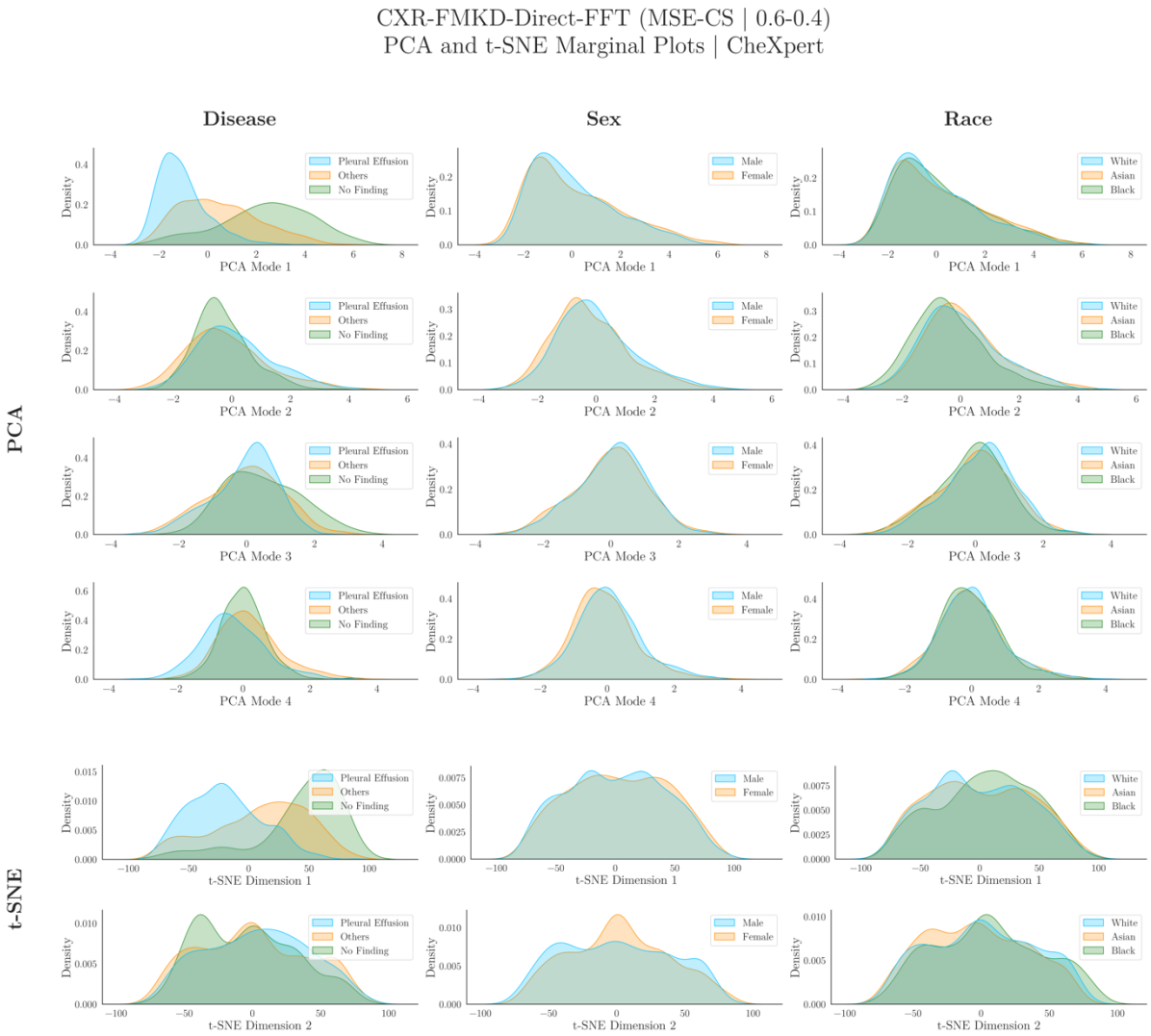

Figure 94. Marginal Plots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4) Tested on CheXpert.
This figure displays the marginal distributions for relevant subgroup comparisons across the first four PCA modes (shown in the first four rows of plots) and the two t-SNE dimensions (shown in the last two rows of plots) applied to the feature embeddings extracted from CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4) tested on CheXpert. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), and race (White, Asian, and Black), arranged from left to right in the figure—in the first, second, and third columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

CheXpert

Model (CheXpert)	Mode	Explained Variance	Disease Detection: Pleural Effusion vs No Finding	Race Attribute: White vs Asian	Race Attribute: White vs Black	Race Attribute: Asian vs Black	Sex Attribute: Male vs Female
CXR-FM-LP	PCA Mode 1	16.71%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 4895 [97.90%] TRUE : 104 [2.08%] TRUE+ : 1 [0.02%]	FALSE : 3424 [68.48%] TRUE : 1416 [28.32%] TRUE+ : 160 [3.20%]	FALSE : 1754 [35.08%] TRUE : 2834 [56.68%] TRUE+ : 412 [8.24%]	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]
	PCA Mode 2	8.63%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 3162 [63.24%] TRUE : 1635 [32.70%] TRUE+ : 203 [4.06%]	FALSE : 2724 [54.48%] TRUE : 1970 [39.40%] TRUE+ : 306 [6.12%]	FALSE : 9 [0.18%] TRUE : 507 [10.14%] TRUE+ : 4484 [89.68%]	FALSE : 8 [0.16%] TRUE : 274 [5.48%] TRUE+ : 4718 [94.36%]
	PCA Mode 3	7.16%	FALSE : 166 [3.32%] TRUE : 1471 [29.42%] TRUE+ : 3363 [67.26%]	FALSE : 0 [0.00%] TRUE : 3 [0.06%] TRUE+ : 4997 [99.94%]	FALSE : 2552 [51.04%] TRUE : 2035 [40.70%] TRUE+ : 413 [8.26%]	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 0 [0.00%] TRUE : 32 [0.64%] TRUE+ : 4968 [99.36%]
	PCA Mode 4	4.21%	FALSE : 4962 [99.24%] TRUE : 37 [0.74%] TRUE+ : 1 [0.02%]	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 4824 [96.48%] TRUE : 173 [3.46%] TRUE+ : 3 [0.06%]	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]
CXR-Model-FFT	PCA Mode 1	18.83%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 4839 [96.78%] TRUE : 156 [3.12%] TRUE+ : 5 [0.10%]	FALSE : 4917 [98.34%] TRUE : 82 [1.64%] TRUE+ : 1 [0.02%]	FALSE : 4982 [99.64%] TRUE : 18 [0.36%] TRUE+ : 0 [0.00%]	FALSE : 4993 [99.86%] TRUE : 7 [0.14%] TRUE+ : 0 [0.00%]
	PCA Mode 2	9.31%	FALSE : 0 [0.00%] TRUE : 4 [0.08%] TRUE+ : 4996 [99.92%]	FALSE : 4811 [96.22%] TRUE : 184 [3.68%] TRUE+ : 5 [0.10%]	FALSE : 3885 [77.70%] TRUE : 1020 [20.40%] TRUE+ : 95 [1.90%]	FALSE : 4436 [88.72%] TRUE : 544 [10.88%] TRUE+ : 20 [0.40%]	FALSE : 3762 [75.24%] TRUE : 1133 [22.66%] TRUE+ : 105 [2.10%]
	PCA Mode 3	7.45%	FALSE : 0 [0.00%] TRUE : 12 [0.24%] TRUE+ : 4988 [99.76%]	FALSE : 4821 [96.42%] TRUE : 172 [3.44%] TRUE+ : 7 [0.14%]	FALSE : 9 [0.18%] TRUE : 409 [8.18%] TRUE+ : 4582 [91.64%]	FALSE : 0 [0.00%] TRUE : 22 [0.44%] TRUE+ : 4978 [99.56%]	FALSE : 4975 [99.50%] TRUE : 25 [0.50%] TRUE+ : 0 [0.00%]
	PCA Mode 4	6.23%	FALSE : 4889 [97.78%] TRUE : 110 [2.20%] TRUE+ : 1 [0.02%]	FALSE : 4365 [87.30%] TRUE : 593 [11.86%] TRUE+ : 42 [0.84%]	FALSE : 50 [1.00%] TRUE : 710 [14.20%] TRUE+ : 4240 [84.80%]	FALSE : 590 [11.80%] TRUE : 2385 [47.70%] TRUE+ : 2025 [40.50%]	FALSE : 1012 [20.24%] TRUE : 2497 [49.94%] TRUE+ : 1491 [29.82%]
CXR-FMKD-Direct-FFT (MSE)	PCA Mode 1	19.47%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 4961 [99.22%] TRUE : 39 [0.78%] TRUE+ : 0 [0.00%]	FALSE : 4938 [98.76%] TRUE : 61 [1.22%] TRUE+ : 1 [0.02%]	FALSE : 4875 [97.50%] TRUE : 121 [2.42%] TRUE+ : 4 [0.08%]	FALSE : 4938 [98.76%] TRUE : 61 [1.22%] TRUE+ : 1 [0.02%]
	PCA Mode 2	9.95%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 4761 [95.22%] TRUE : 231 [4.62%] TRUE+ : 8 [0.16%]	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 4474 [89.48%] TRUE : 499 [9.98%] TRUE+ : 27 [0.54%]
	PCA Mode 3	7.16%	FALSE : 0 [0.00%] TRUE : 41 [0.82%] TRUE+ : 4959 [99.18%]	FALSE : 4974 [99.48%] TRUE : 26 [0.52%] TRUE+ : 0 [0.00%]	FALSE : 3956 [79.12%] TRUE : 996 [19.92%] TRUE+ : 48 [0.96%]	FALSE : 4367 [87.34%] TRUE : 609 [12.18%] TRUE+ : 24 [0.48%]	FALSE : 4792 [95.84%] TRUE : 202 [4.04%] TRUE+ : 6 [0.12%]
	PCA Mode 4	4.96%	FALSE : 0 [0.00%] TRUE : 5 [0.10%] TRUE+ : 4995 [99.90%]	FALSE : 4382 [87.64%] TRUE : 570 [11.40%] TRUE+ : 48 [0.96%]	FALSE : 4786 [95.72%] TRUE : 207 [4.14%] TRUE+ : 7 [0.14%]	FALSE : 4982 [99.64%] TRUE : 18 [0.36%] TRUE+ : 0 [0.00%]	FALSE : 4428 [88.56%] TRUE : 542 [10.84%] TRUE+ : 30 [0.60%]
CXR-FMKD-Direct-FFT (CS)	PCA Mode 1	11.86%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 4970 [99.40%] TRUE : 30 [0.60%] TRUE+ : 0 [0.00%]	FALSE : 4857 [97.14%] TRUE : 141 [2.82%] TRUE+ : 2 [0.04%]	FALSE : 4889 [97.78%] TRUE : 109 [2.18%] TRUE+ : 2 [0.04%]	FALSE : 4966 [99.32%] TRUE : 34 [0.68%] TRUE+ : 0 [0.00%]
	PCA Mode 2	5.18%	FALSE : 76 [1.52%] TRUE : 1167 [23.34%] TRUE+ : 3757 [75.14%]	FALSE : 4971 [99.42%] TRUE : 29 [0.58%] TRUE+ : 0 [0.00%]	FALSE : 0 [0.00%] TRUE : 2 [0.04%] TRUE+ : 4998 [99.96%]	FALSE : 0 [0.00%] TRUE : 8 [0.16%] TRUE+ : 4992 [99.84%]	FALSE : 403 [8.06%] TRUE : 2220 [44.40%] TRUE+ : 2377 [47.54%]
	PCA Mode 3	4.59%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 1298 [25.96%] TRUE : 2515 [50.30%] TRUE+ : 1187 [23.74%]	FALSE : 4957 [99.14%] TRUE : 43 [0.86%] TRUE+ : 0 [0.00%]	FALSE : 1393 [27.86%] TRUE : 2706 [54.12%] TRUE+ : 901 [18.02%]	FALSE : 41 [0.82%] TRUE : 696 [13.92%] TRUE+ : 4263 [85.26%]
	PCA Mode 4	3.80%	FALSE : 83 [1.66%] TRUE : 2264 [45.28%] TRUE+ : 2653 [53.06%]	FALSE : 0 [0.00%] TRUE : 27 [0.54%] TRUE+ : 4973 [99.46%]	FALSE : 3129 [62.58%] TRUE : 1693 [33.86%] TRUE+ : 178 [3.56%]	FALSE : 262 [5.24%] TRUE : 1905 [38.10%] TRUE+ : 2833 [56.66%]	FALSE : 4599 [91.98%] TRUE : 398 [7.96%] TRUE+ : 3 [0.06%]
CXR-FMKD-Direct-FFT (MSE-CS | 0.6-0.4)	PCA Mode 1	17.20%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 4600 [92.00%] TRUE : 382 [7.64%] TRUE+ : 18 [0.36%]	FALSE : 3972 [79.44%] TRUE : 948 [18.96%] TRUE+ : 80 [1.60%]	FALSE : 4704 [94.08%] TRUE : 289 [5.78%] TRUE+ : 7 [0.14%]	FALSE : 4968 [99.36%] TRUE : 32 [0.64%] TRUE+ : 0 [0.00%]
	PCA Mode 2	9.01%	FALSE : 0 [0.00%] TRUE : 23 [0.46%] TRUE+ : 4977 [99.54%]	FALSE : 4964 [99.28%] TRUE : 35 [0.70%] TRUE+ : 1 [0.02%]	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 2886 [57.72%] TRUE : 1719 [34.38%] TRUE+ : 395 [7.90%]
	PCA Mode 3	6.15%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 4771 [95.42%] TRUE : 217 [4.34%] TRUE+ : 12 [0.24%]	FALSE : 3431 [68.62%] TRUE : 1386 [27.72%] TRUE+ : 183 [3.66%]	FALSE : 4544 [90.88%] TRUE : 443 [8.86%] TRUE+ : 13 [0.26%]	FALSE : 3823 [76.46%] TRUE : 1057 [21.14%] TRUE+ : 120 [2.40%]
	PCA Mode 4	5.04%	FALSE : 0 [0.00%] TRUE : 2 [0.04%] TRUE+ : 4998 [99.96%]	FALSE : 4948 [98.96%] TRUE : 51 [1.02%] TRUE+ : 1 [0.02%]	FALSE : 4925 [98.50%] TRUE : 74 [1.48%] TRUE+ : 1 [0.02%]	FALSE : 4973 [99.46%] TRUE : 27 [0.54%] TRUE+ : 0 [0.00%]	FALSE : 1414 [28.28%] TRUE : 2374 [47.48%] TRUE+ : 1212 [24.24%]

Table 21. Results from Bootstrapping-like Simulations for Statistical Comparison of Pairwise Subgroup Marginal Distributions across PCA Modes for Selected Models Tested on CheXpert.
Utilising a bootstrapping-like approach, this table presents the outcomes of two-sample Kolmogorov-Smirnov tests performed to compare all relevant pairwise subgroup marginal distributions for disease detection, biological sex, and race across the first four PCA modes, as indicated in the last five columns, for our five selected models tested on CheXpert. Each test was replicated across 5000 simulations, each based on a balanced stratified sample of 3000 patients—1000 from each racial group with no duplicates—to ensure robustness and mitigate sampling variability. These tests yielded p-values testing the null hypothesis that distributions between each subgroup pair are identical. The Benjamini-Yekutieli procedure was applied to adjust the p-values for multiple testing, with significance determined at a p-value < 0.05 (95% confidence level). Results of these simulations were categorised into 'FALSE' ($p \geq 0.05$), 'TRUE' ($0.001 \leq p < 0.05$), and 'TRUE+' ($p < 0.001$) to help quantify bias, indicating varying levels of statistical significance.

S.7. Bias Analysis | Bias Inspection – MIMIC

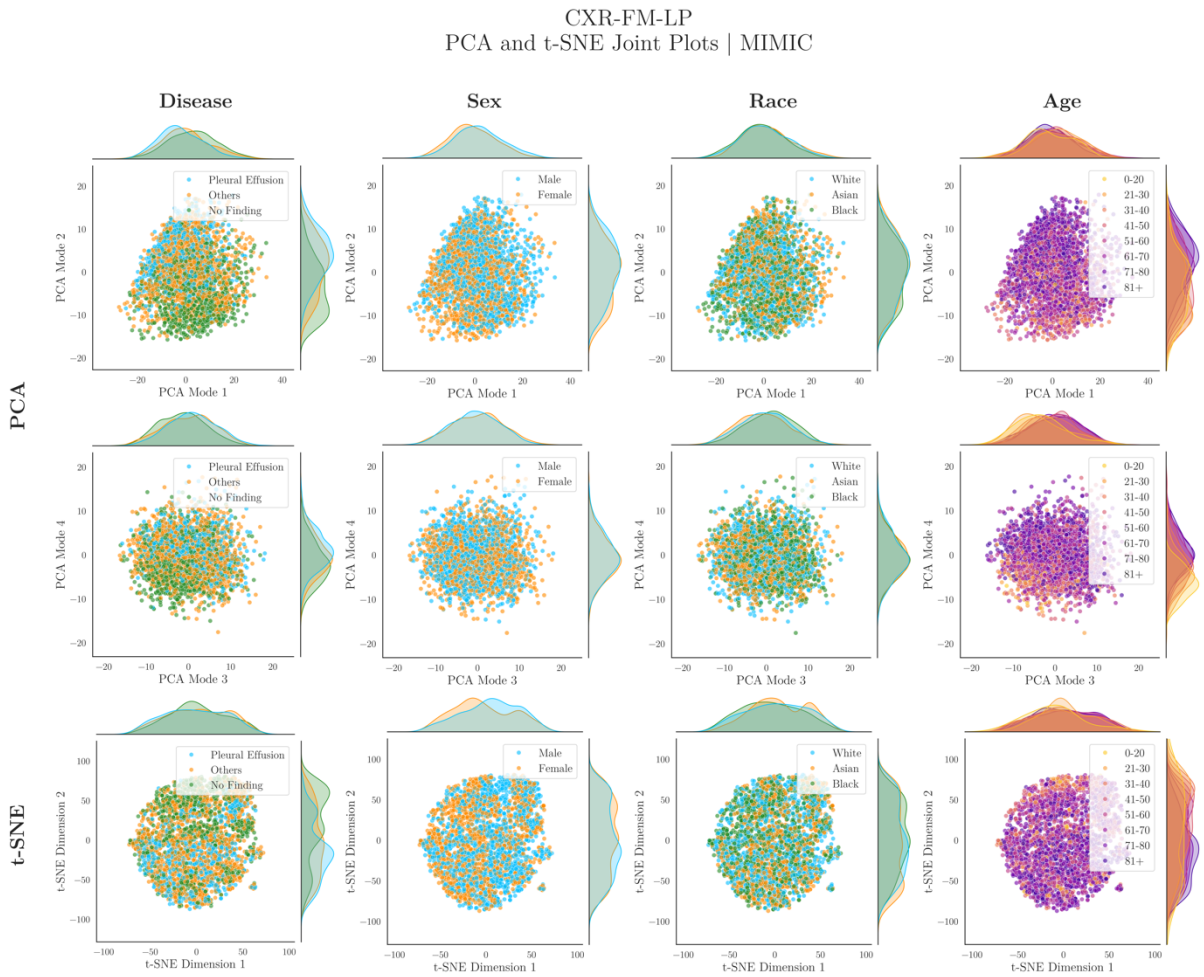

Figure 95. Joint Scatterplots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FM Tested on MIMIC.
This figure displays joint scatterplots with corresponding marginal distributions represented at the axes for relevant subgroup comparisons across the first four PCA modes (shown in the first two rows of plots) and the two t-SNE dimensions (shown in the last row of plots) applied to the feature embeddings extracted from CXR-FM tested on MIMIC. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), race (White, Asian, and Black), and age (grouped into eight age bins) arranged from left to right in the figure—in the first, second, third, and fourth columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

Model (MIMIC)	Mode	Explained Variance	P-Values				
			Pleural Effusion vs No Finding	White vs Asian Patients	White vs Black Patients	Asian vs Black Patients	Male vs Female Patients
CXR-FM	PCA Mode 1	19.38%	3.94E-21	1.00	2.95E-01	1.62E-02	1.72E-07
	PCA Mode 2	9.99%	2.82E-79	1.00	1.08E-01	2.38E-02	8.22E-05
	PCA Mode 3	6.77%	1.88E-08	3.71E-02	1.46E-03	4.18E-09	3.71E-02
	PCA Mode 4	4.59%	2.98E-30	4.37E-01	1.00	1.00	2.70E-01

Table 22. Statistical Comparison of Pairwise Subgroup Marginal Distributions across PCA Modes for CXR-FM Tested on MIMIC.
Two-sample Kolmogorov-Smirnov tests were conducted to compare all relevant pairwise subgroup marginal distributions for disease detection, biological sex, and race across the first four PCA modes, as indicated in the last five columns. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. These tests yielded p-values testing the null hypothesis that the distributions for each subgroup pair are identical. The Benjamini-Yekutieli procedure was then applied to adjust the p-values for multiple testing, with the significance determined at a p-value $<$ 0.05 (95% confidence level). P-values are colour-coded in the table: green if $p \geq 0.05$ (not significant), yellow if $0.001 \leq p < 0.05$ (significant), and red if $p < 0.001$ (highly significant), indicating increasing levels of statistical significance.

Figure 96. Marginal Plots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FM Tested on MIMIC [Repeated for Appendix].
This figure displays the marginal distributions for relevant subgroup comparisons across the first four PCA modes (shown in the first four rows of plots) and the two t-SNE dimensions (shown in the last two rows of plots) applied to the feature embeddings extracted from CXR-FM tested on MIMIC. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), and race (White, Asian, and Black), arranged from left to right in the figure—in the first, second, and third columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

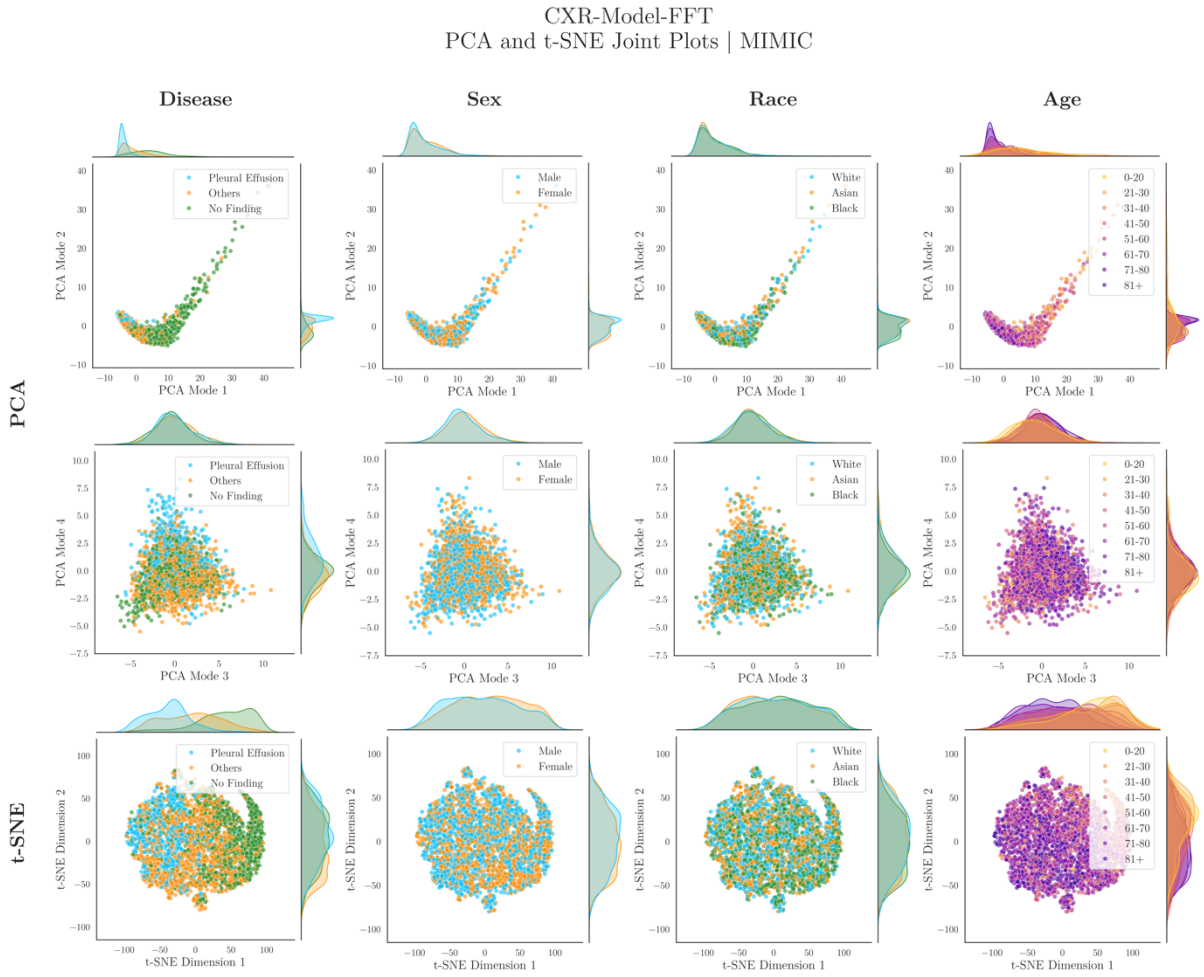

Figure 97. Joint Scatterplots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-Model FFT Tested on MIMIC.
This figure displays joint scatterplots with corresponding marginal distributions represented at the axes for relevant subgroup comparisons across the first four PCA modes (shown in the first two rows of plots) and the two t-SNE dimensions (shown in the last row of plots) applied to the feature embeddings extracted from CXR-Model FFT tested on MIMIC. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), race (White, Asian, and Black), and age (grouped into eight age bins) arranged from left to right in the figure—in the first, second, third, and fourth columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

Model (MIMIC)	Mode	Explained Variance	P-Values				
			Pleural Effusion vs No Finding	White vs Asian Patients	White vs Black Patients	Asian vs Black Patients	Male vs Female Patients
CXR-Model FFT	PCA Mode 1	42.23%	1.18E-236	1.00	5.75E-02	1.32E-02	1.80E-05
	PCA Mode 2	11.38%	4.66E-153	1.00	3.51E-04	1.80E-05	2.25E-06
	PCA Mode 3	6.07%	1.14E-01	1.00	3.91E-04	5.16E-03	2.03E-07
	PCA Mode 4	4.35%	2.85E-28	1.00	3.99E-05	5.59E-05	1.00E-01

Table 23. Statistical Comparison of Pairwise Subgroup Marginal Distributions across PCA Modes for CXR-Model FFT Tested on MIMIC.
Two-sample Kolmogorov-Smirnov tests were conducted to compare all relevant pairwise subgroup marginal distributions for disease detection, biological sex, and race across the first four PCA modes, as indicated in the last five columns. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. These tests yielded p-values testing the null hypothesis that the distributions for each subgroup pair are identical. The Benjamini-Yekutieli procedure was then applied to adjust the p-values for multiple testing, with the significance determined at a p-value $<$ 0.05 (95% confidence level). P-values are colour-coded in the table: green if $p \geq 0.05$ (not significant), yellow if $0.001 \leq p < 0.05$ (significant), and red if $p < 0.001$ (highly significant), indicating increasing levels of statistical significance.

