A FAIRNESS-AWARE FRAMEWORK FOR BIAS MITIGATION IN DIAGNOSTIC SKIN IMAGING USING AN ADVERSARIAL CNN APPROACH

WAITHIRA JOSHUA MWAURA

BSC. COMPUTER SCIENCE (KISII UNIVERSITY)

A RESEARCH THESIS SUBMITTED TO THE BOARD OF POSTGRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE MASTER OF INFORMATION SYSTEMS, SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY, KISII UNIVERSITY

2026

DECLARATION AND RECOMMENDATIONS

Declaration by the Student

I declare that this thesis is original and has not been submitted for any other degree or qualification at this university or any other institution. Any contribution others make to this work is explicitly acknowledged in the text through citation.

Joshua Waithira **Signature** _____ **Date:** 16/04/2026

Reg No: MIN12/00001/23

Recommendation by the Supervisors

This thesis has been submitted for examination with our approval as University supervisors.

Dr. Ruth Chweya **Signature** _____ **Date** 16/04/2026

Lecturer,

Department of Computing Sciences,

Kisii University.

Prof. Cyprian Makiya Ratemo **Signature** _____ **Date** 16/04/2026

Department of Computing and Information Sciences,

Maasai Mara University.

ii

KSU/SPGS/PG/03

Telephone : +254 720 668900
Email : spgs@kisiiuniversity.ac.ke

P. O. Box 408-40200
KISII, KENYA.
www.kisiiuniversity.ac.ke

KISII UNIVERSITY

OFFICE OF THE DIRECTOR POST GRADUTATE STUDIES

PLAGIARISM DECLARATION

Definition of plagiarism

Is academic dishonesty which involves; taking and using the thoughts, writings, and inventions of another person as one's own.

DECLARATION BY STUDENT

i. I declare I have read and understood Kisii University rules and regulations, and other documents concerning academic dishonesty
ii. I do understand that ignorance of these rules and regulations is not an excuse for a violation of the said rules.
iii. If I have any questions or doubts, I realize that it is my responsibility to keep seeking an answer until I understand.
iv. I understand I must do my own work.
v. I also understand that if I commit any act of academic dishonesty like plagiarism, my thesis/project can be assigned a fail grade ("F")
vi. I further understand I may be suspended or expelled from the University for Academic Dishonesty.

Name: Joshua Mwaura Waithira Signature _____

Reg. No: MIN12/00001/23 Date: 15/04/2026

DECLARATION BY SUPERVISOR (S)

i. I/we declare that this thesis/project has been submitted to plagiarism detection service.
ii. **The thesis/project contains less than 20% of plagiarized work**.
iii. I/we hereby give consent for marking.

1. Name: Dr. Ruth Chweya Signature _____

   Affiliation Kisii University Date 15/04/2026

2. Name: Prof. Cyprian Makiya Ratemo Signature _____

   Affiliation Masai Mara University Date 15/04/2026

3. Name _____ Signature _____

   Affiliation _____ Date _____

KISII UNIVERSITY
SCHOOL OF POSTGRADUATE STUDIES
DECLARATION OF NUMBER OF WORDS FOR MASTERS/PROJECT/PHD THESES

DECLARATION OF NUMBER OF WORDS FOR MASTERS/PROJECT/ PHD THESES

This form should be signed by the candidate and the candidate's supervisor (s) and returned to the Director of Postgraduate Studies at the same time as you submit copies of your thesis/project.

Please note at Kisii University Masters and PhD thesis shall comprise a piece of scholarly writing of not less than 20,000 words for the Masters degree and 50 000 words for the PhD degree. In both cases this length includes references, but excludes the bibliography and any appendices.

Where a candidate wishes to exceed or reduce the word limit for a thesis specified in the regulations, the candidate must enquire with the Director of Postgraduate about the procedures to be followed. Any such enquiries must be made at least 2 months before the submission of the thesis.

Please note in cases where students exceed/reduce the prescribed word limit set out, Director of Postgraduate may refer the thesis for resubmission requiring it to be shortened or lengthened.

Name of Candidate: Joshua Mwaura Waithira ADM NO: MIN12/00001/23

Faculty: Information Science And Technology Department: Computing

Thesis Title: A FAIRNESS-AWARE FRAMEWORK FOR BIAS MITIGATION IN DIAGNOSTIC SKIN

IMAGING USING AN ADVERSARIAL CNN APPROACH

I confirm that the word length of:

1) the thesis, including footnotes, is 35,373 2) the bibliography is 4387 and, if applicable, 3) the appendices are 797……

I also declare the electronic version is identical to the final, hard-bound copy of the thesis and corresponds with those on which the examiners based their recommendation for the award of the degree.

Signed: …… Date:… 16/4/2026…… …
(Candidate)

I confirm that the thesis submitted by the above-named candidate complies with the relevant word length specified in the School of Postgraduate and Commission of University Education regulations for the Masters and PhD Degrees.

Signed: Dr. Ruth Chweya Email Rchweya@kisiiuniversity.ac.ke Tel 0791796026 Date: 16/4/2026
(Supervisor 1)

Signed: Prof. Makiya Ratemo Email Ratemomakiyacyprian@gmail.com Tel 0722741666 Date: 16/4/2026
(Supervisor 2)

REPEAT NAME(S) OF SUPERVISORS AS MAY BE NECESSARY

COPYRIGHT

DEDICATION

This work is dedicated to my dear family and friends.

ACKNOWLEDGEMENTS

First and foremost, I am deeply indebted to God for granting me good health and the focus needed to pursue and achieve this goal in life. I extend my heartfelt gratitude to my dear wife, mother, children, and my brothers and sisters for their tireless effort and unwavering support throughout this journey. Your love, sacrifice, and encouragement have been my greatest strength. I am equally grateful to my friends for standing by me and offering their support during my studies. I wish to express my sincere appreciation to my supervisors, Dr. Ruth Chweya and Prof. Ratemo Makiya Cyprian, for their invaluable academic guidance and encouragement. Your mentorship has been instrumental to the completion of this work. Finally, I offer special thanks to the Dean of the School of Information Science and Technology, colleagues, and coursemates for their support, wisdom, and camaraderie. God bless you all.

ABSTRACT

The integration of Artificial Intelligence (AI) into dermatological diagnostics offers great potential for improving clinical decision- making and patient outcomes. However, persistent biases in training data and model structures continue to increase disparities in diagnostic accuracy, especially across different skin tones and lesion types. These inequalities emphasize the urgent need for fairness - aware methods that provide equitable performance in clinical settings. Therefore, this study aimed to develop and validate a fairness - aware AI framework for dermatological imaging, designed to reduce bias while maintaining diagnostic accuracy. The framework was guided by three goals: (i) to assess how well adversarial debiasing techniques can remove bias- related features, (ii) to analyze current diagnostic systems and identify fairness gaps, and (iii) to introduce a customized bias mitigation strategy for skin lesion classification. Adversarial debiasing was applied through a Gradient Reversal Layer (GRL) within a Convolutional Neural Network (CNN). The GRL works by reversing gradient signals during training, discouraging the network from encoding bias- related attributes while still optimizing for diagnosis accuracy. This mechanism formed the core of the fairness - aware framework. Performance was evaluated by comparing the framework to a baseline CNN using the ISIC 2020 skin lesion dataset, which includes 33, 126 dermoscopic images from over 2, 000 patients. Preprocessing and augmentation techniques were used to improve data quality and model robustness. External validation utilized the Fitzpatrick 17 k dataset, containing 16, 577 clinical images annotated by skin type, to test demographic generalizability. Expert validation was also performed, with dermatologists reviewing interpretability outputs to confirm clinical relevance and practical use. Model interpretability was improved through SHAP- based feature attribution, helping clinicians visualize decision boundaries and better understand the framework's diagnostic reasoning. The fairness- aware framework significantly reduced the Statistical Parity Difference from. .35 to. .05, while maintaining high diagnostic accuracy, sensitivity, and AUC--ROC scores above. .87. Fairness improvements were supported by Equalized Odds metrics, and statistical testing showed that these equity gains did not compromise reliability. These results highlight the clinical potential of fairness- aware AI in dermatology. By combining adversarial debiasing, expert validation, and practical application, the proposed framework offers actionable paths for the ethical use of AI in precision medicine. More broadly, it advances the discussion on bias mitigation in healthcare AI, demonstrating the transformative potential of equity- focused approaches to improve diagnosis outcomes across diverse populations.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
AIF360	AI Fairness 360 toolkit
AUC-ROC	Area Under the Curve-Receiver Operating Characteristic
CNN	Convolutional Neural Networks
CT scan	Computed Tomography Scan
DICOM	Digital Imaging and Communications in Medicine
DIR	Disparate Impact Ratio
EMR	Electronic Media Record
EO	Equalized Odds
FST	Fitzpatrick Skin Type
ISIC	International Skin Imaging Collaboration
LIME	Local Interpretable Model-agnostic Explanations
LYNA	Lymph Node Assistant
MIDRC	Medical Imaging and Data Resource Center
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NIfTI	Neuroimaging Informatics Technology Initiative
NLP	Natural Language Processing
SHAP	SHapley Additive ExPlanations
SPD	Statistical Parity Difference
UNETR	UNEt Transformers

CHAPTER ONE

INTRODUCTION

1.1 Background to the Study

Artificial Intelligence (AI) has emerged as one of the most transformative technologies of the 21st century, reshaping industries, economies, and daily life. From automating manufacturing processes to personalizing educational curricula, AI's ability to analyze vast datasets, recognize patterns, and make predictions has unlocked unprecedented efficiencies and innovations (Brynjolfsson & Mcafee, 2017). Its applications span sectors as diverse as agriculture, finance, transportation, and criminal justice, each leveraging AI to solve complex challenges. However, nowhere is its impact more profound, and its ethical stakes higher, than in healthcare, particularly in diagnostic imaging, where AI's promise to enhance accuracy and accessibility is tempered by risks of bias, inequity, and opacity.

Artificial Intelligence is a branch of computer science that focuses on creating intelligent machines capable of performing tasks that typically require human-like intelligence. At its core, AI involves the development of algorithms and systems that can learn, reason, and adapt to new situations, much like the human mind (Stone et al., 2022). Key concepts in AI include machine learning, which enables systems to learn from data without being explicitly programmed; neural networks, which are inspired by the structure and function of the human brain and consist of interconnected nodes that process information (Rumelhart et al., 1986), and deep learning, a more advanced form of machine learning that uses multi-layered neural networks to learn from vast amounts of data (Goodfellow et al., 2016). AI has numerous real-world applications, from virtual assistants and self-driving cars to healthcare and finance. Still, it also raises important ethical considerations

regarding privacy, bias, accountability, and the impact on jobs, which researchers and policymakers are actively working to address.

AI has revolutionized financial systems by enabling real-time decision-making and predictive analytics. Machine learning models analyze market trends, optimize trading strategies, and detect fraudulent transactions with unparalleled speed. JPMorgan Chase's coin platform, for instance, uses Natural Language Processing (NLP) to review legal documents, reducing 360,000 hours of manual labor annually (Chui & Francisco, 2017). Similarly, AI-driven credit scoring systems assess borrowers' risk profiles using non-traditional data, such as social media activity, and expand financial inclusion in underserved regions (Björkegren & Grissen, 2020). However, these systems risk encoding historical biases; for instance, algorithms trained on data from racially segregated neighborhoods may deny loans to marginalized groups (Bartlett et al., 2022).

Self-driving cars, powered by AI, promise to reduce accidents caused by human error and optimize fuel efficiency. Companies like Tesla and Waymo use Convolutional Neural Networks (CNNs) to process real-time sensor data, enabling vehicles to navigate complex environments (Litman, 2017). Yet, ethical dilemmas persist. Autonomous vehicles face "trolley problem" scenarios, choosing between passenger and pedestrian safety, raising questions about algorithmic morality (Nyholm & Smids, 2020).

Healthcare has undergone a significant shift with the advent of AI's ability to process complex medical data. Machine learning models now rival human experts in diagnosing conditions such as diabetic retinopathy (Gulshan et al., 2016) and skin cancer (Esteva et al., 2017). IBM Watson for Oncology, for instance, cross-references patient data with 30 million medical papers to recommend personalized treatment plans (Somashekhar et al., 2018). These advancements are particularly transformative in low-resource settings, where AI-powered, portable devices, such as smartphone-

based ultrasound tools, enable prenatal screenings in rural areas with limited access to specialists. AI streamlines administrative tasks, reducing burnout among healthcare professionals. Natural Language Processing (NLP) systems automate clinical documentation, cutting charting time by 30% (Rajkomar et al., 2018a). Predictive analytics tools leverage historical and real-time data to forecast patient admissions, enabling healthcare providers to optimize staff allocation and bed management. These tools enhance operational efficiency and improve patient care outcomes (Clack, 2023).

However, over-reliance on automation risks depersonalizing care, as seen in cases where AI triage systems prioritize cost-cutting over patient needs (Topol, 2019). AI accelerates drug development by predicting molecular interactions and identifying candidate compounds. DeepMind's AlphaFold, for example, has mapped over 200 million protein structures, expediting research into diseases like Alzheimer's (Jumper et al., 2021). In genomics, AI algorithms analyze genetic sequences to predict disease susceptibility, enabling preventive interventions (Zou et al., 2019). These breakthroughs, however, raise ethical concerns about data privacy and equitable access to cutting-edge therapies.

Diagnostic imaging, encompassing X-rays, MRIs, CT scans, and ultrasounds, has been profoundly transformed by AI. Deep learning models detect anomalies with superhuman precision. For example, AI systems like Google Health's Lymph Node Assistant (LYNA) identify metastatic breast cancer in lymph nodes with 99% accuracy, reducing pathologists' workload (Liu et al., 2017). In cardiology, AI algorithms analyze echocardiograms to predict heart failure risk, outperforming traditional biomarkers (Zhang et al., 2018b). In neurology, CNNs segment brain tumors in MRI scans, guiding surgeons in real time (Pereira et al., 2016). In resource-limited regions, AI democratizes access to diagnostics. For example, the startup Butterfly Network

deploys handheld AI-powered ultrasound devices in sub-Saharan Africa, enabling non-specialists to perform fetal screenings.

