

Master_Thesis.pdf



Document Details

Submission ID

trn:oid:::29034:104133953

Submission Date

Jul 11, 2025, 11:14 AM GMT+5

Download Date

Jul 11, 2025, 11:17 AM GMT+5

File Name

Master_Thesis.pdf

File Size

924.0 KB

45 Pages

7,082 Words

42,494 Characters



27% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

-  **16 AI-generated only 27%**
Likely AI-generated text from a large-language model.
-  **0 AI-generated text that was AI-paraphrased 0%**
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.



Improving Generalization and Robustness of Chest X-ray AI Model: Preprocessing Methods to Mitigate Racial Bias

Master's Thesis in Computer Science

submitted
by

Dishantkumar Sutariya

born 24.10.1999 in Bela

Work produced at
Fraunhofer Institute for Digital Medicine MEVIS

and evaluated at
Chair of Digital Health (FAU)
Department Medical Informatics
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU).

Advisor: Prof. Dr. Andreas Rowald¹, Dr. Eike Petersen²

¹ Chair of Digital Health, FAU,

² Fraunhofer Institute for Digital Medicine MEVIS

Started: 15.02.2025

Finished: 15.08.2025

ii

iii

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Die Richtlinien des Lehrstuhls für Bachelor- und Masterarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Erlangen, den 15. August 2025

iv

Übersicht

Rassenbedingte Verzerrungen in Modellen der künstlichen Intelligenz (KI) für die Röntgendiagnostik stellen ein erhebliches Risiko für die Gleichbehandlung im Gesundheitswesen dar, da diese Modelle bei bestimmten demografischen Gruppen schlecht abschneiden können, was zu möglichen Fehldiagnosen führt. Diese Arbeit reproduziert das Training eines Krankheitsklassifikators und zeigt, dass Ethnie immer noch genau aus der Einbettung des Modells vorhergesagt werden kann - obwohl das Modell nicht explizit dafür trainiert wurde und Ethnie kein relevantes Merkmal für die Krankheitsvorhersage ist. Dieses Ergebnis deutet darauf hin, dass es eine demografische Abkürzung geben könnte, bei der sich die Modelle auf falsche, nicht-klinische, mit der Ethnie korrelierende Hinweise stützen können. Um dieses Problem anzugehen, schlagen wir eine Reihe von Vorverarbeitungsmethoden vor, die darauf abzielen, rassistische Verzerrungen abzuschwächen. Dazu gehören die Maskierung der Lunge, um die Aufmerksamkeit des Modells auf klinisch relevante Regionen zu beschränken, und die kontrastbegrenzte adaptive Histogramm-Entzerrung (CLAHE), um den lokalen Kontrast zu erhöhen und die Sichtbarkeit der Merkmale zu verbessern. Diese Ansätze zielen darauf ab, nicht-klinische Hinweise zu unterdrücken, die möglicherweise rassistische Informationen kodieren, und gleichzeitig die Gesamtleistung des Modells zu erhalten oder sogar zu verbessern. Die Ergebnisse tragen zur Entwicklung robuster, gerechter KI-gesteuerter CXR-Diagnosesysteme bei und bieten praktische Einblicke in Strategien zur Abschwächung von Verzerrungen in der KI für medizinische Bildgebung.

Abstract

Racial bias in artificial intelligence (AI) models for chest X-ray (CXR) diagnostics pose significant risks to healthcare equity, as these models can perform poorly across certain demographic groups, leading to potential misdiagnoses. This thesis reproduces the training of a disease classifier and demonstrates that race can still be accurately predicted from the model's embedding—even though the model was not explicitly trained to do so, and race is not a relevant feature for disease prediction. This finding suggests there might be presence of a demographic shortcut, where models may rely on spurious, non-clinical cues correlated with race. To address this issue, we propose a set of preprocessing methods aimed at mitigating racial bias. These include lung masking to restrict the model's attention to clinically relevant regions, and Contrast Limited Adaptive Histogram Equalization (CLAHE) to enhance local contrast and improve feature visibility. **These approaches aim to suppress non-clinical cues that may encode racial information while preserving—or even**

vi

enhancing—overall model performance. The findings contribute to the development of robust, equitable AI-driven CXR diagnostic systems and offer practical insights into bias mitigation strategies in medical imaging AI.

Contents

1	Introduction	1
1.1	Motivation and Context	1
1.2	Problem Statement	2
1.3	Research Objectives and Hypotheses	2
1.3.1	Bias definition	2
1.3.2	Research Questions	3
1.3.3	Hypothesis	3
2	Literature Review	5
2.1	Chest X-ray Imaging	5
2.2	Bias and Fairness in AI-based CXR diagnosis	6
2.2.1	Sources of Bias	6
2.2.2	AI recognition of patient race in CXR Images	7
2.2.3	Diagnostic Bias	9
2.3	Mitigation Strategies for Bias in Medical Imaging	10
3	Methodology	13
3.1	Datasets	13
3.1.1	CheXpert Dataset	13
3.1.2	MIMIC Dataset	14
3.1.3	Data Splitting & Sampling	14
3.1.4	CheXmask Dataset	15
3.2	Model Architectures	16
3.2.1	Foundational Concepts	16
3.2.2	DenseNet 121	17
3.2.3	HybridGNet	19

viii

CONTENTS

3.3	Experimental Setup	19
3.3.1	Implementation Details	19
4	Results	21
4.1	Overview	21
4.2	Baseline Performance	21
	List of Abbreviations	27
	List of Figures	29
	List of Tables	31
	Bibliography	33

Chapter 1

Introduction

1.1 Motivation and Context

The integration of artificial intelligence into medical imaging holds transformative promise, particularly in automating diagnostics and improving clinical efficiency and decision making. However, the deployment of these systems in real-world healthcare settings has revealed critical concerns about their fairness and reliability across diverse patient populations. One of the most alarming findings in recent research of gichoya AI recognition of patient race [7] is the presence of racial bias in AI models trained on chest X-ray (CXR) data. These models have been shown to infer race from medical images—despite race being imperceptible to human radiologists—and exhibit unequal diagnostic performance across racial and ethnic groups [7, 20]. Such disparities threaten to reinforce existing healthcare inequities and undermine the trust and safety of AI in clinical practice.

The motivation behind this thesis arises from the urgent need to address these biases not merely as a technical challenge but as a matter of healthcare ethics and equity. While algorithm-level solutions such as fairness-aware learning, adversarial debiasing and data augmentation have been explored [yan24][21], they often lack robustness when applied across different clinical environments[14]. This highlights the need for a more fundamental, data-centric approach that addresses bias at the source—by controlling what the model learns from during training.