Figure 98. Marginal Plots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-Model FFT Tested on MIMIC.
This figure displays the marginal distributions for relevant subgroup comparisons across the first four PCA modes (shown in the first four rows of plots) and the two t-SNE dimensions (shown in the last two rows of plots) applied to the feature embeddings extracted from CXR-Model FFT tested on MIMIC. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), and race (White, Asian, and Black), arranged from left to right in the figure—in the first, second, and third columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

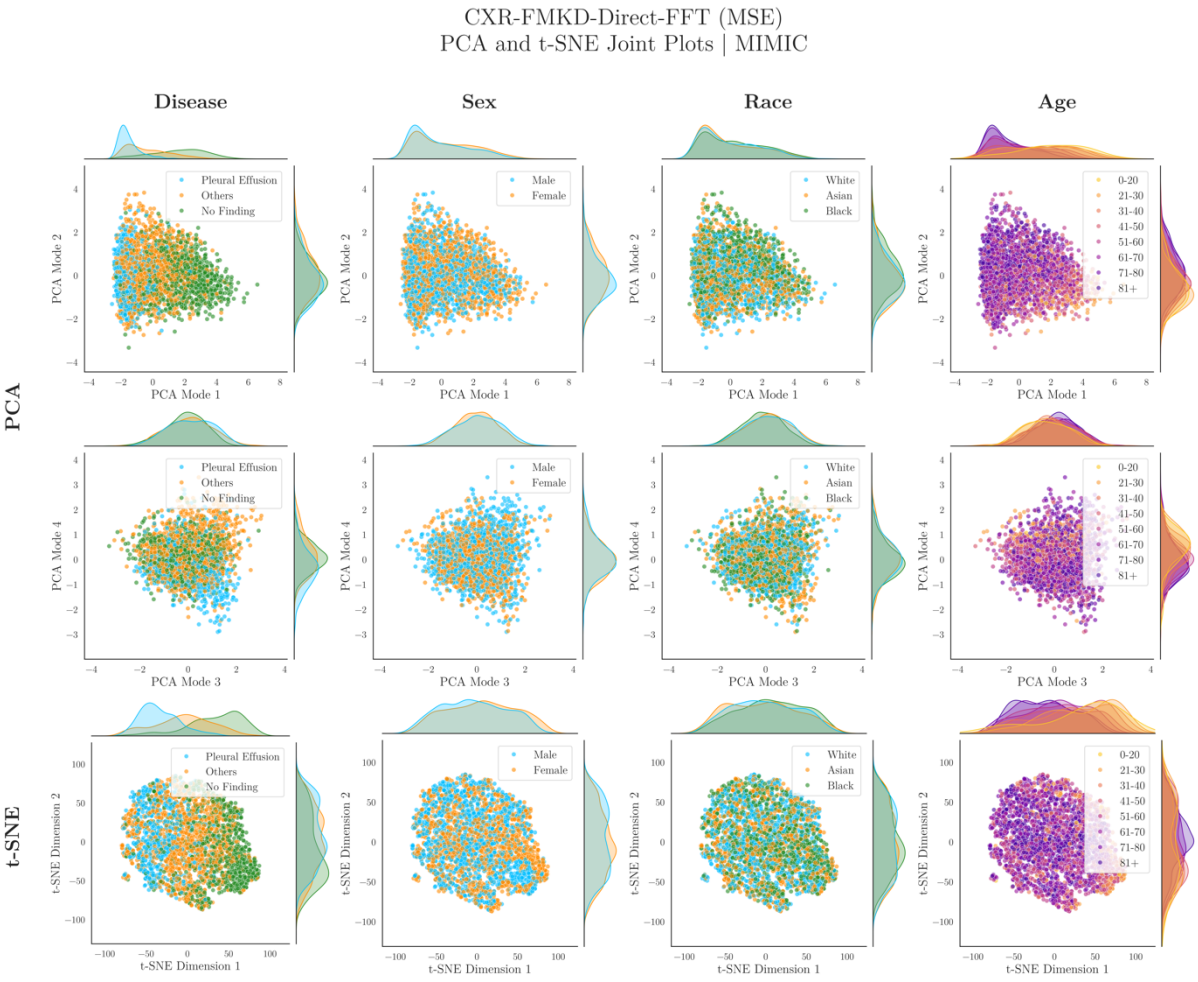

Figure 99. Joint Scatterplots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FMKD-Direct FFT (MSE) Tested on MIMIC.
This figure displays joint scatterplots with corresponding marginal distributions represented at the axes for relevant subgroup comparisons across the first four PCA modes (shown in the first two rows of plots) and the two t-SNE dimensions (shown in the last row of plots) applied to the feature embeddings extracted from CXR-FMKD-Direct FFT (MSE) tested on MIMIC. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), race (White, Asian, and Black), and age (grouped into eight age bins) arranged from left to right in the figure—in the first, second, third, and fourth columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

			P-Values				
Model (MIMIC)	**Mode**	**Explained Variance**	**Pleural Effusion vs No Finding**	**White vs Asian Patients**	**White vs Black Patients**	**Asian vs Black Patients**	**Male vs Female Patients**
CXR-FMKD-Direct FFT (MSE)	PCA Mode 1	24.79%	3.51E-236	1.00	1.04E-02	1.60E-05	1.60E-05
	PCA Mode 2	8.05%	1.00	1.00	8.91E-12	2.04E-17	1.25E-06
	PCA Mode 3	6.31%	6.99E-07	1.00	1.22E-04	2.82E-05	3.17E-03
	PCA Mode 4	4.82%	4.76E-28	1.00	6.76E-03	1.56E-03	1.72E-01

Table 24. Statistical Comparison of Pairwise Subgroup Marginal Distributions across PCA Modes for CXR-FMKD-Direct FFT (MSE) Tested on MIMIC.
Two-sample Kolmogorov-Smirnov tests were conducted to compare all relevant pairwise subgroup marginal distributions for disease detection, biological sex, and race across the first four PCA modes, as indicated in the last five columns. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. These tests yielded p-values testing the null hypothesis that the distributions for each subgroup pair are identical. The Benjamini-Yekutieli procedure was then applied to adjust the p-values for multiple testing, with the significance determined at a p-value $<$ 0.05 (95% confidence level). P-values are colour-coded in the table: green if $p \geq 0.05$ (not significant), yellow if $0.001 \leq p < 0.05$ (significant), and red if $p < 0.001$ (highly significant), indicating increasing levels of statistical significance.

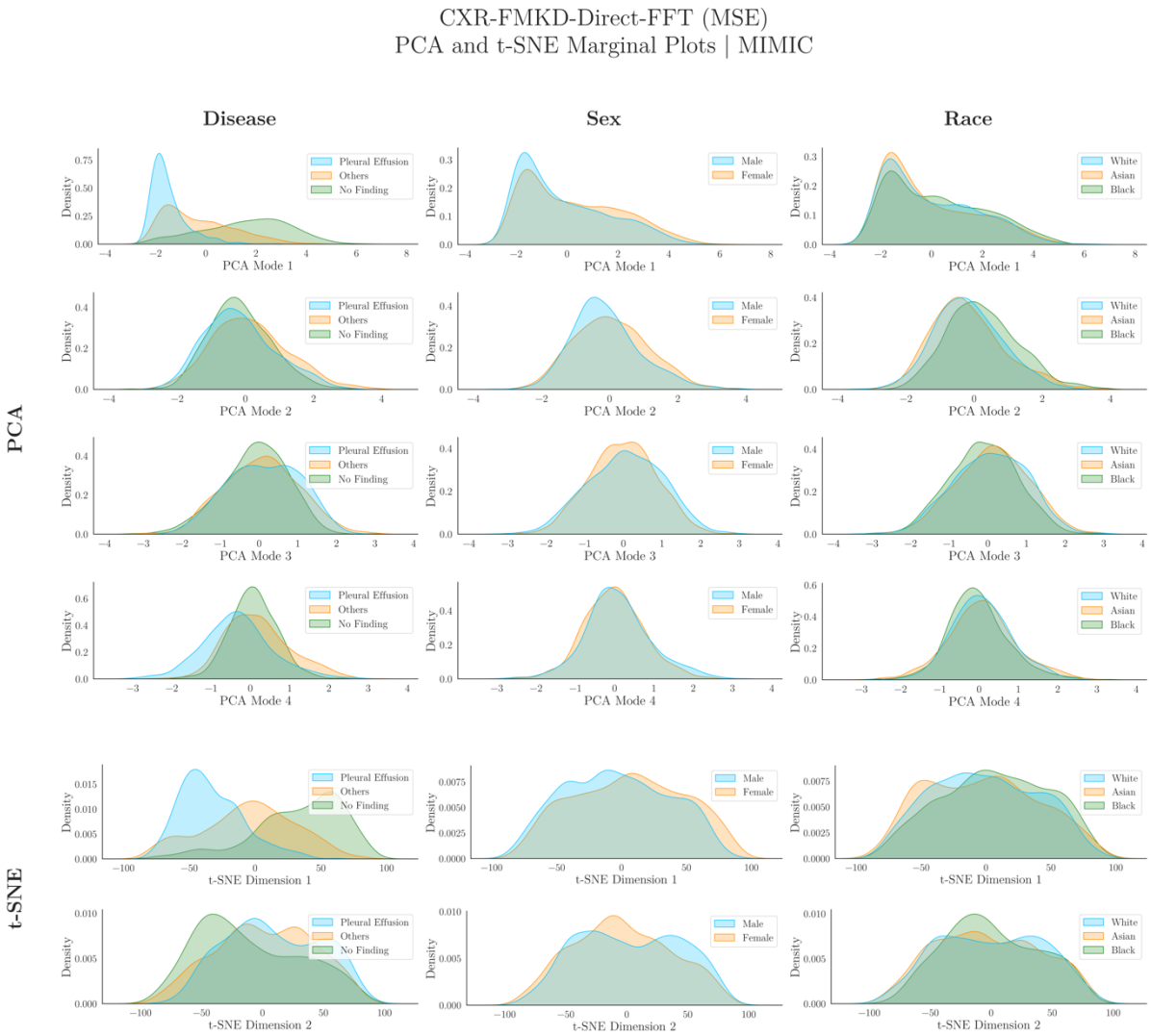

Figure 100. Marginal Plots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FMKD-Direct FFT (MSE) Tested on MIMIC [Repeated for Appendix]. This figure displays the marginal distributions for relevant subgroup comparisons across the first four PCA modes (shown in the first four rows of plots) and the two t-SNE dimensions (shown in the last two rows of plots) applied to the feature embeddings extracted from CXR-FMKD-Direct FFT (MSE) tested on MIMIC. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), and race (White, Asian, and Black), arranged from left to right in the figure—in the first, second, and third columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

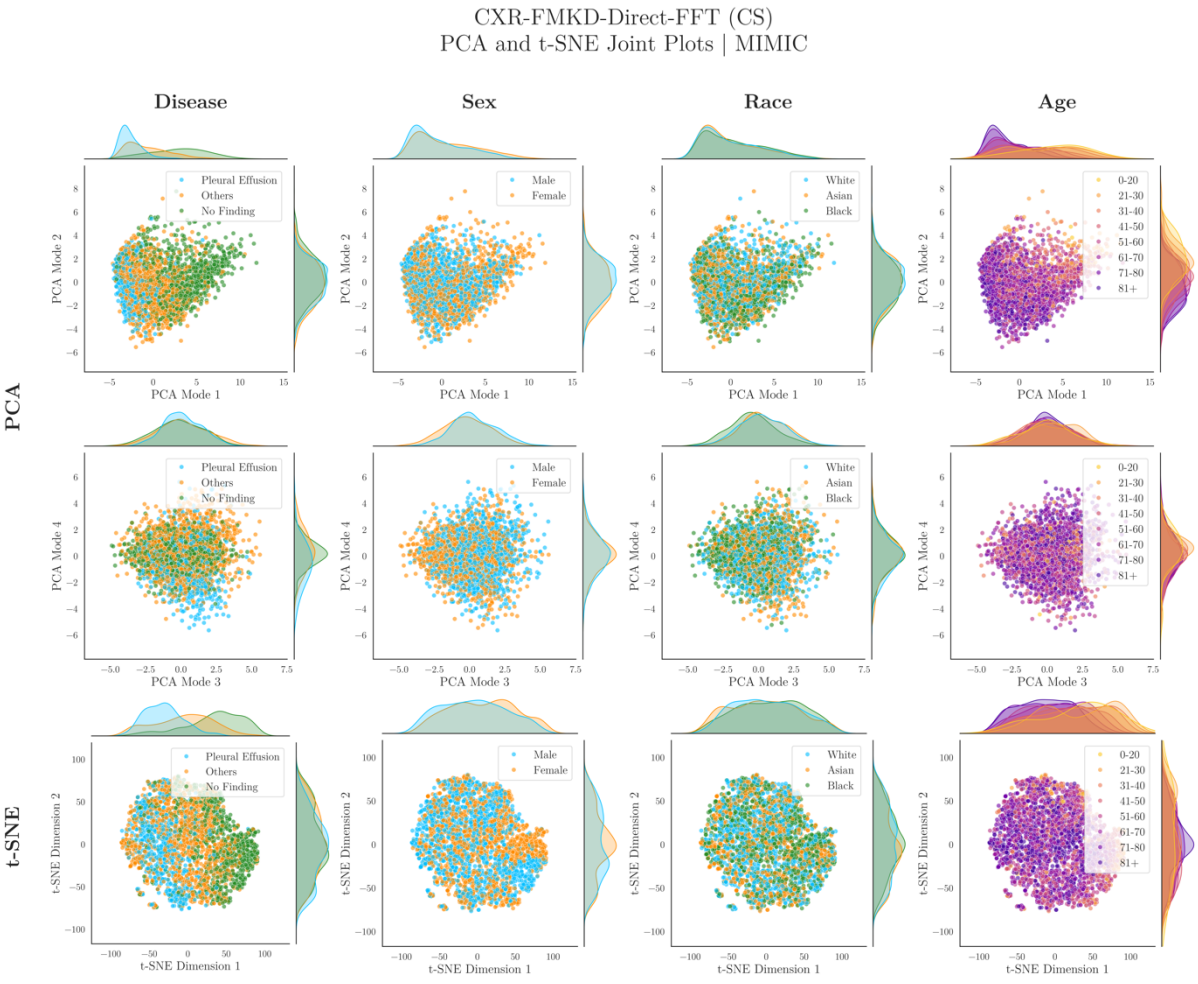

Figure 101. Joint Scatterplots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FMKD-Direct FFT (CS) Tested on MIMIC.
This figure displays joint scatterplots with corresponding marginal distributions represented at the axes for relevant subgroup comparisons across the first four PCA modes (shown in the first two rows of plots) and the two t-SNE dimensions (shown in the last row of plots) applied to the feature embeddings extracted from CXR-FMKD-Direct FFT (CS) tested on MIMIC. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), race (White, Asian, and Black), and age (grouped into eight age bins) arranged from left to right in the figure—in the first, second, third, and fourth columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

Model (MIMIC)	Mode	Explained Variance	P-Values				
			Pleural Effusion vs No Finding	White vs Asian Patients	White vs Black Patients	Asian vs Black Patients	Male vs Female Patients
CXR-FMKD-Direct FFT (CS)	PCA Mode 1	17.89%	1.25E-227	1.00	7.15E-02	2.27E-03	4.96E-08
	PCA Mode 2	5.07%	1.56E-04	6.86E-01	7.17E-04	1.40E-01	2.27E-03
	PCA Mode 3	4.52%	1.86E-05	4.58E-01	2.22E-12	7.89E-16	8.73E-11
	PCA Mode 4	3.89%	2.56E-28	1.40E-01	7.32E-03	6.56E-01	2.04E-01

Table 25. Statistical Comparison of Pairwise Subgroup Marginal Distributions across PCA Modes for CXR-FMKD-Direct FFT (CS) Tested on MIMIC.
Two-sample Kolmogorov-Smirnov tests were conducted to compare all relevant pairwise subgroup marginal distributions for disease detection, biological sex, and race across the first four PCA modes, as indicated in the last five columns. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. These tests yielded p-values testing the null hypothesis that the distributions for each subgroup pair are identical. The Benjamini-Yekutieli procedure was then applied to adjust the p-values for multiple testing, with the significance determined at a p-value $<$ 0.05 (95% confidence level). P-values are colour-coded in the table: green if $p \geq 0.05$ (not significant), yellow if $0.001 \leq p < 0.05$ (significant), and red if $p < 0.001$ (highly significant), indicating increasing levels of statistical significance.

Figure 102. Marginal Plots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FMKD-Direct FFT (CS) Tested on MIMIC.
This figure displays the marginal distributions for relevant subgroup comparisons across the first four PCA modes (shown in the first four rows of plots) and the two t-SNE dimensions (shown in the last two rows of plots) applied to the feature embeddings extracted from CXR-FMKD-Direct FFT (CS) tested on MIMIC. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), and race (White, Asian, and Black), arranged from left to right in the figure—in the first, second, and third columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

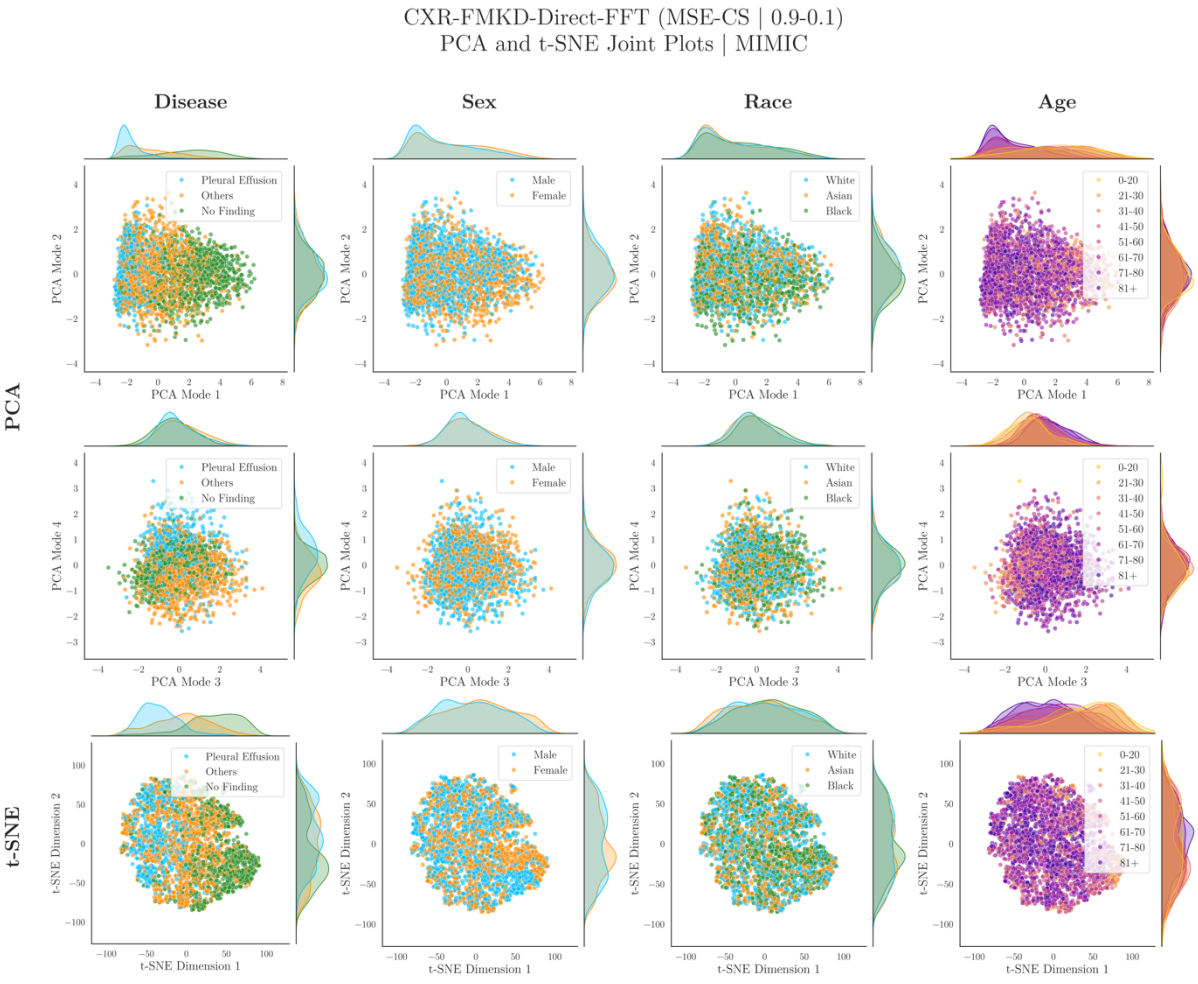

Figure 103. Joint Scatterplots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1) Tested on MIMIC.
This figure displays joint scatterplots with corresponding marginal distributions represented at the axes for relevant subgroup comparisons across the first four PCA modes (shown in the first two rows of plots) and the two t-SNE dimensions (shown in the last row of plots) applied to the feature embeddings extracted from CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1) tested on MIMIC. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), race (White, Asian, and Black), and age (grouped into eight age bins) arranged from left to right in the figure—in the first, second, third, and fourth columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

Model (MIMIC)	Mode	Explained Variance	P-Values				
			Pleural Effusion vs No Finding	White vs Asian Patients	White vs Black Patients	Asian vs Black Patients	Male vs Female Patients
CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1)	PCA Mode 1	31.70%	1.24E-244	1.00	1.23E-02	2.34E-04	5.85E-08
	PCA Mode 2	7.57%	9.80E-03	1.00	1.76E-11	1.76E-11	1.16E-03
	PCA Mode 3	7.52%	9.80E-03	1.00	1.17E-06	8.12E-07	1.43E-02
	PCA Mode 4	4.31%	2.98E-31	1.00	1.49E-02	1.07E-03	3.44E-03

Table 26. Statistical Comparison of Pairwise Subgroup Marginal Distributions across PCA Modes for CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1) Tested on MIMIC.
Two-sample Kolmogorov-Smirnov tests were conducted to compare all relevant pairwise subgroup marginal distributions for disease detection, biological sex, and race across the first four PCA modes, as indicated in the last five columns. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. These tests yielded p-values testing the null hypothesis that the distributions for each subgroup pair are identical. The Benjamini-Yekutieli procedure was then applied to adjust the p-values for multiple testing, with the significance determined at a p-value $<$ 0.05 (95% confidence level). P-values are colour-coded in the table: green if $p \geq 0.05$ (not significant), yellow if $0.001 \leq p < 0.05$ (significant), and red if $p < 0.001$ (highly significant), indicating increasing levels of statistical significance.

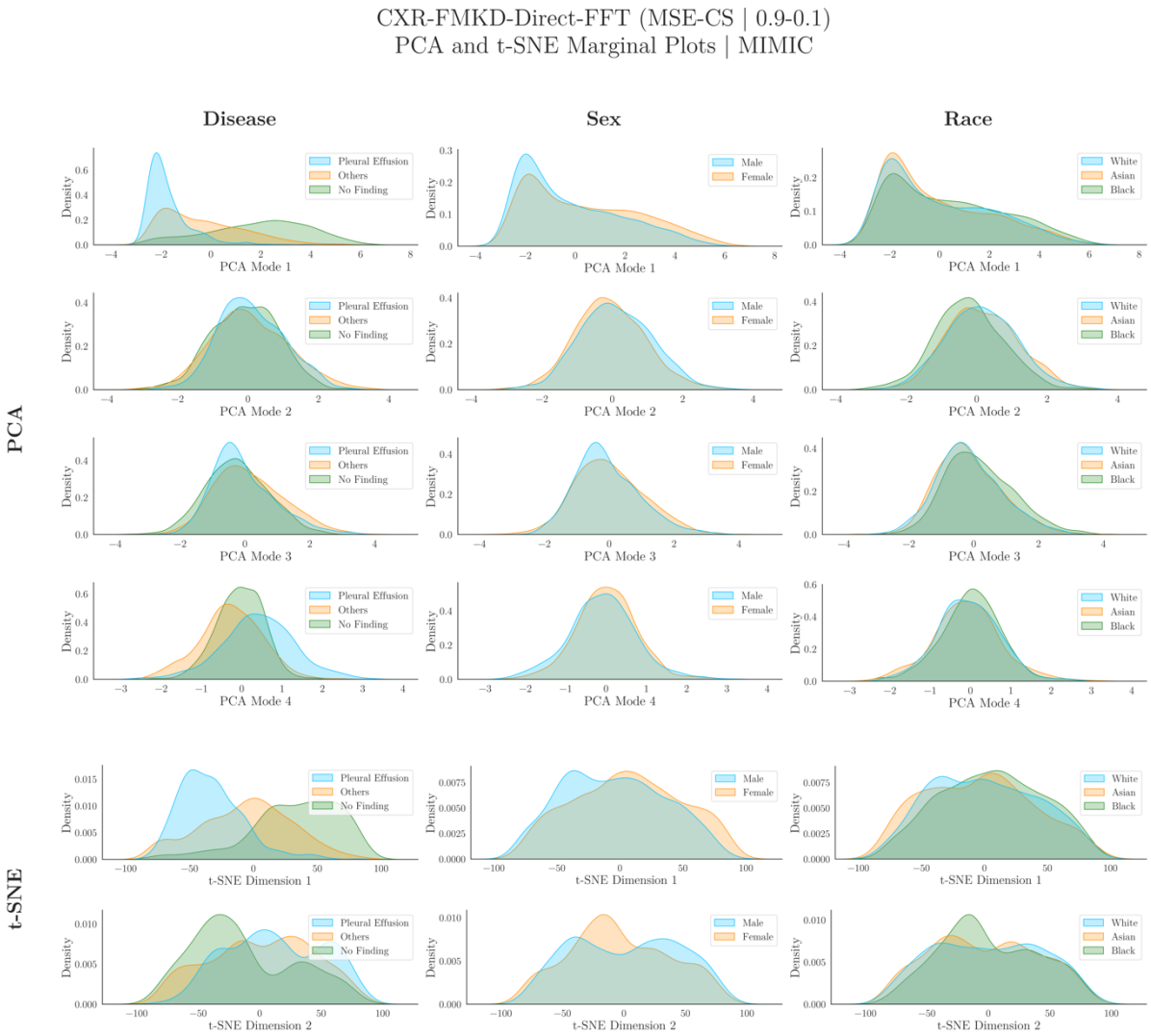

Figure 104. Marginal Plots for Inspection of Subgroup Distribution Shifts in PCA and t-SNE Feature Space Projections for CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1) Tested on MIMIC.
This figure displays the marginal distributions for relevant subgroup comparisons across the first four PCA modes (shown in the first four rows of plots) and the two t-SNE dimensions (shown in the last two rows of plots) applied to the feature embeddings extracted from CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1) tested on MIMIC. These distributions are based on a balanced random sample of 3000 patients, with 1000 selected from each racial group, ensuring no duplicates are included. The characteristics inspected include the presence of disease (categorised into 'Pleural Effusion', 'No Finding', and 'Others', which includes all 12 remaining disease labels), biological sex (Male and Female), and race (White, Asian, and Black), arranged from left to right in the figure—in the first, second, and third columns of plots, respectively. To address differences in subgroup base rates, each marginal distribution was normalised independently.