Despite its promise, AI in diagnostic imaging faces significant hurdles. Models trained on homogeneous datasets underperform for underrepresented groups. A landmark study found that chest X-ray algorithms misdiagnosed pneumonia in women and rural populations due to gender and geographic imbalances in training data (Larrazabal et al., 2020). Many AI systems lacks interpretability. An FDA-approved stroke detection tool, for instance, failed to explain why it flagged certain CT scans, eroding clinicians' trust. Current frameworks, like the EU's Medical Device Regulation (MDR), emphasize safety but lack mandates for bias audits or diversity in validation datasets (Gerke et al., 2020). AI's limitations are starkly evident in dermatology. Commercial tools like DermaSensor achieve 95% accuracy in detecting melanoma, but only for light-skinned patients. A 2023 audit revealed a 34% error rate for darker skin tones due to underrepresentation in training data (Groh et al., 2021a). This disparity underscores the need for inclusive AI development.

To mitigate bias through inclusive design, initiatives like the Medical Imaging and Data Resource Center (MIDRC) curate global imaging datasets, ensuring representation of ethnic, gender, and socioeconomic diversity. Explainability tools like Local Interpretable Model-agnostic Explanations (LIME) demystify AI decisions, showing clinicians which image regions influenced diagnoses (Ribeiro et al., 2016b). Policy and collaboration are also crucial. The FDA's 2023 draft guidance on AI/ML-based software mandates ongoing bias monitoring post-deployment, a critical step toward accountability (Administration, 2021). The WHO's Global Initiative on AI for Health fosters collaboration between developers and low-income countries to co-design equitable tools (WHO, 2023).

Mitigating bias in AI systems is essential for ensuring fairness and inclusivity, particularly in diverse healthcare settings. According to Secinaro et al. (2021), addressing bias in AI not only improves diagnostic accuracy but also enhances trust in AI systems, fostering their adoption in clinical practice. Similarly, Alowais et al. (2023) highlight that fairness-aware AI models lead to better clinical outcomes by reducing disparities in diagnostic accuracy and treatment recommendations. These findings underscore the importance of bias mitigation as a cornerstone for developing equitable and effective AI tools in healthcare.

Artificial Intelligence (AI) offers transformative benefits for healthcare, including unparalleled diagnostic speed, accuracy, and scalability. By automating image analysis, AI reduces radiologists workloads, accelerates early disease detection, and expands access to care in underserved regions where specialists are scarce (Rajpurkar et al., 2022). However, these benefits are undermined when AI systems perpetuate biases, as seen in dermatology tools that misdiagnose melanoma in darker-skinned patients or chest X-ray algorithms that underperform for women (Larrazabal et al., 2020).

Addressing bias amplifies AI's potential by ensuring equitable outcomes across demographics, fostering trust among clinicians and patients, and preventing harm to marginalized populations. Bias-free AI systems could reduce diagnostic errors by up to 40% in underrepresented groups, close healthcare access gaps, and align with ethical mandates for justice in medical innovation (Gichoya et al., 2022). Furthermore, equitable AI strengthens compliance with emerging regulations, mitigates legal risks for healthcare institutions, and positions AI as a tool for global health equity rather than exclusion (Gerke et al., 2020). Koçak et al. (2024) discussed the risks of bias in AI systems for medical imaging. The study emphasized that biases in training data could lead to systematic errors in AI predictions, potentially resulting in misdiagnoses, and underscored the importance of identifying and mitigating these biases to prevent adverse patient outcomes.

This study, therefore, aimed to develop and validate a comprehensive framework specifically designed to minimize bias in diagnostic imaging AI.

1.2 Statement of the Problem

The increasing adoption of artificial intelligence (AI) in diagnostic imaging has transformed clinical practice by improving the speed, consistency, and scalability of medical decision-making. Despite these advancements, concerns have emerged regarding the presence of bias in AI-driven diagnostic systems. When bias is embedded in training datasets or model development processes, diagnostic systems may produce unequal outcomes across different patient populations. Such bias can lead to uneven diagnostic performance, increasing the risk of misdiagnosis or delayed diagnosis among marginalized groups. These outcomes raise serious concerns about equity in healthcare delivery and the reliability of AI-assisted clinical decision-making.

The challenge becomes particularly significant in dermatological imaging, where diagnostic systems must interpret visual patterns that vary across skin tones, lesion characteristics, and imaging conditions. When the datasets used to train these systems do not adequately represent diverse patient populations, models may learn patterns that perform well for some groups while producing less reliable results for others. Evidence from recent studies indicates that many medical imaging models are developed using datasets that lack sufficient demographic representation or fail to report important demographic characteristics of the training data (Gaube et al., 2021; Willemink et al., 2020). As a result, AI systems risk reinforcing existing disparities in healthcare outcomes rather than helping to reduce them.

Although researchers have proposed various approaches for addressing bias in AI systems, existing solutions remain fragmented. Many bias mitigation techniques are applied only at specific stages of the machine learning pipeline or rely on post-hoc evaluation after the model has already been

developed (Mehrabi et al., 2021). Such approaches do not fully address the need to embed fairness considerations throughout the entire development process. In addition, many available fairness frameworks have been designed primarily for structured datasets and therefore provide limited guidance for handling the complexities associated with unstructured medical image data. These complexities include annotation variability, differences in imaging devices, and evolving patterns of disease presentation (Glocker et al., 2023; Zech et al., 2018). Consequently, current approaches do not adequately support the development of fairness-aware diagnostic imaging systems.

Hence, there is a need for a comprehensive framework that integrates bias detection, mitigation, and evaluation throughout the AI development lifecycle. Such a framework would guide the development of diagnostic imaging systems that maintain strong predictive performance while promoting equitable outcomes across diverse patient populations. This study, therefore, addresses this gap by developing and validating a fairness-by-design framework tailored for dermatological diagnostic imaging.

1.3 Research Objectives

The main aim of this study was to develop a fairness-aware framework for bias mitigation in skin diagnostic imaging using adversarial CNN models.

1.3.1 Specific Objectives

The study was guided by the following specific objectives:

i. To examine adversarial debiasing mechanisms in diagnostic imaging and assess their applicability within diagnostic skin imaging.
ii. To analyze existing fairness frameworks to identify gaps that limit their applicability to medical image datasets.

iii. To develop and validate a fairness-aware framework for diagnostic Skin imaging.

1.4 Research Questions

This study was guided by the following research questions:

1. How effective are adversarial debiasing mechanisms used in diagnostic imaging?
2. What are the existing frameworks used in diagnostic imaging?
3. How will the framework for mitigating bias in diagnostic skin imaging be developed and evaluated?

1.5 Scope of the Study

This research aimed to develop a fairness-aware framework designed to minimize bias in AI-driven diagnostic imaging while maintaining clinical reliability. This study introduced a custom-built fairness framework, integrating adversarial debiasing, data augmentation, and explainability tools to actively suppress bias during model training and improve transparency in AI decision-making.

The study utilized the ISIC 2020 skin lesion dataset downloaded from *https://www.isic-archive.com/*, which comprised over 33,000 images, including demographic attributes such as Fitzpatrick skin types. Two distinct models were developed: a baseline Convolutional Neural Network (CNN) and an experimental CNN incorporating fairness-aware constraints to mitigate disparities in melanoma detection across demographic subgroups. This was to enable a direct comparison between standard and fairness-aware approaches, allowing for precise evaluation of how fairness constraints impact both diagnostic accuracy and bias mitigation across demographic subgroups.

Both models underwent training on the same preprocessed dataset, employing data augmentation techniques to enhance the representation of underrepresented skin tones. The use of two models,

a base model and a fairness-aware model, ensures a controlled benchmark for evaluating bias mitigation without introducing unnecessary complexity or conflicting optimization goals. This approach aligns with standard fairness validation methodologies (Duff et al., 2024; Tu, 2023), allowing clear quantification of fairness trade-offs while preserving predictive stability.

The study assessed diagnostic accuracy using conventional metrics, such as sensitivity, specificity, and AUC-ROC. At the same time, fairness was quantitatively evaluated through Statistical Parity Difference (SPD), Equalized Odds (EO), and Disparate Impact Ratio (DIR) to determine the effectiveness of bias reduction.

Given prior research findings indicating up to a 34% reduction in melanoma detection accuracy for Black patients compared to White patients (Groh et al., 2021a), additional validation was conducted using the Fitzpatrick 17k dataset to confirm that fairness interventions generalized effectively to new data distributions.

1.6 Significance of the Study

In a world where AI increasingly dictates who receives care and how, this study is a moral and practical imperative. It advances healthcare equity by ensuring AI-driven diagnostics serve all patients, regardless of race, gender, or zip code. It empowers clinicians with transparent, trustworthy tools. It equips policymakers to regulate AI responsibly. And it challenges the AI industry to prioritize justice as a cornerstone of innovation.

Left unchecked, biased AI will deepen existing healthcare disparities, erode trust in medical institutions, and undermine the potential of one of history's most promising technologies. This research offers a path forward not just to mitigate harm but to reimagine AI as a force for global health equity. By addressing the intersectional, technical, and ethical dimensions of bias in diagnostic imaging, this study makes four pivotal contributions to the field.

Framework for Intersectional Bias Analysis: While prior work has focused on single-axis fairness metrics, for instance, race or gender, this study operationalizes an intersectional lens (Buolamwini & Gebru, 2018) to quantify how overlapping identities disproportionately bear the risks of algorithmic misdiagnosis. This advanced methodological rigor in fairness research bridges critical theory with computational modeling.

Clinical Validation Protocol: Existing bias-mitigation strategies often prioritize technical accuracy over clinical utility (Ghassemi et al., 2021). By co-developing evaluation metrics with clinicians and ethicists, this framework embeds population health outcomes such as reduced diagnostic delays in marginalized groups as success criteria, redefining how AI systems are validated for real-world impact.

Temporal Adaptability: Unlike static fairness toolkits, the framework introduces dynamic monitoring mechanisms to detect bias arising from evolving scanner technologies, shifting disease prevalence, and demographic changes. This addresses a critical gap identified by Chen et al. (2023), who notes that 72% of imaging AI models degrade within three years of deployment.

Ethical-Technical Synthesis*:* The study proposes a governance model aimed at integrating technical interventions, like data augmentation, with ethical considerations. This approach seeks to contribute to ongoing discussions about translational fairness, which emphasizes linking algorithmic design to broader systemic reform (Parikh et al., 2019).

By centering the needs of underserved communities in both design and evaluation, this work goes beyond the limitations of fairness-as-afterthought approaches. It lights the way toward a future where technology heals rather than divides, not by naively ignoring bias, but by dismantling its roots in data, design, and deployment.

1.7 Limitations of the Study

Several limitations defined the scope of this research. First, although the study utilized the ISIC dataset, only 1,459 images contained Fitzpatrick skin type annotations, with darker skin tones underrepresented. This reflects a broader limitation in dermatological datasets. To mitigate this constraint, the study applied data augmentation and incorporated adversarial debiasing to reduce demographic bias during model training.

Second, the study focused exclusively on dermatological imaging. This scope allowed for a detailed investigation of fairness issues in skin lesion classification but did not assess the applicability of the proposed framework to other medical imaging domains such as X-rays or MRIs.

Additionally, the study concentrated on algorithmic fairness and model validation within an experimental setting. Factors related to real-world deployment, including economic feasibility, regulatory requirements, and integration into healthcare systems, were beyond the scope of this research.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This section presents a review of scholarly work relating to the area of study. It explores the mechanisms currently employed to address bias, examines widely used frameworks along with their strengths and limitations, and highlights areas where further research is needed.

2.2 Artificial Intelligence in Diagnostic Skin Imaging

Within the broader landscape of AI in healthcare, diagnostic imaging represents a cornerstone of modern medical practice. This field encompasses various modalities, including X-rays, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and ultrasound, which enable non-invasive visualization of internal structures and facilitate early disease detection (Litjens et al., 2017). These technologies have significantly advanced healthcare by improving diagnostic accuracy and guiding treatment decisions. However, diagnostic imaging is not without its challenges.

One of the primary issues in diagnostic imaging is the variability in image quality, which can be influenced by factors such as patient movement, equipment calibration, and operator expertise (Marty, 2024). Additionally, the interpretation of imaging results often relies on the subjective judgment of radiologists, which can lead to inconsistencies and diagnostic errors (Mark, 2025). The increasing demand for imaging services, driven by an aging population and the growing prevalence of chronic diseases, has also placed significant strain on radiology departments, resulting in delays and potential burnout among radiologists (Marty, 2024). Furthermore, access to advanced imaging technologies remains limited in low-resource settings, exacerbating healthcare disparities.

Integrating AI into diagnostic imaging has emerged as a potential solution to address some of these challenges. AI algorithms, particularly those based on deep learning, have demonstrated remarkable capabilities in tasks such as image classification, segmentation, and anomaly detection (Smarter Technology, n.d.). These AI systems, for instance, have been shown to detect conditions like breast cancer and lung nodules with accuracy comparable to or exceeding that of human radiologists (Ardila et al., 2019; McKinney et al., 2020; Science X staff, 2025).However, deploying AI in diagnostic imaging introduces its own set of challenges, particularly concerning bias.

Studies have revealed that AI models trained on non-representative datasets can exhibit significant disparities in performance across demographic groups, such as underdiagnosing diseases in women or minority populations (Larrazabal et al., 2020; Trafton, 2024; Seyyed-Kalantari et al., 2021). These biases often stem from imbalances in training data, differences in disease presentation across populations, and the lack of diversity in the development and validation of AI systems. Additionally, the "black-box" nature of many AI models raises concerns about interpretability and accountability, making it difficult for clinicians to trust and adopt these tools (Future of AI in Medical Imaging: Challenges and Opportunities - Quibim, n.d.; Lipton, 2018). Addressing these issues is critical to ensuring that AI-driven diagnostic tools are equitable, reliable, and effective for all patients.

By highlighting these challenges, this study underscored the importance of developing governance models that align technical interventions, such as data augmentation, with ethical imperatives. This dual focus aimed to bridge the gap between algorithmic design and systemic reform, ensuring that AI technologies contribute to fair and inclusive healthcare outcomes.