This research is situated in the broader context of advancing fair and generalizable medical AI. It proposes a targeted preprocessing strategy for CXR images that limits the influence of non-clinical features associated with demographic bias. By combining anatomical region masking, image normalization, and adaptive histogram equalization techniques like

CLAHE [18], the goal is to shift model attention toward clinically meaningful patterns while minimizing confounding demographic bias factors. This work contributes to the growing field of ethical AI in medicine by exploring practical, scalable solutions that enhance both performance and fairness in diagnostic systems.

1.2 Problem Statement

Despite the promising advancements in AI-based diagnostic tools for chest X-rays (CXR), there remains a critical challenge: these models often learn to make predictions based on non-clinical, demographically linked signals rather than actual pathological features. This problem, known as shortcut learning, enables models to rely on racial or institutional artifacts—such as differences in pixel intensity, image framing, or site-specific markers—as proxies for disease, instead of focusing on medically meaningful areas of the image.

Studies have shown that CXR models can accurately infer patient race, despite race being a non-visual characteristic [7], which raises serious concerns about fairness and model generalization. When a model internalizes race as a shortcut during training, its predictions risk being biased, especially toward underrepresented racial groups [20]. This has been linked to underdiagnosis and diagnostic disparities, particularly in minority populations, potentially leading to harmful clinical outcomes.

This thesis addresses the specific problem of racial encoding shortcut learning in chest X-ray models by proposing a preprocessing-based bias mitigation methods. Instead of modifying model architecture or training objectives, the approach focuses on data-level interventions—such as anatomical masking, and CLAHE-based contrast adjustment—to suppress non-clinical cues and enforce attention on relevant anatomical regions. By doing so, the work aims to reduce racial bias in model predictions and improve fairness without sacrificing diagnostic performance.

1.3 Research Objectives and Hypotheses

1.3.1 Bias definition

Racial/Ethnic Encoding Bias: This refers to the unintended encoding or reflection of racial or ethnic information in model representations or predictions, even when such information is not explicitly provided as input.

1.3. RESEARCH OBJECTIVES AND HYPOTHESES

3

Diagnostic Accuracy Bias: This occurs when the model exhibits different levels of diagnostic performance (e.g., AUROC) across demographic groups (e.g., by race).

1.3.2 Research Questions

The research seeks to answer the following questions:

- (1) Can preprocessing techniques reduce Racial/Ethnic Encoding bias in AI-driven chest X-ray disease diagnosis?
- (2) Do models trained with such preprocessing-based methods improve fairness in diagnostic accuracy bias without significant difference in overall performance?

1.3.3 Hypothesis

For this thesis, following hypotheses are proposed:

- H1: Preprocessing methods can reduce the ability of models to detect race-related shortcuts, thereby mitigating Racial/Ethnic Encoding bias
- H2: Models trained with such preprocessing-based bias mitigation techniques will show improved fairness in diagnostic accuracy bias without significant loss in overall performance, or improve overall performance

The structure of this thesis is organized as follows:

This thesis is organized into five chapters. Chapter 2 presents a comprehensive literature review, examining prior studies that expose the presence of racial and demographic bias in AI models for chest X-ray (CXR) diagnosis and their proposed solutions. Chapter 3 outlines the methodology, describing the datasets used, the design of preprocessing strategies aimed at bias mitigation, as well as the deep learning architectures and experimental setup. Chapter 4 presents the results and discussion, comparing model performance and fairness metrics before and after applying bias mitigation techniques, and analyzing generalization across racial and ethnic subgroups. Chapter 5 concludes the thesis with a summary of key findings and contributions, discusses the limitations of the current approach, and outlines potential directions for future work to develop more equitable AI systems in medical imaging.

CHAPTER 1. INTRODUCTION

Chapter 2

Literature Review

In recent years, the integration of artificial intelligence (AI) into medical imaging has shown significant promise in enhancing diagnostic accuracy, efficiency, and accessibility. However, alongside this progress, growing concerns have emerged regarding the fairness, transparency, and generalizability of AI models, particularly in relation to demographic biases. As chest X-ray datasets play a foundational role in training and evaluating diagnostic algorithms, it has become increasingly evident that systemic disparities in data representation and labeling practices can inadvertently lead to unequal performance across patient subgroups[20] [7]. Among these concerns, racial and ethnic bias has gained particular attention, raising critical questions about how AI systems interpret underlying patient characteristics and whether such systems might propagate or even amplify existing healthcare inequalities. This literature review explores the current state of research on demographic bias in medical imaging, with a particular focus on racial disparities in chest X-ray-based diagnostic models.

2.1 Chest X-ray Imaging

Chest X-ray (CXR) imaging is one of the most commonly used and cost-effective diagnostic tools in clinical practice for detecting thoracic abnormalities such as pneumonia, Pneumothorax, lung lesions, and cardiomegaly. Its non-invasive nature, low radiation exposure, and rapid acquisition make it an essential modality in routine medical diagnostics, particularly in resource-constrained settings. However, interpreting chest X-rays can be complex, requiring significant expertise due to overlapping anatomical structures and subtle pathological changes.

In recent years, the advent of deep learning has transformed the landscape of medical image analysis, with numerous studies demonstrating the potential of convolutional neural networks (CNNs) to perform at or even above the level of radiologists in certain diagnostic tasks. For example, Rajpurkar et al. introduced CheXNet, a 121-layer DenseNet model trained on the NIH ChestX-ray14 dataset, which achieved radiologist-level performance in pneumonia detection [17]. Following this, large-scale public datasets such as MIMIC-CXR [10] and CheXpert [9] enabled the development and benchmarking of multi-label classification models capable of identifying a wide range of chest diseases from frontal CXR images.

Despite these advances, concerns have emerged around the reliability and fairness of such models in real-world applications. The variability in imaging protocols, device manufacturers, patient positioning, and demographic factors can affect model performance, often leading to generalization issues when applied to unseen populations or institutions [12]. Additionally, label quality remains a challenge, as many datasets rely on natural language processing (NLP) [10][9] to extract findings from radiology reports, introducing potential noise and ambiguity in the ground truth.

Overall, while CXR-based AI models have shown significant promise in automating disease detection and triage, their real-world deployment necessitates careful evaluation for robustness, interpretability, and demographic fairness.