MIMIC

			Disease Detection	Race Attribute			Sex Attribute
Model (MIMIC)	**Mode**	**Explained Variance**	**Pleural Effusion vs No Finding**	**White vs Asian**	**White vs Black**	**Asian vs Black**	**Male vs Female**
CXR-FM-LP	PCA Mode 1	19.38%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 2865 [57.30%] TRUE : 1875 [37.50%] TRUE+ : 260 [5.20%]	FALSE : 4783 [95.66%] TRUE : 211 [4.22%] TRUE+ : 6 [0.12%]	FALSE : 699 [13.98%] TRUE : 2600 [52.00%] TRUE+ : 1701 [34.02%]	FALSE : 0 [0.00%] TRUE : 2 [0.04%] TRUE+ : 4998 [99.96%]
	PCA Mode 2	9.99%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 4788 [95.76%] TRUE : 206 [4.12%] TRUE+ : 6 [0.12%]	FALSE : 4556 [91.12%] TRUE : 423 [8.46%] TRUE+ : 21 [0.42%]	FALSE : 2910 [58.20%] TRUE : 1803 [36.06%] TRUE+ : 287 [5.74%]	FALSE : 18 [0.36%] TRUE : 645 [12.90%] TRUE+ : 4337 [86.74%]
	PCA Mode 3	6.77%	FALSE : 0 [0.00%] TRUE : 17 [0.34%] TRUE+ : 4983 [99.66%]	FALSE : 3902 [78.04%] TRUE : 1024 [20.48%] TRUE+ : 74 [1.48%]	FALSE : 1111 [22.22%] TRUE : 2456 [49.12%] TRUE+ : 1433 [28.66%]	FALSE : 1 [0.02%] TRUE : 217 [4.34%] TRUE+ : 4782 [95.64%]	FALSE : 3924 [78.48%] TRUE : 989 [19.78%] TRUE+ : 87 [1.74%]
	PCA Mode 4	4.59%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 3425 [68.50%] TRUE : 1420 [28.40%] TRUE+ : 155 [3.10%]	FALSE : 4636 [92.72%] TRUE : 350 [7.00%] TRUE+ : 14 [0.28%]	FALSE : 4798 [95.96%] TRUE : 198 [3.96%] TRUE+ : 4 [0.08%]	FALSE : 2678 [53.56%] TRUE : 2006 [40.12%] TRUE+ : 316 [6.32%]
CXR-Model-FFT	PCA Mode 1	42.23%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 4938 [98.76%] TRUE : 61 [1.22%] TRUE+ : 1 [0.02%]	FALSE : 4298 [85.96%] TRUE : 644 [12.88%] TRUE+ : 58 [1.16%]	FALSE : 4156 [83.12%] TRUE : 806 [16.12%] TRUE+ : 38 [0.76%]	FALSE : 789 [15.78%] TRUE : 2495 [49.90%] TRUE+ : 1716 [34.32%]
	PCA Mode 2	11.38%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 4983 [99.66%] TRUE : 17 [0.34%] TRUE+ : 0 [0.00%]	FALSE : 1687 [33.74%] TRUE : 2441 [48.82%] TRUE+ : 872 [17.44%]	FALSE : 1458 [29.16%] TRUE : 2695 [53.90%] TRUE+ : 847 [16.94%]	FALSE : 1285 [25.70%] TRUE : 2706 [54.12%] TRUE+ : 1009 [20.18%]
	PCA Mode 3	6.07%	FALSE : 2456 [49.12%] TRUE : 2441 [48.82%] TRUE+ : 103 [2.06%]	FALSE : 4987 [99.74%] TRUE : 13 [0.26%] TRUE+ : 0 [0.00%]	FALSE : 416 [8.32%] TRUE : 2191 [43.82%] TRUE+ : 2393 [47.86%]	FALSE : 649 [12.98%] TRUE : 2649 [52.98%] TRUE+ : 1702 [34.04%]	FALSE : 146 [2.92%] TRUE : 1642 [32.84%] TRUE+ : 3212 [64.24%]
	PCA Mode 4	4.35%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 4991 [99.82%] TRUE : 9 [0.18%] TRUE+ : 0 [0.00%]	FALSE : 241 [4.82%] TRUE : 1919 [38.38%] TRUE+ : 2840 [56.80%]	FALSE : 75 [1.50%] TRUE : 1696 [33.92%] TRUE+ : 3229 [64.58%]	FALSE : 2854 [57.08%] TRUE : 1944 [38.88%] TRUE+ : 202 [4.04%]
CXR-FMKD-Direct-FFT (MSE)	PCA Mode 1	24.79%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 4980 [99.60%] TRUE : 20 [0.40%] TRUE+ : 0 [0.00%]	FALSE : 3573 [71.46%] TRUE : 1243 [24.86%] TRUE+ : 184 [3.68%]	FALSE : 2844 [56.88%] TRUE : 1883 [37.66%] TRUE+ : 273 [5.46%]	FALSE : 393 [7.86%] TRUE : 2230 [44.60%] TRUE+ : 2377 [47.54%]
	PCA Mode 2	8.05%	FALSE : 4919 [98.38%] TRUE : 80 [1.60%] TRUE+ : 1 [0.02%]	FALSE : 4321 [86.42%] TRUE : 653 [13.06%] TRUE+ : 26 [0.52%]	FALSE : 0 [0.00%] TRUE : 1 [0.02%] TRUE+ : 4999 [99.98%]	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 662 [13.24%] TRUE : 2571 [51.42%] TRUE+ : 1767 [35.34%]
	PCA Mode 3	6.31%	FALSE : 12 [0.24%] TRUE : 375 [7.50%] TRUE+ : 4613 [92.26%]	FALSE : 4313 [86.26%] TRUE : 658 [13.16%] TRUE+ : 29 [0.58%]	FALSE : 3317 [66.34%] TRUE : 1469 [29.38%] TRUE+ : 214 [4.28%]	FALSE : 317 [6.34%] TRUE : 2149 [42.98%] TRUE+ : 2534 [50.68%]	FALSE : 2054 [41.08%] TRUE : 2396 [47.92%] TRUE+ : 550 [11.00%]
	PCA Mode 4	4.82%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 4714 [94.28%] TRUE : 282 [5.64%] TRUE+ : 4 [0.08%]	FALSE : 4752 [95.04%] TRUE : 242 [4.84%] TRUE+ : 6 [0.12%]	FALSE : 2819 [56.38%] TRUE : 1944 [38.88%] TRUE+ : 237 [4.74%]	FALSE : 4541 [90.82%] TRUE : 449 [8.98%] TRUE+ : 10 [0.20%]
CXR-FMKD-Direct-FFT (CS)	PCA Mode 1	17.89%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 4984 [99.68%] TRUE : 16 [0.32%] TRUE+ : 0 [0.00%]	FALSE : 3918 [78.36%] TRUE : 970 [19.40%] TRUE+ : 112 [2.24%]	FALSE : 3570 [71.40%] TRUE : 1313 [26.26%] TRUE+ : 117 [2.34%]	FALSE : 7 [0.14%] TRUE : 494 [9.88%] TRUE+ : 4499 [89.98%]
	PCA Mode 2	5.07%	FALSE : 2 [0.04%] TRUE : 448 [8.96%] TRUE+ : 4550 [91.00%]	FALSE : 4965 [99.30%] TRUE : 35 [0.70%] TRUE+ : 0 [0.00%]	FALSE : 563 [11.26%] TRUE : 2465 [49.30%] TRUE+ : 1972 [39.44%]	FALSE : 2244 [44.88%] TRUE : 2372 [47.44%] TRUE+ : 384 [7.68%]	FALSE : 2751 [55.02%] TRUE : 1979 [39.58%] TRUE+ : 270 [5.40%]
	PCA Mode 3	4.52%	FALSE : 836 [16.72%] TRUE : 3153 [63.06%] TRUE+ : 1011 [20.22%]	FALSE : 2440 [48.80%] TRUE : 2254 [45.08%] TRUE+ : 306 [6.12%]	FALSE : 51 [1.02%] TRUE : 964 [19.28%] TRUE+ : 3985 [79.70%]	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 0 [0.00%] TRUE : 2 [0.04%] TRUE+ : 4998 [99.96%]
	PCA Mode 4	3.89%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 3393 [67.86%] TRUE : 1460 [29.20%] TRUE+ : 147 [2.94%]	FALSE : 62 [1.24%] TRUE : 798 [15.96%] TRUE+ : 4140 [82.80%]	FALSE : 2636 [52.72%] TRUE : 2116 [42.32%] TRUE+ : 248 [4.96%]	FALSE : 4739 [94.78%] TRUE : 256 [5.12%] TRUE+ : 5 [0.10%]
CXR-FMKD-Direct-FFT (MSE-CS | 0.9-0.1)	PCA Mode 1	31.70%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 4984 [99.68%] TRUE : 16 [0.32%] TRUE+ : 0 [0.00%]	FALSE : 2394 [47.88%] TRUE : 2070 [41.40%] TRUE+ : 536 [10.72%]	FALSE : 1936 [38.72%] TRUE : 2518 [50.36%] TRUE+ : 546 [10.92%]	FALSE : 13 [0.26%] TRUE : 471 [9.42%] TRUE+ : 4516 [90.32%]
	PCA Mode 2	7.57%	FALSE : 3832 [76.64%] TRUE : 1130 [22.60%] TRUE+ : 38 [0.76%]	FALSE : 4703 [94.06%] TRUE : 287 [5.74%] TRUE+ : 10 [0.20%]	FALSE : 126 [2.52%] TRUE : 1168 [23.36%] TRUE+ : 3706 [74.12%]	FALSE : 0 [0.00%] TRUE : 61 [1.22%] TRUE+ : 4939 [98.78%]	FALSE : 651 [13.02%] TRUE : 2376 [47.52%] TRUE+ : 1973 [39.46%]
	PCA Mode 3	7.52%	FALSE : 190 [3.80%] TRUE : 2936 [58.72%] TRUE+ : 1874 [37.48%]	FALSE : 4793 [95.86%] TRUE : 204 [4.08%] TRUE+ : 3 [0.06%]	FALSE : 20 [0.40%] TRUE : 653 [13.06%] TRUE+ : 4327 [86.54%]	FALSE : 0 [0.00%] TRUE : 39 [0.78%] TRUE+ : 4961 [99.22%]	FALSE : 4193 [83.86%] TRUE : 778 [15.56%] TRUE+ : 29 [0.58%]
	PCA Mode 4	4.31%	FALSE : 0 [0.00%] TRUE : 0 [0.00%] TRUE+ : 5000 [100.00%]	FALSE : 4480 [89.60%] TRUE : 504 [10.08%] TRUE+ : 16 [0.32%]	FALSE : 4899 [97.98%] TRUE : 100 [2.00%] TRUE+ : 1 [0.02%]	FALSE : 3721 [74.42%] TRUE : 1195 [23.90%] TRUE+ : 84 [1.68%]	FALSE : 1360 [27.20%] TRUE : 2765 [55.30%] TRUE+ : 875 [17.50%]

Table 27. Results from Bootstrapping-like Simulations for Statistical Comparison of Pairwise Subgroup Marginal Distributions across PCA Modes for Selected Models Tested on MIMIC.

Utilising a bootstrapping-like approach, this table presents the outcomes of two-sample Kolmogorov-Smirnov tests performed to compare all relevant pairwise subgroup marginal distributions for disease detection, biological sex, and race across the first four PCA modes, as indicated in the last five columns, for our five selected models tested on MIMIC. Each test was replicated across 5000 simulations, each based on a balanced stratified sample of 3000 patients—1000 from each racial group with no duplicates—to ensure robustness and mitigate sampling variability. These tests yielded p-values testing the null hypothesis that distributions between each subgroup pair are identical. The Benjamini-Yekutieli procedure was applied to adjust the p-values for multiple testing, with significance determined at a p-value < 0.05 (95% confidence level). Results of these simulations were categorised into 'FALSE' ($p \geq 0.05$), 'TRUE' ($0.001 \leq p < 0.05$), and 'TRUE+' ($p < 0.001$) to help quantify bias, indicating varying levels of statistical significance.

S.8. Bias Analysis | Subgroup Performance Analysis – CheXpert

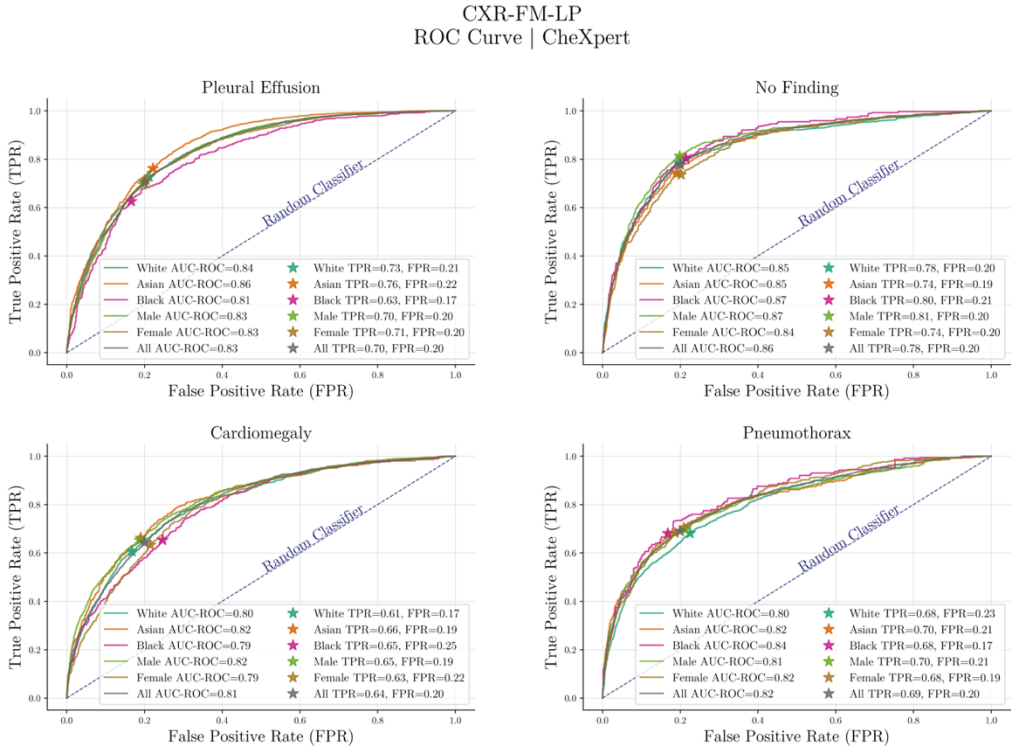

Figure 105. ROC Performance Across Subgroups for CXR-FM Tested on CheXpert.
This figure displays ROC curves for the detection of 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax' using CXR-FM tested on the resampled CheXpert test set. TPR and FPR for each subgroup are determined at a fixed decision threshold, optimised to achieve an FPR of 20% across the entire patient sample. The shifts in TPR/FPR across subgroups highlight disparities in disease detection performance.

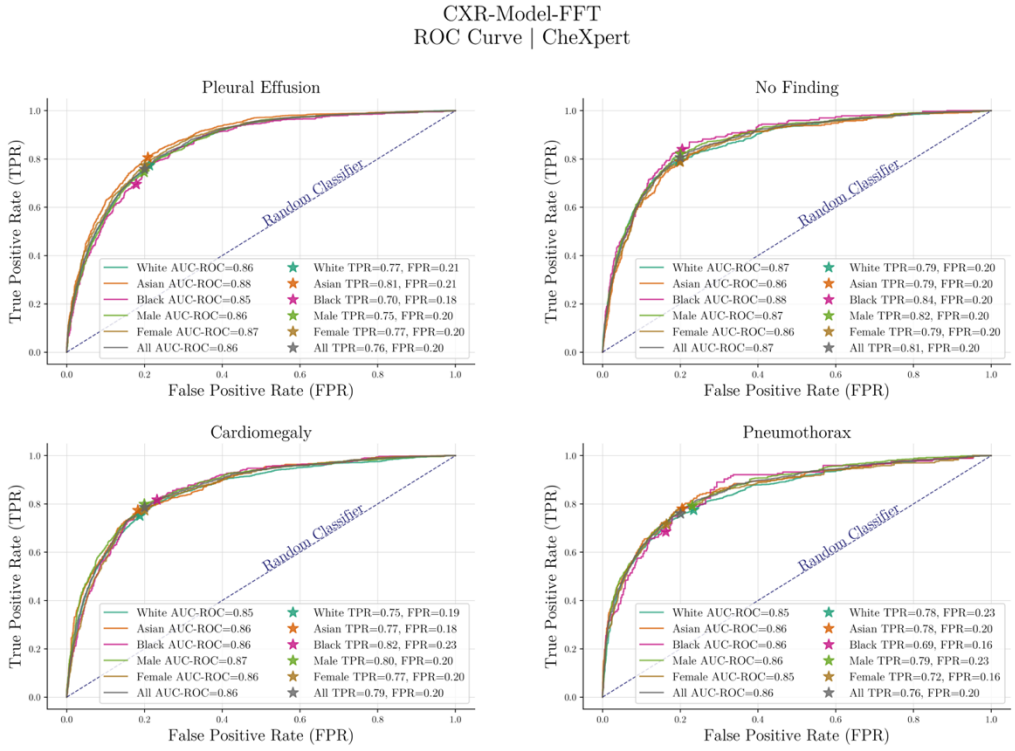

Figure 106. ROC Performance Across Subgroups for CXR-Model FFT Tested on CheXpert.
This figure displays ROC curves for the detection of 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax' using CXR-Model FFT tested on the resampled CheXpert test set. TPR and FPR for each subgroup are determined at a fixed decision threshold, optimised to achieve an FPR of 20% across the entire patient sample. The shifts in TPR/FPR across subgroups highlight disparities in disease detection performance.

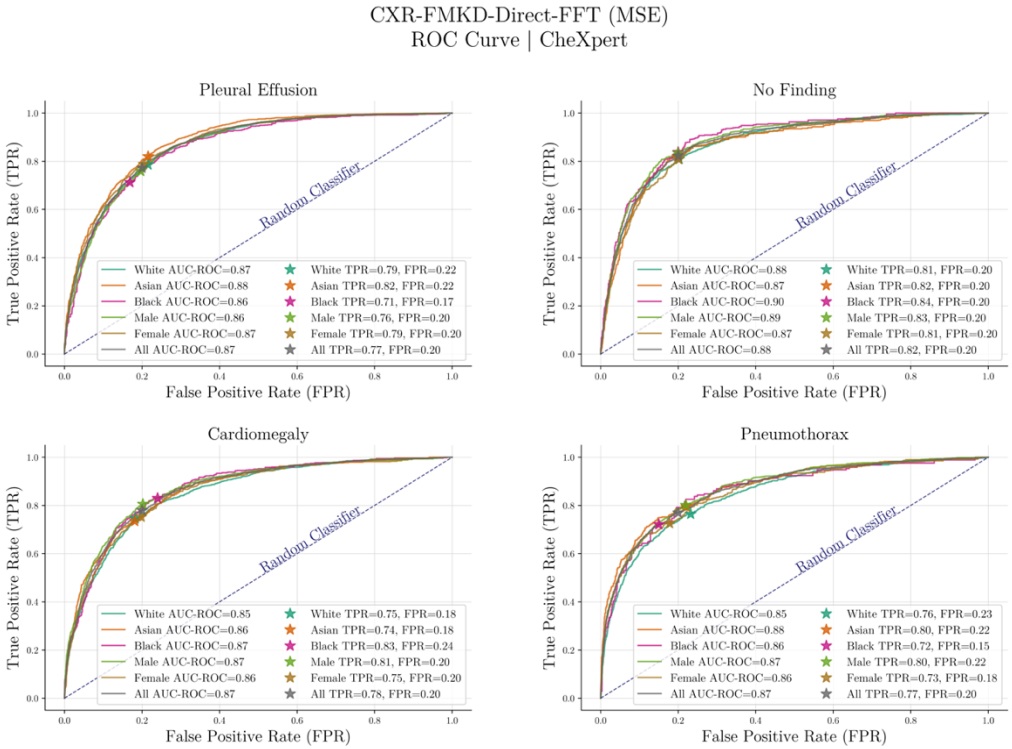

Figure 107. ROC Performance Across Subgroups for the Selected (MSE)-Student Tested on CheXpert. This figure displays ROC curves for the detection of 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax' using CXR-FMKD-Direct FFT (MSE) tested on the resampled CheXpert test set. TPR and FPR for each subgroup are determined at a fixed decision threshold, optimised to achieve an FPR of 20% across the entire patient sample. The shifts in TPR/FPR across subgroups highlight disparities in disease detection performance.

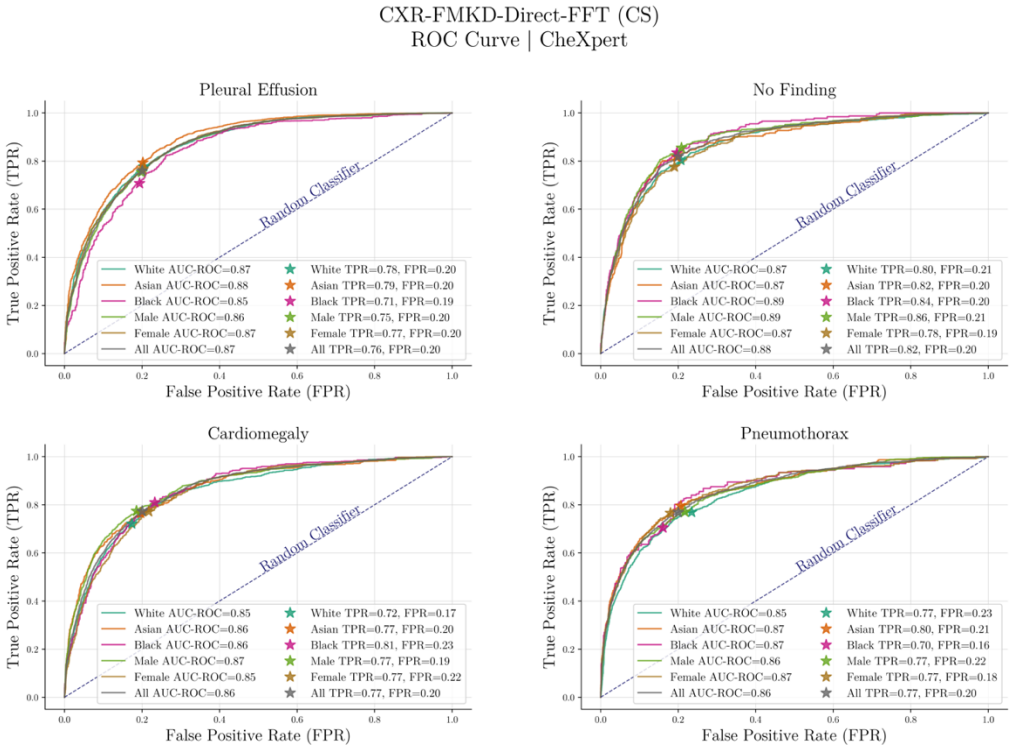

Figure 108. ROC Performance Across Subgroups for the Selected (CS)-Student Tested on CheXpert. This figure displays ROC curves for the detection of 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax' using CXR-FMKD-Direct FFT (CS) tested on the resampled CheXpert test set. TPR and FPR for each subgroup are determined at a fixed decision threshold, optimised to achieve an FPR of 20% across the entire patient sample. The shifts in TPR/FPR across subgroups highlight disparities in disease detection performance.