2.3 Theoretical Framework

This study is guided by theoretical perspectives from fairness-aware machine learning, particularly adversarial learning approaches used to mitigate bias in predictive models. The theoretical framework provides the foundation for understanding how machine learning models can be designed to reduce demographic bias while maintaining predictive performance. In this study, the concept of adversarial fairness learning informs the development of the fairness-by-design framework used to improve equity in dermatological diagnostic imaging systems.

2.3.1 Adversarial Fairness Learning

Adversarial Fairness Learning (AFL) is a machine learning approach designed to mitigate bias through adversarial training mechanisms. The concept originates from adversarial machine learning principles introduced by (Goodfellow et al., 2014) through Generative Adversarial Networks (GANs). In fairness-aware AI systems, this approach has been adapted to train predictive models while simultaneously minimizing the ability of the model to encode sensitive demographic attributes. As a result, AFL has emerged as an important strategy for addressing bias in automated decision-making systems, including applications in healthcare diagnostics and fairness-aware artificial intelligence (Liu & Wong, 2024).

The core mechanism of AFL involves introducing an adversarial component during model training, which attempts to predict sensitive attributes, such as race, gender, or skin tone, from the model's outputs (Hardt et al., 2016). The primary model is then trained to minimize the adversary's ability to make such predictions, reducing the risk of biased learned representations (Li et al., 2021). AFL incorporates several fundamental constructs: fairness constraints such as statistical parity and equalized odds, a debiasing loss function that neutralizes discriminatory patterns (Liu

& Wong, 2024), and an adversarial classifier that exposes bias tendencies. These components collectively ensure that the model remains fairness-aware while preserving predictive accuracy.

AFL is founded on several key assumptions. First, it assumes that bias within AI models can be mathematically quantified and corrected through adversarial optimization. Second, it recognizes that models inherently learn false correlations between sensitive attributes and target variables, necessitating explicit fairness constraints (Zhang et al., 2024). Furthermore, as outlined by Hardt et al. (2016), AFL assumes that fairness-aware models can maintain high predictive accuracy if adversarial training is properly optimized. Finally, it posits that data representation plays a crucial role in fairness outcomes, requiring fairness considerations from dataset preprocessing to final AI deployment.

While AFL is recognized for its effectiveness in bias mitigation, it is not without limitations. One of its primary strengths is its ability to directly embed fairness constraints within model training, unlike post hoc debiasing approaches that adjust predictions after inference. Additionally, AFL draws upon adversarial machine learning principles that enhance fairness-aware optimization, making it particularly useful in high-risk applications such as medical imaging. However, AFL also presents challenges, including high computational demands due to the dual-model optimization required, decreased interpretability of learned feature representations, and limited applicability in small-scale datasets where adversarial training may over-regularize the model (Madras et al., 2019). Researchers have suggested addressing these limitations by integrating AFL with data augmentation techniques or ensemble fairness strategies (Zhang et al., 2024).

AFL has been successfully applied across multiple domains, particularly in healthcare AI. Zhang et al. (2024) demonstrated AFL's ability to improve fairness in medical imaging models, ensuring that classification performance remained consistent across demographic subgroups. In

dermatology applications, Liu and Wong (2024) developed GIFair, an AFL-based framework that balances group and individual fairness in diagnostic predictions, reducing biased decision-making in skin disease detection. Furthermore, studies presented at CVPR (2024) highlighted AFL's effectiveness in mitigating bias in radiology and melanoma detection models, where adversarial debiasing techniques reduced false negative rates for darker-skinned individuals.

In the context of diagnostic skin imaging, AFL is particularly relevant due to the documented bias in dermatological AI models. Skin disease classification systems often struggle with fairness due to the underrepresentation of darker skin tones in medical datasets, leading to disparities in model performance. Traditional fairness-aware approaches, such as post hoc debiasing, fail to correct these biases at the model level, making AFL a superior alternative. By embedding fairness constraints directly into training, AFL prevents models from learning spurious correlations between skin tone and disease probability. Additionally, AFL ensures that classification performance remains robust across diverse populations, reducing disparities in diagnostic outcomes.

Unlike auditing-based fairness frameworks that evaluate biases post-deployment, AFL proactively integrates bias mitigation into the AI lifecycle, making it particularly suitable for fairness-aware dermatological AI development. Studies on fairness-driven adversarial training in healthcare confirm AFL's ability to produce clinically reliable and ethically fair models, reinforcing its role as a foundational methodology for mitigating bias in medical AI applications. As AI-driven healthcare continues to evolve, incorporating AFL into dermatological AI models offers a path toward equitable, trustworthy, and clinically validated diagnostics (Zhang et al., 2024). By leveraging AFL within this study, fairness-aware machine learning applications in skin imaging

can advance, ensuring that bias is addressed systematically through adversarial debiasing techniques.

In this study, AFL provides the theoretical foundation for the fairness-aware diagnostic framework proposed for dermatological AI. By integrating an adversarial debiasing component into the model architecture, the study operationalizes AFL to minimize the influence of sensitive attributes, particularly skin tone and gender, on predictive outcomes. This approach directly addresses the disparities commonly observed in dermatological classifiers and aligns with the broader goal of ensuring equitable diagnostic performance across demographic groups. AFL's ability to embed fairness constraints during training makes it particularly suited for this context, offering both conceptual rigor and practical efficacy in mitigating algorithmic bias in skin disease detection.

2.4 Adversarial Debiasing

In healthcare applications, biased artificial intelligence (AI) systems can lead to serious consequences, including misdiagnosis, unequal allocation of medical resources, and the reinforcement of existing disparities among marginalized populations. Machine learning models trained on imbalanced datasets often learn patterns that are correlated with demographic attributes such as race, gender, or socioeconomic status. As a result, these models may perform unevenly across different patient groups, leading to unequal healthcare outcomes.

One approach that has been proposed to address this challenge is adversarial debiasing, a machine learning technique designed to reduce the influence of sensitive attributes during model training. Adversarial debiasing typically involves two interacting components: a predictor and an adversary. The predictor performs the main task, such as identifying or classifying medical conditions from images. The adversary attempts to infer sensitive attributes from the predictor's outputs or internal

representations. During training, the predictor learns to perform the diagnostic task effectively while minimizing the information available to the adversary about sensitive attributes. This process encourages the model to focus on medically relevant features rather than demographic correlations that may introduce bias.

Previous studies have demonstrated the usefulness of adversarial debiasing in several domains. For example, the technique has been applied in facial recognition systems to reduce racial and gender bias (Buolamwini & Gebru, 2018), in credit scoring systems to prevent discrimination in financial decisions (Aggarwal, 2021), and in healthcare diagnostics to improve fairness in disease detection (Chen et al., 2019; Groh et al., 2021). Similarly, research in natural language processing has used adversarial learning to reduce gender bias in clinical text analysis (De-Arteaga et al., 2019).

Despite these advances, several challenges remain. Many existing approaches focus on single-attribute bias, such as race or gender, while real-world disparities often arise from intersectional factors involving multiple demographic attributes (Crenshaw, 2021). In addition, adversarial debiasing often requires access to sensitive attribute labels, which are frequently incomplete or unavailable in medical datasets due to privacy regulations or inconsistent data collection (Holstein et al., 2019). Attempts to suppress bias-related features may also affect model accuracy, creating a trade-off between fairness and predictive performance (Madras et al., 2018). Furthermore, many debiased models provide limited interpretability, which can reduce clinicians' confidence in automated diagnostic systems (Amann et al., 2020). Another challenge is that many studies rely on curated datasets that do not adequately represent the diversity of real-world patient populations (Gichoya et al., 2022).

These limitations highlight an important gap in the application of adversarial debiasing to dermatological diagnostic imaging, particularly in addressing disparities related to skin tone representation. Existing diagnostic models have often shown reduced accuracy for darker skin tones due to dataset imbalance and limited demographic diversity.

To address this gap, this study incorporated adversarial debiasing within a fairness-by-design framework for dermatological diagnosis. A Convolutional Neural Network (CNN) was trained on the ISIC dataset to perform skin lesion classification, while an adversarial component was introduced to identify and suppress correlations associated with sensitive attributes such as Fitzpatrick skin types.

A custom loss function was developed to balance diagnostic accuracy with fairness objectives by minimizing the predictability of sensitive attributes while preserving clinically relevant image features. Additional steps included preprocessing dermatological images, applying feature disentanglement techniques to separate clinical indicators from demographic characteristics, and using balanced mini-batch sampling to improve subgroup representation during training. These methodological choices were intended to address disparities observed in previous dermatological diagnostic models, particularly the reduced performance for darker skin tones.

Through this approach, adversarial debiasing was integrated as a core component of a fairness-aware AI framework aimed at improving equity in dermatological diagnostic systems. By embedding bias mitigation directly within the training process, the framework sought to support the development of more reliable and equitable AI-assisted diagnostic tools.

2.5 Frameworks for Mitigating Bias in AI Systems

The review of existing fairness frameworks shows that most tools were designed primarily for structured datasets and general machine learning tasks. Although these frameworks provide valuable bias detection and mitigation techniques, they often lack mechanisms tailored to the complexities of medical image datasets. In dermatological imaging, fairness challenges arise from factors such as skin tone variation, lesion appearance differences, and dataset imbalance across demographic groups. These limitations highlight the need for fairness-aware frameworks specifically designed for medical imaging contexts. Addressing this gap formed the basis for the second objective of this study, which examined existing fairness frameworks to identify limitations that informed the design of a fairness-aware framework for diagnostic skin imaging.

2.5.1 IBM AI Fairness 360 (AIF360) Framework

IBM's AI Fairness 360 (AIF360) framework, introduced by Bellamy et al. (2019), has been widely used as an open-source toolkit for detecting and mitigating bias in machine learning models. The framework provides more than 70 fairness metrics together with several bias mitigation algorithms, making it a valuable tool for evaluating fairness in algorithmic decision-making systems. AIF360 addresses bias across three stages of the machine learning pipeline: pre-processing, where datasets are adjusted to reduce bias before training; in-processing, where learning algorithms incorporate fairness constraints during model training; and post-processing, where model predictions are modified to reduce discriminatory outcomes. This modular architecture allows the framework to be applied across a wide range of machine learning applications.

Despite its usefulness, the applicability of AIF360 to medical imaging tasks such as dermatological diagnosis remains limited. Most of the framework's tools were originally designed for structured

tabular datasets rather than image-based deep learning models (Mehrabi et al., 2022). Diagnostic skin imaging involves complex visual patterns, including variations in pigmentation, lesion morphology, and disease presentation. These characteristics require fairness evaluations at the image and lesion level, which are not directly supported by AIF360's default fairness metrics. Furthermore, medical imaging datasets introduce additional challenges such as scanner variability, annotation inconsistencies, and variations in disease prevalence across populations (Ghassemi et al., 2021). Such factors can significantly influence model performance across demographic groups, as demonstrated by studies reporting disparities in dermatological diagnostic accuracy across different skin tones (Larrazabal et al., 2020).

Another limitation of AIF360 lies in its emphasis on post hoc fairness evaluation. The framework primarily identifies and corrects bias after models have already been developed. However, recent research suggests that fairness considerations should be integrated throughout the entire AI development lifecycle rather than applied as a corrective step after training (Willemink et al., 2020). In medical imaging applications, biases introduced during dataset collection or annotation may persist throughout model training and deployment if fairness is not addressed during the learning process (Gaube et al., 2021).

In addition, AIF360 does not incorporate mechanisms for clinical validation or stakeholder involvement, which are important considerations in healthcare AI systems. Dermatological diagnostic models must perform reliably across diverse patient populations characterized by differences in skin tone, age, and medical history. Studies have shown that AI frameworks that lack clinical feedback and intersectional fairness analysis may produce models with uneven diagnostic accuracy across demographic groups (Obermeyer et al., 2019; Parikh et al., 2019). This

limitation highlights the importance of fairness frameworks that combine technical bias mitigation techniques with domain-specific medical knowledge.

Although AIF360 provides valuable tools for fairness assessment, its focus on structured datasets and post-training bias correction limits its effectiveness for dermatological diagnostic imaging. These limitations highlight the need for fairness approaches that can operate directly within deep learning pipelines and account for the unique characteristics of medical image datasets. Identifying these gaps formed an important step toward achieving the second objective of this study, which sought to analyze existing fairness frameworks in order to determine their limitations in medical image applications.

2.5.2 Fairlearn Framework

Fairlearn, introduced by Bird et al. (2020), is an open-source toolkit developed to assess and mitigate bias in machine learning models. The framework was designed to integrate fairness evaluation directly into existing machine learning workflows, allowing researchers and practitioners to measure disparities in model performance across different demographic groups. Fairlearn uses a modular architecture that enables users to apply fairness metrics, visualize disparities, and implement mitigation strategies within standard machine learning pipelines.

A key feature of Fairlearn is its reduction-based approach to bias mitigation, which treats fairness as an optimization problem. In this approach, predictive accuracy is balanced against fairness constraints across demographic groups. Algorithms such as grid search and exponentiated gradient reduction allow users to explore trade-offs between model performance and fairness. This enables developers to adjust models so that predictive outcomes remain accurate while reducing disparities between population subgroups. Previous studies have highlighted the usefulness of Fairlearn in

auditing algorithmic decision-making systems and improving transparency in machine learning models (Mehrabi et al., 2022; Microsoft Research, 2019).

Despite its adaptability, the framework presents several limitations when applied to diagnostic skin imaging. Similar to AI Fairness 360, many of Fairlearn's fairness metrics were originally designed for structured tabular data, making them less suitable for evaluating bias in image-based deep learning models (Mehrabi et al., 2022). Dermatological imaging models must account for complex visual characteristics such as variations in skin pigmentation, lesion texture, and disease manifestation. These factors require fairness assessments at the image and lesion level, yet Fairlearn does not provide built-in metrics specifically designed to evaluate such image-based disparities (Fairlearn Documentation, 2024).

In addition, the framework does not explicitly address challenges commonly encountered in medical imaging datasets, including scanner variability, annotation inconsistencies, and variations in disease prevalence across different populations (Liu et al., 2025). These factors can influence how machine learning models interpret dermatological images and may contribute to unequal diagnostic performance across demographic groups. As a result, general-purpose fairness auditing tools such as Fairlearn may identify disparities but lack the domain-specific mechanisms needed to address them effectively.