2.2 Bias and Fairness in AI-based CXR diagnosis

In the context of medical AI, bias refers to systematic differences in model performance across subgroups defined by demographic attributes such as race, gender, age, or socioeconomic status. Fairness refers to the goal of minimizing or eliminating these disparities to ensure equitable healthcare outcomes for all groups.

2.2.1 Sources of Bias

Bias in medical AI systems can originate from several sources:

Dataset Imbalance: Many medical imaging datasets are not demographically representative. For instance, widely used datasets like NIH ChestX-ray14 or CheXpert are predominantly composed of patients from specific racial or geographic groups. This imbalance can lead to models that perform well on majority groups but poorly on underrepresented populations [20]

2.2. BIAS AND FAIRNESS IN AI-BASED CXR DIAGNOSIS

7

Label Noise and Subjectivity: Systematic label noise, or label bias, differs crucially from random label noise in that—if not addressed properly—it results in a biased decision boundary being learned [16]. Diagnostic labels in public datasets are often extracted using natural language processing (NLP) from radiology reports [10][9], which can introduce inaccuracies. Moreover, subjective interpretations by radiologists can be inconsistent across demographic groups [9]

Sample Selection Bias: Selection biases occur when the dataset used for training does not accurately represent the true target population. Intersectional effects—such as age distribution within gender groups—can confound subgroup performance analysis. A poor model performance in elderly patients may be misinterpreted as sex bias if one sex is overrepresented among older individuals [16]

Shortcut Learning: Deep models often rely on easily learnable but non-causal patterns — a phenomenon known as shortcut learning [6]. In medical imaging, this can mean models may exploit spurious correlations (e.g., hospital markers, image contrast) that are inadvertently tied to demographic features, rather than truly learning disease-relevant signals [4]

2.2.2 AI recognition of patient race in CXR Images

Recent studies have demonstrated that AI models can infer a patient's racial or ethnic identity from medical images, such as chest radiographs, with high accuracy—even when clinician cannot. In particular, it has been shown that deep learning models can predict self-reported race from X-rays with an AUROC score above 0.9 across multiple modalities, imaging vendors, and clinical tasks [7] See in the below fig 2.1. This occurs even when the image is heavily degraded or cropped means distorted image.

According to Seyyed-Kalantari et al. [20], subgroups such as female patients, patients under 20, Black and Hispanic individuals, and Medicaid-insured patients were disproportionately underdiagnosed by AI models trained on CXR datasets. This study shows that false positive rate (FPR) and false negative rate (FNR) have an inverse relationship in these populations, indicating that underserved groups are often falsely labeled as healthy, rather than over-diagnosed. This points to a systemic issue where certain demographics are aggressively flagged as healthy, leading to potentially harmful missed diagnoses.

A study by Burns et al. [1] explored whether the ability of AI to identify patient race from chest X-rays is rooted not in anatomical structure but in pixel intensity dis-

Area under the receiver operating characteristics curve	
Race detection in radiology imaging	
Chest x-ray (internal validation)*	
MXR (Resnet34, Densenet121)	0.97, 0.94
CXP (Resnet 34)	0.98
EMX (Resnet34, Densenet121, EfficientNet-B0)	0.98, 0.97, 0.99
Chest x-ray (external validation)*	
MXR to CXP, MXR to EMX	0.97, 0.97
CXP to EMX, CXP to MXR	0.97, 0.96
EMX to MXR, EMX to CXP	0.98, 0.98
Chest x-ray (comparison of models)†	
MXR, CXP, EMX	Multiple results (appendix p 26)
CT chest (internal validation)*	
NLST (slice, study)	0.92, 0.96
CT chest (external validation)*	
NLST to EM-CT (slice, study)	0.80, 0.87
NLST to RSPECT (slice, study)	0.83, 0.90
Limb x-ray (internal validation)*	
DHA	0.91
Mammography*	
EM-Mammo (image, study)	0.78, 0.81
Cervical spine x-ray*	
EM-CS	0.92

Figure 2.1: Race detection in radiology imaging [7]

tributions alone. These histograms were normalized to percent-per-image (PPI) values and analyzed using multivariate analysis of variance (MANOVA), which rejected the null hypothesis of no difference in distributions across racial groups ($F=7.38$, $p<0.0001$) with 95% confidence. Even under class balancing, race-specific differences persisted ($F=2.02$, $p<0.0001$)[1]. Machine learning models trained only on grayscale intensity counts—without any spatial or anatomical information—achieved non-trivial race classification performance, with gradient-boosted decision trees reaching an AUROC of 77.24% and feed-forward neural networks achieving 69.18%. Stratified analyses across factors such as BMI, age, sex, scanner model, and acquisition settings confirmed the robustness of these findings. This study provides strong evidence that race information is statistically embedded in the grayscale intensity distributions of chest X-rays, even when all structural cues are removed.[1].

Another paper found that technical parameters of image acquisition and processing — such as machine type, resolution, and view positioning — are major contributors to this capability [12]. Importantly, Lotter et al [12] demonstrated that mitigating these acquisition differences via a demographics-independent calibration strategy significantly

2.2. BIAS AND FAIRNESS IN AI-BASED CXR DIAGNOSIS

9

reduces underdiagnosis bias. This suggests that part of the diagnostic disparity is not merely due to model limitations or imbalanced data, but is rooted in the technical characteristics of how medical images are collected and processed.

These findings imply that there are subtle, high-dimensional racial signals in medical imaging that are learnable by AI but invisible to human clinicians. Such signals could be incorporated into bone structure, tissue density, or other patterns at the pixel level [1]. There are important ramifications: if race can be inferred from pictures, AI models that have been trained on clinical tasks might unintentionally come to depend on these proxies, which would bias prognostic or diagnostic predictions. This means that race is implicitly encoded in medical images, even when not explicitly labeled [6][7]. Such hidden use of race can result in biased diagnostic accuracy, poorer generalization to diverse populations, and unequal treatment recommendations. Even if overall model performance appears acceptable, group-specific harms may persist beneath the surface. Therefore, understanding and addressing racial encoding is critical for developing trustworthy, equitable medical AI systems. Preprocessing methods are commonly used in fairness-focused AI model[15], motivated by this study, we propose targeted preprocessing strategies to mitigate racial encoding, and produce fairer diagnostic outcomes. These methods are detailed in the following Methodology section.

2.2.3 Diagnostic Bias

Understanding diagnostic bias in CXR:

it is critical, as deep learning models may show unequal performance across demographic groups. This bias often arises from dataset imbalances, label inaccuracies, or models relying on spurious correlations like race encoding. Such issues can lead to misdiagnosis or reduced accuracy for underrepresented populations.