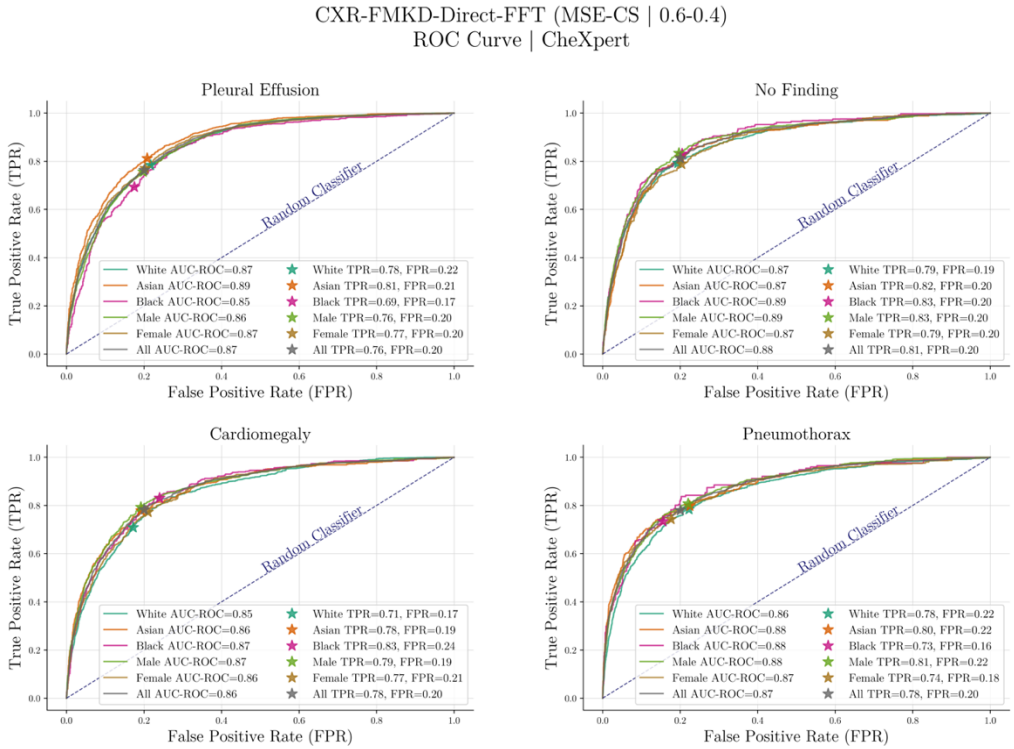

Figure 109. ROC Performance Across Subgroups for the Selected (MSE-CS | 0.6-0.4)-Student Tested on CheXpert.

This figure displays ROC curves for the detection of 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax' using CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4) tested on the resampled CheXpert test set. TPR and FPR for each subgroup are determined at a fixed decision threshold, optimised to achieve an FPR of 20% across the entire patient sample. The shifts in TPR/FPR across subgroups highlight disparities in disease detection performance.

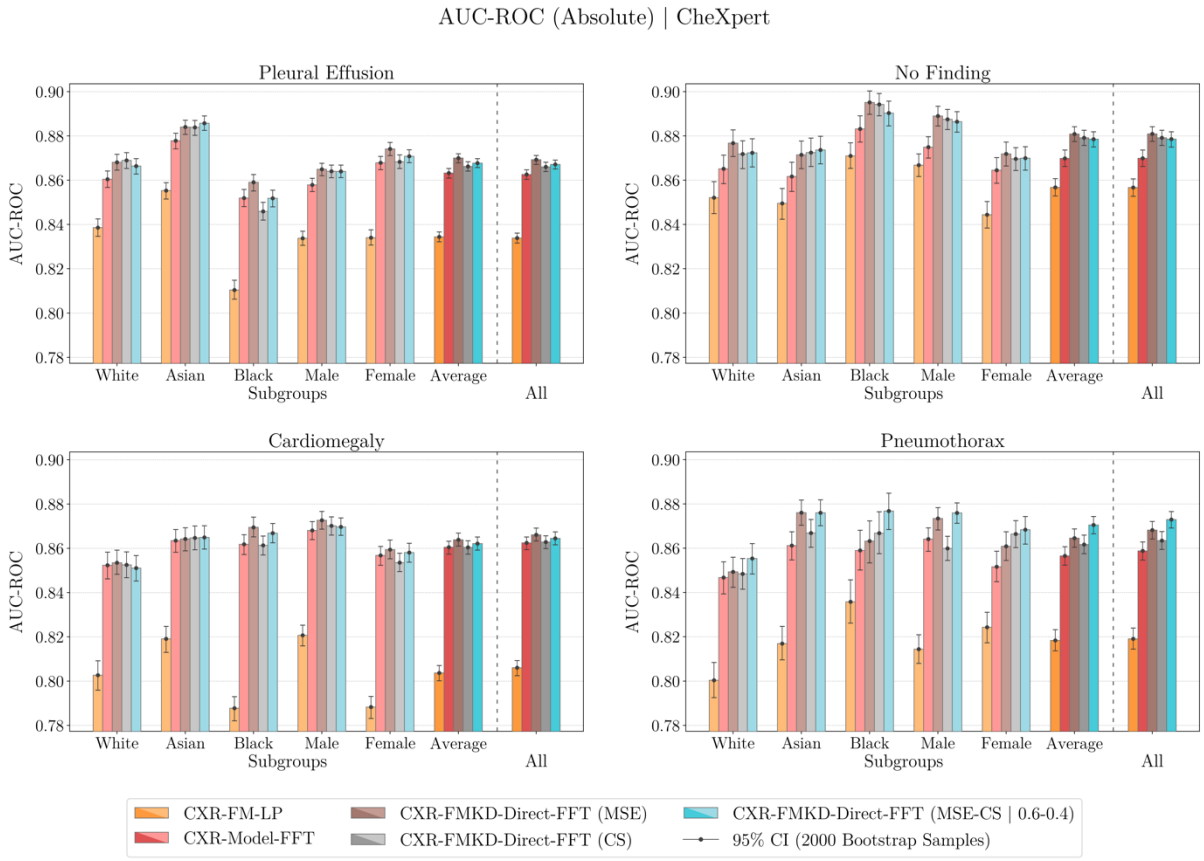

Figure 110. Comparison of AUC-ROC Disease Detection Performance Across Patient Subgroups for Selected Models Tested on CheXpert.

This figure illustrates the mean AUC-ROC values, depicted by dots, and their corresponding 95% confidence intervals (CIs), shown with whiskers and derived from bootstrapping with 2000 samples, across the relevant patient subgroups (for race and biological sex), their average, and the entire patient sample—denoted by 'All'—for our five selected models developed and tested on CheXpert. These models include the teacher CXR-FM; the CXR-FMKD-Direct FFT variants of the three selected student model types (MSE, CS, and MSE-CS | 0.6-0.4); and the traditional baseline CXR-Model FFT, which shares the same DenseNet169 architecture as the students but was developed without Knowledge Distillation (KD) from CXR-FM. The models were assessed for their ability (average absolute classification performance) to detect 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax'. The teacher CXR-FM consistently underperforms compared to the other models, with the student models exhibiting the highest performances. Performance disparities can be observed across subgroups, particularly pronounced for the teacher CXR-FM which shows a notable underperformance in the Black patient subgroup for 'Pleural Effusion' and in both Black and Female subgroups for 'Cardiomegaly'. Additionally, the teacher model shows a significant drop in performance for 'Cardiomegaly' compared to other disease labels.

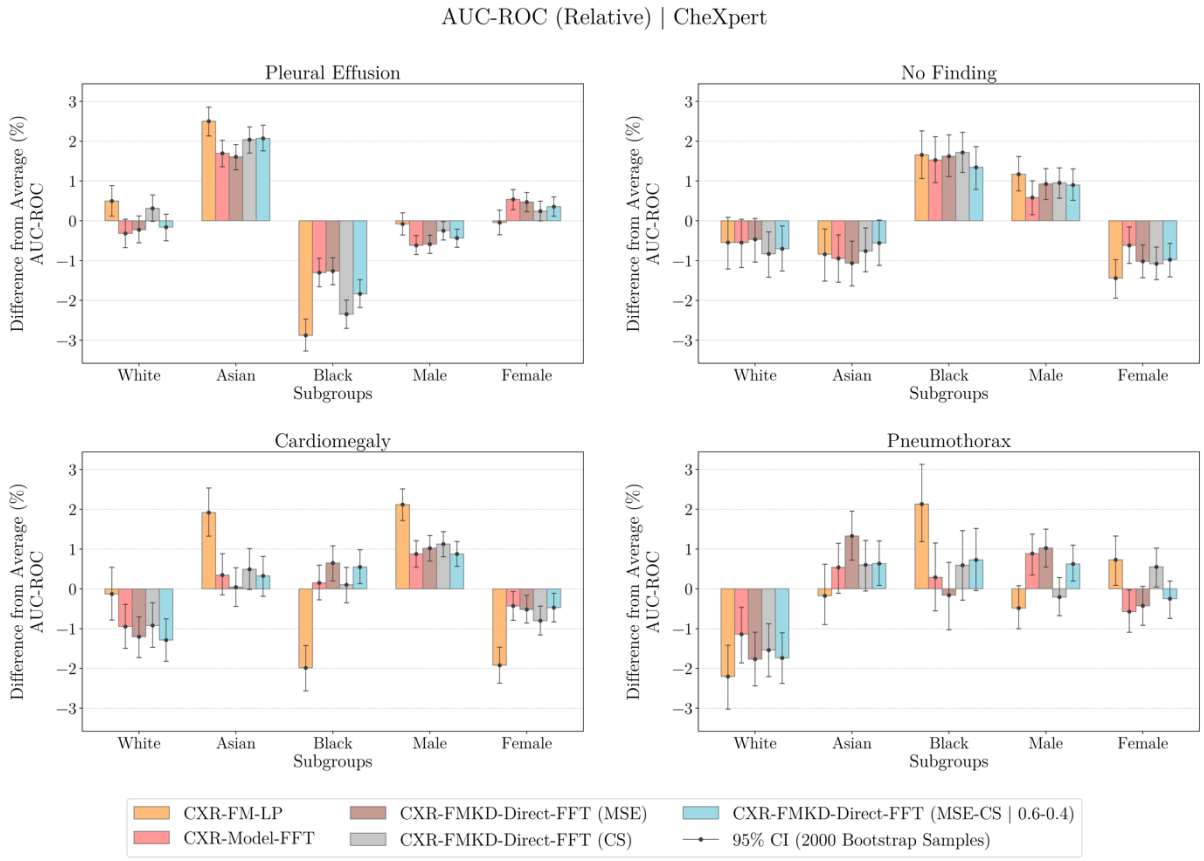

Figure 111. Relative Change in AUC-ROC Disease Detection Performance Across Patient Subgroups for Selected Models Tested on CheXpert.

This figure illustrates the mean relative changes in AUC-ROC performance, depicted by dots, and their corresponding 95% confidence intervals (CIs), shown with whiskers and derived from bootstrapping with 2000 samples, across the relevant patient subgroups (for race and biological sex) for our five selected models developed and tested on CheXpert. For each model, the relative performance change for each subgroup for a specific disease label was computed by comparing the subgroup's performance (Subgroup Value) with the average performance across all subgroups (Average Value) for that label using the formula: (Subgroup Value – Average Value)/ Average Value × 100%. The models evaluated include the teacher CXR-FM; the CXR-FMKD-Direct FFT variants of the three selected student model types (MSE, CS, and MSE-CS | 0.6-0.4); and the traditional baseline CXR-Model FFT, which shares the same DenseNet169 architecture as the students but was developed without Knowledge Distillation (KD) from CXR-FM. These models were assessed for their ability (average relative classification performance) to detect 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax'. Disparities in relative performance can be observed across subgroups, particularly pronounced for the teacher CXR-FM.

CheXpert – Pleural Effusion

Model	Metric	Subgroups						All
		White	Asian	Black	Male	Female	*Average**	
CXR-FM	AUC-ROC	0.84 (0.83-0.84)	0.86 (0.85-0.86)	0.81 (0.81-0.81)	0.83 (0.83-0.84)	0.83 (0.83-0.84)	0.83 (0.83-0.84)	0.83 (0.83-0.84)
	TPR at global threshold	0.73 (0.72-0.73)	0.76 (0.75-0.77)	0.63 (0.62-0.64)	0.70 (0.69-0.70)	0.71 (0.71-0.72)	0.71 (0.70-0.71)	0.70 (0.70-0.71)
	FPR at global threshold	0.21 (0.21-0.22)	0.22 (0.22-0.23)	0.17 (0.16-0.17)	0.20 (0.19-0.20)	0.20 (0.20-0.21)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.52 (0.51-0.52)	0.54 (0.53-0.55)	0.46 (0.45-0.47)	0.50 (0.49-0.51)	0.51 (0.50-0.52)	0.51 (0.50-0.51)	0.50 (0.50-0.51)
	Relative AUC-ROC	0.50 (0.12-0.88)	2.50 (2.13-2.85)	-2.88 (-3.27--2.47)	-0.08 (-0.36-0.20)	-0.05 (-0.35-0.27)	0.00 (0.00-0.00)	-0.06 (-0.09--0.03)
	Relative TPR at global threshold	3.13 (2.22-3.93)	8.05 (7.20-8.84)	-11.23 (-12.02--10.27)	-1.28 (-1.86--0.66)	1.34 (0.67-1.93)	0.00 (0.00-0.00)	-0.04 (-0.09-0.01)
	Relative FPR at global threshold	5.85 (3.68-7.96)	10.93 (8.63-13.01)	-17.04 (-19.09--14.78)	-1.73 (-3.35--0.35)	1.99 (0.45-3.80)	0.00 (0.00-0.00)	-0.04 (-0.13-0.03)
	Relative Youden's Index at global threshold	2.05 (0.56-3.46)	6.90 (5.49-8.34)	-8.93 (-10.23--7.43)	-1.10 (-2.07-0.06)	1.08 (-0.19-2.11)	0.00 (0.00-0.00)	-0.04 (-0.11-0.04)
CXR-Model FFT	AUC-ROC	0.86 (0.86-0.86)	0.88 (0.87-0.88)	0.85 (0.85-0.86)	0.86 (0.85-0.86)	0.87 (0.86-0.87)	0.86 (0.86-0.87)	0.86 (0.86-0.86)
	TPR at global threshold	0.77 (0.77-0.78)	0.81 (0.80-0.81)	0.70 (0.69-0.70)	0.75 (0.74-0.75)	0.77 (0.77-0.78)	0.76 (0.75-0.76)	0.76 (0.75-0.76)
	FPR at global threshold	0.21 (0.21-0.22)	0.21 (0.20-0.21)	0.18 (0.17-0.18)	0.20 (0.20-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.56 (0.55-0.57)	0.60 (0.59-0.61)	0.52 (0.51-0.53)	0.55 (0.54-0.55)	0.57 (0.56-0.58)	0.56 (0.55-0.56)	0.56 (0.55-0.56)
	Relative AUC-ROC	-0.32 (-0.68-0.04)	1.70 (1.36-2.02)	-1.30 (-1.66--0.94)	-0.62 (-0.85--0.37)	0.54 (0.28-0.79)	0.00 (0.00-0.00)	-0.07 (-0.09--0.05)
	Relative TPR at global threshold	1.98 (1.23-2.68)	6.23 (5.59-7.02)	-8.30 (-9.11--7.57)	-1.75 (-2.27--1.31)	1.84 (1.37-2.39)	0.00 (0.00-0.00)	-0.05 (-0.08--0.01)
	Relative FPR at global threshold	6.31 (4.29-8.48)	4.30 (2.10-6.68)	-10.71 (-13.18--8.56)	-0.41 (-1.90-1.03)	0.51 (-1.10-2.17)	0.00 (0.00-0.00)	0.01 (-0.06-0.08)
	Relative Youden's Index at global threshold	0.43 (-0.84-1.64)	6.92 (5.75-8.22)	-7.44 (-8.78--6.14)	-2.23 (-3.11--1.39)	2.31 (1.38-3.24)	0.00 (0.00-0.00)	-0.07 (-0.12--0.02)
CXR-FMKD-Direct FFT (MSE)	AUC-ROC	0.87 (0.86-0.87)	0.88 (0.88-0.89)	0.86 (0.86-0.86)	0.86 (0.86-0.87)	0.87 (0.87-0.88)	0.87 (0.87-0.87)	0.87 (0.87-0.87)
	TPR at global threshold	0.79 (0.78-0.79)	0.82 (0.81-0.83)	0.71 (0.70-0.72)	0.76 (0.75-0.76)	0.79 (0.78-0.79)	0.77 (0.77-0.78)	0.77 (0.77-0.78)
	FPR at global threshold	0.22 (0.21-0.22)	0.22 (0.21-0.22)	0.17 (0.17-0.17)	0.20 (0.19-0.20)	0.20 (0.20-0.21)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.57 (0.56-0.58)	0.61 (0.60-0.61)	0.54 (0.53-0.55)	0.56 (0.55-0.57)	0.58 (0.58-0.59)	0.57 (0.57-0.58)	0.57 (0.57-0.58)
	Relative AUC-ROC	-0.22 (-0.55-0.12)	1.61 (1.28-1.92)	-1.26 (-1.60--0.93)	-0.59 (-0.82--0.36)	0.47 (0.23-0.71)	0.00 (0.00-0.00)	-0.09 (-0.11--0.07)
	Relative TPR at global threshold	1.66 (0.94-2.34)	6.10 (5.46-6.89)	-7.85 (-8.68--7.19)	-1.84 (-2.39--1.41)	1.94 (1.49-2.52)	0.00 (0.00-0.00)	-0.05 (-0.08--0.02)
	Relative FPR at global threshold	7.68 (5.12-9.52)	7.74 (5.50-9.86)	-15.69 (-17.38--13.16)	-1.68 (-3.01--0.13)	1.94 (0.19-3.42)	0.00 (0.00-0.00)	-0.04 (-0.10-0.05)
	Relative Youden's Index at global threshold	-0.44 (-1.56-0.90)	5.53 (4.45-6.84)	-5.12 (-6.57--4.09)	-1.90 (-2.84--1.21)	1.93 (1.17-2.94)	0.00 (0.00-0.00)	-0.05 (-0.11--0.00)
CXR-FMKD-Direct FFT (CS)	AUC-ROC	0.87 (0.87-0.87)	0.88 (0.88-0.89)	0.85 (0.84-0.85)	0.86 (0.86-0.87)	0.87 (0.87-0.87)	0.87 (0.86-0.87)	0.87 (0.86-0.87)
	TPR at global threshold	0.78 (0.77-0.78)	0.79 (0.79-0.80)	0.71 (0.70-0.72)	0.75 (0.75-0.76)	0.77 (0.76-0.77)	0.76 (0.76-0.77)	0.76 (0.75-0.77)
	FPR at global threshold	0.20 (0.20-0.21)	0.20 (0.20-0.21)	0.19 (0.19-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.21)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.57 (0.56-0.58)	0.59 (0.58-0.60)	0.51 (0.51-0.53)	0.56 (0.55-0.56)	0.56 (0.56-0.57)	0.56 (0.55-0.57)	0.56 (0.55-0.57)
	Relative AUC-ROC	0.31 (-0.02-0.65)	2.04 (1.70-2.36)	-2.34 (-2.70--1.99)	-0.25 (-0.48--0.02)	0.24 (-0.00-0.49)	0.00 (0.00-0.00)	-0.02 (-0.04--0.00)
	Relative TPR at global threshold	2.33 (1.45-2.90)	4.44 (3.82-5.28)	-6.80 (-7.55--6.02)	-0.72 (-1.33--0.27)	0.75 (0.27-1.39)	0.00 (0.00-0.00)	-0.02 (-0.05-0.01)
	Relative FPR at global threshold	2.18 (0.02-4.49)	0.94 (-1.22-3.14)	-3.26 (-5.47--1.22)	-0.96 (-2.52-0.48)	1.09 (-0.53-2.81)	0.00 (0.00-0.00)	-0.03 (-0.09-0.03)
	Relative Youden's Index at global threshold	2.38 (0.77-3.47)	5.69 (4.53-7.14)	-8.07 (-9.28--6.66)	-0.63 (-1.62-0.20)	0.63 (-0.27-1.69)	0.00 (0.00-0.00)	-0.02 (-0.06-0.03)
CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4)	AUC-ROC	0.87 (0.86-0.87)	0.89 (0.88-0.89)	0.85 (0.85-0.86)	0.86 (0.86-0.87)	0.87 (0.87-0.87)	0.87 (0.87-0.87)	0.87 (0.87-0.87)
	TPR at global threshold	0.78 (0.78-0.79)	0.81 (0.80-0.82)	0.69 (0.69-0.70)	0.76 (0.75-0.76)	0.77 (0.77-0.78)	0.76 (0.76-0.77)	0.76 (0.76-0.77)
	FPR at global threshold	0.22 (0.21-0.22)	0.21 (0.20-0.21)	0.17 (0.17-0.18)	0.20 (0.19-0.20)	0.20 (0.20-0.21)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.57 (0.56-0.57)	0.60 (0.60-0.61)	0.52 (0.51-0.53)	0.56 (0.55-0.56)	0.57 (0.56-0.57)	0.56 (0.56-0.57)	0.56 (0.56-0.57)
	Relative AUC-ROC	-0.16 (-0.50-0.17)	2.07 (1.76-2.40)	-1.83 (-2.18--1.48)	-0.43 (-0.67--0.21)	0.35 (0.12-0.60)	0.00 (0.00-0.00)	-0.07 (-0.09--0.05)
	Relative TPR at global threshold	2.72 (2.04-3.41)	6.41 (5.69-7.11)	-9.17 (-9.87--8.40)	-0.96 (-1.54--0.48)	1.00 (0.50-1.61)	0.00 (0.00-0.00)	-0.03 (-0.07-0.01)
	Relative FPR at global threshold	8.40 (5.96-10.44)	3.97 (1.85-6.15)	-12.58 (-14.61--10.28)	-1.20 (-2.63-0.32)	1.41 (-0.28-3.00)	0.00 (0.00-0.00)	-0.02 (-0.09-0.06)
	Relative Youden's Index at global threshold	0.70 (-0.50-1.91)	7.27 (6.10-8.50)	-7.95 (-9.24--6.67)	-0.87 (-1.81--0.11)	0.85 (0.02-1.84)	0.00 (0.00-0.00)	-0.04 (-0.09-0.02)

**Relative values are computed with respect to the Subgroups' Average*

***All data are presented as: value (95% CI), derived from bootstrapping with 2000 samples*

Table 28. Detailed Absolute and Relative Subgroup Performances for 'Pleural Effusion' in CheXpert. This table reports disease detection performance for our five selected models across race and sex subgroups, their average, and the total patient sample, denoted 'All'. Data is presented as mean and 95% confidence interval (CI). TPR and FPR for each subgroup are determined at a fixed (global) decision threshold, optimised to achieve an FPR of 20% across the entire patient sample.