Another challenge concerns the framework's post hoc bias mitigation approach. Fairlearn primarily evaluates and adjusts models after they have already been developed rather than embedding fairness considerations directly within the model training process (Bird et al., 2020). Research has shown that biases introduced during dataset collection and model development often persist throughout deployment if fairness mechanisms are not incorporated during training. Furthermore, Fairlearn does not integrate clinician-driven validation processes, which are essential

for ensuring that fairness evaluations reflect real-world dermatological practice and patient diversity (Jaiyeoba et al., 2025).

Although Fairlearn provides valuable transparency tools such as fairness dashboards and visualization features, its structured-data orientation limits its effectiveness in medical image analysis. Applying the framework to dermatological imaging would require substantial customization, including the development of lesion-specific fairness metrics, integration of fairness monitoring during model training, and incorporation of clinical expertise during evaluation. These limitations highlight the need for fairness approaches specifically designed for medical imaging systems. Identifying these gaps contributed to the second objective of this study, which sought to analyze existing fairness frameworks to determine their limitations when applied to medical image datasets.

2.5.3 TensorFlow Fairness Indicators

TensorFlow Fairness Indicators, developed by Google, is a framework designed to evaluate and monitor fairness in machine learning models, particularly those built using the TensorFlow ecosystem. The framework provides scalable fairness metrics that examine model performance across different subpopulations defined by sensitive attributes such as race, gender, or age (TensorFlow, 2025). It integrates with TensorFlow tools such as TensorBoard, enabling practitioners to visualize fairness metrics and identify potential disparities in model predictions across demographic groups.

Methodologically, TensorFlow Fairness Indicators compute key performance metrics including accuracy, false positive rates, and false negative rates across user-defined data slices. This slice-based evaluation approach allows developers to examine how machine learning models perform across specific subgroups, helping to uncover disparities that may not be visible in aggregated

performance metrics. By embedding fairness evaluation into the model development workflow, the framework enables ongoing fairness monitoring during model training and validation.

One of the key strengths of TensorFlow Fairness Indicators lies in its integration with deep learning pipelines. The framework provides interactive dashboards that allow practitioners to visualize model performance across multiple demographic slices, facilitating the early detection of bias during model development. In addition, its compatibility with tools such as TensorFlow Model Analysis (TFMA) and TensorFlow Data Validation (TFDV) enables large-scale fairness evaluation across complex datasets. These capabilities make the framework useful for monitoring fairness in machine learning systems that rely heavily on TensorFlow-based development environments.

Despite these strengths, several limitations arise when applying TensorFlow Fairness Indicators to diagnostic skin imaging. The framework was originally designed to evaluate fairness in structured datasets and classification models rather than in high-dimensional medical imaging tasks. Dermatological diagnostic systems require fairness assessments that consider complex visual patterns, including variations in skin pigmentation, lesion morphology, and disease presentation. These characteristics require fairness metrics capable of capturing disparities at the image and lesion levels, yet TensorFlow Fairness Indicators does not provide built-in metrics specifically tailored to these imaging characteristics.

Furthermore, the framework does not directly address challenges commonly encountered in medical imaging datasets, such as scanner variability, annotation inconsistencies, and variations in disease prevalence across patient populations. These factors can influence how machine learning models interpret dermatological images and may contribute to unequal diagnostic performance across demographic groups. As a result, while TensorFlow Fairness Indicators can help identify

disparities in model outcomes, it does not inherently provide mechanisms for addressing domain-specific biases within dermatological diagnostic pipelines.

Adapting TensorFlow Fairness Indicators for dermatological imaging would therefore require significant customization, including the development of domain-specific fairness metrics capable of capturing pixel-level and lesion-level disparities. Additional enhancements could include improved visualization tools for interpreting fairness outcomes in medical images and the integration of fairness constraints directly into model training processes. These limitations highlight the need for fairness frameworks specifically designed for medical imaging environments. Identifying such gaps contributed to the second objective of this study, which involved analyzing existing fairness frameworks to determine their limitations when applied to medical image datasets.

2.5.4 The MONAI Framework

Medical Open Network for AI (MONAI) is an open-source framework specifically developed for deep learning applications in healthcare, with a strong focus on medical imaging (MONAI Consortium, 2020). Built on the PyTorch platform, MONAI provides domain-optimized tools, libraries, and workflows tailored to the unique characteristics of medical data. The framework supports a range of tasks, including medical image segmentation, classification, and registration, enabling researchers to build and evaluate deep learning models for clinical applications (Jorge Cardoso et al., 2022). Its design aims to bridge the gap between AI research and clinical practice by promoting reproducibility, reliability, and safe deployment of AI models in healthcare environments.

One of MONAI's key strengths is its specialized support for medical imaging. The framework offers advanced image processing capabilities and optimized deep learning architectures, such as

UNETR, for three-dimensional segmentation tasks (Pinaya et al., 2023). In addition, MONAI benefits from contributions by leading research institutions, including Stanford University and the National Institutes of Health (NIH), which have strengthened its credibility and adoption within the medical AI research community. MONAI also emphasizes reproducibility through standardized workflows and benchmarking tools that allow researchers to compare model performance against state-of-the-art implementations.

Despite these strengths, several limitations arise when applying MONAI to diagnostic skin imaging. The framework was primarily designed to operate with structured medical imaging formats such as DICOM and NIfTI, which are commonly used in radiology. Dermatological imaging datasets, however, often consist of unstructured clinical photographs captured under varying lighting conditions and imaging devices. As a result, adapting dermatological datasets to MONAI's workflow may require extensive preprocessing, which introduces additional complexity to the model development process.

Another limitation is that MONAI focuses primarily on improving model performance and reproducibility rather than explicitly addressing algorithmic fairness. While the framework provides powerful tools for medical image analysis, it does not incorporate built-in fairness metrics or bias mitigation techniques designed to evaluate disparities across demographic groups. In dermatological diagnostic systems, fairness considerations are particularly important due to variations in skin tone representation and disease presentation across populations. Without dedicated mechanisms for detecting and mitigating demographic bias, models developed within MONAI may still produce unequal diagnostic outcomes.

Additionally, MONAI requires familiarity with deep learning frameworks such as PyTorch, which may present a steep learning curve for researchers who are new to deep learning environments. Its

advanced architectures and computational requirements can also limit accessibility for researchers operating in resource-constrained settings where high-performance computing infrastructure is unavailable (Hatamizadeh et al., 2021).

Although MONAI provides a strong foundation for developing medical imaging models, its limited focus on fairness-aware model design highlights an important gap in existing frameworks. Addressing this gap requires the integration of fairness-aware mechanisms capable of detecting and mitigating demographic bias within dermatological diagnostic systems. Identifying these limitations supported the second objective of this study, which sought to analyze existing frameworks to determine their limitations when applied to medical image datasets.

2.5.5 Summary of Existing Fairness Frameworks

Table 2.1

Comparison of Existing Fairness Frameworks and Their Limitations

Framework	Key Strengths	Limitations	Relevance to Dermatological Imaging
AI Fairness 360 (AIF360)	Provides over 70 fairness metrics and multiple bias mitigation techniques across pre-processing, in-processing, and post-processing stages.	Primarily designed for structured datasets, it lacks support for image-level fairness analysis and real-time fairness monitoring.	Limited applicability to dermatological imaging due to a lack of lesion-level fairness metrics and medical imaging considerations.

Framework	Key Strengths	Limitations	Relevance to Dermatological Imaging
Fairlearn	Flexible toolkit for fairness auditing and mitigation; integrates fairness evaluation within machine learning pipelines using reduction-based optimization methods.	Fairness metrics are mainly designed for structured data; does not account for pixel-level disparities or medical imaging complexities.	Requires significant customization to evaluate fairness in dermatological image datasets.
TensorFlow Fairness Indicators	Provides scalable slice-based fairness evaluation and interactive dashboards for monitoring model performance across demographic groups.	Focuses mainly on performance auditing rather than bias mitigation; limited support for image-specific fairness metrics.	Useful for fairness monitoring but lacks mechanisms for addressing dermatology-specific biases.
MONAI	Specialized framework for medical imaging with strong support for deep learning workflows, reproducibility, and clinical AI development.	Focuses on model performance rather than fairness; lacks built-in bias detection and mitigation mechanisms.	Suitable for medical imaging pipelines but requires additional fairness-aware components for equitable dermatological diagnosis.

As recorded in Table 2.1, the comparison of existing fairness frameworks highlights important progress in the development of tools for detecting and mitigating bias in machine learning systems.

Frameworks such as AI Fairness 360, Fairlearn, and TensorFlow Fairness Indicators provide valuable mechanisms for evaluating fairness across demographic groups, while MONAI offers specialized support for medical imaging workflows. However, the analysis also reveals several limitations when these frameworks are applied to dermatological diagnostic systems.

Most existing frameworks were originally developed for structured datasets and general machine learning applications. Consequently, their fairness metrics and mitigation strategies are not fully suited for evaluating bias in high-dimensional medical image datasets where model decisions depend on complex visual features such as skin pigmentation, lesion morphology, and disease presentation. In addition, several frameworks emphasize post hoc fairness auditing rather than embedding bias mitigation mechanisms directly within model training pipelines.

Furthermore, many frameworks lack domain-specific considerations required for healthcare applications, including the integration of clinical expertise and fairness metrics capable of addressing variations in skin tone and disease manifestation across diverse patient populations. These limitations highlight the need for fairness-aware frameworks specifically designed to address the unique challenges of dermatological diagnostic imaging.

Identifying these gaps informed the second objective of this study, which sought to analyze existing fairness frameworks and determine their limitations in medical image datasets. The insights obtained from this analysis subsequently guided the development of a fairness-aware framework for diagnostic skin imaging, as presented in the following sections.

2.6 Case Studies on Successes and Failures in Diagnostic Imaging

To further illustrate the practical implications of bias in medical artificial intelligence systems, several studies have examined the effectiveness of different approaches to reducing bias in

diagnostic imaging. These case studies provide valuable insights into both the successes and limitations of fairness-aware AI techniques in real-world medical contexts. Examining such examples helps demonstrate how bias mitigation strategies perform in practice and highlights the challenges that remain in achieving equitable diagnostic outcomes across diverse patient populations.

2.6.1 Successful Bias Mitigation Approaches

Several studies have demonstrated that bias mitigation techniques can improve equity in diagnostic imaging systems when properly implemented. For example, Weng et al. (2023) investigated the impact of gender imbalance in diagnostic imaging datasets and showed that combining gender-balanced training data with adversarial debiasing significantly improved model performance across gender groups. By augmenting a male-dominated dataset with synthetic female X-ray images, the researchers achieved approximately a 15% improvement in diagnostic accuracy for female patients. Their findings highlight the importance of dataset diversity and fairness-aware training techniques in reducing gender bias within medical AI systems.

Similarly, Melisande (2020) addressed racial disparities in mammography by applying transfer learning and domain adaptation techniques. The study revealed that models trained primarily on images from Caucasian patients performed poorly when tested on images from African-American patients. By incorporating domain adaptation methods and introducing more diverse training data, the researchers achieved a 12% improvement in diagnostic accuracy for minority groups. This case demonstrates how domain adaptation techniques can help reduce racial disparities in diagnostic imaging models.

In the context of dermatological imaging, Ansari et al. (2024) explored skin lesion classification using datasets balanced across Fitzpatrick skin types. Their study applied adversarial debiasing

techniques within convolutional neural networks and achieved a 10% reduction in diagnostic disparities between lighter and darker skin tones. The results emphasize the importance of fairness-aware learning strategies when developing AI models for dermatological diagnosis.

Other studies have also addressed bias in different diagnostic imaging contexts. For instance, Guan et al. (2023) examined age-related bias in brain MRI analysis using multimodal learning approaches. By incorporating both diagnostic and age-related features into the model, they reduced the influence of age-related bias and achieved an 8% improvement in fairness across age groups. Similarly, Gencer & Gencer (2025) demonstrated improvements in fairness within retinal disease detection models by using diverse retinal imaging datasets and hybrid learning approaches. Their model reduced diagnostic disparities across ethnic groups by approximately 14%, highlighting the role of dataset diversity in promoting equitable diagnostic outcomes.

2.6.2 Persistent Failures and Challenges

Despite these successes, several studies have demonstrated that bias in diagnostic imaging systems remains a significant challenge. For example, Groh et al. (2022) evaluated commercial AI dermatology tools and found substantial racial disparities in their diagnostic performance. The study reported that these tools misdiagnosed melanoma 34% more frequently in darker-skinned patients compared to lighter-skinned patients. This disparity was largely attributed to the lack of diversity in the training datasets, where less than 5% of images represented darker skin tones. Attempts to apply post hoc debiasing techniques were insufficient to fully correct these disparities, illustrating the limitations of fairness interventions that are applied after model development.

Similarly, research conducted by the University of Cambridge and Simon Fraser University (University of Cambridge, 2020) revealed that certain AI techniques used in medical image reconstruction can produce false positives and false negatives due to algorithmic instability. In

these cases, small variations in input data were found to produce significant differences in model outputs. Such instability raises concerns about the reliability and robustness of AI-based diagnostic systems, particularly when deployed in clinical environments where consistent performance is critical.

These cases highlight that while bias mitigation strategies have shown promise in improving fairness in diagnostic imaging, significant challenges remain. Many current approaches rely heavily on dataset balancing or post hoc bias correction, which may not adequately address deeper structural biases within medical datasets and model training processes. These limitations reinforce the need for fairness-aware frameworks that integrate bias mitigation mechanisms directly into the AI development lifecycle. Addressing these challenges forms an important motivation for the fairness-aware framework proposed in this study for diagnostic skin imaging.

2.7 Addressing Intersectional and Systemic Biases

A significant challenge in achieving fairness in healthcare artificial intelligence systems is the presence of intersectional bias, where individuals belonging to multiple marginalized groups experience compounded disparities. Intersectionality theory, originally introduced by Crenshaw (2013), explains how overlapping social identities such as race, gender, and socioeconomic status can interact to produce unique forms of discrimination. In medical AI systems, these intersecting factors can influence how diagnostic models perform across different patient populations, often leading to unequal outcomes for individuals who fall within multiple vulnerable groups (Petsko et al., 2022; Williams, 2014).