Empirical Evidence of Diagnostic Bias:

Several studies have documented fairness gaps in widely used medical AI systems:

- Seyyed-Kalantari et al. (2021) demonstrated that chest X-ray classification models trained on large datasets exhibited systematic underdiagnosis for Black patients across multiple disease labels, even when overall performance metrics appeared acceptable [20].

- Banerjee et al. (2022), in the CheXclusion study, revealed that fairness gaps persisted across multiple datasets and tasks in multi-label chest X-ray classification, suggesting that merely increasing dataset diversity is insufficient to eliminate diagnostic disparities [19].
- Wang et al. (2022) found that models could maintain high diagnostic performance while minimizing the ability to infer race, indicating that racial signal in chest X-rays is not necessary for disease classification and may instead contribute to biased decision-making when learned as a shortcut [21].
- Lotter et al. (2023) showed that diagnostic bias arises in models due to variation in acquisition parameters across institutions, which can encode race-related information. Their findings emphasized that mitigating race-related artifacts reduced underdiagnosis bias in marginalized groups [12].

2.3 Mitigation Strategies for Bias in Medical Imaging

A variety of techniques have been proposed to mitigate these biases. These can generally be categorized into three groups:

(1) Data Balancing and Reweighting:

Some researchers propose ensuring that datasets are demographically balanced or applying reweighting strategies during training. While these methods can reduce disparities in training data representation, they often do not generalize well to new domains or test sets from other institutions [20].

(2) Fairness-Aware Learning:

Algorithmic solutions such as adversarial debiasing have also been explored. These techniques involve training models that minimize the prediction of protected attributes (e.g., race) while maximizing task performance. [yan24]

(3) Shortcut Mitigation via Image Preprocessing:

An emerging and promising strategy involves addressing the input space directly by preprocessing the chest X-ray images to remove or reduce non-clinical cues that may encode race or hospital identity. For instance, the study “Drop the shortcuts” [21] demonstrates that basic image augmentations—like cropping out hospital tags and randomizing brightness—can significantly reduce a model’s ability to predict

2.3. MITIGATION STRATEGIES FOR BIAS IN MEDICAL IMAGING

11

demographic attributes from medical images. A very recent paper [2] of J. Gichoya further suggests that DICOM acquisition parameters themselves are a major source of encoded bias. Their study, shows that choices in windowing, LUT application, and modality transformations significantly influence the generalizability and fairness of AI models. Preprocessing, therefore, should consider these acquisition artifacts.

Challenges and Limitations :

(1) Poor Generalization to External Datasets:

One of the most consistent issues across bias mitigation strategies is their limited ability to generalize beyond the training environment. Techniques such as data reweighting or adversarial training may show improved fairness on internal validation data but often fail to maintain performance across datasets from different hospitals or populations[23].

(2) DICOM-Specific Artifacts:

The DICOM file format introduces additional complexity. Variations in LUT (Look-Up Table) transformations, windowing parameters, and modality presentation can all encode site-specific or demographic biases [2]. Most public datasets convert DICOMs to PNG without standardized LUT correction, which may preserve unwanted artifacts that contribute to racial inference.

(3) Lack of Diverse and Transparent Datasets:

A major contributor to racial bias in chest X-ray AI models is the lack of demographic diversity in widely used datasets such as CheXpert, and MIMIC-CXR which are predominantly composed of White patients (about more than half set) . Underrepresentation of minority groups like: Asian limits the model's ability to generalize and leads to disparities in diagnostic performance.

Seyyed-Kalantari et al. [20] demonstrated that models trained on imbalanced datasets systematically underdiagnose diseases in underserved populations, particularly Black patients. This bias persists across multiple datasets and clinical settings.

Chapter 3

Methodology

This chapter outlines the methodology adopted to investigate racial bias in chest X-ray AI models and improve their generalization and fairness through preprocessing techniques. It includes a detailed description of the datasets used, preprocessing steps applied to mitigate bias, model architecture, and the experimental setup used for evaluation.

3.1 Datasets

In this study, two large-scale publicly available chest X-ray datasets are used: CheXpert and MIMIC-CXR. These datasets provide high-quality radiographic images along with clinical labels, making them suitable for training and evaluating deep learning models for disease detection.

3.1.1 CheXpert Dataset

CheXpert [9] is a labeled chest radiograph dataset developed by Stanford University, comprising 224,316 chest X-ray images from 65,240 patients. They retrospectively collected chest radiographic studies from Stanford Hospital, performed between October 2002 and July 2017 in both inpatient and outpatient centers, along with their associated radiology reports. The dataset includes both frontal and lateral views and provides labels for 14 common chest pathologies, such as pneumonia, cardiomegaly, and pleural effusion. The images are provided in JPEG and have a resolution of 390 * 320 pixels. The labels were extracted using a rule-based natural language processing system applied to radiology reports.

These labels include positive, negative, and uncertain categories, which reflect the confidence or ambiguity present in clinical interpretation. CheXpert is particularly well-

suited for supervised learning tasks due to its labeled structure and has been used as a benchmark in numerous studies.

3.1.2 MIMIC Dataset

The MIMIC-CXR [10](Medical Information Mart for Intensive Care Chest X-Ray) dataset is a large-scale chest radiograph dataset released by the MIT Lab for Computational Physiology. It includes 377,110 images across 227,835 studies, covering 65,379 unique patients from the Beth Israel Deaconess Medical Center between 2011 and 2016.

Original DICOM images typically around 2048×2048 pixels, though may vary depending on equipment and acquisition settings. Each image is linked to a free-text radiology report, which can be used for label extraction or clinical language modeling. MIMIC-CXR includes a broad range of CXR view positions, including AP, PA, and lateral, and the dataset is stored in DICOM format, which preserves rich metadata, including acquisition parameters such as exposure, resolution, and scanner model.

To understand the demographic distribution within the datasets, particularly in terms of racial composition, we examined the reported breakdown of patient race for both MIMIC and CheXpert. As shown in Figure 3.1 [23], both datasets have a clear racial imbalance, with the majority of chest X-ray images belonging to White patients—61.0% in MIMIC and 56.4% in CheXpert. Other racial groups, including Black (15.6%), Asian (3.1%), and Other (20.3%) are represented in MIMIC, while CheXpert includes Asian (10.5%), Black (5.4%), and Other (27.8%) populations. This disproportionate representation suggests a skew in the dataset composition, which may introduce demographic bias in AI models trained on these datasets. This observed imbalance underscores the importance of careful preprocessing and fairness-aware evaluation to ensure equitable diagnostic performance across racial groups.