CheXpert – No Finding

Model	Metric	Subgroups						All
		White	Asian	Black	Male	Female	*Average**	
CXR-FM	AUC-ROC	0.85 (0.84-0.86)	0.85 (0.84-0.86)	0.87 (0.87-0.88)	0.87 (0.86-0.87)	0.84 (0.84-0.85)	0.86 (0.85-0.86)	0.86 (0.85-0.86)
	TPR at global threshold	0.78 (0.77-0.80)	0.74 (0.73-0.76)	0.80 (0.79-0.82)	0.81 (0.80-0.82)	0.74 (0.72-0.75)	0.78 (0.77-0.79)	0.78 (0.77-0.79)
	FPR at global threshold	0.20 (0.19-0.20)	0.19 (0.19-0.19)	0.21 (0.21-0.22)	0.20 (0.20-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.59 (0.57-0.60)	0.55 (0.54-0.57)	0.59 (0.58-0.60)	0.61 (0.60-0.62)	0.53 (0.52-0.55)	0.58 (0.57-0.59)	0.58 (0.57-0.59)
	Relative AUC-ROC	-0.54 (-1.21-0.09)	-0.84 (-1.51--0.21)	1.66 (1.06-2.26)	1.17 (0.75-1.62)	-1.44 (-1.94--0.97)	0.00 (0.00-0.00)	-0.01 (-0.04-0.03)
	Relative TPR at global threshold	1.08 (-0.47-2.51)	-4.13 (-5.64--2.53)	3.53 (2.01-5.03)	4.69 (3.59-5.67)	-5.17 (-6.22--3.92)	0.00 (0.00-0.00)	0.17 (0.11-0.24)
	Relative FPR at global threshold	-1.87 (-3.64--0.10)	-5.14 (-6.83--3.40)	6.93 (5.07-8.70)	-0.90 (-2.11-0.26)	0.98 (-0.28-2.31)	0.00 (0.00-0.00)	-0.03 (-0.08-0.01)
	Relative Youden's Index at global threshold	2.10 (-0.07-4.10)	-3.78 (-5.87--1.58)	2.35 (0.29-4.46)	6.63 (5.08-8.02)	-7.30 (-8.85--5.59)	0.00 (0.00-0.00)	0.24 (0.15-0.34)
CXR-Model FFT	AUC-ROC	0.87 (0.86-0.87)	0.86 (0.85-0.87)	0.88 (0.88-0.89)	0.87 (0.87-0.88)	0.86 (0.86-0.87)	0.87 (0.87-0.87)	0.87 (0.87-0.87)
	TPR at global threshold	0.79 (0.78-0.81)	0.79 (0.77-0.80)	0.84 (0.83-0.85)	0.82 (0.81-0.83)	0.79 (0.78-0.80)	0.81 (0.80-0.81)	0.81 (0.80-0.81)
	FPR at global threshold	0.20 (0.19-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.21)	0.20 (0.20-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.60 (0.58-0.61)	0.59 (0.57-0.60)	0.64 (0.62-0.65)	0.62 (0.61-0.63)	0.59 (0.58-0.60)	0.61 (0.60-0.61)	0.61 (0.60-0.61)
	Relative AUC-ROC	-0.55 (-1.17-0.04)	-0.95 (-1.54--0.35)	1.53 (0.96-2.11)	0.58 (0.16-1.00)	-0.62 (-1.07--0.15)	0.00 (0.00-0.00)	0.00 (-0.02-0.03)
	Relative TPR at global threshold	-1.81 (-3.17--0.37)	-2.22 (-3.70--0.87)	4.23 (2.93-5.56)	1.79 (0.84-2.80)	-1.98 (-3.07--0.95)	0.00 (0.00-0.00)	0.06 (0.01-0.11)
	Relative FPR at global threshold	-1.95 (-3.87--0.48)	-0.49 (-2.19-1.43)	2.48 (0.79-4.35)	0.28 (-0.91-1.60)	-0.32 (-1.77-0.99)	0.00 (0.00-0.00)	0.00 (-0.04-0.05)
	Relative Youden's Index at global threshold	-1.77 (-3.60-0.29)	-2.79 (-4.83--0.91)	4.80 (3.00-6.66)	2.28 (0.92-3.64)	-2.53 (-4.04--1.03)	0.00 (0.00-0.00)	0.08 (0.01-0.15)
CXR-FMKD-Direct FFT (MSE)	AUC-ROC	0.88 (0.87-0.88)	0.87 (0.87-0.88)	0.90 (0.89-0.90)	0.89 (0.88-0.89)	0.87 (0.87-0.88)	0.88 (0.88-0.88)	0.88 (0.88-0.88)
	TPR at global threshold	0.81 (0.80-0.82)	0.82 (0.81-0.84)	0.84 (0.82-0.85)	0.83 (0.83-0.85)	0.81 (0.79-0.82)	0.82 (0.81-0.83)	0.82 (0.81-0.83)
	FPR at global threshold	0.20 (0.20-0.21)	0.20 (0.19-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.61 (0.59-0.62)	0.62 (0.61-0.64)	0.64 (0.62-0.65)	0.64 (0.63-0.65)	0.61 (0.59-0.62)	0.62 (0.61-0.63)	0.62 (0.61-0.63)
	Relative AUC-ROC	-0.47 (-1.04-0.06)	-1.06 (-1.63--0.51)	1.62 (1.11-2.16)	0.93 (0.54-1.31)	-1.02 (-1.43--0.60)	0.00 (0.00-0.00)	0.01 (-0.01-0.03)
	Relative TPR at global threshold	-1.51 (-2.92--0.14)	-0.20 (-1.35-1.29)	1.88 (0.44-2.97)	1.53 (0.76-2.56)	-1.71 (-2.83--0.84)	0.00 (0.00-0.00)	0.05 (0.01-0.09)
	Relative FPR at global threshold	1.37 (-0.33-3.17)	-1.23 (-2.70-0.72)	-0.20 (-2.16-1.21)	-0.54 (-1.72-0.63)	0.60 (-0.67-1.88)	0.00 (0.00-0.00)	-0.01 (-0.05-0.03)
	Relative Youden's Index at global threshold	-2.44 (-4.42--0.60)	0.13 (-1.59-2.12)	2.55 (0.63-4.21)	2.20 (1.09-3.63)	-2.45 (-4.03--1.21)	0.00 (0.00-0.00)	0.07 (0.02-0.13)
CXR-FMKD-Direct FFT (CS)	AUC-ROC	0.87 (0.87-0.88)	0.87 (0.87-0.88)	0.89 (0.89-0.90)	0.89 (0.88-0.89)	0.87 (0.86-0.87)	0.88 (0.88-0.88)	0.88 (0.88-0.88)
	TPR at global threshold	0.80 (0.79-0.82)	0.82 (0.80-0.83)	0.84 (0.82-0.85)	0.86 (0.85-0.86)	0.78 (0.76-0.79)	0.82 (0.81-0.83)	0.82 (0.81-0.83)
	FPR at global threshold	0.21 (0.20-0.21)	0.20 (0.19-0.20)	0.20 (0.19-0.20)	0.21 (0.21-0.21)	0.19 (0.19-0.19)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.60 (0.58-0.61)	0.62 (0.61-0.64)	0.64 (0.63-0.65)	0.65 (0.64-0.66)	0.59 (0.57-0.60)	0.62 (0.61-0.63)	0.62 (0.61-0.63)
	Relative AUC-ROC	-0.83 (-1.42--0.28)	-0.76 (-1.28--0.18)	1.72 (1.22-2.22)	0.95 (0.57-1.33)	-1.08 (-1.48--0.66)	0.00 (0.00-0.00)	0.00 (-0.02-0.02)
	Relative TPR at global threshold	-1.64 (-3.02--0.40)	0.00 (-1.28-1.37)	2.13 (0.89-3.41)	4.60 (3.65-5.49)	-5.10 (-6.05--4.05)	0.00 (0.00-0.00)	0.15 (0.10-0.21)
	Relative FPR at global threshold	3.71 (1.81-5.31)	-1.79 (-3.70--0.26)	-1.53 (-2.89-0.64)	4.30 (3.16-5.47)	-4.69 (-5.97--3.44)	0.00 (0.00-0.00)	0.15 (0.11-0.19)
	Relative Youden's Index at global threshold	-3.36 (-5.21--1.62)	0.58 (-1.17-2.55)	3.32 (1.46-5.01)	4.70 (3.39-5.92)	-5.23 (-6.53--3.79)	0.00 (0.00-0.00)	0.16 (0.09-0.23)
CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4)	AUC-ROC	0.87 (0.87-0.88)	0.87 (0.87-0.88)	0.89 (0.88-0.90)	0.89 (0.88-0.89)	0.87 (0.86-0.88)	0.88 (0.88-0.88)	0.88 (0.87-0.88)
	TPR at global threshold	0.79 (0.78-0.81)	0.82 (0.80-0.83)	0.83 (0.82-0.84)	0.83 (0.82-0.84)	0.79 (0.78-0.80)	0.81 (0.81-0.82)	0.81 (0.81-0.82)
	FPR at global threshold	0.19 (0.19-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.21)	0.20 (0.19-0.20)	0.20 (0.20-0.21)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.60 (0.58-0.61)	0.62 (0.60-0.63)	0.63 (0.61-0.64)	0.64 (0.63-0.65)	0.59 (0.57-0.60)	0.61 (0.61-0.62)	0.61 (0.61-0.62)
	Relative AUC-ROC	-0.70 (-1.26--0.13)	-0.56 (-1.12-0.02)	1.34 (0.79-1.87)	0.90 (0.51-1.31)	-0.98 (-1.41--0.57)	0.00 (0.00-0.00)	0.00 (-0.02-0.02)
	Relative TPR at global threshold	-2.46 (-3.85--1.13)	0.49 (-0.73-2.04)	2.27 (0.74-3.54)	2.62 (1.62-3.51)	-2.91 (-3.87--1.81)	0.00 (0.00-0.00)	0.08 (0.03-0.13)
	Relative FPR at global threshold	-2.94 (-4.52--1.15)	0.50 (-1.26-2.30)	2.29 (0.41-3.92)	-1.69 (-2.92--0.54)	1.84 (0.58-3.20)	0.00 (0.00-0.00)	-0.06 (-0.10--0.02)
	Relative Youden's Index at global threshold	-2.31 (-4.27--0.52)	0.49 (-1.22-2.64)	2.26 (0.24-4.01)	4.02 (2.66-5.30)	-4.46 (-5.90--2.98)	0.00 (0.00-0.00)	0.13 (0.06-0.19)

**Relative values are computed with respect to the Subgroups' Average*

***All data are presented as: value (95% CI), derived from bootstrapping with 2000 samples*

Table 29. Detailed Absolute and Relative Subgroup Performances for 'No Finding' in CheXpert. This table reports disease detection performance for our five selected models across race and sex subgroups, their average, and the total patient sample, denoted 'All'. Data is presented as mean and 95% confidence interval (CI). TPR and FPR for each subgroup are determined at a fixed (global) decision threshold, optimised to achieve an FPR of 20% across the entire patient sample.

CheXpert – Cardiomegaly

Model	Metric	Subgroups						All
		White	Asian	Black	Male	Female	*Average**	
CXR-FM	AUC-ROC	0.80 (0.80-0.81)	0.82 (0.81-0.82)	0.79 (0.78-0.79)	0.82 (0.82-0.83)	0.79 (0.78-0.79)	0.80 (0.80-0.81)	0.81 (0.80-0.81)
	TPR at global threshold	0.61 (0.59-0.62)	0.66 (0.65-0.68)	0.65 (0.64-0.66)	0.65 (0.64-0.66)	0.63 (0.62-0.64)	0.64 (0.63-0.65)	0.64 (0.64-0.65)
	FPR at global threshold	0.17 (0.16-0.17)	0.19 (0.19-0.19)	0.25 (0.24-0.25)	0.19 (0.18-0.19)	0.22 (0.21-0.22)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.44 (0.42-0.45)	0.47 (0.46-0.49)	0.41 (0.40-0.42)	0.47 (0.46-0.48)	0.42 (0.41-0.43)	0.44 (0.43-0.45)	0.44 (0.44-0.45)
	Relative AUC-ROC	-0.13 (-0.78-0.54)	1.92 (1.33-2.54)	-1.99 (-2.56--1.42)	2.12 (1.72-2.51)	-1.92 (-2.37--1.46)	0.00 (0.00-0.00)	0.29 (0.24-0.34)
	Relative TPR at global threshold	-5.54 (-7.16--3.71)	3.22 (1.57-4.78)	1.78 (0.37-3.19)	1.82 (0.75-2.97)	-1.29 (-2.52--0.15)	0.00 (0.00-0.00)	0.29 (0.14-0.44)
	Relative FPR at global threshold	-16.65 (-18.32--14.89)	-5.74 (-7.53--3.96)	22.56 (20.57-24.61)	-6.96 (-8.21--5.73)	6.78 (5.41-8.17)	0.00 (0.00-0.00)	-0.69 (-0.77--0.60)
	Relative Youden's Index at global threshold	-0.46 (-2.99-2.15)	7.32 (4.85-9.78)	-7.71 (-9.99--5.46)	5.83 (4.14-7.52)	-4.99 (-6.80--3.18)	0.00 (0.00-0.00)	0.73 (0.52-0.96)
CXR-Model FFT	AUC-ROC	0.85 (0.85-0.86)	0.86 (0.86-0.87)	0.86 (0.86-0.87)	0.87 (0.86-0.87)	0.86 (0.85-0.86)	0.86 (0.86-0.86)	0.86 (0.86-0.87)
	TPR at global threshold	0.75 (0.74-0.76)	0.77 (0.76-0.78)	0.82 (0.81-0.83)	0.80 (0.79-0.81)	0.77 (0.76-0.78)	0.78 (0.78-0.79)	0.79 (0.78-0.79)
	FPR at global threshold	0.19 (0.18-0.19)	0.18 (0.18-0.19)	0.23 (0.23-0.24)	0.20 (0.20-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.56 (0.55-0.57)	0.59 (0.58-0.60)	0.59 (0.58-0.59)	0.60 (0.59-0.61)	0.57 (0.56-0.58)	0.58 (0.58-0.59)	0.59 (0.58-0.59)
	Relative AUC-ROC	-0.95 (-1.49--0.39)	0.35 (-0.15-0.89)	0.15 (-0.27-0.60)	0.88 (0.54-1.21)	-0.43 (-0.79--0.07)	0.00 (0.00-0.00)	0.23 (0.19-0.27)
	Relative TPR at global threshold	-4.09 (-5.33--2.81)	-1.17 (-2.48--0.08)	4.37 (3.43-5.43)	2.09 (1.39-2.90)	-1.18 (-2.04--0.40)	0.00 (0.00-0.00)	0.48 (0.38-0.60)
	Relative FPR at global threshold	-6.24 (-7.78--4.34)	-8.93 (-10.54--7.08)	15.68 (13.53-17.33)	-0.52 (-1.65-0.83)	0.01 (-1.47-1.25)	0.00 (0.00-0.00)	-0.28 (-0.34--0.20)
	Relative Youden's Index at global threshold	-3.36 (-5.17--1.59)	1.50 (-0.41-3.07)	0.47 (-0.87-2.08)	2.99 (1.92-4.11)	-1.60 (-2.80--0.42)	0.00 (0.00-0.00)	0.74 (0.60-0.90)
CXR-FMKD-Direct FFT (MSE)	AUC-ROC	0.85 (0.85-0.86)	0.86 (0.86-0.87)	0.87 (0.87-0.87)	0.87 (0.87-0.88)	0.86 (0.86-0.86)	0.86 (0.86-0.87)	0.87 (0.86-0.87)
	TPR at global threshold	0.75 (0.73-0.76)	0.74 (0.72-0.75)	0.83 (0.82-0.84)	0.81 (0.80-0.81)	0.75 (0.74-0.76)	0.77 (0.77-0.78)	0.78 (0.77-0.79)
	FPR at global threshold	0.18 (0.18-0.19)	0.18 (0.18-0.18)	0.24 (0.24-0.24)	0.20 (0.20-0.20)	0.20 (0.19-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.56 (0.55-0.57)	0.55 (0.54-0.57)	0.59 (0.58-0.60)	0.60 (0.59-0.61)	0.56 (0.54-0.56)	0.57 (0.57-0.58)	0.58 (0.57-0.59)
	Relative AUC-ROC	-1.20 (-1.73--0.70)	0.05 (-0.44-0.53)	0.65 (0.20-1.08)	1.02 (0.70-1.34)	-0.52 (-0.86--0.16)	0.00 (0.00-0.00)	0.26 (0.22-0.30)
	Relative TPR at global threshold	-3.61 (-5.12--2.49)	-5.05 (-6.23--3.85)	7.31 (6.41-8.43)	4.09 (3.33-4.92)	-2.74 (-3.57--1.91)	0.00 (0.00-0.00)	0.73 (0.62-0.86)
	Relative FPR at global threshold	-9.03 (-10.99--7.46)	-9.75 (-11.61--8.12)	19.58 (17.88-21.94)	0.63 (-0.55-1.91)	-1.42 (-2.86--0.15)	0.00 (0.00-0.00)	-0.31 (-0.39--0.23)
	Relative Youden's Index at global threshold	-1.71 (-3.68-0.01)	-3.40 (-5.03--1.66)	3.02 (1.61-4.55)	5.30 (4.17-6.43)	-3.20 (-4.37--1.97)	0.00 (0.00-0.00)	1.09 (0.95-1.27)
CXR-FMKD-Direct FFT (CS)	AUC-ROC	0.85 (0.85-0.86)	0.86 (0.86-0.87)	0.86 (0.86-0.87)	0.87 (0.87-0.87)	0.85 (0.85-0.86)	0.86 (0.86-0.86)	0.86 (0.86-0.87)
	TPR at global threshold	0.72 (0.71-0.73)	0.77 (0.75-0.78)	0.81 (0.80-0.82)	0.77 (0.76-0.78)	0.77 (0.76-0.78)	0.77 (0.76-0.77)	0.77 (0.77-0.78)
	FPR at global threshold	0.17 (0.17-0.18)	0.20 (0.19-0.20)	0.23 (0.23-0.24)	0.19 (0.18-0.19)	0.22 (0.21-0.22)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.55 (0.53-0.56)	0.57 (0.56-0.58)	0.58 (0.57-0.59)	0.59 (0.58-0.60)	0.56 (0.55-0.56)	0.57 (0.56-0.57)	0.57 (0.57-0.58)
	Relative AUC-ROC	-0.92 (-1.47--0.35)	0.49 (-0.02-1.01)	0.10 (-0.35-0.54)	1.13 (0.81-1.44)	-0.80 (-1.16--0.43)	0.00 (0.00-0.00)	0.28 (0.24-0.32)
	Relative TPR at global threshold	-6.15 (-7.66--4.89)	-0.43 (-1.60-0.90)	5.38 (4.36-6.39)	0.75 (-0.18-1.46)	0.45 (-0.31-1.43)	0.00 (0.00-0.00)	0.60 (0.49-0.72)
	Relative FPR at global threshold	-13.69 (-15.38--12.01)	-2.22 (-4.01--0.44)	15.75 (13.92-17.63)	-7.73 (-8.86--6.43)	7.90 (6.43-9.15)	0.00 (0.00-0.00)	-0.60 (-0.68--0.52)
	Relative Youden's Index at global threshold	-3.48 (-5.50--1.62)	0.21 (-1.52-2.11)	1.70 (0.14-3.18)	3.76 (2.44-4.79)	-2.19 (-3.35--0.76)	0.00 (0.00-0.00)	1.03 (0.87-1.19)
CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4)	AUC-ROC	0.85 (0.85-0.86)	0.86 (0.86-0.87)	0.87 (0.86-0.87)	0.87 (0.87-0.87)	0.86 (0.85-0.86)	0.86 (0.86-0.87)	0.86 (0.86-0.87)
	TPR at global threshold	0.71 (0.70-0.73)	0.78 (0.77-0.79)	0.83 (0.82-0.84)	0.79 (0.79-0.80)	0.77 (0.76-0.78)	0.78 (0.77-0.78)	0.78 (0.78-0.79)
	FPR at global threshold	0.17 (0.17-0.18)	0.19 (0.19-0.20)	0.24 (0.24-0.24)	0.19 (0.19-0.19)	0.21 (0.21-0.21)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.54 (0.52-0.55)	0.59 (0.57-0.60)	0.59 (0.58-0.60)	0.60 (0.59-0.61)	0.56 (0.55-0.57)	0.58 (0.57-0.58)	0.58 (0.58-0.59)
	Relative AUC-ROC	-1.28 (-1.82--0.75)	0.33 (-0.18-0.82)	0.55 (0.13-0.98)	0.88 (0.57-1.19)	-0.47 (-0.83--0.11)	0.00 (0.00-0.00)	0.28 (0.24-0.32)
	Relative TPR at global threshold	-8.75 (-10.08--7.14)	0.24 (-1.00-1.35)	6.95 (5.91-7.92)	2.06 (1.37-2.88)	-0.49 (-1.41-0.25)	0.00 (0.00-0.00)	0.80 (0.68-0.92)
	Relative FPR at global threshold	-14.61 (-16.30--12.81)	-4.33 (-6.07--2.48)	19.22 (17.22-21.06)	-4.78 (-5.99--3.45)	4.50 (3.03-5.82)	0.00 (0.00-0.00)	-0.54 (-0.63--0.46)
	Relative Youden's Index at global threshold	-6.71 (-8.54--4.54)	1.83 (0.05-3.47)	2.67 (1.17-4.06)	4.44 (3.45-5.59)	-2.23 (-3.57--1.16)	0.00 (0.00-0.00)	1.27 (1.09-1.43)

**Relative values are computed with respect to the Subgroups' Average*

***All data are presented as: value (95% CI), derived from bootstrapping with 2000 samples*

Table 30. Detailed Absolute and Relative Subgroup Performances for 'Cardiomegaly' in CheXpert.
This table reports disease detection performance for our five selected models across race and sex subgroups, their average, and the total patient sample, denoted 'All'. Data is presented as mean and 95% confidence interval (CI). TPR and FPR for each subgroup are determined at a fixed (global) decision threshold, optimised to achieve an FPR of 20% across the entire patient sample.

CheXpert – Pneumothorax

Model	Metric	Subgroups						All
		White	Asian	Black	Male	Female	*Average**	
CXR-FM	AUC-ROC	0.80 (0.79-0.81)	0.82 (0.81-0.82)	0.84 (0.83-0.85)	0.81 (0.81-0.82)	0.82 (0.82-0.83)	0.82 (0.81-0.82)	0.82 (0.81-0.82)
	TPR at global threshold	0.68 (0.67-0.70)	0.70 (0.69-0.72)	0.68 (0.66-0.70)	0.70 (0.69-0.71)	0.68 (0.67-0.70)	0.69 (0.68-0.70)	0.69 (0.68-0.70)
	FPR at global threshold	0.23 (0.22-0.23)	0.21 (0.21-0.21)	0.17 (0.16-0.17)	0.21 (0.21-0.21)	0.19 (0.18-0.19)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.46 (0.44-0.47)	0.49 (0.48-0.51)	0.51 (0.49-0.54)	0.49 (0.47-0.50)	0.50 (0.48-0.51)	0.49 (0.48-0.50)	0.49 (0.48-0.50)
	Relative AUC-ROC	-2.20 (-3.02--1.42)	-0.18 (-0.90-0.62)	2.13 (1.19-3.13)	-0.48 (-1.00-0.08)	0.73 (0.09-1.33)	0.00 (0.00-0.00)	0.09 (-0.01-0.19)
	Relative TPR at global threshold	-1.06 (-3.04-0.72)	2.06 (0.27-3.80)	-1.22 (-3.47-1.29)	1.15 (-0.08-2.44)	-0.92 (-2.44-0.54)	0.00 (0.00-0.00)	0.28 (-0.04-0.58)
	Relative FPR at global threshold	12.58 (10.85-14.39)	4.31 (2.75-6.17)	-16.09 (-17.92--14.64)	5.98 (4.73-7.10)	-6.79 (-8.04--5.45)	0.00 (0.00-0.00)	0.04 (-0.03-0.10)
	Relative Youden's Index at global threshold	-6.63 (-9.64--3.95)	1.13 (-1.47-3.71)	4.85 (1.58-8.48)	-0.82 (-2.62-1.08)	1.48 (-0.76-3.54)	0.00 (0.00-0.00)	0.38 (-0.08-0.80)
CXR-Model FFT	AUC-ROC	0.85 (0.84-0.85)	0.86 (0.85-0.87)	0.86 (0.85-0.87)	0.86 (0.86-0.87)	0.85 (0.84-0.86)	0.86 (0.85-0.86)	0.86 (0.85-0.86)
	TPR at global threshold	0.78 (0.76-0.79)	0.78 (0.77-0.80)	0.69 (0.66-0.71)	0.79 (0.78-0.80)	0.72 (0.70-0.73)	0.75 (0.74-0.76)	0.76 (0.75-0.77)
	FPR at global threshold	0.23 (0.23-0.24)	0.20 (0.20-0.21)	0.16 (0.16-0.17)	0.23 (0.23-0.23)	0.16 (0.16-0.17)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.54 (0.53-0.56)	0.58 (0.56-0.59)	0.52 (0.50-0.54)	0.56 (0.55-0.57)	0.55 (0.54-0.57)	0.55 (0.54-0.56)	0.56 (0.55-0.57)
	Relative AUC-ROC	-1.14 (-1.86--0.46)	0.54 (-0.11-1.14)	0.29 (-0.55-1.15)	0.89 (0.35-1.38)	-0.57 (-1.09--0.03)	0.00 (0.00-0.00)	0.26 (0.17-0.36)
	Relative TPR at global threshold	3.40 (1.80-4.98)	4.06 (2.46-5.55)	-8.62 (-10.91--6.31)	5.68 (4.49-6.82)	-4.52 (-5.87--3.19)	0.00 (0.00-0.00)	1.38 (1.07-1.68)
	Relative FPR at global threshold	17.50 (15.85-19.45)	2.59 (0.65-4.15)	-18.41 (-20.08--16.82)	15.94 (14.82-17.10)	-17.60 (-18.90--16.37)	0.00 (0.00-0.00)	0.32 (0.25-0.41)
	Relative Youden's Index at global threshold	-1.70 (-4.04-0.50)	4.59 (2.37-6.86)	-5.07 (-8.11--1.91)	1.97 (0.29-3.57)	0.21 (-1.70-2.16)	0.00 (0.00-0.00)	1.77 (1.35-2.17)
CXR-FMKD-Direct FFT (MSE)	AUC-ROC	0.85 (0.84-0.86)	0.88 (0.87-0.88)	0.86 (0.85-0.87)	0.87 (0.87-0.88)	0.86 (0.85-0.87)	0.86 (0.86-0.87)	0.87 (0.86-0.87)
	TPR at global threshold	0.76 (0.75-0.78)	0.80 (0.78-0.81)	0.72 (0.70-0.74)	0.80 (0.79-0.81)	0.73 (0.71-0.74)	0.76 (0.75-0.77)	0.77 (0.76-0.78)
	FPR at global threshold	0.23 (0.23-0.24)	0.22 (0.22-0.22)	0.15 (0.15-0.15)	0.22 (0.22-0.22)	0.18 (0.18-0.18)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.53 (0.52-0.55)	0.57 (0.56-0.59)	0.57 (0.55-0.59)	0.58 (0.57-0.59)	0.55 (0.53-0.56)	0.56 (0.55-0.57)	0.57 (0.56-0.58)
	Relative AUC-ROC	-1.77 (-2.44--1.09)	1.33 (0.72-1.95)	-0.16 (-1.03-0.66)	1.02 (0.55-1.50)	-0.43 (-0.91-0.06)	0.00 (0.00-0.00)	0.42 (0.32-0.53)
	Relative TPR at global threshold	0.41 (-1.15-1.98)	4.38 (2.99-5.88)	-5.34 (-7.61--3.26)	5.22 (4.14-6.37)	-4.67 (-5.98--3.40)	0.00 (0.00-0.00)	1.05 (0.79-1.35)
	Relative FPR at global threshold	16.08 (14.31-17.78)	10.36 (8.39-11.92)	-25.10 (-26.53--23.36)	9.64 (8.34-10.66)	-10.98 (-12.08--9.56)	0.00 (0.00-0.00)	0.04 (-0.04-0.12)
	Relative Youden's Index at global threshold	-5.16 (-7.27--2.93)	2.26 (0.38-4.39)	1.69 (-1.49-4.55)	3.64 (2.17-5.35)	-2.43 (-4.32--0.76)	0.00 (0.00-0.00)	1.41 (1.05-1.81)
CXR-FMKD-Direct FFT (CS)	AUC-ROC	0.85 (0.84-0.86)	0.87 (0.86-0.87)	0.87 (0.86-0.88)	0.86 (0.85-0.87)	0.87 (0.86-0.87)	0.86 (0.86-0.87)	0.86 (0.86-0.87)
	TPR at global threshold	0.77 (0.76-0.78)	0.80 (0.78-0.81)	0.70 (0.68-0.73)	0.77 (0.76-0.78)	0.77 (0.75-0.78)	0.76 (0.75-0.77)	0.77 (0.76-0.78)
	FPR at global threshold	0.23 (0.23-0.24)	0.21 (0.20-0.21)	0.16 (0.16-0.16)	0.22 (0.22-0.22)	0.18 (0.18-0.18)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.54 (0.52-0.55)	0.59 (0.58-0.60)	0.54 (0.52-0.57)	0.55 (0.54-0.57)	0.59 (0.57-0.60)	0.56 (0.55-0.57)	0.57 (0.56-0.58)
	Relative AUC-ROC	-1.54 (-2.20--0.87)	0.60 (-0.05-1.21)	0.59 (-0.29-1.46)	-0.21 (-0.67-0.28)	0.55 (0.05-1.03)	0.00 (0.00-0.00)	0.21 (0.10-0.31)
	Relative TPR at global threshold	1.11 (-0.50-2.67)	4.53 (2.86-5.84)	-7.47 (-9.63--4.84)	1.26 (0.04-2.39)	0.57 (-0.68-1.78)	0.00 (0.00-0.00)	0.97 (0.63-1.25)
	Relative FPR at global threshold	17.34 (15.52-19.06)	3.32 (1.39-4.97)	-19.58 (-21.16--17.77)	8.77 (7.63-9.91)	-9.85 (-11.09--8.57)	0.00 (0.00-0.00)	0.10 (0.03-0.18)
	Relative Youden's Index at global threshold	-4.66 (-6.89--2.49)	4.96 (2.73-6.86)	-3.16 (-6.23-0.20)	-1.42 (-3.06-0.15)	4.28 (2.51-5.94)	0.00 (0.00-0.00)	1.28 (0.82-1.66)
CXR-FMKD-Direct FFT (MSE-CS | 0.6-0.4)	AUC-ROC	0.86 (0.85-0.86)	0.88 (0.87-0.88)	0.88 (0.87-0.88)	0.88 (0.87-0.88)	0.87 (0.86-0.87)	0.87 (0.87-0.87)	0.87 (0.87-0.88)
	TPR at global threshold	0.78 (0.77-0.80)	0.80 (0.78-0.81)	0.73 (0.72-0.76)	0.81 (0.80-0.82)	0.74 (0.73-0.76)	0.77 (0.76-0.78)	0.78 (0.77-0.79)
	FPR at global threshold	0.22 (0.22-0.23)	0.22 (0.22-0.23)	0.16 (0.15-0.16)	0.22 (0.22-0.22)	0.18 (0.17-0.18)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.56 (0.55-0.58)	0.57 (0.56-0.58)	0.58 (0.56-0.60)	0.59 (0.58-0.60)	0.57 (0.55-0.58)	0.57 (0.56-0.58)	0.58 (0.57-0.59)
	Relative AUC-ROC	-1.74 (-2.38--1.10)	0.63 (0.09-1.20)	0.73 (-0.04-1.52)	0.62 (0.20-1.10)	-0.25 (-0.74-0.20)	0.00 (0.00-0.00)	0.28 (0.20-0.37)
	Relative TPR at global threshold	1.38 (-0.09-2.92)	3.18 (1.61-4.40)	-5.10 (-7.04--2.81)	4.42 (3.34-5.56)	-3.87 (-5.22--2.77)	0.00 (0.00-0.00)	0.93 (0.62-1.18)
	Relative FPR at global threshold	11.27 (9.38-12.81)	12.34 (10.67-14.29)	-22.22 (-23.74--20.57)	10.33 (9.15-11.46)	-11.72 (-12.98--10.45)	0.00 (0.00-0.00)	0.07 (-0.01-0.14)
	Relative Youden's Index at global threshold	-2.06 (-4.05-0.15)	-0.02 (-2.25-1.60)	0.86 (-1.73-4.06)	2.35 (0.86-3.96)	-1.14 (-3.07-0.36)	0.00 (0.00-0.00)	1.23 (0.81-1.56)

**Relative values are computed with respect to the Subgroups' Average*

***All data are presented as: value (95% CI), derived from bootstrapping with 2000 samples*

Table 31. Detailed Absolute and Relative Subgroup Performances for 'Pneumothorax' in CheXpert.
This table reports disease detection performance for our five selected models across race and sex subgroups, their average, and the total patient sample, denoted 'All'. Data is presented as mean and 95% confidence interval (CI). TPR and FPR for each subgroup are determined at a fixed (global) decision threshold, optimised to achieve an FPR of 20% across the entire patient sample.