Traditional fairness metrics often evaluate bias along a single demographic dimension, such as race or gender. However, this approach may fail to capture more complex patterns of discrimination that emerge when multiple attributes interact. Kearns et al. (2019) addressed this

limitation by proposing methods for rich subgroup fairness, which aim to ensure equitable treatment across a large number of demographic subgroups defined by combinations of sensitive attributes. These approaches recognize that fairness cannot always be achieved through simple demographic comparisons and require more advanced analytical techniques.

In addition to intersectional bias, systemic biases rooted in historical and structural inequalities can also influence the performance of AI systems in healthcare. Medical datasets, including electronic health records, clinical trial data, and medical imaging repositories, often reflect disparities in healthcare access, treatment availability, and disease reporting across different populations (Obermeyer et al., 2019). When machine learning models are trained on such data, they may inadvertently learn and reinforce these underlying inequalities.

Addressing systemic bias requires both technical and institutional interventions. From a technical perspective, strategies such as dataset diversification, adversarial debiasing, and fairness-aware model training can help reduce the influence of demographic bias during model development. From a broader perspective, improvements in data collection practices, transparency in dataset documentation, and increased representation of underserved populations in medical datasets are essential for promoting equitable AI systems. These considerations highlight the importance of developing fairness-aware approaches that can mitigate demographic disparities in medical imaging datasets. Such insights informed the design of the fairness-aware framework proposed in this study for dermatological diagnostic imaging.

2.8 Emerging Approaches for Fairness in AI Systems

In addition to established fairness frameworks, several emerging approaches have been explored to improve equity in artificial intelligence systems. These approaches seek to address bias at

different stages of the AI development lifecycle, including data collection, model training, and system deployment.

One promising approach is federated learning, which enables machine learning models to be trained across multiple decentralized datasets without requiring the sharing of raw data (Li et al., 2020). In healthcare settings, federated learning allows hospitals and medical institutions to collaboratively train models while preserving patient privacy. This approach can increase dataset diversity by incorporating data from multiple sources, thereby reducing biases associated with limited or homogeneous datasets (Sheller et al., 2020). In medical imaging applications, federated learning has shown potential for improving model generalizability across different clinical environments.

Another emerging technique involves the integration of causal inference methods into machine learning systems. Traditional machine learning models often rely on correlations within data, which can inadvertently reinforce biased associations. Causal inference approaches attempt to distinguish correlation from causation, enabling researchers to identify and control for confounding factors that contribute to biased predictions (Pearl, 2019). By modeling causal relationships within datasets, AI systems can be designed to reduce reliance on spurious correlations that may lead to unfair outcomes.

Active learning has also been explored as a strategy for improving dataset quality and reducing bias. In this approach, models iteratively select the most informative data samples for annotation, allowing training datasets to be expanded in a targeted manner (Settles, 2012). Although active learning can improve model performance by focusing on uncertain predictions, it does not necessarily guarantee diversity within the training dataset. As a result, additional strategies may

be required to ensure that underrepresented demographic groups are adequately included during data collection.

Finally, participatory design and community engagement have been proposed as complementary strategies for improving fairness in AI systems. These approaches emphasize the involvement of diverse stakeholders, including clinicians, patients, and representatives from marginalized communities, in the development and evaluation of AI technologies (Holstein et al., 2019). Incorporating multiple perspectives into the design process can help ensure that AI systems reflect the needs and values of diverse patient populations.

While these emerging approaches offer promising directions for improving fairness in AI systems, many of them function as complementary strategies rather than direct solutions for bias in diagnostic imaging models. Consequently, integrating fairness-aware mechanisms within model training remains essential for addressing demographic disparities in medical imaging applications. This study, therefore, focuses on adversarial debiasing as a practical approach for mitigating bias in dermatological diagnostic systems.

2.9 Research Gap

Despite the significant progress made in applying artificial intelligence to diagnostic imaging, several challenges remain in ensuring that these systems operate fairly across diverse patient populations. Existing studies have demonstrated that machine learning models trained on medical imaging datasets often inherit biases present in the data, leading to unequal diagnostic performance across demographic groups. In dermatological imaging, this issue is particularly evident due to the underrepresentation of darker skin tones in widely used datasets, which can result in reduced diagnostic accuracy for certain populations.

Several fairness frameworks, including AI Fairness 360, Fairlearn, TensorFlow Fairness Indicators, and MONAI, have been developed to detect and mitigate bias in machine learning systems. While these frameworks provide useful tools for fairness evaluation and model auditing, most were originally designed for structured datasets and general machine learning tasks. As a result, their fairness metrics and bias mitigation strategies are not fully suited to address the complexities of high-dimensional medical image datasets, where diagnostic decisions rely on subtle visual features such as pigmentation patterns, lesion morphology, and disease presentation.

Furthermore, many existing approaches treat fairness as a post hoc evaluation problem rather than integrating bias mitigation directly within the model training process. Although techniques such as dataset balancing and fairness auditing can reveal disparities in model performance, they may not sufficiently address deeper structural biases embedded in training data and model architectures. This limitation is particularly critical in dermatological imaging, where variations in skin tone and disease manifestation require fairness-aware learning strategies capable of reducing demographic bias during model development.

In addition, current research has not sufficiently explored the integration of adversarial debiasing techniques within fairness-aware frameworks specifically tailored to dermatological diagnostic systems. While adversarial learning has been successfully applied in other domains to mitigate demographic bias, its potential for improving fairness in dermatological image classification remains underexplored.

Therefore, there is a need for a fairness-aware framework that integrates bias mitigation mechanisms directly into the model training pipeline while accounting for the unique characteristics of dermatological imaging datasets. Addressing this gap formed the foundation of

the study, which developed and validated a fairness-aware framework for diagnostic skin imaging using adversarial debiasing techniques.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

This chapter presents the research methodology adopted in this study. It describes the research design, study area, datasets, model development procedures, and analytical techniques used to achieve the research objectives. In addition, the chapter outlines the tools and technologies used for developing and evaluating the fairness-aware framework for bias mitigation in dermatological diagnostic imaging.

3.2 Research Design

This study employed an experimental research design to systematically assess the impact of fairness-aware interventions in mitigating bias within AI-driven diagnostic imaging models. The experimental approach was chosen due to its strength in establishing causal relationships by directly manipulating model architecture through fairness-enhancing techniques, specifically, adversarial debiasing via gradient reversal layers. The primary objective was to determine whether fairness constraints could effectively reduce biases associated with skin tone, gender, or other demographic attributes in dermatological diagnostics while maintaining clinical reliability. To achieve this, two distinct model architectures were developed and evaluated using the International Skin Imaging Collaboration (ISIC 2020) dataset: a baseline Convolutional Neural Network (CNN) trained with conventional optimization strategies, and an experimental CNN integrating adversarial debiasing and data augmentation to actively suppress demographic biases during model training.

3.2.1 Selection of CNNs as the Core Model Architecture

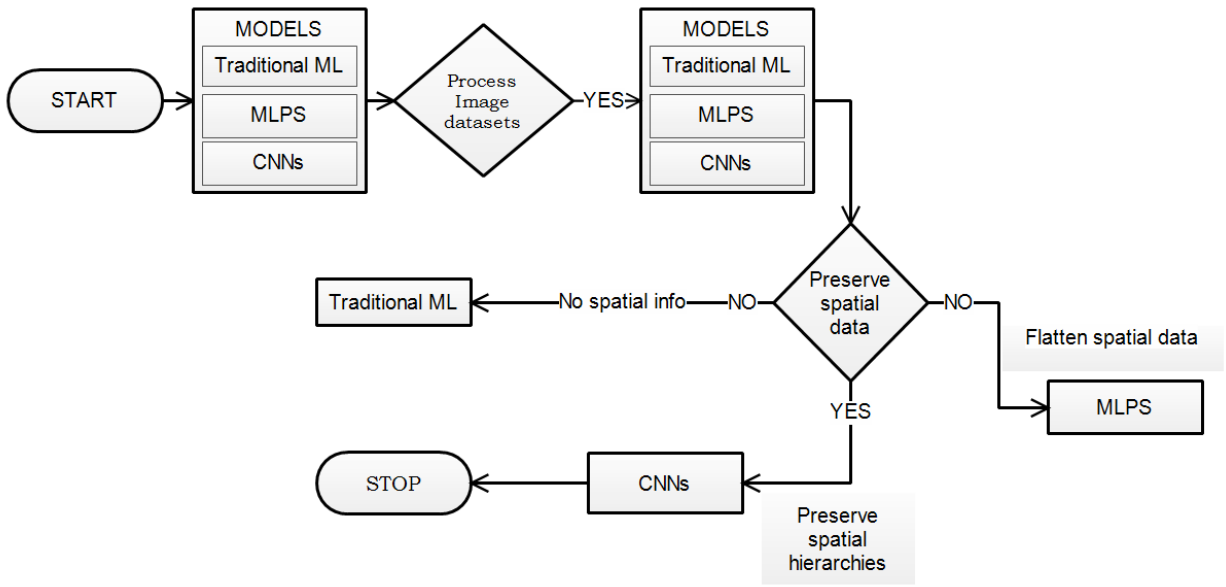
The choice of Convolutional Neural Networks (CNNs) as the primary model for AI-driven dermatological diagnostics followed a systematic evaluation of various machine learning approaches. Initially, traditional models such as Support Vector Machines (SVMs) and Random Forest classifiers were considered due to their well-established performance in tabular classification tasks. These models excelled at learning structured data representations, making them valuable for predictive modeling in many fields. However, their inability to capture spatial dependencies within medical images made them unsuitable for dermatological applications, leading to their exclusion from consideration. Spatial dependencies are crucial in dermatological imaging because skin lesions often exhibit subtle variations in texture, shape, and color that are distributed across neighboring pixels. Capturing these relationships allows a model to understand the structural patterns and contextual cues necessary for distinguishing between conditions, such as the irregular borders or asymmetry of melanoma versus benign lesions. Without accounting for spatial dependencies, models may miss clinically relevant features embedded in the image's layout, leading to inaccurate or unreliable diagnoses.

After SVMs, Fully Connected Neural Networks (Multi-Layer Perceptron, or MLPs) were explored as potential alternatives, given their strong feature representation capabilities. Unlike SVMs and Random Forest models, MLPs could process high-dimensional image inputs and extract complex patterns. Despite this advantage, MLPs failed to preserve spatial hierarchies, flattening the image structure and making it difficult to retain localized medical features. This limitation significantly impacted their ability to distinguish intricate dermatological characteristics, rendering them ineffective for bias-sensitive skin lesion classification.

CNNs emerged as the most suitable model architecture due to their ability to retain spatial dependencies within dermatological images, ensuring that critical medical features were effectively captured and leveraged for diagnostic accuracy. Their inherent ability to process localized patterns, textures, and structures made them superior in detecting bias trends in lesion classification, while also supporting fairness-aware optimization strategies. As a result, CNNs were selected as the optimal approach, providing the necessary balance between predictive reliability and fairness constraints in the development of AI-driven dermatological diagnostics. Figure 3.1 illustrates the decision-making process for selecting machine learning models based on whether image datasets are being processed and whether spatial data needs to be preserved.

Figure 3.1

Model Selection Flowchart for Medical Imaging Bias Mitigation (Researcher, 2025)



In Figure 3.1, Convolutional Neural Networks (CNNs) are chosen when spatial hierarchies are to be preserved.

The experimental setup was designed to measure the impact of debiasing interventions on both diagnostic accuracy and fairness across demographic subgroups. To ensure a controlled evaluation, both models underwent identical preprocessing procedures and training strategies, preventing external dataset variations from influencing fairness assessments. This approach allowed biases to be directly attributed to architectural modifications rather than dataset composition alone.

Fairness evaluations were conducted using multiple subgroup analyses, assessing model performance across Fitzpatrick skin types, gender subgroups, and lesion categories. Bias mitigation was quantified using fairness metrics such as statistical parity, equalized odds, and disparate impact ratio, which served as benchmarks to evaluate the effectiveness of adversarial debiasing.

By employing a controlled experimental design, this study supported causal inference, ensuring high internal validity in fairness evaluations. The integration of adversarial debiasing within the model development pipeline contributed to the advancement of fairness-aware AI methodologies, offering practical solutions for equitable dermatological diagnostics and reducing disparities in medical imaging applications.

3.3 Data Extraction and Selection

This study utilized the ISIC 2020 dataset, a widely recognized repository of dermatological images. The dataset was acquired directly from the ISIC archive (*https://www.isic-archive.com/*), ensuring access to high-resolution skin lesion images annotated with histopathological diagnoses. The validation, Fitzpatrick 17k dataset, was downloaded from (*GitHub – mattgroh/fitzpatrick17k*), where it is publicly available under a Creative Commons license. This dataset comprises 16,577 clinical images sourced from two dermatology atlases, DermaAmin and Atlas Dermatologico, and

annotated with Fitzpatrick skin type labels by expert and crowd annotators. It was chosen for validation due to its rich demographic diversity and explicit skin tone annotations, which are critical for assessing fairness in dermatological AI. Unlike many other datasets, Fitzpatrick 17k enables subgroup analysis across skin types I–VI, making it ideal for evaluating bias mitigation strategies and ensuring equitable diagnostic performance.

The selection process of the primary dataset, ISIC 2020, was guided by strict inclusion criteria to ensure demographic diversity and clinical relevance in fairness assessments. The ISIC 2020 dataset contained 33,126 images labeled with skin disease categories such as melanoma, benign nevi, and actinic keratosis. Each image was accompanied by metadata, including details on patient demographics (age, gender), lesion characteristics (diameter, asymmetry), and diagnostic labels validated through histopathology reports. Given the dataset's original skew toward lighter skin tones (Fitzpatrick I–III), efforts were made to enhance the representation of Fitzpatrick IV–VI skin types through synthetic augmentation. Figure 3.2 presents a decision-making flowchart for selecting dermatological datasets based on demographic diversity and coverage of skin conditions. The diagram visually outlines the inclusion criteria applied to four datasets, HAM10000, PermNet, PAD-UFES-20, and ISIC 2020, highlighting the selection process for models requiring population diversity, for example, gender, skin type, and comprehensive disease representation. This structured approach ensures that dataset choices align with fairness-aware AI methodologies, optimizing model performance for equitable dermatological diagnostics.