3.1.3 Data Splitting & Sampling

We use the MIMIC-CXR-JPG [10], and CheXpert [9] database. Applied same data processing, sampling strategies, and augmentation techniques for the CheXpert and MIMIC-CXR-JPG datasets. For the MIMIC-CXR-JPG dataset, we discard lateral recordings and retain only frontal (AP/PA) chest X-rays. For cheXpert we only select frontal view only for maintain consistency. we exclude the support devices', fracture', and pleural other' labels, focusing our analysis on the remaining 10 disease labels along with the No Finding' label.

3.1. DATASETS

15

		MIMIC	CheXpert
	Location	Boston, MA	Stanford, CA
	No. of images	357,167	222,792
	Percent frontal	64.5	85.5
Sex (%)	Female	47.8	40.7
	Male	52.2	59.3
Race (%)	Asian	3.1	10.5
	Black	15.6	5.4
	White	61.0	56.4
	Other	20.3	27.8

Figure 3.1: Demographic characteristics of the datasets [23]

Following the approach of Weng et al.[22], we eliminate multiple recordings for the same patient, retaining only one image—specifically, the one with the most annotated disease labels—to reduce potential label bias, particularly within the ‘No Finding’ category. This results in mimic dataset of 41,168 unique recordings. From this, we construct a test set of 1,757. The remaining data are randomly split into a training and validation set of 37,439 (95%) and 1,972 (5%) samples, respectively. We ensure that there is no patient overlap between any of the three sets. Chexpert dataset contain 64,522 unique recordings. From this, we construct a test set of 1,879. The remaining data are randomly split into a training and validation set of 59,509 (95%) and 3,134 (5%) samples, respectively.¹

3.1.4 CheXmask Dataset

To support lung-region-focused training for chest disease prediction, we utilize the CheXmask dataset [5], a publicly available resource that provides high-quality lung segmentation masks

¹In cases where fewer than 35 samples were available for a given race-label combination, all available samples were included. Due to the multi-label nature of the dataset, some samples may contribute to more than one label group.

Table 3.1: Dataset Splitting and Sampling Strategy for CheXpert and MIMIC-CXR

Subset	CheXpert	MIMIC-CXR
Training Set	59,509	37,439
Validation Set	3,134	1,972
Test Set (Total)	1,879	1,757
Total Images Used	64,522	41,168

for chest X-ray images. These masks are used to generate lung-masked inputs, helping reduce model dependence on non-relevant visual features and improving fairness across demographic subgroups.

The CheXmask dataset aggregates a total of 657,566 anatomical lung segmentation masks sourced from the following public chest X-ray databases: ChestX-ray8, CheXpert, MIMIC-CXR-JPG, Padchest, VinDr-CXR. In this thesis, we specifically use the CheXpert and MIMIC subset of CheXmask to generate lung-masked images.

The segmentation masks in CheXmask were generated using HybridGNet[5], a hybrid architecture combining convolutional layers and vision transformers for improved anatomical segmentation accuracy. Unlike many segmentation pipelines that lack validation, CheXmask provides an individual Reverse Classification Accuracy (RCA) score for each segmentation, enabling users to assess mask reliability at scale.

3.2 Model Architectures

This study undertakes a comparative analysis of several prominent neural network architectures applied to chest x-ray image to classify diseases. We primary use Densenet-121 architecture to adapt and compare results to Seyyed-Kalantari et al. [19] and Yang et al. [23]. We also already tried other architectures: ResNet50, EfficientNetV2, ResNeXt50 and ConvNeXtTiny. This all model gave results similar while there computation power and complexity is higher than Densenet-121.

3.2.1 Foundational Concepts

Before detailing the specific architectures evaluated, it is pertinent to introduce the core technologies upon which they are built: Convolutional Neural Networks (CNNs).

Convolutional Neural Networks

3.2. MODEL ARCHITECTURES

17

CNNs represent a cornerstone of modern computer vision, demonstrating exceptional proficiency in learning hierarchical feature representations directly from grid-like data such as images [13]. At their core, convolution operations extract spatial features through learnable filters (kernels) that scan across the input, computing element-wise multiplications followed by summation. This operation can be mathematically expressed as:

$$(f * g)(p) = \sum_{s+t=p} f(s)g(t) = \sum_{s \in \mathbb{Z}^2} f(s)g(p-s) \quad (3.1)$$

where f represents the input feature map, g denotes the kernel, and p indicates the spatial position. This formulation enables CNNs to capture local patterns with parameter sharing and translation equivariance—critical properties that dramatically reduce model complexity compared to fully connected architectures while maintaining spatial relationships. The CNN [13] architecture typically comprises several convolutional layers that progressively learn more abstract representations, pooling layers (e.g., max-pooling) that reduce spatial dimensions while preserving salient information, and non-linear activation functions (predominantly ReLU) that enable modelling of complex patterns. The hierarchical structure allows CNNs to transition from detecting simple low-level attributes (edges, textures) to complex high-level semantic concepts as information progresses through the network.

For medical image analysis specifically, CNNs have proven invaluable due to their ability to learn task-specific features directly from data, reducing reliance on handcrafted features while achieving superior performance across various modalities and applications [11].

3.2.2 DenseNet 121

DenseNet-121 (Densely Connected Convolutional Network), introduced by Huang et al. in 2017 [8], is a convolutional neural network designed to encourage feature reuse, reduce the number of parameters, and strengthen gradient flow. This architecture is well-suited for medical imaging tasks such as chest X-ray analysis due to its efficiency and capacity to extract rich, hierarchical features even with limited training data.