S.9. Bias Analysis | Subgroup Performance Analysis – MIMIC

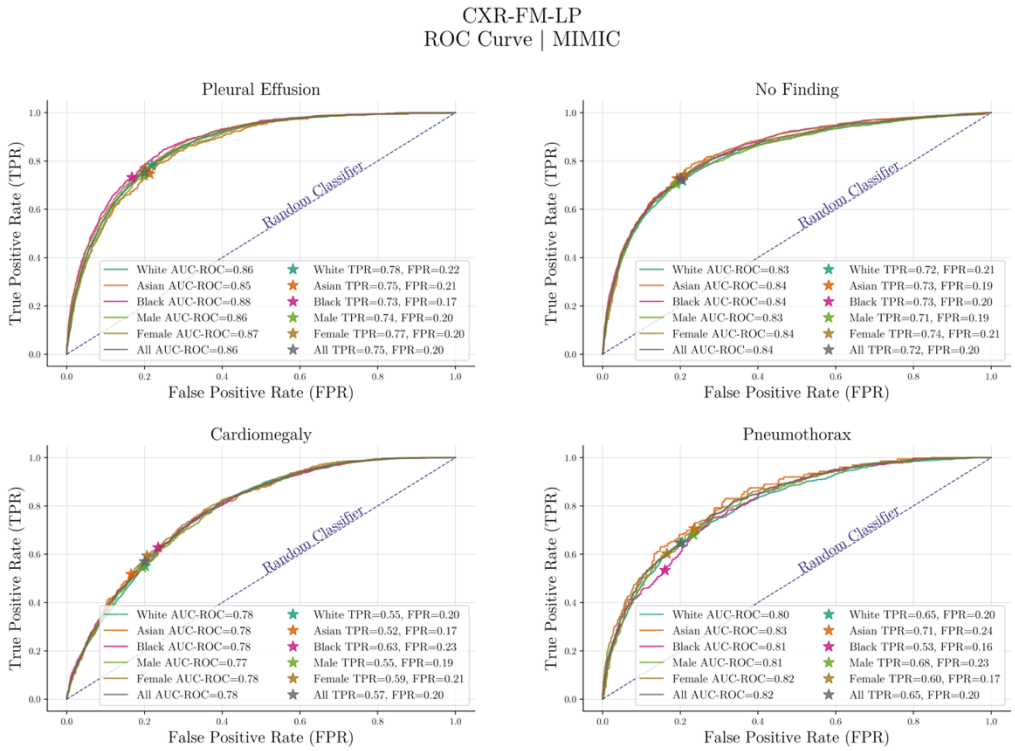

Figure 112. ROC Performance Across Subgroups for CXR-FM Tested on MIMIC.
This figure displays ROC curves for the detection of 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax' using CXR-FM tested on the resampled MIMIC test set. TPR and FPR for each subgroup are determined at a fixed decision threshold, optimised to achieve an FPR of 20% across the entire patient sample. The shifts in TPR/FPR across subgroups highlight disparities in disease detection performance.

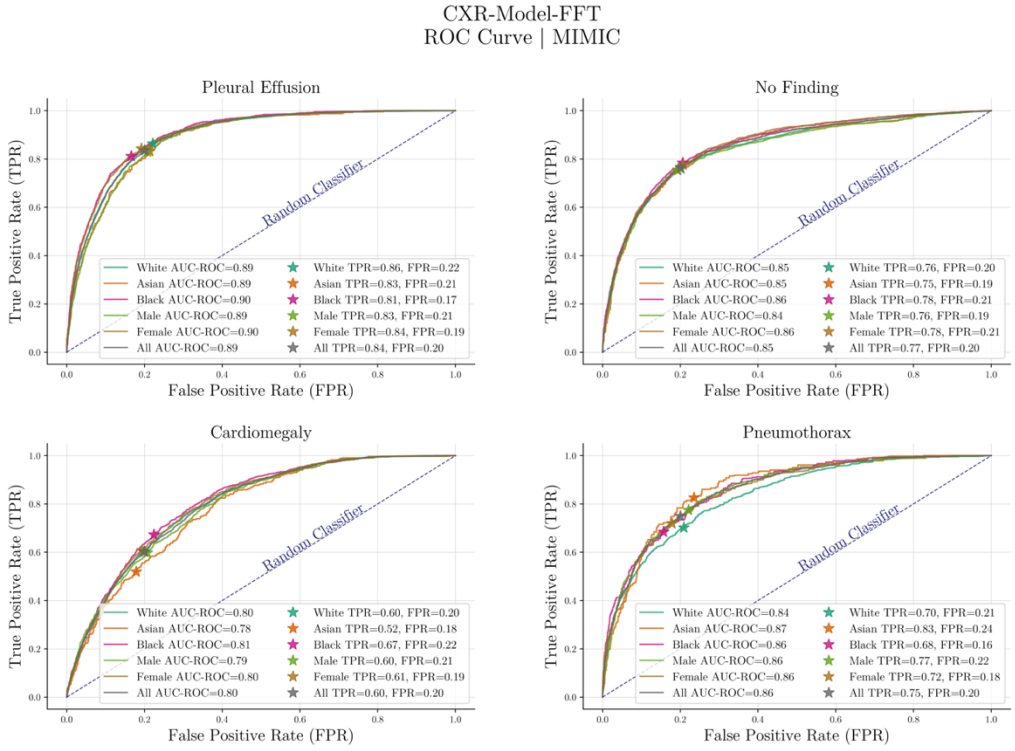

Figure 113. ROC Performance Across Subgroups for CXR-Model FFT Tested on MIMIC.
This figure displays ROC curves for the detection of 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax' using CXR-Model FFT tested on the resampled MIMIC test set. TPR and FPR for each subgroup are determined at a fixed decision threshold, optimised to achieve an FPR of 20% across the entire patient sample. The shifts in TPR/FPR across subgroups highlight disparities in disease detection performance.

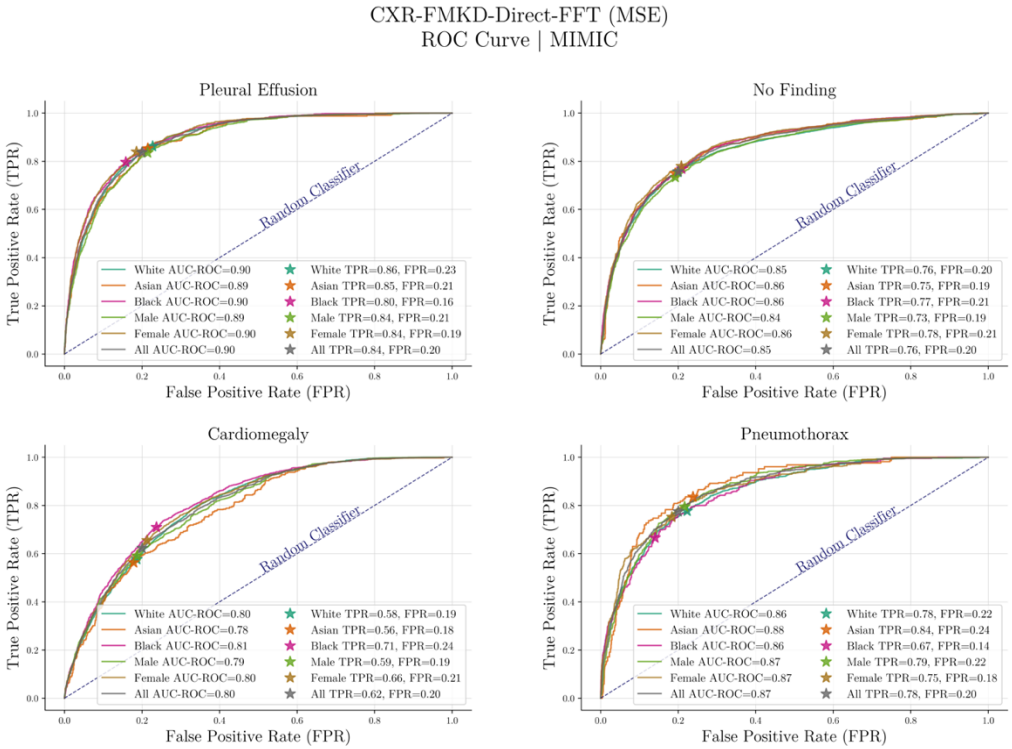

Figure 114. ROC Performance Across Subgroups for the Selected (MSE)-Student Tested on MIMIC. This figure displays ROC curves for the detection of 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax' using CXR-FMKD-Direct FFT (MSE) tested on the resampled MIMIC test set. TPR and FPR for each subgroup are determined at a fixed decision threshold, optimised to achieve an FPR of 20% across the entire patient sample. The shifts in TPR/FPR across subgroups highlight disparities in disease detection performance.

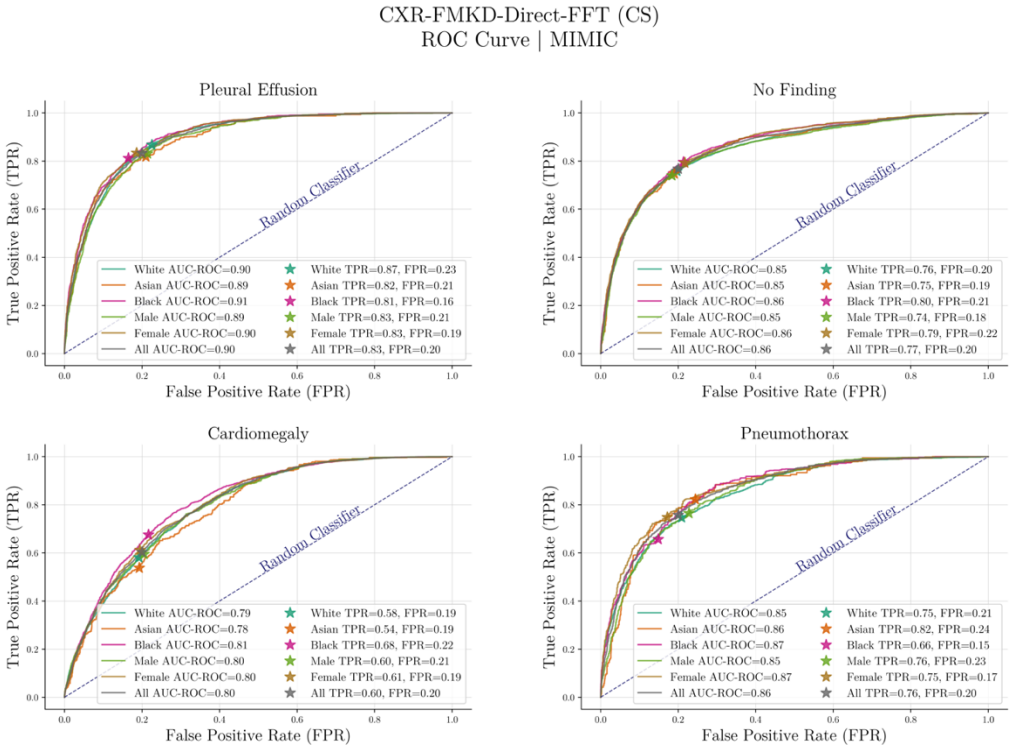

Figure 115. ROC Performance Across Subgroups for the Selected (CS)-Student Tested on MIMIC. This figure displays ROC curves for the detection of 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax' using CXR-FMKD-Direct FFT (CS) tested on the resampled MIMIC test set. TPR and FPR for each subgroup are determined at a fixed decision threshold, optimised to achieve an FPR of 20% across the entire patient sample. The shifts in TPR/FPR across subgroups highlight disparities in disease detection performance.

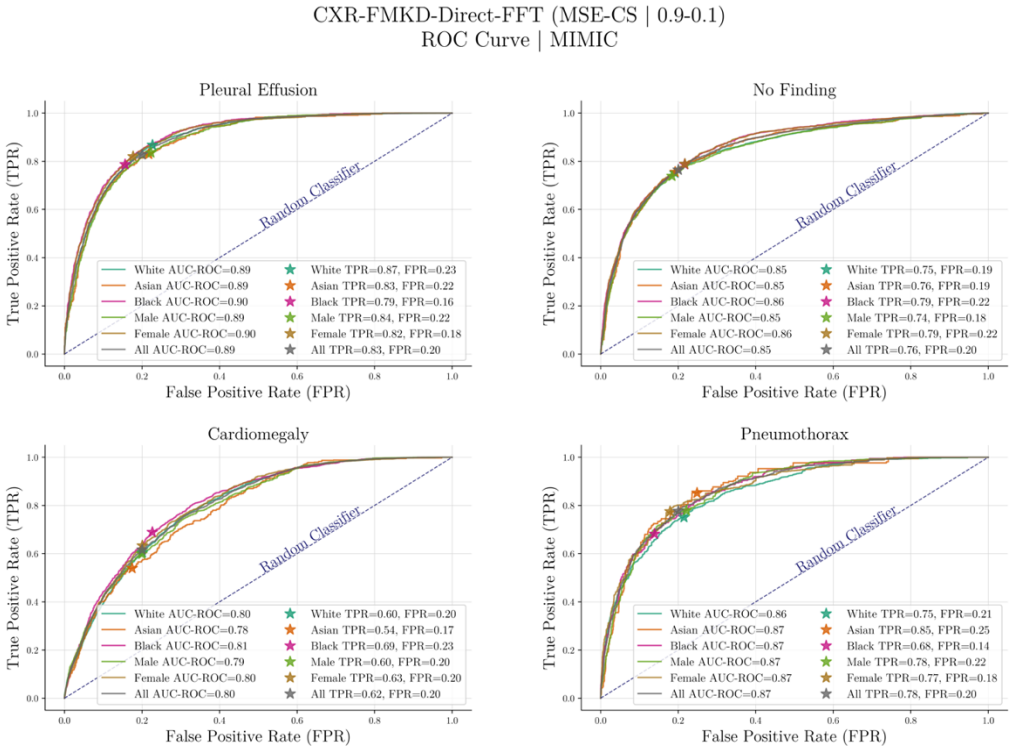

Figure 116. ROC Performance Across Subgroups for the Selected (MSE-CS | 0.9-0.1)-Student Tested on MIMIC.

This figure displays ROC curves for the detection of 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax' using CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1) tested on the resampled MIMIC test set. TPR and FPR for each subgroup are determined at a fixed decision threshold, optimised to achieve an FPR of 20% across the entire patient sample. The shifts in TPR/FPR across subgroups highlight disparities in disease detection performance.

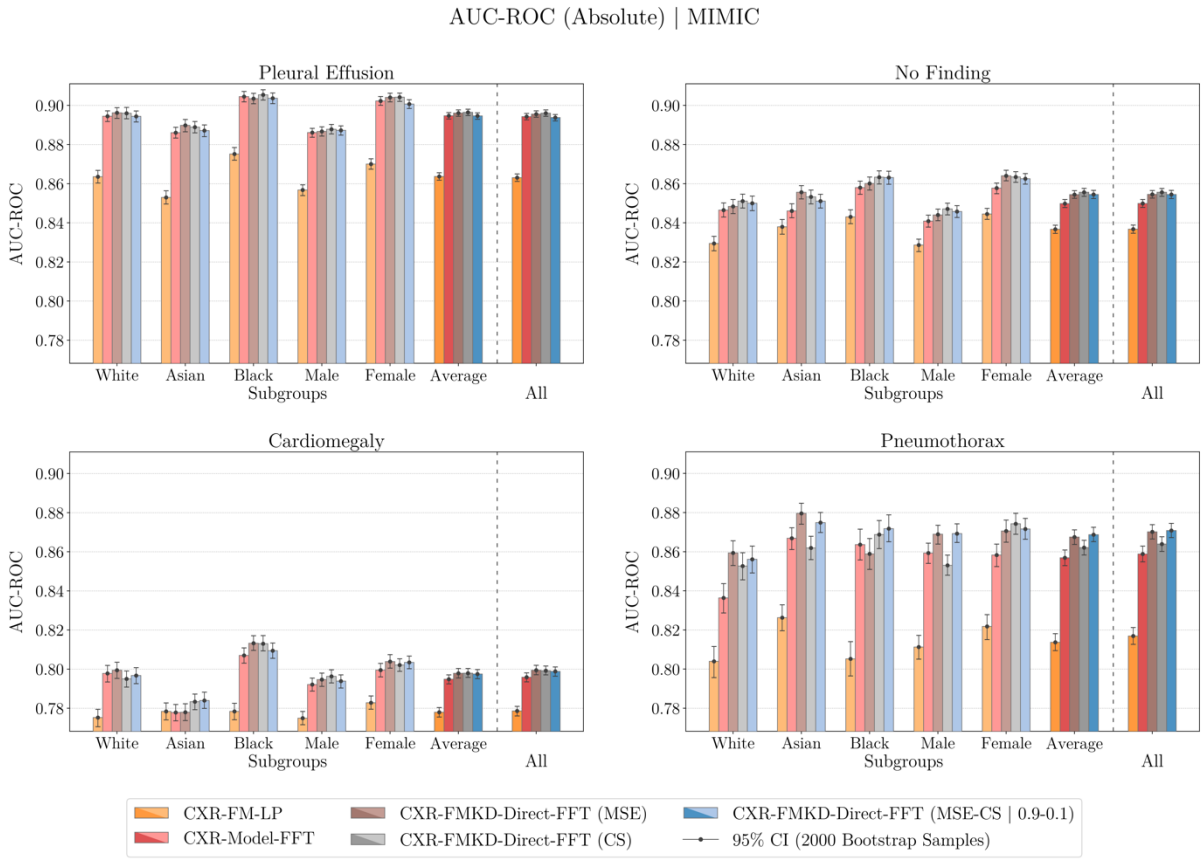

Figure 117. Comparison of AUC-ROC Disease Detection Performance Across Patient Subgroups for Selected Models Tested on MIMIC.

This figure illustrates the mean AUC-ROC values, depicted by dots, and their corresponding 95% confidence intervals (CIs), shown with whiskers and derived from bootstrapping with 2000 samples, across the relevant patient subgroups (for race and biological sex), their average, and the entire patient sample—denoted by 'All'—for our five selected models developed and tested on MIMIC. These models include the teacher CXR-FM; the CXR-FMKD-Direct FFT variants of the three selected student model types (MSE, CS, and MSE-CS | 0.9-0.1); and the traditional baseline CXR-Model FFT, which shares the same DenseNet169 architecture as the students but was developed without Knowledge Distillation (KD) from CXR-FM. The models were assessed for their ability (average absolute classification performance) to detect 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax'. The teacher CXR-FM consistently underperforms compared to the other models, with the student models exhibiting the highest performances, closely followed by the baseline CXR-Model FFT. Performance disparities can be observed across subgroups and there is also a significant drop in overall performance across all subgroups for our models in detecting 'Cardiomegaly'.

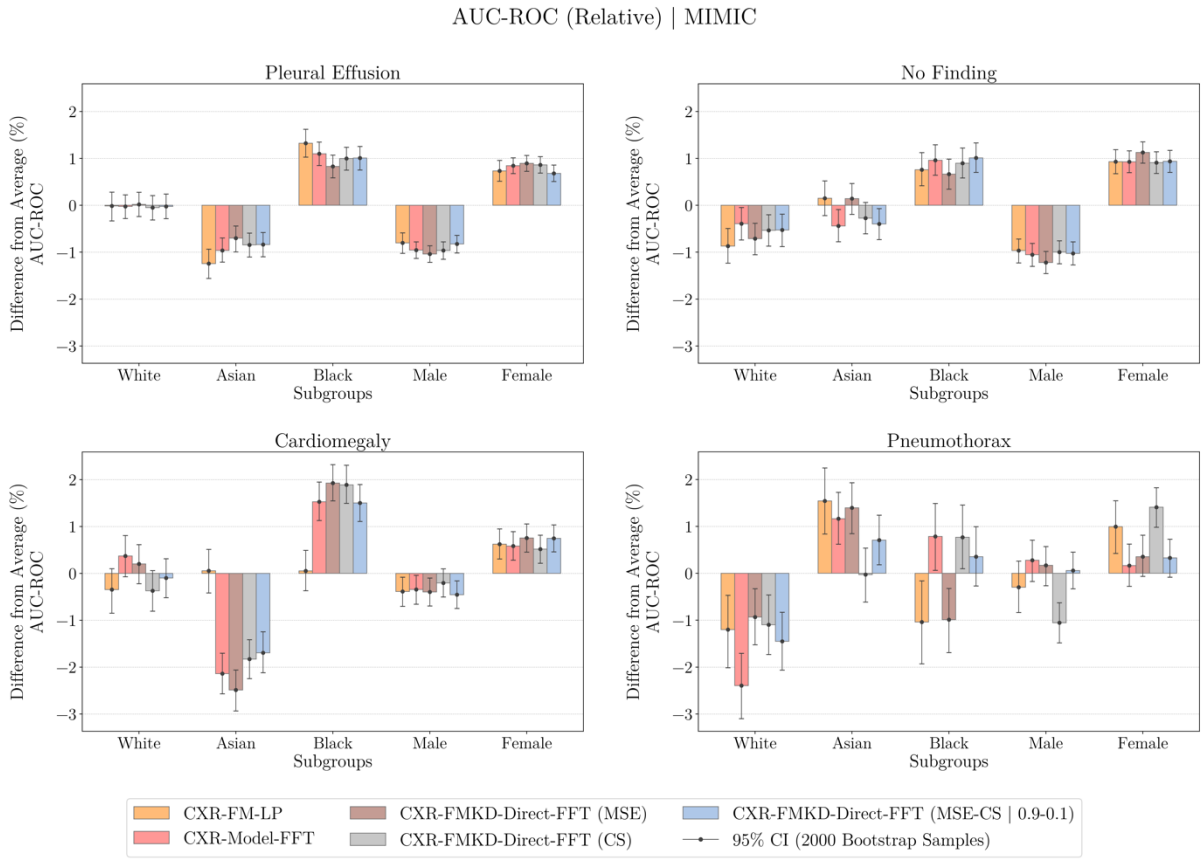

Figure 118. Relative Change in AUC-ROC Disease Detection Performance Across Patient Subgroups for Selected Models Tested on MIMIC.

This figure illustrates the mean relative changes in AUC-ROC performance, depicted by dots, and their corresponding 95% confidence intervals (CIs), shown with whiskers and derived from bootstrapping with 2000 samples, across the relevant patient subgroups (for race and biological sex) for our five selected models developed and tested on MIMIC. For each model, the relative performance change for each subgroup for a specific disease label was computed by comparing the subgroup's performance (Subgroup Value) with the average performance across all subgroups (Average Value) for that label using the formula: (Subgroup Value – Average Value)/ Average Value × 100%. The models evaluated include the teacher CXR-FM; the CXR-FMKD-Direct FFT variants of the three selected student model types (MSE, CS, and MSE-CS | 0.9-0.1); and the traditional baseline CXR-Model FFT, which shares the same DenseNet169 architecture as the students but was developed without Knowledge Distillation (KD) from CXR-FM. These models were assessed for their ability (average relative classification performance) to detect 'Pleural Effusion', 'No Finding', 'Cardiomegaly', and 'Pneumothorax'. Disparities in relative performance can be observed across subgroups for the models being evaluated.