Figure 3.2

Dataset Selection Flow for Dermatological Image Analysis (Researcher, 2025)

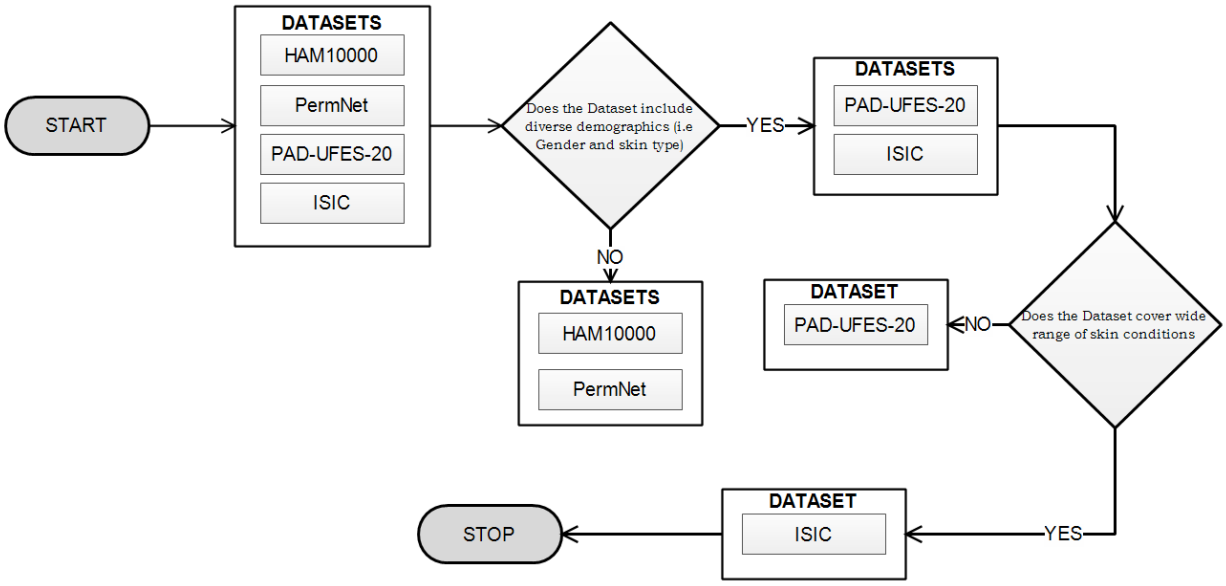


Figure 3.2 above illustrates the systematic evaluation and selection of dermatology image datasets based on demographic diversity and lesion coverage. ISIC 2020 is selected as the final dataset due to its global representation and broad diagnostic categories.

A key assumption was that histopathologically confirmed diagnoses served as reliable ground truth labels. Additionally, it was assumed that imaging conditions across different clinical settings introduced minor variations in lighting and resolution, which could be mitigated through preprocessing. The dataset was stratified to ensure balanced representation across demographic groups, preventing biased model training.

To develop and evaluate the AI models, the dataset was divided into training, validation, and test sets using an 80-10-10 split. This ensured that the model was trained on a sufficiently large portion, validated for hyperparameter tuning, and tested on unseen data for performance assessment. The

stratified split preserved demographic proportions to maintain fairness in evaluating skin type and gender-based disparities.

3.4 Data Preprocessing

Data preprocessing played a crucial role in ensuring model accuracy and fairness. A multi-step pipeline was implemented to standardize, clean, and optimize data for AI model training.

3.4.1 Data Cleaning and Purification

To ensure image consistency, all images were resized to 224×224 pixels, aligning with standard deep learning architectures for medical imaging. The resizing process applied an affine transformation, mathematically represented as: [I' = T(I)] where (I) represents the original image, (T) denotes the transformation matrix, and (I') is the resized output. This transformation ensured uniform scaling across the dataset, preventing discrepancies due to varying resolutions. The transformation matrix used for resizing was:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix} \quad (1)$$

Where, $\begin{bmatrix} x \\ y \end{bmatrix}$ are the original coordinates, $\begin{bmatrix} x' \\ y' \end{bmatrix}$are the transformed (resized) coordinates, $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is the linear transformation matrix (handles scaling, rotation, shearing) and $\begin{bmatrix} e \\ f \end{bmatrix}$ is the translation vector (shifts the image position). Equation 1 presents the transformation matrix formulation used in this study, adapted from standard principles in image coordinate mapping. The scaling and translation parameters adjusted each image to the required dimensions.

Histogram equalization was applied to normalize brightness and contrast levels to enhance image quality and reduce discrepancies caused by lighting variations. This was formulated as:

$$s_k = \frac{(L-1)}{N}\sum_{j=0}^{k} h_j \quad (2)$$

where (s_k) represents the equalized pixel intensity, *(L)* is the number of intensity levels, *(N)* is the total number of pixels, and *(h_j)* is the histogram count at intensity *(j).* The affine transformation matrix (Equation 2) follows the standard geometric formulation for image coordinate mapping, as described by Gonzalez (2009).

Additionally, Contrast Limited Adaptive Histogram Equalization (CLAHE) was employed to enhance local contrast by operating on small contextual regions and limiting noise amplification. A simplified representation of its local contrast adjustment can be expressed as:

$$I' = I + \alpha(I - I_{avg}) \quad (3)$$

where (I') is the enhanced pixel intensity, (I) represents the original pixel, (I_avg) is the local mean intensity, and (α) is the contrast scaling factor. While this formulation abstracts the full CLAHE process, it captures the principle of contrast amplification relative to local luminance (Pizer et al., 1987).

Metadata cleaning addressed missing demographic attributes, particularly skin type labels, which were absent in approximately 95.6% of ISIC images. To infer Fitzpatrick skin types, a pre-trained skin tone classifier was utilized, using a softmax activation function to generate probabilities for each skin type:

$$P(y_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \quad (4)$$

where *($P(y_i)$)* is the probability of skin type *(i), (zi)* represents the logit (raw score) output from a model for class I, and the denominator ensures probability normalization. To validate these

inferred skin classifications, predictions were compared against a subset of known labeled samples, ensuring consistency in subgroup representation.

For missing age and lesion location data, the k-nearest neighbors (k-NN) algorithm was used to impute values based on the closest matching instances in the dataset. The similarity between instances was determined using Euclidean distance, defined as:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_{i-} y_i)^2} \tag{5}$$

Where *(d(x,y))* represents Euclidean distance between two feature vectors *x* and *y*, and $(x_{i-} y_i)$ denote feature values of neighbouring data points, while *n* is the total number of features. Once the nearest neighbors were identified, missing values were estimated using the mean of k-nearest samples:

$$x = \frac{1}{k} \sum_{i=1}^{k} x_i \tag{6}$$

Where x is the imputed (estimated) value for a missing feature, x_i represents Feature values from the k-nearest neighbors, and *k* represents the Number of neighbors considered. This imputation ensured demographic consistency and enhanced fairness-aware model training by reducing bias in missing data handling.

These preprocessing steps established a standardized input pipeline, ensuring that bias mitigation techniques within the adversarial debiasing framework operated on uniformly processed images and structured metadata. By addressing disparities in image quality and demographic

representation, the dataset was optimized for fairness-aware AI modeling, reducing systemic biases linked to varying imaging conditions and incomplete metadata attributes.

3.4.2 Feature Engineering and Selection

To optimize model learning, key features were extracted and engineered:

Lesion shape and texture characteristics played a crucial role in assessing malignancy. Asymmetry was calculated by analyzing differences between the two halves of a lesion, as malignant tumors often exhibit irregular growth patterns compared to benign nevi. Edge irregularity was evaluated by mapping lesion borders, identifying jagged or blurred outlines, which are typical indicators of malignant transformations. Additionally, lesion diameter was measured since larger lesions tend to carry a higher risk of malignancy. These shape and texture metrics provided essential predictive signals, allowing the model to learn clinically relevant patterns for melanoma detection.

To further refine feature selection, pigmentation patterns were examined using color histograms. Skin lesions often display varying pigmentation densities, with malignant melanomas showing uneven coloration and darker, irregular spots. By computing color distribution histograms, the model was able to quantify pigmentation variability, identifying tonal differences that distinguish benign from malignant lesions. This approach minimized reliance on subjective visual assessments by dermatologists, ensuring objective and reproducible pattern recognition in AI-driven diagnostics.

In addition to lesion-specific attributes, demographic encodings were incorporated to enhance model interpretability and fairness. Gender and Fitzpatrick skin type, originally stored as categorical metadata, were encoded into numerical formats to facilitate machine learning processing, gender as binary values, for instance, 0 for male, 1 for female, and Fitzpatrick skin

types as discrete integer categories ranging from Type I to Type VI. This standardization allowed the model to identify potential biases in diagnostic accuracy across demographic groups, laying the groundwork for fairness-aware adjustments during adversarial debiasing. Recognizing that class imbalance was a significant concern, with benign cases making up 98.2% of the dataset versus only 1.8% of malignant cases, and that Fitzpatrick IV–VI images were underrepresented, balanced sampling techniques were implemented during training. In addition to traditional augmentations such as rotation, flipping, and controlled color adjustments, we specifically leveraged advanced augmentation techniques like GAN-based synthesis, color normalization, and geometric transformations to rebalance the dataset. These methods expanded the diversity of lesion appearances, prevented overfitting by exposing the model to a broader range of realistic scenarios, and ensured that the AI developed robust and equitable diagnostic capabilities across all demographic groups while preserving clinical relevance.

3.4.3 Hyperparameter Tuning

To achieve optimal model performance, Bayesian optimization and grid search were used to fine-tune key hyperparameters; these key hyperparameters were carefully fine-tuned to ensure stable convergence, efficiency, and generalization while preventing overfitting. The learning rate, ranging from .001 to .0001, governed how quickly the model updated its weights after each training iteration. A higher learning rate initially allowed the model to make faster progress toward optimal performance, but as training progressed, a lower rate ensured smooth refinements, preventing unstable fluctuations or overshooting optimal values. This gradual reduction in learning rate stabilized convergence, allowing the AI to improve predictability without erratic behavior.

The batch size, set at 16, 32, or 64, controlled the number of training samples processed before updating model weights. Smaller batch sizes helped the model capture fine-grained details, but

they introduced greater variability in weight adjustments due to the limited sample scope. Conversely, larger batch sizes improved computational efficiency and provided smoother updates, but they reduced the model's adaptability to subtle variations in lesion characteristics. By balancing batch size selection, the training process-maintained efficiency while ensuring that the model preserved its ability to generalize across different dermatological presentations.

To prevent overfitting, a dropout rate between .2 and .5 was applied during training. This technique randomly disabled a portion of neurons in the network, forcing the model to rely on a broader set of features rather than over-emphasizing specific patterns. A dropout rate of .2 retained most neurons, allowing the model to focus on distinct lesion characteristics while minimizing redundancy. In contrast, a dropout rate of .5 introduced a stronger regularization effect, enhancing generalization by discouraging the model from relying excessively on specific pixel-based correlations that might not be clinically meaningful.

Additionally, weight decay (L2 regularization) was implemented to enhance model generalization. By penalizing excessively large weight values, this technique discouraged the model from assigning disproportionate importance to specific features, thus promoting fairness and robustness in predictions. The regularization prevented the model from becoming overly sensitive to dataset biases, ensuring that AI-driven dermatological diagnostics remained stable across both the ISIC training dataset and the Fitzpatrick 17k validation dataset. Together, these hyperparameters played a crucial role in refining the model, balancing diagnostic accuracy with fairness considerations, and ensuring that adversarial debiasing efforts yielded equitable results.

Preprocessing validation procedures confirmed that standardization did not introduce new biases or remove clinically significant features. This ensured that fairness-aware AI models effectively addressed demographic disparities without compromising diagnostic accuracy.

3.5 Model Development and Training

In this study, a fairness-aware deep learning model was developed to classify dermatological lesions from dermoscopic images. The model was designed to detect subtle visual patterns associated with skin lesions while minimizing the influence of demographic characteristics that could introduce bias. A convolutional neural network architecture based on ResNet-50 was adopted for its demonstrated effectiveness in medical image classification and its ability to learn deep, hierarchical feature representations from complex image data.

The ResNet-50 architecture was employed as the backbone of the classification system. The network consisted of multiple convolutional layers that extracted spatial features from dermoscopic images, followed by pooling layers that reduced dimensionality while preserving critical lesion characteristics. The extracted features were then passed to fully connected layers responsible for performing the final lesion classification. The use of residual connections within the ResNet architecture enabled efficient training of deep networks by mitigating the vanishing gradient problem, thereby improving model stability and performance.

The model development process followed a structured pipeline consisting of dataset preparation, feature learning, adversarial bias mitigation, and model optimization.

First, dermoscopic images from the ISIC 2020 dataset were preprocessed to ensure uniform input dimensions and consistent image quality. All images were resized to a fixed resolution suitable for the ResNet-50 architecture and normalized to standardize pixel intensity values. The dataset was then divided into training, validation, and testing subsets using an 80-10-10 split. The training set was used for model learning, the validation set was used for hyperparameter tuning and monitoring model performance during training, and the testing set was reserved for final model evaluation.

Second, the CNN model was initialized using the ResNet-50 architecture and trained to learn discriminative visual features associated with dermatological lesions. Feature extraction occurred through successive convolutional layers that captured lesion boundaries, color patterns, and structural irregularities commonly associated with skin conditions. These learned features were subsequently used by the fully connected classification layer to produce diagnostic predictions.

Third, fairness considerations were incorporated through the integration of an adversarial debiasing mechanism. The adversarial framework consisted of two components: a primary predictor network and an adversary network. The predictor network was responsible for classifying skin lesions, while the adversary network attempted to predict sensitive attributes such as Fitzpatrick skin type from the features learned by the predictor. During training, the predictor was optimized to improve lesion classification accuracy while simultaneously minimizing the adversary's ability to infer sensitive attributes. This adversarial training process encouraged the model to learn diagnostic features that were independent of demographic characteristics.

Fourth, model optimization was conducted using the Adam optimizer with a learning rate of 0.001. Training was performed for 50 epochs with a batch size of 32, allowing the model to iteratively update its parameters based on mini-batch gradient descent. Hyperparameters were refined using validation experiments, and early stopping was implemented to prevent overfitting when validation performance stopped improving.