The network contains approximately 8 million parameters, which is relatively compact compared to other deep CNN architectures such as ResNet-152, making it suitable for efficient deployment and training. Below figure(3.2) [3] show the DenseNet-121 architecture:

Architecture Overview

DenseNet-121 consists of an initial convolutional stem followed by four densely connected blocks, interleaved with transition layers. Each layer in a dense block receives as input the

18

CHAPTER 3. METHODOLOGY

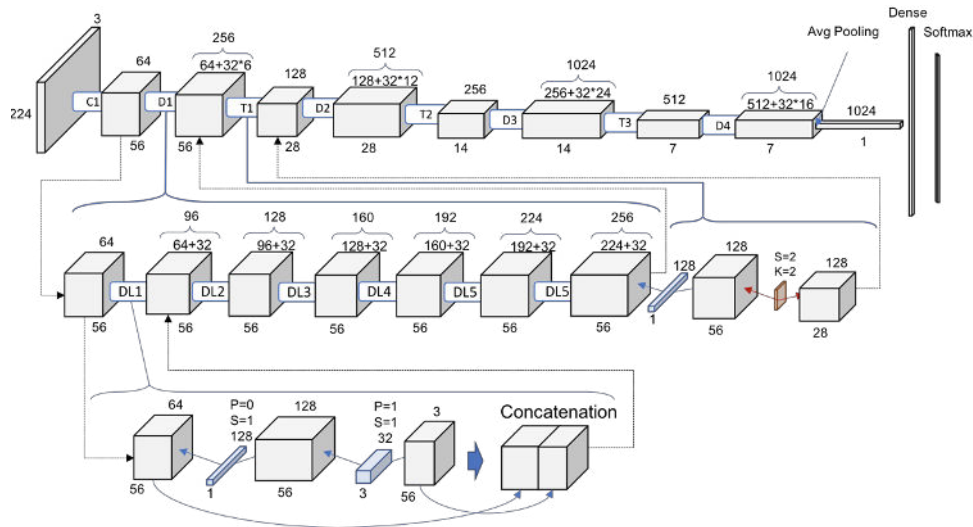


Figure 3.2: DenseNet 121 Model Architecture [3]

concatenation of all feature maps from preceding layers, ensuring maximum information flow between layers.

Initial Convolution and Pooling: A 7×7 convolution with stride 2 is followed by a 3×3 max pooling operation, which reduces the spatial dimensions of the input image.

Dense Blocks and Transition Layers: The model comprises four dense blocks with 6, 12, 24, and 16 layers respectively. Between dense blocks, transition layers consisting of 1×1 convolutions and 2×2 average pooling are used to downsample the feature maps and reduce dimensionality.

Final Layers: The network concludes with a global average pooling layer and a fully connected classification layer. In our case, the final layer is adapted for multi-label classification to predict multiple thoracic pathologies.

Pre-training and Performance on ImageNet

To leverage transfer learning, DenseNet-121 is initialized with weights pre-trained on the ImageNet-1K dataset. This initialization improves convergence and performance, especially in medical tasks where annotated data is limited or imbalanced. On the ImageNet benchmark, DenseNet-121 achieves: Top-1 Accuracy: 74.43% and Top-5 Accuracy: 91.27%

3.3. EXPERIMENTAL SETUP

19

3.2.3 HybridGNet

To further assess the impact of preprocessing techniques on model generalization and fairness, we also experiment with HybridGNet [5], a hybrid architecture designed specifically for robust medical image analysis. HybridGNet integrates both convolutional and transformer-based components to capture both local spatial patterns and global contextual relationships in medical images, making it well-suited for the challenges posed by chest X-ray interpretation across demographically diverse populations.

HybridGNet adopts a dual-path design:

A CNN branch extracts low-level spatial features with local receptive fields, preserving fine-grained anatomical structures. A Vision Transformer (ViT) branch captures long-range dependencies and semantic-level global context by operating on image patches.

To enhance the focus of the disease prediction model and mitigate potential biases from non-relevant regions, we employ lung region masking as a preprocessing step. Specifically, we use the CheXmask[5] dataset, a publicly available resource that provides lung segmentation masks for chest X-ray images. These masks are generated using the HybridGNet architecture, which is designed for accurate and robust lung segmentation across diverse populations and imaging conditions.

3.3 Experimental Setup

To ensure the reproducibility of the findings presented herein and to facilitate equitable comparisons across the diverse architectures, design methodologies, and learning paradigms investigated within this thesis, a standardized experimental configuration was rigorously maintained, except where explicitly noted deviations are described. This section elucidates the common parameters and procedures applied concerning dataset management, software implementation, model training protocols, and quantitative evaluation.

3.3.1 Implementation Details

All architectures were implemented utilizing the PyTorch deep learning framework. Model training and subsequent inference procedures were executed on NVIDIA A100 Graphics Processing Units (GPUs), each provisioned with 40GB of Video Random Access Memory (VRAM). In adherence to best practices for scientific reproducibility, deterministic behaviour

was promoted through the utilization of fixed random seed values for the initialization of model parameters and other stochastic elements within the training pipeline.

Chapter 4

Results

4.1 Overview

This chapter presents the results of the baseline model and model after applying the proposed preprocessing methods. The results are reported for both the MIMIC and CheXpert datasets, covering diagnostic accuracy, race prediction performance, and subgroup analyzes between racial groups.

4.2 Baseline Performance

Table 4.1: AUC of Diagnostic Classification

Datasets	AUROC
Chexpert	0.98
MIMIC	0.97

Disease	Race	Baseline	CLAHE	Lung Masking
Atelectasis	ASIAN	0.787689	0.798335	0.768912
	BLACK	0.752812	0.801216	0.771359
	WHITE	0.812641	0.825141	0.793184
	all	0.804218	0.817463	0.797078
	hisp/lat/SA	0.847796	0.849899	0.865014
	unknown/other	0.814840	0.812179	0.784840

Cardiomegaly	ASIAN	0.771612	0.780350	0.740738
	BLACK	0.810725	0.817261	0.783868
	WHITE	0.745938	0.784103	0.768001
	all	0.790192	0.809998	0.786788
	hisp/lat/SA	0.805169	0.823184	0.822205
	unknown/other	0.822868	0.840180	0.806697
Consolidation	ASIAN	0.788968	0.810463	0.799004
	BLACK	0.739722	0.756734	0.754829
	WHITE	0.745310	0.753391	0.724387
	all	0.753080	0.760886	0.755046
	hisp/lat/SA	0.770757	0.791100	0.762556
	unknown/other	0.736938	0.719010	0.754098
Edema	ASIAN	0.888496	0.908329	0.868287
	BLACK	0.863204	0.884723	0.866528
	WHITE	0.882616	0.879780	0.857291
	all	0.869934	0.883331	0.864573
	hisp/lat/SA	0.884000	0.907956	0.894418
	unknown/other	0.835741	0.837889	0.835630
EnlargedCardiomediatinum	ASIAN	0.710322	0.711392	0.702166
	BLACK	0.736255	0.738621	0.730171
	WHITE	0.686241	0.726324	0.671807
	all	0.698887	0.710464	0.687581
	hisp/lat/SA	0.695120	0.689416	0.661924
	unknown/other	0.661037	0.689573	0.671745
Lung Lesion	ASIAN	0.626620	0.684114	0.697438
	BLACK	0.652683	0.668241	0.732363
	WHITE	0.756639	0.782764	0.781842
	all	0.693354	0.716842	0.737899
	hisp/lat/SA	0.672559	0.687530	0.711738
	unknown/other	0.789413	0.781873	0.766177
Lung Opacity	ASIAN	0.615894	0.644390	0.620307
	BLACK	0.654410	0.680513	0.662752
	WHITE	0.642788	0.688971	0.646917