MIMIC – Pleural Effusion

Model	Metric	Subgroups						All
		White	Asian	Black	Male	Female	*Average**	
CXR-FM	AUC-ROC	0.86 (0.86-0.87)	0.85 (0.85-0.86)	0.88 (0.87-0.88)	0.86 (0.85-0.86)	0.87 (0.87-0.87)	0.86 (0.86-0.87)	0.86 (0.86-0.86)
	TPR at global threshold	0.78 (0.78-0.79)	0.75 (0.74-0.76)	0.73 (0.72-0.74)	0.74 (0.73-0.75)	0.77 (0.76-0.78)	0.75 (0.75-0.76)	0.75 (0.75-0.76)
	FPR at global threshold	0.22 (0.22-0.22)	0.21 (0.21-0.22)	0.17 (0.16-0.17)	0.20 (0.20-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.56 (0.56-0.57)	0.53 (0.53-0.54)	0.56 (0.55-0.57)	0.54 (0.53-0.55)	0.57 (0.56-0.57)	0.55 (0.55-0.56)	0.55 (0.55-0.56)
	Relative AUC-ROC	-0.01 (-0.33-0.28)	-1.24 (-1.56--0.94)	1.32 (1.03-1.62)	-0.80 (-1.02--0.59)	0.73 (0.51-0.96)	0.00 (0.00-0.00)	-0.07 (-0.08--0.06)
	Relative TPR at global threshold	3.74 (3.03-4.53)	-0.88 (-1.61--0.03)	-2.91 (-3.82--2.18)	-1.98 (-2.54--1.47)	2.02 (1.50-2.59)	0.00 (0.00-0.00)	-0.02 (-0.04--0.00)
	Relative FPR at global threshold	9.64 (8.08-11.26)	6.57 (4.65-8.00)	-16.21 (-17.54--14.36)	-0.56 (-1.88-0.43)	0.56 (-0.45-1.90)	0.00 (0.00-0.00)	-0.00 (-0.04-0.03)
	Relative Youden's Index at global threshold	1.62 (0.43-2.82)	-3.56 (-4.63--2.16)	1.89 (0.47-2.91)	-2.50 (-3.26--1.62)	2.55 (1.67-3.34)	0.00 (0.00-0.00)	-0.03 (-0.05-0.00)
CXR-Model FFT	AUC-ROC	0.89 (0.89-0.90)	0.89 (0.88-0.89)	0.90 (0.90-0.91)	0.89 (0.88-0.89)	0.90 (0.90-0.90)	0.89 (0.89-0.90)	0.89 (0.89-0.90)
	TPR at global threshold	0.86 (0.86-0.87)	0.83 (0.82-0.84)	0.81 (0.81-0.82)	0.83 (0.82-0.84)	0.84 (0.84-0.85)	0.84 (0.83-0.84)	0.84 (0.83-0.84)
	FPR at global threshold	0.22 (0.22-0.22)	0.21 (0.21-0.22)	0.17 (0.16-0.17)	0.21 (0.21-0.21)	0.19 (0.19-0.19)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.64 (0.64-0.65)	0.62 (0.61-0.63)	0.65 (0.64-0.65)	0.62 (0.61-0.63)	0.65 (0.65-0.66)	0.64 (0.63-0.64)	0.64 (0.63-0.64)
	Relative AUC-ROC	-0.02 (-0.28-0.22)	-0.96 (-1.21--0.70)	1.10 (0.85-1.35)	-0.96 (-1.13--0.78)	0.85 (0.68-1.01)	0.00 (0.00-0.00)	-0.05 (-0.06--0.04)
	Relative TPR at global threshold	3.31 (2.64-3.95)	-0.44 (-1.30-0.30)	-2.88 (-3.50--2.21)	-0.82 (-1.39--0.41)	0.83 (0.41-1.42)	0.00 (0.00-0.00)	-0.01 (-0.03-0.00)
	Relative FPR at global threshold	10.43 (8.95-12.32)	6.78 (4.89-8.15)	-17.20 (-18.72--15.67)	3.95 (2.62-4.92)	-3.96 (-4.93--2.62)	0.00 (0.00-0.00)	0.01 (-0.04-0.05)
	Relative Youden's Index at global threshold	1.07 (-0.00-1.95)	-2.71 (-3.87--1.52)	1.62 (0.64-2.63)	-2.32 (-3.07--1.65)	2.34 (1.66-3.12)	0.00 (0.00-0.00)	-0.01 (-0.04-0.01)
CXR-FMKD-Direct FFT (MSE)	AUC-ROC	0.90 (0.89-0.90)	0.89 (0.89-0.89)	0.90 (0.90-0.91)	0.89 (0.88-0.89)	0.90 (0.90-0.91)	0.90 (0.89-0.90)	0.90 (0.89-0.90)
	TPR at global threshold	0.86 (0.86-0.87)	0.85 (0.85-0.86)	0.80 (0.79-0.80)	0.84 (0.83-0.84)	0.84 (0.83-0.84)	0.84 (0.83-0.84)	0.84 (0.83-0.84)
	FPR at global threshold	0.23 (0.22-0.23)	0.21 (0.21-0.22)	0.16 (0.15-0.16)	0.21 (0.21-0.22)	0.19 (0.18-0.19)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.64 (0.63-0.64)	0.64 (0.63-0.64)	0.64 (0.63-0.65)	0.62 (0.62-0.63)	0.65 (0.65-0.66)	0.64 (0.63-0.64)	0.64 (0.63-0.64)
	Relative AUC-ROC	0.02 (-0.24-0.28)	-0.70 (-0.99--0.44)	0.83 (0.59-1.07)	-1.04 (-1.22--0.86)	0.89 (0.73-1.06)	0.00 (0.00-0.00)	-0.06 (-0.07--0.04)
	Relative TPR at global threshold	2.97 (2.42-3.60)	1.88 (1.31-2.49)	-4.85 (-5.53--4.26)	-0.25 (-0.67-0.16)	0.25 (-0.17-0.67)	0.00 (0.00-0.00)	-0.00 (-0.03-0.02)
	Relative FPR at global threshold	13.60 (11.90-15.16)	7.47 (5.96-9.26)	-21.04 (-22.60--19.59)	7.16 (6.00-8.21)	-7.19 (-8.27--6.05)	0.00 (0.00-0.00)	0.01 (-0.05-0.06)
	Relative Youden's Index at global threshold	-0.36 (-1.22-0.63)	0.13 (-0.86-1.09)	0.22 (-0.78-1.15)	-2.58 (-3.19--1.93)	2.59 (1.93-3.21)	0.00 (0.00-0.00)	-0.01 (-0.04-0.02)
CXR-FMKD-Direct FFT (CS)	AUC-ROC	0.90 (0.89-0.90)	0.89 (0.89-0.89)	0.91 (0.90-0.91)	0.89 (0.89-0.89)	0.90 (0.90-0.91)	0.90 (0.89-0.90)	0.90 (0.89-0.90)
	TPR at global threshold	0.87 (0.86-0.87)	0.82 (0.81-0.83)	0.81 (0.80-0.82)	0.83 (0.83-0.84)	0.83 (0.83-0.84)	0.83 (0.83-0.84)	0.83 (0.83-0.84)
	FPR at global threshold	0.23 (0.22-0.23)	0.21 (0.21-0.21)	0.16 (0.16-0.17)	0.21 (0.21-0.22)	0.19 (0.18-0.19)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.64 (0.63-0.65)	0.61 (0.60-0.62)	0.65 (0.64-0.66)	0.62 (0.61-0.62)	0.65 (0.64-0.65)	0.63 (0.63-0.64)	0.63 (0.63-0.64)
	Relative AUC-ROC	-0.05 (-0.31-0.21)	-0.85 (-1.10--0.59)	1.00 (0.75-1.24)	-0.96 (-1.15--0.78)	0.86 (0.69-1.04)	0.00 (0.00-0.00)	-0.04 (-0.06--0.03)
	Relative TPR at global threshold	4.16 (3.50-4.69)	-1.70 (-2.33--1.04)	-2.45 (-3.06--1.79)	-0.17 (-0.59-0.27)	0.17 (-0.28-0.61)	0.00 (0.00-0.00)	-0.01 (-0.02-0.01)
	Relative FPR at global threshold	12.79 (11.06-14.37)	4.91 (3.32-6.60)	-17.67 (-19.22--16.12)	7.04 (5.79-8.23)	-7.07 (-8.24--5.83)	0.00 (0.00-0.00)	0.01 (-0.04-0.06)
	Relative Youden's Index at global threshold	1.43 (0.43-2.33)	-3.79 (-4.78--2.84)	2.36 (1.40-3.33)	-2.45 (-3.12--1.73)	2.46 (1.72-3.14)	0.00 (0.00-0.00)	-0.01 (-0.04-0.02)
CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1)	AUC-ROC	0.89 (0.89-0.90)	0.89 (0.88-0.89)	0.90 (0.90-0.91)	0.89 (0.88-0.89)	0.90 (0.90-0.90)	0.89 (0.89-0.90)	0.89 (0.89-0.90)
	TPR at global threshold	0.87 (0.86-0.87)	0.83 (0.82-0.84)	0.79 (0.78-0.80)	0.84 (0.83-0.84)	0.82 (0.81-0.83)	0.83 (0.82-0.83)	0.83 (0.82-0.83)
	FPR at global threshold	0.23 (0.22-0.23)	0.22 (0.21-0.22)	0.16 (0.15-0.16)	0.22 (0.22-0.23)	0.18 (0.17-0.18)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.64 (0.63-0.65)	0.61 (0.60-0.62)	0.63 (0.62-0.64)	0.61 (0.61-0.62)	0.64 (0.64-0.65)	0.63 (0.62-0.63)	0.63 (0.62-0.63)
	Relative AUC-ROC	-0.02 (-0.28-0.24)	-0.84 (-1.10--0.58)	1.01 (0.75-1.26)	-0.83 (-1.02--0.64)	0.68 (0.51-0.86)	0.00 (0.00-0.00)	-0.09 (-0.11--0.08)
	Relative TPR at global threshold	4.77 (4.12-5.33)	0.21 (-0.53-0.87)	-4.94 (-5.58--4.10)	0.90 (0.47-1.37)	-0.93 (-1.41--0.50)	0.00 (0.00-0.00)	0.00 (-0.02-0.03)
	Relative FPR at global threshold	13.36 (11.72-14.92)	8.62 (6.97-10.42)	-21.93 (-23.53--20.29)	11.50 (10.52-12.74)	-11.54 (-12.78--10.54)	0.00 (0.00-0.00)	0.01 (-0.05-0.08)
	Relative Youden's Index at global threshold	2.03 (1.05-2.96)	-2.47 (-3.56--1.53)	0.47 (-0.44-1.60)	-2.47 (-3.17--1.77)	2.44 (1.73-3.15)	0.00 (0.00-0.00)	-0.00 (-0.04-0.04)

**Relative values are computed with respect to the Subgroups' Average*

***All data are presented as: value (95% CI), derived from bootstrapping with 2000 samples*

Table 32. Detailed Absolute and Relative Subgroup Performances for 'Pleural Effusion' in MIMIC.
This table reports disease detection performance for our five selected models across race and sex subgroups, their average, and the total patient sample, denoted 'All'. Data is presented as mean and 95% confidence interval (CI). TPR and FPR for each subgroup are determined at a fixed (global) decision threshold, optimised to achieve an FPR of 20% across the entire patient sample.

MIMIC – No Finding

Model	Metric	Subgroups						All
		White	Asian	Black	Male	Female	*Average**	
CXR-FM	AUC-ROC	0.83 (0.83-0.83)	0.84 (0.83-0.84)	0.84 (0.84-0.85)	0.83 (0.83-0.83)	0.84 (0.84-0.85)	0.84 (0.83-0.84)	0.84 (0.83-0.84)
	TPR at global threshold	0.72 (0.71-0.73)	0.73 (0.72-0.74)	0.73 (0.72-0.73)	0.71 (0.70-0.71)	0.74 (0.74-0.75)	0.72 (0.72-0.73)	0.72 (0.72-0.73)
	FPR at global threshold	0.21 (0.20-0.21)	0.19 (0.19-0.20)	0.20 (0.20-0.21)	0.19 (0.19-0.19)	0.21 (0.21-0.21)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.51 (0.50-0.52)	0.54 (0.53-0.55)	0.52 (0.52-0.53)	0.52 (0.51-0.52)	0.53 (0.53-0.54)	0.52 (0.52-0.53)	0.52 (0.52-0.53)
	Relative AUC-ROC	-0.87 (-1.23--0.49)	0.15 (-0.22-0.52)	0.76 (0.42-1.12)	-0.97 (-1.23--0.72)	0.93 (0.68-1.19)	0.00 (0.00-0.00)	0.01 (-0.00-0.02)
	Relative TPR at global threshold	-0.77 (-1.64-0.00)	0.48 (-0.24-1.53)	0.40 (-0.47-1.05)	-2.64 (-3.18--2.07)	2.52 (1.98-3.04)	0.00 (0.00-0.00)	0.04 (0.03-0.05)
	Relative FPR at global threshold	2.63 (0.86-4.27)	-4.11 (-5.66--2.28)	1.32 (-0.53-2.97)	-5.11 (-6.42--3.95)	5.28 (4.07-6.61)	0.00 (0.00-0.00)	-0.06 (-0.08--0.04)
	Relative Youden's Index at global threshold	-2.06 (-3.40--0.84)	2.24 (1.07-3.83)	0.05 (-1.31-1.22)	-1.70 (-2.58--0.74)	1.47 (0.54-2.34)	0.00 (0.00-0.00)	0.08 (0.06-0.10)
CXR-Model FFT	AUC-ROC	0.85 (0.84-0.85)	0.85 (0.84-0.85)	0.86 (0.85-0.86)	0.84 (0.84-0.84)	0.86 (0.85-0.86)	0.85 (0.85-0.85)	0.85 (0.85-0.85)
	TPR at global threshold	0.76 (0.75-0.77)	0.75 (0.75-0.76)	0.78 (0.78-0.79)	0.76 (0.75-0.76)	0.78 (0.77-0.78)	0.77 (0.76-0.77)	0.77 (0.76-0.77)
	FPR at global threshold	0.20 (0.20-0.20)	0.19 (0.19-0.20)	0.21 (0.20-0.21)	0.19 (0.19-0.20)	0.21 (0.20-0.21)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.56 (0.55-0.57)	0.56 (0.55-0.57)	0.58 (0.57-0.59)	0.56 (0.56-0.57)	0.57 (0.56-0.58)	0.57 (0.56-0.57)	0.57 (0.56-0.57)
	Relative AUC-ROC	-0.39 (-0.74--0.05)	-0.44 (-0.78--0.09)	0.96 (0.64-1.29)	-1.06 (-1.30--0.81)	0.93 (0.70-1.16)	0.00 (0.00-0.00)	0.01 (-0.00-0.02)
	Relative TPR at global threshold	-0.71 (-1.38--0.00)	-1.64 (-2.27--0.88)	2.41 (1.66-3.05)	-1.37 (-1.94--0.91)	1.32 (0.87-1.85)	0.00 (0.00-0.00)	0.02 (0.01-0.04)
	Relative FPR at global threshold	0.59 (-0.96-2.38)	-3.66 (-5.63--2.10)	2.96 (1.30-4.67)	-3.33 (-4.66--2.20)	3.44 (2.28-4.81)	0.00 (0.00-0.00)	-0.04 (-0.06--0.02)
	Relative Youden's Index at global threshold	-1.17 (-2.30--0.11)	-0.93 (-1.94-0.39)	2.22 (1.00-3.29)	-0.68 (-1.48-0.12)	0.56 (-0.21-1.33)	0.00 (0.00-0.00)	0.04 (0.03-0.07)
CXR-FMKD-Direct FFT (MSE)	AUC-ROC	0.85 (0.84-0.85)	0.86 (0.85-0.86)	0.86 (0.86-0.86)	0.84 (0.84-0.85)	0.86 (0.86-0.87)	0.85 (0.85-0.86)	0.85 (0.85-0.86)
	TPR at global threshold	0.76 (0.75-0.76)	0.75 (0.74-0.76)	0.77 (0.76-0.78)	0.73 (0.73-0.74)	0.78 (0.78-0.79)	0.76 (0.75-0.76)	0.76 (0.75-0.76)
	FPR at global threshold	0.20 (0.20-0.20)	0.19 (0.19-0.20)	0.21 (0.21-0.21)	0.19 (0.19-0.19)	0.21 (0.21-0.21)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.56 (0.55-0.56)	0.56 (0.55-0.57)	0.56 (0.55-0.57)	0.54 (0.53-0.55)	0.57 (0.57-0.58)	0.56 (0.55-0.56)	0.56 (0.55-0.56)
	Relative AUC-ROC	-0.71 (-1.06--0.38)	0.14 (-0.19-0.47)	0.66 (0.35-0.98)	-1.22 (-1.46--0.98)	1.13 (0.90-1.35)	0.00 (0.00-0.00)	0.01 (0.00-0.02)
	Relative TPR at global threshold	-0.28 (-0.96-0.51)	-1.06 (-1.95--0.39)	1.47 (0.78-2.20)	-3.09 (-3.71--2.66)	2.95 (2.54-3.52)	0.00 (0.00-0.00)	0.05 (0.03-0.06)
	Relative FPR at global threshold	-0.54 (-2.24-1.10)	-4.04 (-5.82--2.20)	4.44 (2.90-6.15)	-3.88 (-5.01--2.61)	4.01 (2.69-5.17)	0.00 (0.00-0.00)	-0.04 (-0.06--0.02)
	Relative Youden's Index at global threshold	-0.18 (-1.30-1.01)	0.01 (-1.22-1.08)	0.41 (-0.70-1.49)	-2.81 (-3.74--2.13)	2.57 (1.92-3.49)	0.00 (0.00-0.00)	0.08 (0.06-0.10)
CXR-FMKD-Direct FFT (CS)	AUC-ROC	0.85 (0.85-0.85)	0.85 (0.85-0.86)	0.86 (0.86-0.87)	0.85 (0.84-0.85)	0.86 (0.86-0.87)	0.86 (0.85-0.86)	0.86 (0.85-0.86)
	TPR at global threshold	0.76 (0.75-0.77)	0.75 (0.74-0.75)	0.80 (0.79-0.80)	0.74 (0.73-0.75)	0.79 (0.78-0.80)	0.77 (0.76-0.77)	0.77 (0.76-0.77)
	FPR at global threshold	0.20 (0.19-0.20)	0.19 (0.19-0.19)	0.21 (0.21-0.22)	0.18 (0.18-0.19)	0.22 (0.21-0.22)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.56 (0.55-0.57)	0.56 (0.55-0.57)	0.58 (0.57-0.59)	0.56 (0.55-0.56)	0.58 (0.57-0.58)	0.57 (0.56-0.57)	0.57 (0.56-0.57)
	Relative AUC-ROC	-0.53 (-0.87--0.20)	-0.27 (-0.61-0.06)	0.90 (0.58-1.22)	-1.00 (-1.25--0.76)	0.91 (0.68-1.14)	0.00 (0.00-0.00)	-0.00 (-0.02-0.01)
	Relative TPR at global threshold	-1.08 (-1.73--0.25)	-2.52 (-3.40--1.87)	3.75 (3.09-4.49)	-3.50 (-3.94--2.84)	3.34 (2.71-3.77)	0.00 (0.00-0.00)	0.06 (0.03-0.07)
	Relative FPR at global threshold	-1.72 (-3.32-0.01)	-5.22 (-6.85--3.52)	6.66 (4.96-8.26)	-8.22 (-9.26--6.91)	8.49 (7.12-9.57)	0.00 (0.00-0.00)	-0.09 (-0.12--0.06)
	Relative Youden's Index at global threshold	-0.86 (-1.92-0.44)	-1.57 (-2.94--0.53)	2.73 (1.71-3.88)	-1.83 (-2.57--0.89)	1.53 (0.63-2.25)	0.00 (0.00-0.00)	0.11 (0.07-0.14)
CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1)	AUC-ROC	0.85 (0.85-0.85)	0.85 (0.85-0.85)	0.86 (0.86-0.87)	0.85 (0.84-0.85)	0.86 (0.86-0.87)	0.85 (0.85-0.86)	0.85 (0.85-0.86)
	TPR at global threshold	0.75 (0.74-0.76)	0.76 (0.75-0.76)	0.79 (0.78-0.79)	0.74 (0.73-0.75)	0.79 (0.78-0.79)	0.76 (0.76-0.77)	0.76 (0.76-0.77)
	FPR at global threshold	0.19 (0.19-0.20)	0.19 (0.19-0.20)	0.22 (0.21-0.22)	0.18 (0.18-0.19)	0.22 (0.21-0.22)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.56 (0.55-0.57)	0.56 (0.55-0.57)	0.57 (0.56-0.58)	0.56 (0.55-0.56)	0.57 (0.56-0.58)	0.56 (0.56-0.57)	0.56 (0.56-0.57)
	Relative AUC-ROC	-0.53 (-0.88--0.19)	-0.40 (-0.73--0.07)	1.01 (0.70-1.33)	-1.03 (-1.27--0.78)	0.94 (0.70-1.17)	0.00 (0.00-0.00)	0.00 (-0.01-0.01)
	Relative TPR at global threshold	-1.64 (-2.38--0.94)	-1.17 (-1.97--0.38)	2.95 (2.22-3.66)	-3.31 (-3.85--2.81)	3.17 (2.68-3.68)	0.00 (0.00-0.00)	0.05 (0.04-0.07)
	Relative FPR at global threshold	-4.31 (-5.68--2.50)	-4.07 (-5.94--2.47)	8.12 (6.31-9.75)	-8.21 (-9.56--7.14)	8.47 (7.37-9.87)	0.00 (0.00-0.00)	-0.09 (-0.13--0.06)
	Relative Youden's Index at global threshold	-0.69 (-1.91-0.37)	-0.14 (-1.39-1.19)	1.12 (-0.04-2.36)	-1.57 (-2.33--0.73)	1.28 (0.46-2.01)	0.00 (0.00-0.00)	0.11 (0.08-0.14)

**Relative values are computed with respect to the Subgroups' Average*

***All data are presented as: value (95% CI), derived from bootstrapping with 2000 samples*

Table 33. Detailed Absolute and Relative Subgroup Performances for 'No Finding' in MIMIC.
This table reports disease detection performance for our five selected models across race and sex subgroups, their average, and the total patient sample, denoted 'All'. Data is presented as mean and 95% confidence interval (CI). TPR and FPR for each subgroup are determined at a fixed (global) decision threshold, optimised to achieve an FPR of 20% across the entire patient sample.