By jointly optimizing classification accuracy and fairness objectives, the resulting model achieved reliable dermatological lesion classification while reducing the influence of demographic bias. This fairness-aware training process ensured that the model maintained consistent diagnostic performance across diverse patient populations.

3.6 Experiment Procedure

This section describes the experimental implementation of adversarial debiasing used to mitigate demographic bias in dermatological lesion classification.

The experiment was conducted using dermoscopic images obtained from the ISIC 2020 dataset, which served as the primary dataset for model training and evaluation. The objective of the experiment was to train a fairness-aware convolutional neural network capable of performing accurate lesion classification while minimizing the influence of demographic attributes such as Fitzpatrick skin type.

The experimental architecture consisted of two interacting components: a primary predictor network and an adversarial debiasing module. The predictor network was implemented using a ResNet-50 convolutional neural network, which was responsible for classifying skin lesions as benign or malignant based on dermoscopic image features. The adversarial module functioned as a secondary network trained to predict sensitive demographic attributes, including Fitzpatrick skin type and gender, from the internal feature representations learned by the predictor network.

During the experiment, the two networks were trained simultaneously through an adversarial learning process. The predictor network aimed to improve diagnostic classification accuracy, while the adversarial network attempted to infer demographic attributes from the predictor's learned features. A custom loss function was implemented to balance these competing objectives. The predictor minimized lesion classification error while simultaneously minimizing the adversary's ability to accurately predict sensitive attributes.

To implement this adversarial interaction, gradient reversal layers (GRL) were incorporated into the model architecture. During forward propagation, the gradient reversal layer passes feature

representations normally to the adversary network. However, during backpropagation, the gradients were reversed, forcing the predictor network to learn feature representations that were less informative about demographic attributes. This mechanism allowed the model to retain clinically relevant lesion features while suppressing bias-correlated information.

Hyperparameter tuning was conducted to optimize the adversarial training process. Parameters such as adversarial loss weight, regularization strength, and adversary network complexity were adjusted using validation experiments to ensure that fairness constraints did not significantly reduce classification accuracy.

Following model training, experimental evaluation was conducted using fairness and performance metrics. Model performance was assessed using traditional classification metrics such as accuracy and recall, while fairness was evaluated using Statistical Parity Difference (SPD), Equalized Odds (EO), and Disparate Impact Ratio (DIR). These metrics were used to measure disparities in diagnostic performance across different Fitzpatrick skin types and demographic groups.

Through this experimental procedure, adversarial debiasing was integrated directly into the model training process, enabling the development of a dermatological diagnostic model that maintained strong classification performance while reducing demographic bias.

3.7 Data Analysis and Presentation

This study evaluated the effectiveness of the fairness-aware AI framework in reducing bias within dermatological diagnostics while maintaining clinical reliability. The analysis was structured into three key stages: detecting bias before fairness interventions, improving fairness using adversarial debiasing and synthetic data augmentation, and comparing results before and after fairness-aware

adjustments to ensure measurable improvements in both fairness and diagnostic accuracy, as outlined below.

3.7.1 Detecting Bias Before Fairness Adjustments

The first stage involved quantifying bias trends within the model's predictions, particularly with Fitzpatrick skin types, gender subgroups, and lesion categories. To assess bias, several key evaluations were conducted.

Error rate disparities were examined, with a focus on whether darker skin tones, specifically Fitzpatrick IV–VI, exhibited higher false negative rates, suggesting that melanoma cases in these subgroups were missed more frequently. Intersectional bias analysis was also performed, measuring model accuracy across combined demographic attributes, such as skin type and gender, to determine if the AI was less effective for certain subpopulations, such as women with darker skin tones.

Additionally, feature representation analysis was used to investigate whether the model learned biased patterns, analyzing how learned features clustered across demographic groups to detect representation disparities. To further validate these findings, the Coefficient of Determination (R^2) was used to assess how effectively the model explained variation in melanoma detection across different skin types, ensuring that fairness interventions did not distort diagnostic accuracy.

3.7.2 Improving Fairness with Adversarial Debiasing and Data Augmentation

Once bias was quantified, fairness constraints were embedded directly into the model training process to proactively mitigate disparities rather than applying corrections post-hoc. The fairness-aware framework integrated several optimization strategies to ensure that bias reduction was inherent to the AI's learning process.

Fairness-constrained training involved adjustments that ensured demographic attributes such as skin type or gender did not disproportionately influence model predictions, preventing biases from being reinforced during classification. Additionally, fine-tuning optimization balanced accuracy with fairness constraints, ensuring that bias mitigation measures did not come at the cost of predictive stability.

To further compensate for demographic imbalances, advanced synthetic data augmentation techniques were employed. This included SMOTE, contrast normalization, and GAN-based image synthesis, which enhanced the representation of underrepresented Fitzpatrick skin types, ensuring a more equitable distribution of learning samples.

3.7.3 Comparing Model Performance Before and After Fairness Adjustments

The final stage of analysis focused on measuring the impact of fairness-aware interventions by comparing model performance before and after bias mitigation strategies were applied. This evaluation ensured that the improvements in fairness did not come at the expense of diagnostic accuracy.

The F2 Score was prioritized to enhance recall, ensuring that melanoma cases in underrepresented subgroups were detected with high reliability while reducing false negatives. This was critical in addressing disparities where certain demographic groups, particularly those with darker skin tones, experienced lower sensitivity rates in diagnosis. The Root Mean Square Error of Approximation (RMSEA) was applied to evaluate how well fairness constraints were integrated within the model, ensuring that adjustments did not introduce distortions in the learning process. Similarly, Root Mean Square Error (RMSE) quantified differences between actual and predicted diagnoses, validating that fairness modifications did not result in additional classification errors.

By systematically comparing pre- and post-intervention results, this study confirmed that fairness-aware AI models could mitigate biases in dermatological diagnostics while maintaining clinical reliability. These findings reinforced the feasibility of implementing fairness-aware optimizations in AI-driven healthcare applications, improving equity without compromising medical accuracy.

3.7.4 Visualizing Results

To enhance clarity in fairness improvements, the results were presented through multiple visualization techniques:

Subgroup-specific ROC Curves demonstrated improvements in diagnostic accuracy across different Fitzpatrick skin types, providing a clear view of enhanced sensitivity and specificity.

Bias Correction Trend Graphs tracked fairness improvements over multiple training iterations, showcasing progressive reductions in subgroup disparities. These visualizations provided comprehensive insights into the fairness-aware modifications, enabling a clear understanding of how bias mitigation strategies affected AI-driven dermatological diagnostics.

3.8 Model Validation and Clinical Relevance

Ensuring that the fairness-aware AI model remained reliable in real-world dermatological diagnostics required a structured validation process. The validation phase assessed the model's ability to generalize beyond controlled training conditions, reinforcing its effectiveness for broader clinical applications.

3.8.1 Model Validation

Model validation is a systematic process for assessing whether a machine learning model meets predefined criteria for accuracy, fairness, robustness, and generalizability before deployment. This study employed a multi-layered validation approach to ensure that fairness-aware interventions

maintained predictive reliability while effectively reducing bias. The technical evaluation of the model relied on standardized performance metrics, including the F2 Score, which prioritized recall to minimize false negatives and ensure accurate melanoma detection in underrepresented subgroups. Additionally, the Coefficient of Determination (R^2) measured the model's ability to explain diagnostic variance, confirming that fairness-aware modifications did not compromise predictive stability. To assess the alignment of fairness constraints within the model, the Root Mean Square Error of Approximation (RMSEA) was applied, ensuring adjustments were appropriately balanced without excessive correction, while the Root Mean Square Error (RMSE) validated the consistency between actual and predicted diagnoses, reinforcing the model's reliability in clinical applications.

Beyond technical performance, fairness metrics were rigorously evaluated, including Statistical Parity Difference (SPD), Equalized Odds (EO), and Disparate Impact Ratio (DIR). These metrics quantified bias reduction, ensuring that fairness-aware modifications achieved measurable improvements across demographic subgroups. Interpretability validation was also conducted using Shapley values and LIME, confirming that fairness-aware interventions did not obscure model transparency and that predictions remained clinically interpretable. To verify the model's generalizability beyond controlled training environments, an external dataset (Fitzpatrick 17k) was used for validation, demonstrating the framework's effectiveness across diverse patient populations. Additionally, stakeholder validation was integrated into the process, engaging dermatologists, AI ethics specialists, and healthcare policymakers to evaluate the practical implications of bias mitigation strategies within real-world clinical workflows.

This structured validation approach ensured that bias mitigation strategies effectively balanced equity with predictive accuracy, reinforcing confidence in the application of fairness-aware AI methodologies in dermatological diagnostics. By demonstrating statistically significant improvements in fairness and maintaining clinical reliability, the study advances efforts toward ethical AI-driven healthcare, making fairness optimization both effective and implementable in real-world medical settings.

3.8.2 Clinical Relevance

For AI-driven dermatological diagnostics to be impactful, fairness optimization must align with clinical needs, medical decision-making, and expert validation. To ensure the reliability of fairness-aware interventions, stakeholder validation was conducted, incorporating insights from dermatologists, AI ethics specialists, and healthcare policymakers. Their feedback confirmed the practicality of fairness adjustments while reinforcing the necessity of maintaining diagnostic integrity and interpretability in clinical workflows.

This study ensured clinical relevance by incorporating the following principles:

Alignment with Medical Standards – The AI model's fairness constraints were designed to preserve diagnostic accuracy, ensuring medical practitioners could trust its outputs without concerns over bias-induced performance degradation.

Minimizing Health Risks for Underrepresented Groups – Fairness optimization directly improved diagnostic accuracy for groups historically overlooked in skin lesion classification, such as individuals with darker skin tones, who often experience delayed melanoma diagnoses due to biased AI models.

Expert Validation of Bias Mitigation – Through structured stakeholder engagement, clinical experts confirmed that bias mitigation strategies did not disrupt established diagnostic workflows but rather enhanced model transparency and fairness in skin lesion classification.

This fairness-aware AI framework was designed to bridge the gap between technical fairness optimization and clinical applicability, ensuring equity-driven AI advancements remain practical, ethical, and implementable in real-world dermatological diagnostics.

3.9 Logistical and Ethical Considerations

As the model architecture was developed and its training methodology refined, ethical considerations played a central role in ensuring fairness in dermatological diagnostics. Recognizing that AI-driven healthcare systems could inadvertently reinforce disparities, this study was grounded in the principle that equitable treatment is essential for clinical reliability. In particular, biased predictions, such as higher false-negative rates for darker-skinned individuals, were evaluated to determine their potential impact on delayed diagnoses and poorer health outcomes. This ethical imperative guided the research approach, ensuring that the AI model was not only technically robust but also aligned with fairness principles that support equitable patient care.

Before the research commenced, the necessary institutional approvals were obtained to ensure compliance with ethical research standards. Once the study proposal was accepted at the school level, the Dean of the faculty submitted a request to the School of Postgraduate Studies for clearance to apply for a research license. Upon approval, the postgraduate office forwarded the request to the Registrar for Research and Extension, initiating the formal process with the National

Commission for Science, Technology, and Innovation (NACOSTI). Ethical clearance from the Scientific and Ethics Committee was required before NACOSTI granted the official research permit. This structured process ensured that the study adhered to regulatory standards and ethical research principles from the outset.

Bias mitigation in the AI model involved a multi-layered approach, proactively addressing fairness at various stages of development. The dataset was examined for imbalances, particularly the underrepresentation of Fitzpatrick IV–VI skin types, which could lead to skewed diagnostic accuracy across demographic groups. To correct these disparities, data augmentation techniques, including color adjustments, rotations, and GAN-based synthesis, were integrated to improve the diversity of training samples. These strategies ensured a richer, more representative dataset, reducing the likelihood of biased model predictions.

Additionally, fairness-aware mechanisms were embedded directly into model training through adversarial debiasing techniques. During this process, the model actively minimized reliance on demographic markers when making diagnostic predictions, ensuring alignment between technical optimization and ethical accountability. While no AI system is entirely free from bias, this approach contributed to the development of a more equitable and clinically trustworthy diagnostic tool. This study delivered AI-driven dermatological diagnostics that were not only accurate but also ethically sound and fair, supporting equitable healthcare outcomes for diverse populations.

CHAPTER FOUR

DATA PRESENTATION, ANALYSIS, INTERPRETATIONS, AND DISCUSSIONS

4.1 Introduction

This chapter presents the experimental results of the study, focusing on how the fairness-aware AI framework performs in mitigating bias and maintaining diagnostic reliability in dermatological imaging. It highlights comparisons between models, key fairness and accuracy metrics, and validation techniques used to assess the impact of fairness interventions.

4.2 Data Cleaning and Preprocessing

Data preprocessing played a critical role in ensuring both accuracy and fairness in AI-driven dermatological diagnostics. The dataset underwent a multi-step refinement process to mitigate biases, handle missing values, and standardize imaging inputs for optimal model training. The first stage, data cleaning, involved the removal of duplicate images, ensuring that redundant samples did not disproportionately influence model learning. Outliers, including images with excessive artifacts or poor resolution, were filtered out to maintain diagnostic reliability, ensuring that the AI model received only clinically relevant data.

To address missing demographic attributes, particularly skin type labels, a pre-trained skin tone classifier was employed. Predictions were validated against a subset of known labeled samples to ensure consistency in subgroup representation, thereby reinforcing demographic fairness. For missing age and lesion location data, the k-nearest neighbors (k-NN) algorithm was applied, imputing values based on the closest matching instances in the dataset. Euclidean distance

calculations guided this imputation process, enabling accurate estimations of missing attributes while preserving demographic consistency.

Standardization techniques were applied to normalize imaging conditions across the dataset, reducing the impact of varying lighting and contrast. Histogram equalization and Contrast Limited Adaptive Histogram Equalization (CLAHE) enhanced image quality by ensuring consistent brightness and contrast levels. These methods minimized disparities caused by different imaging conditions and ensured equitable feature extraction across diverse skin tones. To further enhance fairness, synthetic data augmentation techniques, particularly the Synthetic Minority Oversampling Technique (SMOTE), were employed. SMOTE-generated synthetic samples balanced the representation of underrepresented Fitzpatrick skin types IV–VI, addressing demographic imbalances and improving fairness-aware training.