4.2. BASELINE PERFORMANCE

23

	all	0.665758	0.688687	0.671429
	hisp/lat/SA	0.724857	0.719619	0.721176
	unknown/other	0.689967	0.706830	0.705668
No Finding	ASIAN	0.932819	0.942375	0.931467
	BLACK	0.924429	0.940490	0.899155
	WHITE	0.938398	0.934021	0.930553
	all	0.927685	0.937189	0.923253
	hisp/lat/SA	0.923996	0.928852	0.940336
	unknown/other	0.932154	0.951614	0.927820
Pleural Effusion	ASIAN	0.900334	0.910567	0.880608
	BLACK	0.886334	0.900178	0.880927
	WHITE	0.886215	0.897877	0.862283
	all	0.893744	0.903375	0.877512
	hisp/lat/SA	0.906046	0.915652	0.900790
	unknown/other	0.882434	0.889305	0.855072
Pneumonia	ASIAN	0.629036	0.687703	0.650704
	BLACK	0.614542	0.629167	0.596119
	WHITE	0.701209	0.710907	0.698433
	all	0.649968	0.674481	0.649889
	hisp/lat/SA	0.664963	0.661142	0.650145
	unknown/other	0.624273	0.658430	0.633037
Pneumothorax	ASIAN	0.761512	0.828179	0.830241
	BLACK	0.744919	0.782182	0.712398
	WHITE	0.675510	0.768709	0.635262
	all	0.748741	0.797924	0.740147
	hisp/lat/SA	0.790333	0.805542	0.814700
	unknown/other	0.790318	0.802392	0.740321

Table 4.3: Summary of AUROC per Disease and Race groupBy metrics across preprocessing methods

Race	Macro _{AUROC}	Max _{Diff} AUROC
ASIAN	0.605703	0.118420
BLACK	0.612937	0.036115
WHITE	0.415727	0.050155
all	0.521147	0.029416
hisp/lat/SA	0.460286	0.043604
unknown/other	0.511082	0.053590

Table 4.4: Summary of AUROC per Disease and Race groupBy metrics across preprocessing methods

Race	AUROC	Preprocessing
ASIAN	0.631622	Baseline
BLACK	0.629134	Baseline
WHITE	0.395759	Baseline
hisp/lat/SA	0.486633	Baseline
unknown/other	0.508472	Baseline
all	0.530324	Baseline
ASIAN	0.651955	Lung Masking
BLACK	0.616660	Lung Masking
WHITE	0.405509	Lung Masking
hisp/lat/SA	0.443029	Lung Masking
unknown/other	0.539183	Lung Masking
all	0.531267	Lung Masking
ASIAN	0.533534	CLAHE
BLACK	0.593019	CLAHE
WHITE	0.445914	CLAHE
hisp/lat/SA	0.451196	CLAHE
unknown/other	0.485593	CLAHE
all	0.501851	CLAHE