MIMIC – Cardiomegaly

Model	Metric	Subgroups						All
		White	Asian	Black	Male	Female	*Average**	
CXR-FM	AUC-ROC	0.78 (0.77-0.78)	0.78 (0.77-0.78)	0.78 (0.77-0.78)	0.77 (0.77-0.78)	0.78 (0.78-0.79)	0.78 (0.78-0.78)	0.78 (0.78-0.78)
	TPR at global threshold	0.55 (0.54-0.56)	0.52 (0.51-0.53)	0.63 (0.62-0.64)	0.55 (0.54-0.55)	0.59 (0.59-0.60)	0.57 (0.56-0.57)	0.57 (0.56-0.58)
	FPR at global threshold	0.20 (0.20-0.20)	0.17 (0.16-0.17)	0.23 (0.23-0.24)	0.19 (0.19-0.20)	0.21 (0.20-0.21)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.35 (0.34-0.36)	0.35 (0.34-0.36)	0.39 (0.38-0.40)	0.35 (0.34-0.36)	0.39 (0.38-0.40)	0.37 (0.36-0.37)	0.37 (0.36-0.38)
	Relative AUC-ROC	-0.35 (-0.85-0.10)	0.06 (-0.42-0.51)	0.05 (-0.37-0.49)	-0.39 (-0.70--0.08)	0.62 (0.31-0.95)	0.00 (0.00-0.00)	0.08 (0.06-0.11)
	Relative TPR at global threshold	-2.78 (-4.10--1.42)	-8.77 (-10.15--7.42)	10.69 (9.41-11.86)	-3.89 (-4.82--2.98)	4.74 (3.79-5.71)	0.00 (0.00-0.00)	0.46 (0.38-0.54)
	Relative FPR at global threshold	0.56 (-0.94-2.20)	-17.45 (-19.06--16.04)	17.22 (15.49-19.05)	-2.99 (-4.14--1.97)	2.66 (1.62-3.81)	0.00 (0.00-0.00)	-0.20 (-0.25--0.15)
	Relative Youden's Index at global threshold	-4.60 (-6.91--2.35)	-4.03 (-6.31--1.82)	7.13 (5.07-9.11)	-4.38 (-5.85--2.81)	5.88 (4.29-7.40)	0.00 (0.00-0.00)	0.82 (0.69-0.95)
CXR-Model FFT	AUC-ROC	0.80 (0.79-0.80)	0.78 (0.77-0.78)	0.81 (0.80-0.81)	0.79 (0.79-0.80)	0.80 (0.80-0.80)	0.79 (0.79-0.80)	0.80 (0.79-0.80)
	TPR at global threshold	0.60 (0.59-0.61)	0.52 (0.51-0.53)	0.67 (0.66-0.68)	0.60 (0.59-0.61)	0.61 (0.60-0.62)	0.60 (0.59-0.61)	0.60 (0.60-0.61)
	FPR at global threshold	0.20 (0.20-0.20)	0.18 (0.18-0.18)	0.22 (0.22-0.23)	0.21 (0.20-0.21)	0.19 (0.19-0.19)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.40 (0.39-0.41)	0.34 (0.33-0.35)	0.45 (0.44-0.46)	0.39 (0.38-0.40)	0.41 (0.41-0.42)	0.40 (0.39-0.41)	0.40 (0.40-0.41)
	Relative AUC-ROC	0.37 (-0.07-0.81)	-2.14 (-2.57--1.70)	1.53 (1.13-1.95)	-0.34 (-0.66--0.04)	0.58 (0.28-0.89)	0.00 (0.00-0.00)	0.13 (0.10-0.16)
	Relative TPR at global threshold	0.43 (-0.93-1.50)	-13.51 (-14.59--12.12)	12.17 (11.06-13.30)	-0.35 (-1.20-0.44)	1.26 (0.44-2.13)	0.00 (0.00-0.00)	0.46 (0.38-0.54)
	Relative FPR at global threshold	-0.94 (-2.46-0.61)	-10.79 (-12.30--9.24)	12.05 (10.43-13.72)	3.35 (2.30-4.55)	-3.67 (-4.87--2.64)	0.00 (0.00-0.00)	-0.12 (-0.15--0.08)
	Relative Youden's Index at global threshold	1.12 (-1.11-2.94)	-14.87 (-16.60--12.56)	12.23 (10.40-14.13)	-2.21 (-3.70--0.92)	3.73 (2.38-5.22)	0.00 (0.00-0.00)	0.75 (0.62-0.88)
CXR-FMKD-Direct FFT (MSE)	AUC-ROC	0.80 (0.80-0.80)	0.78 (0.77-0.78)	0.81 (0.81-0.82)	0.79 (0.79-0.80)	0.80 (0.80-0.81)	0.80 (0.80-0.80)	0.80 (0.80-0.80)
	TPR at global threshold	0.58 (0.57-0.59)	0.56 (0.55-0.57)	0.71 (0.70-0.72)	0.59 (0.58-0.60)	0.66 (0.65-0.66)	0.62 (0.61-0.63)	0.62 (0.62-0.63)
	FPR at global threshold	0.19 (0.18-0.19)	0.18 (0.18-0.18)	0.24 (0.23-0.24)	0.19 (0.19-0.19)	0.21 (0.21-0.22)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.39 (0.38-0.40)	0.38 (0.37-0.40)	0.47 (0.47-0.48)	0.40 (0.40-0.41)	0.44 (0.43-0.45)	0.42 (0.41-0.43)	0.42 (0.42-0.43)
	Relative AUC-ROC	0.20 (-0.22-0.61)	-2.49 (-2.94--2.06)	1.93 (1.55-2.32)	-0.40 (-0.69--0.10)	0.75 (0.45-1.06)	0.00 (0.00-0.00)	0.20 (0.16-0.23)
	Relative TPR at global threshold	-6.86 (-8.04--5.61)	-9.03 (-10.47--8.05)	14.63 (13.73-15.92)	-4.52 (-5.38--3.67)	5.78 (4.95-6.64)	0.00 (0.00-0.00)	0.67 (0.58-0.76)
	Relative FPR at global threshold	-6.90 (-8.25--5.08)	-10.91 (-12.62--9.60)	18.16 (16.42-19.71)	-6.49 (-7.64--5.36)	6.14 (4.99-7.29)	0.00 (0.00-0.00)	-0.25 (-0.30--0.20)
	Relative Youden's Index at global threshold	-6.83 (-8.75--4.98)	-8.14 (-10.27--6.43)	12.95 (11.43-14.95)	-3.58 (-4.95--2.17)	5.61 (4.21-6.97)	0.00 (0.00-0.00)	1.10 (0.97-1.26)
CXR-FMKD-Direct FFT (CS)	AUC-ROC	0.79 (0.79-0.80)	0.78 (0.78-0.79)	0.81 (0.81-0.82)	0.80 (0.79-0.80)	0.80 (0.80-0.81)	0.80 (0.80-0.80)	0.80 (0.80-0.80)
	TPR at global threshold	0.58 (0.57-0.59)	0.54 (0.53-0.55)	0.68 (0.67-0.68)	0.60 (0.59-0.60)	0.61 (0.60-0.62)	0.60 (0.59-0.61)	0.60 (0.60-0.61)
	FPR at global threshold	0.19 (0.19-0.19)	0.19 (0.19-0.20)	0.22 (0.21-0.22)	0.21 (0.20-0.21)	0.19 (0.19-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.39 (0.38-0.40)	0.35 (0.34-0.36)	0.46 (0.45-0.47)	0.39 (0.38-0.40)	0.42 (0.41-0.43)	0.40 (0.39-0.41)	0.40 (0.40-0.41)
	Relative AUC-ROC	-0.37 (-0.81-0.06)	-1.83 (-2.25--1.41)	1.89 (1.50-2.31)	-0.21 (-0.50-0.10)	0.52 (0.22-0.82)	0.00 (0.00-0.00)	0.16 (0.13-0.18)
	Relative TPR at global threshold	-3.22 (-4.57--2.01)	-10.40 (-11.56--8.91)	12.58 (11.34-13.75)	-0.87 (-1.73--0.00)	1.90 (1.03-2.77)	0.00 (0.00-0.00)	0.53 (0.44-0.61)
	Relative FPR at global threshold	-4.13 (-5.64--2.59)	-3.80 (-5.30--2.24)	8.17 (6.53-9.70)	2.53 (1.52-3.76)	-2.77 (-4.03--1.75)	0.00 (0.00-0.00)	-0.09 (-0.12--0.06)
	Relative Youden's Index at global threshold	-2.77 (-4.98--0.81)	-13.70 (-15.58--11.29)	14.79 (12.75-16.72)	-2.57 (-4.01--1.14)	4.24 (2.80-5.66)	0.00 (0.00-0.00)	0.84 (0.71-0.96)
CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1)	AUC-ROC	0.80 (0.79-0.80)	0.78 (0.78-0.79)	0.81 (0.81-0.81)	0.79 (0.79-0.80)	0.80 (0.80-0.81)	0.80 (0.80-0.80)	0.80 (0.80-0.80)
	TPR at global threshold	0.60 (0.60-0.61)	0.54 (0.53-0.55)	0.69 (0.68-0.70)	0.60 (0.59-0.61)	0.63 (0.63-0.64)	0.61 (0.61-0.62)	0.62 (0.61-0.62)
	FPR at global threshold	0.20 (0.20-0.20)	0.17 (0.17-0.18)	0.23 (0.22-0.23)	0.20 (0.20-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.40 (0.39-0.41)	0.37 (0.36-0.38)	0.46 (0.45-0.47)	0.40 (0.39-0.41)	0.43 (0.43-0.44)	0.41 (0.41-0.42)	0.42 (0.41-0.42)
	Relative AUC-ROC	-0.10 (-0.52-0.31)	-1.70 (-2.12--1.25)	1.50 (1.11-1.90)	-0.45 (-0.75--0.16)	0.75 (0.46-1.04)	0.00 (0.00-0.00)	0.17 (0.14-0.20)
	Relative TPR at global threshold	-1.40 (-2.49--0.19)	-12.01 (-13.14--10.75)	12.45 (11.27-13.43)	-2.31 (-3.20--1.54)	3.27 (2.48-4.15)	0.00 (0.00-0.00)	0.50 (0.41-0.58)
	Relative FPR at global threshold	0.24 (-1.40-1.72)	-13.04 (-14.46--11.23)	13.08 (11.28-14.60)	-0.34 (-1.47-0.73)	0.06 (-1.01-1.20)	0.00 (0.00-0.00)	-0.14 (-0.18--0.10)
	Relative Youden's Index at global threshold	-2.20 (-3.99--0.24)	-11.50 (-13.50--9.42)	12.14 (10.23-13.82)	-3.27 (-4.62--2.05)	4.83 (3.58-6.20)	0.00 (0.00-0.00)	0.81 (0.68-0.93)

**Relative values are computed with respect to the Subgroups' Average*

***All data are presented as: value (95% CI), derived from bootstrapping with 2000 samples*

Table 34. Detailed Absolute and Relative Subgroup Performances for 'Cardiomegaly' in MIMIC.
This table reports disease detection performance for our five selected models across race and sex subgroups, their average, and the total patient sample, denoted 'All'. Data is presented as mean and 95% confidence interval (CI). TPR and FPR for each subgroup are determined at a fixed (global) decision threshold, optimised to achieve an FPR of 20% across the entire patient sample.

MIMIC – Pneumothorax

Model	Metric	Subgroups						All
		White	Asian	Black	Male	Female	*Average**	
CXR-FM	AUC-ROC	0.80 (0.80-0.81)	0.83 (0.82-0.83)	0.81 (0.80-0.81)	0.81 (0.81-0.82)	0.82 (0.82-0.83)	0.81 (0.81-0.82)	0.82 (0.81-0.82)
	TPR at global threshold	0.65 (0.63-0.67)	0.71 (0.69-0.72)	0.53 (0.51-0.56)	0.68 (0.66-0.69)	0.60 (0.59-0.62)	0.63 (0.62-0.64)	0.65 (0.63-0.66)
	FPR at global threshold	0.20 (0.20-0.21)	0.24 (0.23-0.24)	0.16 (0.16-0.16)	0.23 (0.23-0.24)	0.17 (0.16-0.17)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.44 (0.43-0.46)	0.47 (0.45-0.48)	0.37 (0.35-0.40)	0.45 (0.43-0.46)	0.44 (0.42-0.45)	0.43 (0.42-0.44)	0.45 (0.43-0.46)
	Relative AUC-ROC	-1.20 (-2.01--0.47)	1.54 (0.84-2.25)	-1.04 (-1.93--0.16)	-0.30 (-0.84-0.26)	1.00 (0.43-1.55)	0.00 (0.00-0.00)	0.39 (0.32-0.48)
	Relative TPR at global threshold	2.31 (0.08-4.64)	11.23 (8.82-13.09)	-15.86 (-18.53--12.91)	7.33 (5.58-8.85)	-5.02 (-6.71--3.16)	0.00 (0.00-0.00)	1.71 (1.41-1.99)
	Relative FPR at global threshold	2.37 (0.77-3.59)	17.68 (16.29-19.40)	-19.78 (-21.14--18.49)	17.03 (16.02-17.91)	-17.31 (-18.19--16.31)	0.00 (0.00-0.00)	-0.09 (-0.15--0.03)
	Relative Youden's Index at global threshold	2.28 (-0.93-5.84)	8.25 (4.71-10.98)	-14.05 (-18.16--9.60)	2.86 (0.30-5.15)	0.65 (-1.91-3.45)	0.00 (0.00-0.00)	2.54 (2.10-2.95)
CXR-Model FFT	AUC-ROC	0.84 (0.83-0.84)	0.87 (0.86-0.87)	0.86 (0.86-0.87)	0.86 (0.85-0.86)	0.86 (0.85-0.86)	0.86 (0.85-0.86)	0.86 (0.85-0.86)
	TPR at global threshold	0.70 (0.69-0.72)	0.83 (0.81-0.84)	0.68 (0.66-0.71)	0.77 (0.76-0.79)	0.72 (0.70-0.73)	0.74 (0.73-0.75)	0.75 (0.74-0.76)
	FPR at global threshold	0.21 (0.21-0.21)	0.24 (0.23-0.24)	0.16 (0.15-0.16)	0.22 (0.22-0.22)	0.18 (0.18-0.18)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.49 (0.48-0.51)	0.59 (0.58-0.60)	0.53 (0.51-0.55)	0.55 (0.54-0.57)	0.54 (0.53-0.55)	0.54 (0.53-0.55)	0.55 (0.54-0.56)
	Relative AUC-ROC	-2.39 (-3.10--1.71)	1.17 (0.62-1.73)	0.79 (0.06-1.49)	0.28 (-0.17-0.71)	0.16 (-0.28-0.63)	0.00 (0.00-0.00)	0.22 (0.17-0.29)
	Relative TPR at global threshold	-5.28 (-7.00--3.46)	11.44 (9.75-12.90)	-7.70 (-9.90--5.48)	4.50 (3.25-5.95)	-2.97 (-4.50--1.61)	0.00 (0.00-0.00)	1.11 (0.88-1.35)
	Relative FPR at global threshold	4.38 (3.12-5.95)	17.55 (16.02-18.95)	-21.67 (-23.02--20.31)	10.46 (9.54-11.43)	-10.73 (-11.72--9.79)	0.00 (0.00-0.00)	-0.10 (-0.16--0.04)
	Relative Youden's Index at global threshold	-8.85 (-11.31--6.28)	9.18 (6.80-11.31)	-2.53 (-5.63-0.54)	2.30 (0.55-4.22)	-0.09 (-2.18-1.78)	0.00 (0.00-0.00)	1.55 (1.24-1.89)
CXR-FMKD-Direct FFT (MSE)	AUC-ROC	0.86 (0.85-0.87)	0.88 (0.87-0.88)	0.86 (0.85-0.87)	0.87 (0.86-0.87)	0.87 (0.86-0.88)	0.87 (0.86-0.87)	0.87 (0.87-0.87)
	TPR at global threshold	0.78 (0.76-0.79)	0.84 (0.82-0.85)	0.67 (0.65-0.69)	0.79 (0.78-0.81)	0.75 (0.74-0.77)	0.77 (0.76-0.78)	0.78 (0.77-0.79)
	FPR at global threshold	0.22 (0.22-0.23)	0.24 (0.24-0.24)	0.14 (0.14-0.14)	0.22 (0.21-0.22)	0.18 (0.18-0.19)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.55 (0.54-0.57)	0.60 (0.58-0.61)	0.53 (0.51-0.55)	0.58 (0.57-0.59)	0.57 (0.56-0.58)	0.56 (0.56-0.58)	0.58 (0.57-0.59)
	Relative AUC-ROC	-0.93 (-1.52--0.33)	1.40 (0.85-1.93)	-0.99 (-1.69--0.32)	0.17 (-0.26-0.57)	0.36 (-0.06-0.81)	0.00 (0.00-0.00)	0.31 (0.24-0.37)
	Relative TPR at global threshold	1.57 (-0.02-3.36)	9.24 (7.65-10.61)	-12.94 (-15.12--10.77)	3.82 (2.56-4.88)	-1.68 (-2.83--0.30)	0.00 (0.00-0.00)	1.32 (1.09-1.53)
	Relative FPR at global threshold	11.32 (9.92-12.74)	18.93 (17.45-20.39)	-29.92 (-31.17--28.63)	8.04 (7.05-9.04)	-8.37 (-9.38--7.36)	0.00 (0.00-0.00)	-0.14 (-0.21--0.07)
	Relative Youden's Index at global threshold	-1.90 (-4.09-0.48)	5.80 (3.64-7.80)	-6.92 (-9.88--3.97)	2.32 (0.57-3.79)	0.70 (-0.92-2.58)	0.00 (0.00-0.00)	1.83 (1.53-2.12)
CXR-FMKD-Direct FFT (CS)	AUC-ROC	0.85 (0.85-0.86)	0.86 (0.86-0.87)	0.87 (0.86-0.88)	0.85 (0.85-0.86)	0.87 (0.87-0.88)	0.86 (0.86-0.87)	0.86 (0.86-0.87)
	TPR at global threshold	0.75 (0.73-0.76)	0.82 (0.81-0.84)	0.66 (0.64-0.68)	0.76 (0.75-0.78)	0.75 (0.74-0.76)	0.75 (0.74-0.76)	0.76 (0.75-0.77)
	FPR at global threshold	0.21 (0.21-0.21)	0.24 (0.24-0.25)	0.15 (0.15-0.15)	0.23 (0.23-0.23)	0.17 (0.17-0.17)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.54 (0.52-0.55)	0.58 (0.57-0.59)	0.51 (0.49-0.53)	0.54 (0.52-0.55)	0.58 (0.56-0.59)	0.55 (0.54-0.56)	0.56 (0.55-0.57)
	Relative AUC-ROC	-1.10 (-1.73--0.46)	-0.03 (-0.61-0.54)	0.77 (0.10-1.46)	-1.05 (-1.48--0.63)	1.41 (0.98-1.83)	0.00 (0.00-0.00)	0.21 (0.16-0.26)
	Relative TPR at global threshold	-0.13 (-2.12-1.49)	10.00 (8.42-11.60)	-12.14 (-14.31--9.76)	2.24 (1.04-3.45)	0.03 (-1.34-1.37)	0.00 (0.00-0.00)	1.24 (1.01-1.46)
	Relative FPR at global threshold	4.21 (2.94-5.80)	22.00 (20.38-23.28)	-25.88 (-27.16--24.50)	13.87 (13.01-14.91)	-14.20 (-15.28--13.34)	0.00 (0.00-0.00)	-0.12 (-0.19--0.05)
	Relative Youden's Index at global threshold	-1.72 (-4.60-0.55)	5.61 (3.44-7.97)	-7.12 (-10.12--3.88)	-2.01 (-3.72--0.34)	5.24 (3.36-7.15)	0.00 (0.00-0.00)	1.73 (1.42-2.04)
CXR-FMKD-Direct FFT (MSE-CS | 0.9-0.1)	AUC-ROC	0.86 (0.85-0.86)	0.87 (0.87-0.88)	0.87 (0.87-0.88)	0.87 (0.86-0.87)	0.87 (0.87-0.88)	0.87 (0.87-0.87)	0.87 (0.87-0.87)
	TPR at global threshold	0.75 (0.74-0.77)	0.85 (0.84-0.86)	0.68 (0.66-0.70)	0.78 (0.77-0.79)	0.77 (0.76-0.79)	0.77 (0.76-0.78)	0.78 (0.77-0.79)
	FPR at global threshold	0.21 (0.21-0.22)	0.25 (0.24-0.25)	0.14 (0.14-0.14)	0.22 (0.22-0.22)	0.18 (0.18-0.18)	0.20 (0.20-0.20)	0.20 (0.20-0.20)
	Youden's Index at global threshold	0.54 (0.52-0.55)	0.60 (0.59-0.62)	0.54 (0.52-0.56)	0.56 (0.55-0.57)	0.60 (0.58-0.61)	0.57 (0.56-0.58)	0.58 (0.57-0.59)
	Relative AUC-ROC	-1.45 (-2.06--0.83)	0.71 (0.18-1.24)	0.35 (-0.27-1.00)	0.06 (-0.33-0.45)	0.33 (-0.08-0.73)	0.00 (0.00-0.00)	0.24 (0.19-0.30)
	Relative TPR at global threshold	-2.16 (-3.85--0.39)	11.00 (9.53-12.56)	-11.16 (-13.39--9.15)	1.42 (0.29-2.58)	0.90 (-0.38-2.17)	0.00 (0.00-0.00)	1.18 (0.97-1.40)
	Relative FPR at global threshold	7.16 (5.88-8.75)	24.00 (22.32-25.44)	-30.80 (-32.14--29.45)	10.36 (9.31-11.35)	-10.72 (-11.69--9.67)	0.00 (0.00-0.00)	-0.15 (-0.22--0.07)
	Relative Youden's Index at global threshold	-5.45 (-7.84--3.03)	6.42 (4.46-8.46)	-4.23 (-7.12--1.49)	-1.74 (-3.27--0.09)	5.00 (3.23-6.68)	0.00 (0.00-0.00)	1.65 (1.37-1.96)

**Relative values are computed with respect to the Subgroups' Average*

***All data are presented as: value (95% CI), derived from bootstrapping with 2000 samples*

Table 35. Detailed Absolute and Relative Subgroup Performances for 'Pneumothorax' in MIMIC.
This table reports disease detection performance for our five selected models across race and sex subgroups, their average, and the total patient sample, denoted 'All'. Data is presented as mean and 95% confidence interval (CI). TPR and FPR for each subgroup are determined at a fixed (global) decision threshold, optimised to achieve an FPR of 20% across the entire patient sample.

S.10. Performance vs. Bias Analysis – CheXpert

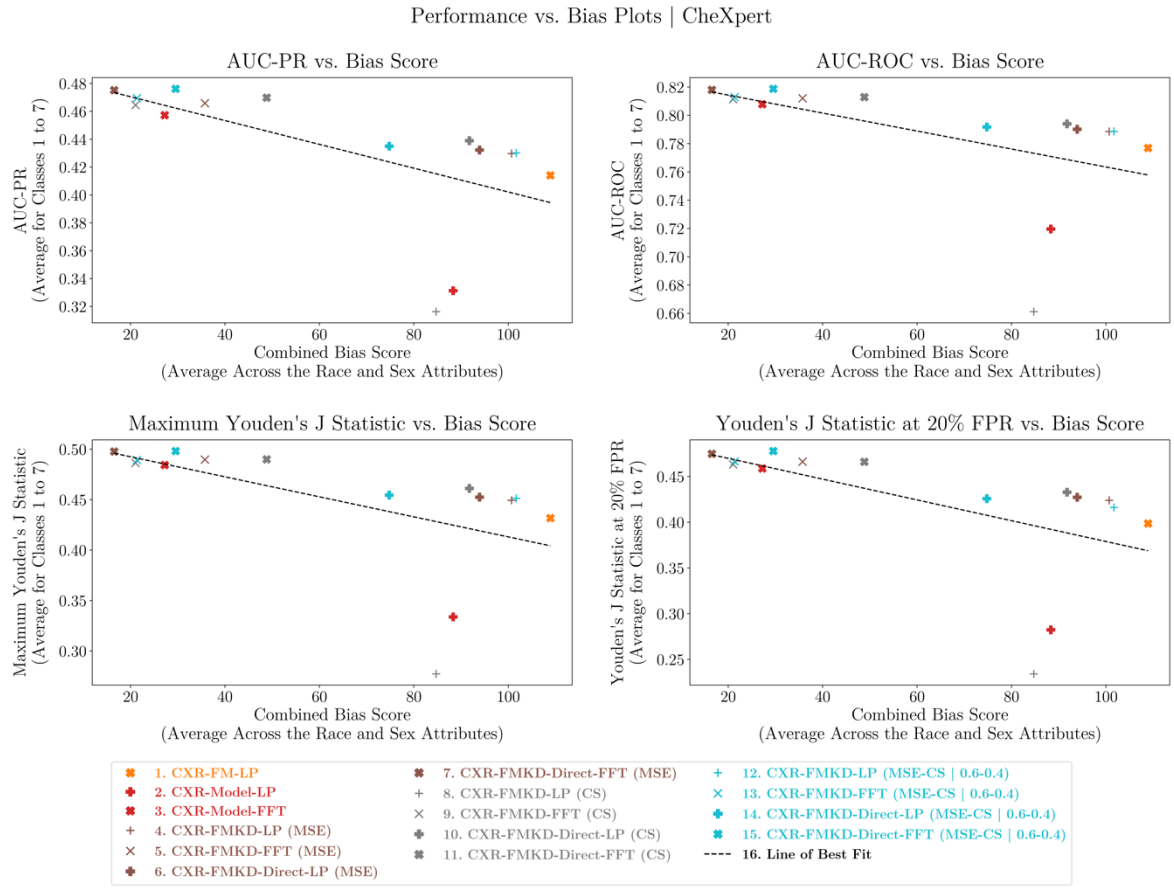

Figure 119. Performance versus Bias Plots for 15 Selected Models Tested on CheXpert.
This figure explores the relationship between performance and bias for the 15 selected models tested on CheXpert. It presents performance data for four metrics—AUC-ROC, AUC-PR, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR—focusing on the average results for the most significant disease labels (Classes 1 to 7). Performance results for each metric are plotted against the computed Combined Bias Score, which averages the bias scores across the race and sex attributes for each model—derived using the bias quantification method proposed in this study. The evaluated models include the teacher CXR-FM; all four variants of the three selected student model types (MSE, CS, and MSE-CS | 0.6-0.4); and the baselines CXR-Model FFT and CXR-Model LP. A general negative correlation is observed, where models with higher performance exhibit lower bias, as indicated by the lines of best fit for each plot. This trend is more pronounced when excluding the CXR-FMKD LP (CS) and CXR-Model LP, which stand out as outliers due to their significant underperformance compared to the other models.

S.11. Performance vs. Bias Analysis – MIMIC

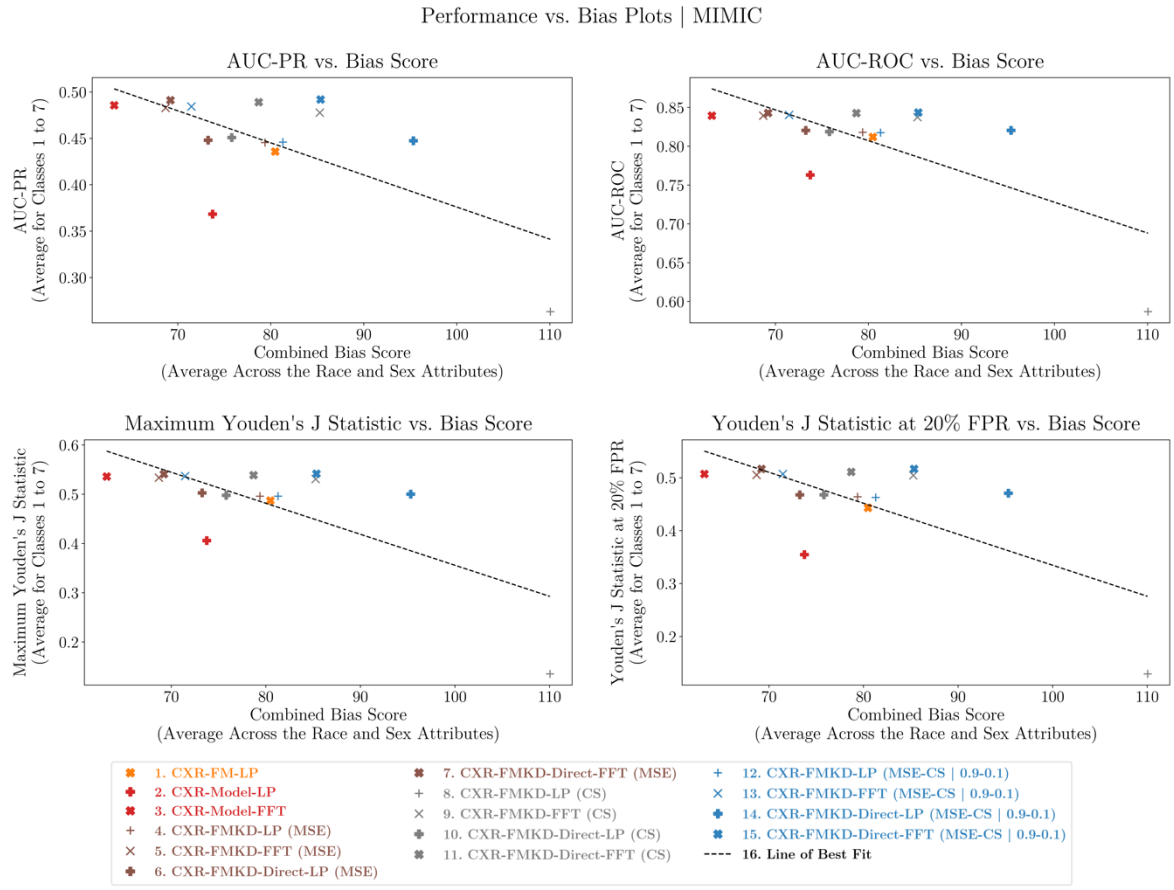

Figure 120. Performance versus Bias Plots for 15 Selected Models Tested on MIMIC.
This figure explores the relationship between performance and bias for the 15 selected models tested on MIMIC. It presents performance data for four metrics—AUC-ROC, AUC-PR, Max Youden's J Statistic, and Youden's J Statistic at 20% FPR—focusing on the average results for the most significant disease labels (Classes 1 to 7). Performance results for each metric are plotted against the computed Combined Bias Score, which averages the bias scores across the race and sex attributes for each model—derived using the bias quantification method proposed in this study. The evaluated models include the teacher CXR-FM; all four variants of the three selected student model types (MSE, CS, and MSE-CS | 0.9-0.1); and the baselines CXR-Model FFT and CXR-Model LP. A general negative correlation is observed, where models with higher performance exhibit lower bias, as indicated by the lines of best fit for each plot. The CXR-FMKD LP (CS) and CXR-Model LP models stand out as outliers due to their significant underperformance compared to the other models, with CXR-FMKD LP (CS) displaying the worst performance, which correlates to the highest measured bias.