4.2. BASELINE PERFORMANCE

25

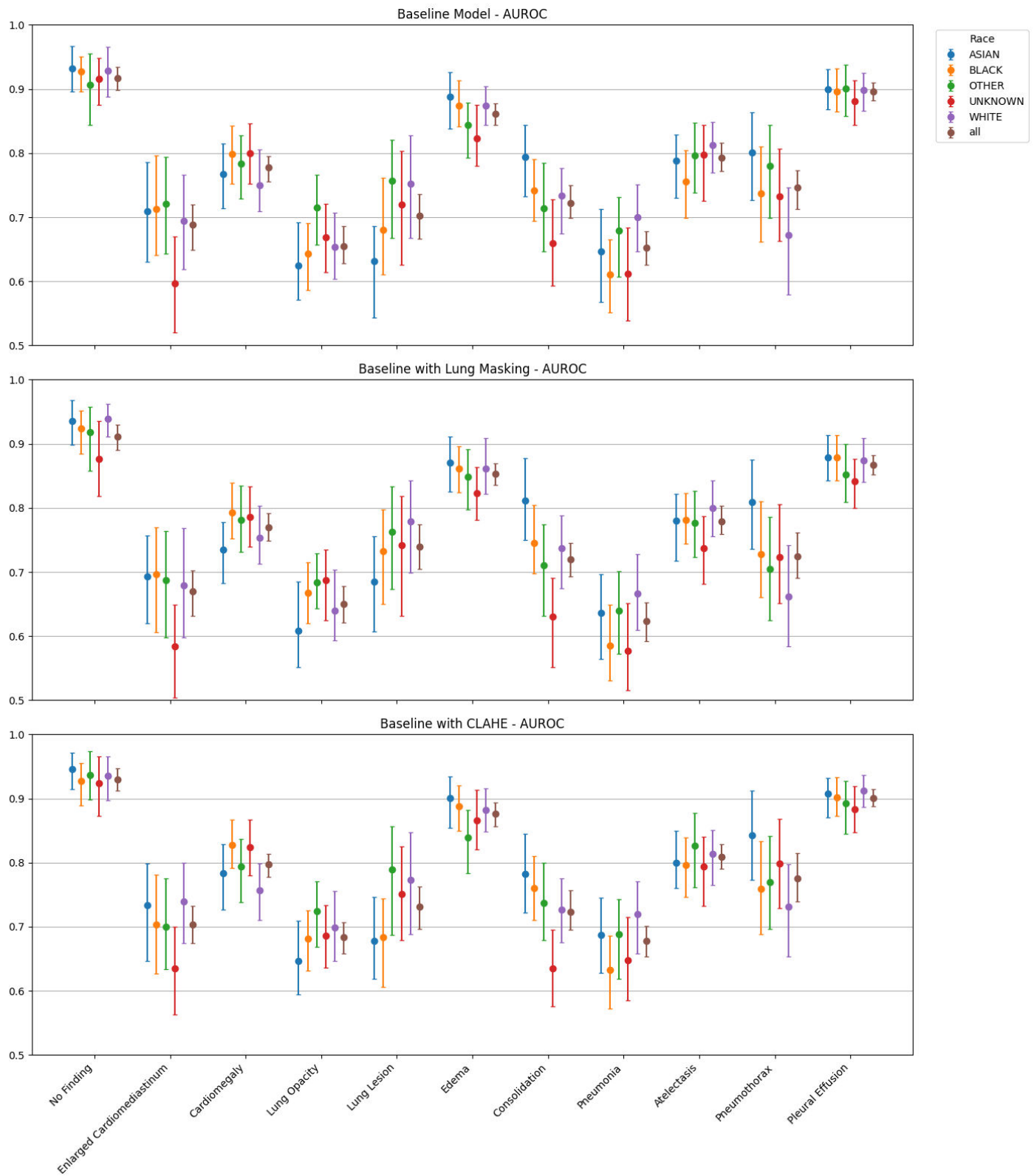


Figure 4.1: Race detection in radiology imaging

List of Abbreviations

AI Artificial Intelligence

CXR Chest X-ray

CLAHE Contrast Limited Adaptive Histogram Equalization

CNN Convolutional Neural Network

AUC Area Under Curve

AUROC Area Under the Receiver Operating Characteristic Curve

MIMIC-CXR Medical Information Mart for Intensive Care Chest X-Ray

AP Anteroposterior

PA Posteroanterior

DenseNet Densely Connected Convolutional Network

List of Figures

2.1	Race detection in radiology imaging [7]	8
3.1	Demographic characteristics of the datasets [23]	15
3.2	DenseNet 121 Model Architecture [3]	18
4.1	Race detection in radiology imaging	25

List of Tables

3.1	Dataset Splitting and Sampling Strategy for CheXpert and MIMIC-CXR .	16
4.1	AUC of Diagnostic Classification	21
4.3	Summary of AUROC per Disease and Race groupBy metrics across preprocessing methods	24
4.4	Summary of AUROC per Disease and Race groupBy metrics across preprocessing methods	24

32

LIST OF TABLES

LIST OF TABLES

33

literature

34

LIST OF TABLES

Bibliography

- [1] John Lee Burns et al. “Ability of artificial intelligence to identify self-reported race in chest x-ray using pixel intensity counts”. In: *Journal of Medical Imaging* 10.6 (2023), pp. 061106–061106.
- [2] Theo Dapamede et al. “DICOM LUT is a Key Step in Medical Image Preprocessing Towards AI Generalizability”. In: *Journal of Imaging Informatics in Medicine* (2025), pp. 1–9.
- [3] Shuvam Das. *Implementing DenseNet-121 in PyTorch: A Step-by-Step Guide*. Accessed: July 8, 2025. 2023. URL: https://miro.medium.com/v2/resize:fit:720/format:webp/1*4kPpyvHv73ypzTvo6y8J9w.png.
- [4] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. “AI for radiographic COVID-19 detection selects shortcuts over signal”. In: *Nature Machine Intelligence* 3.7 (2021), pp. 610–619.
- [5] N Gaggion et al. “CheXmask Database: a large-scale dataset of anatomical segmentation masks for chest x-ray images (version 0.4)”. In: *PhysioNet* <https://doi.org/10.13026/6eky-y831> (2023).
- [6] Robert Geirhos et al. “Shortcut learning in deep neural networks”. In: *Nature Machine Intelligence* 2.11 (2020), pp. 665–673.
- [7] Judy Wawira Gichoya et al. “AI recognition of patient race in medical imaging: a modelling study”. In: *The Lancet Digital Health* 4.6 (2022), e406–e414.
- [8] Gao Huang et al. “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.
- [9] Jeremy Irvin et al. “Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 590–597.

- [10] Alistair EW Johnson et al. “MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs”. In: *arXiv preprint arXiv:1901.07042* (2019).
- [11] Geert Litjens et al. “A survey on deep learning in medical image analysis”. In: *Medical image analysis* 42 (2017), pp. 60–88.
- [12] William Lotter. “Acquisition parameters influence AI recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias”. In: *Nature Communications* 15.1 (2024), p. 7465. DOI: [10.1038/s41467-024-52003-3](https://doi.org/10.1038/s41467-024-52003-3).
- [13] Keiron O’shea and Ryan Nash. “An introduction to convolutional neural networks”. In: *arXiv preprint arXiv:1511.08458* (2015).
- [14] Cathy Ong Ly et al. “Shortcut learning in medical AI hinders generalization: method for estimating AI model generalization without external data”. In: *NPJ Digital Medicine* 7.1 (2024), p. 124. DOI: [10.1038/s41746-024-01118-4](https://doi.org/10.1038/s41746-024-01118-4).
- [15] Dana Pessach and Erez Shmueli. “A review on fairness in machine learning”. In: *ACM Computing Surveys (CSUR)* 55.3 (2022), pp. 1–44.
- [16] Eike Petersen et al. “The path toward equal performance in medical machine learning”. In: *Patterns* 4.7 (2023).
- [17] Pranav Rajpurkar et al. “Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning”. In: *arXiv preprint arXiv:1711.05225* (2017).
- [18] Santanu Roy and Vibhuti Bansal. “Histogram Matching Based Data-Augmentation and Its Impact on CNN Model for Covid-19 and Pneumonia Detection from Radiology Images”. In: *International Conference on Computer Vision and Image Processing*. Springer. 2023, pp. 136–147. DOI: [10.1007/978-3-031-58181-6_12](https://doi.org/10.1007/978-3-031-58181-6_12).
- [19] Laleh Seyyed-Kalantari et al. “CheXclusion: Fairness gaps in deep chest X-ray classifiers”. In: *BIOCOMPUTING 2021: proceedings of the Pacific symposium*. World Scientific. 2020, pp. 232–243.
- [20] Laleh Seyyed-Kalantari et al. “Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations”. In: *Nature medicine* 27.12 (2021), pp. 2176–2182.
- [21] Ryan Wang et al. “Drop the shortcuts: image augmentation improves fairness and decreases AI detection of race and other demographics from medical images”. In: *EBioMedicine* 102 (2024).

BIBLIOGRAPHY

37

- [22] Nina Weng et al. “Are sex-based physiological differences the cause of gender bias for chest x-ray diagnosis?” In: *Workshop on Clinical Image-Based Procedures*. Springer. 2023, pp. 142–152.
- [23] Yuzhe Yang et al. “The limits of fair medical imaging AI in real-world generalization”. In: *Nature Medicine* 30.10 (2024), pp. 2838–2848.