Additional augmentation techniques, including GAN-based synthesis, color normalization, and geometric transformations, were leveraged to expand lesion appearance diversity, reducing model overfitting and enhancing diagnostic generalization. These strategies ensured that AI-driven predictions remained stable across various dermatological presentations while reinforcing fairness-aware modeling. Furthermore, structured feature engineering was implemented to refine predictive signals, incorporating lesion shape, texture, and pigmentation patterns, essential for malignancy assessment.

The preprocessing validation process confirmed that standardization techniques did not introduce new biases or compromise clinically significant features. The resulting dataset was optimized for fairness-aware AI modeling, ensuring that systemic biases linked to imaging inconsistencies and missing demographic attributes were effectively mitigated. Full methodological details regarding

preprocessing procedures can be referenced in Section 3.4, where individual techniques are comprehensively described.

4.3 Analysis of Specific Objectives of the Study

The following subsections provide an analysis, interpretation, and discussion of the three specific objectives that drove this study.

4.4 Findings on Adversarial Debiasing

The first objective of this study was to evaluate the effectiveness of adversarial debiasing mechanisms used in diagnostic imaging. The review of adversarial debiasing in healthcare AI systems revealed its effectiveness in mitigating biases within diagnostic imaging applications. By utilizing a predictor-adversary framework, adversarial debiasing ensures that AI models suppress correlations with sensitive attributes such as race, gender, and socioeconomic status, thereby enhancing fairness in disease detection and medical decision-making. The technique has demonstrated measurable improvements in fairness indicators, including reduced disparities in classification outcomes across demographic groups, particularly in dermatological diagnosis, where biases linked to skin type have historically impacted AI performance.

Studies reviewed in Section 2.4 confirm that adversarial debiasing has been successfully applied in various healthcare domains. The method has shown promising results in radiology, dermatology, cardiovascular risk prediction, and clinical decision-support systems, improving fairness metrics while maintaining diagnostic accuracy (Groh et al., 2021a). Within AI-driven dermatological frameworks, adversarial debiasing has effectively minimized biases associated with Fitzpatrick skin type classifications, ensuring equitable performance across different skin

tones. These applications highlight the potential of fairness-aware AI models in addressing healthcare disparities.

In this study, adversarial debiasing was integrated into a fairness-aware AI framework for dermatological diagnostics. By employing custom loss functions, feature disentanglement techniques, and balanced sampling strategies, the framework successfully reduced disparities while maintaining diagnostic stability. These methodological refinements addressed performance variations observed in previous analyses, reinforcing the role of adversarial debiasing in advancing equitable AI applications in healthcare. For further details on the technical aspects and foundational principles of adversarial debiasing (section 2.4), which provides a comprehensive review of the method and its broader implications in healthcare AI.

4.4.1 Discussion on the Effectiveness of Adversarial Debiasing in Diagnostic Imaging

The findings of the first objective reaffirm that adversarial debiasing is an effective strategy for mitigating bias in AI-driven diagnostic imaging. Grounded in the principles of Adversarial Fairness Learning (AFL), this study operationalized fairness as a core optimization goal rather than a post-hoc adjustment. By embedding a gradient reversal layer into the model architecture, the framework actively suppressed bias-associated features during training, aligning with AFL's central tenet: fairness through representational neutrality (Zhang et al., 2018).

Supportive literature has consistently demonstrated the utility of adversarial debiasing in reducing disparities across demographic groups. In dermatology, bias mitigation targeting Fitzpatrick skin types has improved melanoma detection rates among darker skin tones, addressing long-standing diagnostic inequities (Groh et al., 2021b). Similar gains have been observed in cardiovascular risk

prediction, where adversarial suppression of race-correlated features led to more equitable assessments (Obermeyer et al., 2019). These findings affirm the technical feasibility of AFL-based interventions in clinical AI.

However, some studies caution against overreliance on adversarial methods. Madras et al. (2018) and Schoeffer et al. (2024) highlight that suppressing bias-correlated features can inadvertently reduce model sensitivity, particularly in high-stakes domains where recall and specificity are critical. This trade-off between fairness and accuracy is echoed in Ghassemi et al. (2021), who argue that fairness interventions must be clinically contextualized to avoid unintended harm. Moreover, adversarial training often obscures feature attribution, complicating interpretability and clinician trust (Amann et al., 2020; Ribeiro et al., 2016).

To address these limitations, hybrid debiasing strategies have emerged as promising alternatives. Combining adversarial learning with data augmentation and representation balancing can enhance fairness while preserving diagnostic integrity (Shorten & Khoshgoftaar, 2019). Additionally, multi-objective optimization frameworks offer a principled way to balance fairness constraints against performance thresholds, enabling more nuanced trade-offs (Yu et al., 2022). Interpretability remains a critical frontier. While AFL improves fairness at the representation level, it risks creating opaque decision boundaries. Tools such as SHAP, Grad-CAM, and clinician-validated heatmaps can help restore transparency, ensuring that fairness-aware models remain clinically actionable (Holzinger et al., 2017; Ghosh, 2023).

In summary, this study contributes to the evolving discourse by demonstrating that AFL-based debiasing can achieve substantial fairness gains without compromising diagnostic viability, provided that trade-offs are carefully managed. By integrating fairness into the model's learning

objective and validating across diverse skin types, the framework advances both technical rigor and ethical deployment in dermatological AI.

4.5 Findings on Existing Frameworks in Diagnostic Imaging

The second objective of this study sought to analyze existing fairness frameworks used in diagnostic imaging and identify limitations that restrict their applicability to medical image datasets. A review of major frameworks commonly used in fairness-aware machine learning revealed that although several tools exist for bias detection and mitigation, most were designed primarily for structured datasets and therefore face challenges when applied to unstructured medical imaging data. Table 4.1 summarizes the key characteristics, strengths, and limitations of the reviewed frameworks.

Table 4.1:

Summary of Existing Fairness Frameworks in Diagnostic Imaging

Framework	Key Features	Strengths	Limitations in Diagnostic Imaging
IBM AI Fairness 360 (AIF360)	Provides fairness metrics and bias mitigation algorithms across pre-processing, in-processing, and post-processing stages	Wide range of fairness metrics and mitigation techniques	Designed primarily for structured tabular data; requires extensive customization for image-based bias detection

Framework	Key Features	Strengths	Limitations in Diagnostic Imaging
Fairlearn	Optimization-based fairness auditing and bias mitigation using reduction techniques	Flexible integration into machine learning pipelines	Limited support for high-dimensional medical imaging; fairness assessment mainly performed post-hoc
TensorFlow Fairness Indicators	Slice-based fairness evaluation integrated with TensorFlow and TensorBoard	Real-time fairness monitoring and visualization	Lacks lesion-level fairness metrics required for dermatological imaging
MONAI Framework	Deep learning framework tailored for medical imaging applications	Supports medical imaging workflows and domain-specific pipelines	Focuses mainly on performance optimization rather than explicit fairness constraints

4.5.1 Discussion on Existing Frameworks in Diagnostic Imaging

The findings presented in Table 4.1 reveal several important insights regarding the applicability of existing fairness frameworks in diagnostic imaging. Although numerous frameworks have been developed to detect and mitigate bias in machine learning systems, their effectiveness varies considerably depending on the nature of the data and the specific clinical context in which they are applied.

First, the results indicate that many widely used fairness toolkits, including IBM AI Fairness 360 (AIF360) and Fairlearn, were originally designed to operate on structured tabular datasets. These frameworks provide robust mechanisms for evaluating fairness across demographic groups using metrics such as Statistical Parity Difference and Equalized Odds. However, their applicability to diagnostic imaging remains limited because dermatological images represent high-dimensional, unstructured data where fairness may arise from subtle visual features such as pigmentation patterns, lesion morphology, and variations in skin tone. As a result, these frameworks often require extensive customization to evaluate fairness at the image or lesion level.

Second, the findings highlight that tools such as TensorFlow Fairness Indicators are useful for monitoring disparities across predefined subgroups during model development. Their integration with machine learning workflows allows developers to track fairness metrics in real time. Nevertheless, the analysis suggests that these tools primarily focus on auditing model performance rather than directly mitigating bias during model training. Consequently, they may identify disparities without necessarily providing mechanisms to prevent them from emerging in the first place.

Third, domain-specific frameworks such as MONAI offer stronger support for medical imaging applications because they provide specialized pipelines for handling medical datasets. However, the findings reveal that even these frameworks largely emphasize model performance and reproducibility rather than explicitly embedding fairness constraints within the learning process. This limitation suggests that although medical imaging frameworks are capable of handling complex image data, they still require additional mechanisms to address demographic bias effectively.

Another key observation from the findings is the limited consideration of intersectional bias within most existing frameworks. Many fairness approaches evaluate bias across single attributes such as race or gender. However, in healthcare settings, disparities often arise from the interaction of multiple demographic factors, including race, gender, age, and socioeconomic status. As a result, single-attribute fairness metrics may fail to capture the full extent of bias present in diagnostic systems.

Overall, the discussion of these findings suggests that current fairness frameworks provide valuable tools for bias detection and monitoring but are not fully equipped to address the complexities of dermatological diagnostic imaging. The limitations identified in this study highlight the need for fairness-aware AI models that integrate bias mitigation mechanisms directly into the model training process. These insights informed the development of the fairness-aware adversarial framework proposed in this research, which aims to address bias in dermatological diagnostic systems by embedding fairness constraints within deep learning pipelines.

4.6 Model Development and Training

The third objective of this study was to develop and evaluate a fairness-aware framework capable of mitigating bias in dermatological diagnostic imaging. To achieve this objective, a series of experiments were conducted to compare the performance of a baseline convolutional neural network with a fairness-aware model incorporating adversarial debiasing mechanisms.

The experiments were conducted using dermoscopic images obtained from the International Skin Imaging Collaboration (ISIC) archive. Prior to training, all images were resized to 224 × 224 pixels, which is the standard input size for the ResNet-50 architecture. Image preprocessing techniques, including histogram equalization and adaptive contrast enhancement, were applied to reduce variations in illumination and pigmentation that could influence model predictions. These

preprocessing steps helped ensure consistent feature extraction across images with different lighting conditions and skin tones.

The dataset also required preprocessing of associated metadata to ensure demographic attributes could be used in fairness evaluation. In the original dataset, Fitzpatrick skin type labels were missing for approximately 95.6% of the images. To address this limitation, a pre-trained skin tone classifier was used to infer Fitzpatrick skin types for unlabeled samples. The inferred labels were validated against a subset of images with known skin type annotations to ensure reasonable classification accuracy. In addition, missing demographic variables such as age and lesion location were imputed using a k-nearest neighbors (k-NN) algorithm to maintain demographic consistency within the dataset.

To evaluate the effectiveness of adversarial debiasing, two model configurations were developed and compared. The first model served as a baseline classifier, implemented using a ResNet-50 convolutional neural network. This model was trained using the Adam optimizer with a learning rate of 0.001, a batch size of 32, and a total of 50 training epochs. Early stopping was implemented during training to prevent overfitting and ensure model generalization.

The second model extended the baseline architecture by incorporating an adversarial debiasing component designed to suppress the influence of demographic attributes during feature learning. In this configuration, an adversarial network was trained alongside the primary classifier to predict sensitive attributes such as Fitzpatrick skin type and gender from the learned feature representations. During training, the primary model optimized classification accuracy while simultaneously minimizing the adversary's ability to infer demographic attributes. This adversarial learning strategy encouraged the model to learn diagnostic features that were independent of demographic characteristics.

Through this experimental design, the study evaluated whether embedding fairness constraints directly into the model training process could reduce demographic disparities in lesion classification while maintaining strong diagnostic performance. The results of these experiments are presented and analyzed in the subsequent sections.

4.7 Performance Evaluation of the Baseline and Fairness-Aware Models

Following model development and training, the performance of the two models was evaluated using both traditional classification metrics and fairness-specific metrics. The purpose of this evaluation was to determine whether the fairness-aware interventions improved demographic equity while maintaining clinically relevant predictive performance. The analysis compared the baseline Convolutional Neural Network (CNN) with the fairness-aware model that incorporated adversarial debiasing. Table 4.1 presents the comparative performance results for the two models.

Table 4.1

Comparative Performance Metrics for the Baseline CNN and Fairness-Aware Model

Metric	Baseline CNN	Fairness-aware Model
Accuracy	91.7%	91.3%
Precision	High (qualitatively robust)	High (consistent across subgroups)
Recall	Balanced	Balanced
F1-Score	Robust	Comparable
Statistical Parity Difference (SPD)	.35	.05

Metric	Baseline CNN	Fairness-aware Model
Equalized Odds (EO)	.42	.12
Disparate Impact Ratio (DIR)	.65	.95
Paired t-test (p-value)	--	< .05

As illustrated in Table 4.1, while the baseline CNN achieved an accuracy of 91.7%, the fairness-aware model achieved a comparable accuracy of 91.3%. This negligible difference confirms that incorporating fairness constraints did not compromise the model's clinical performance. However, the impact on fairness metrics is dramatic. The Statistical Parity Difference (SPD) improved from .35, indicating a significant disparity in positive predictions among subgroups, to .05, demonstrating an almost equitable distribution. Similarly, the Equalized Odds (EO) metric, which highlights inconsistencies in false positive and false negative rates, was reduced from .42 to .12. Finally, the Disparate Impact Ratio (DIR) improved from .65 to .95, indicating a near-level playing field across demographic groups.

To further validate that these improvements were not due to random chance, paired samples t-tests were conducted to determine whether there was a statistically significant difference between the baseline CNN and fairness-aware model, which yielded a p-value below .05. This statistical validation confirms that the enhancements in fairness metrics are significant, ensuring that the fairness-aware model reliably reduces subgroup bias while delivering robust predictive performance. Below is a comprehensive, in-depth explanation of each fairness metric used.

4.7.1 Statistical Parity Difference (SPD)

Statistical Parity Difference (SPD) is a fairness metric used to evaluate whether different demographic groups receive similar proportions of positive outcomes from a predictive model. It compares the probability of a favorable prediction, for instance, a positive diagnosis, across sensitive attribute groups, such as race or gender. Ideally, a model should produce similar outcome rates for all groups, indicating parity. A non-zero SPD signals that one group is being favored or disadvantaged relative to another, with larger deviations reflecting greater disparity.

This metric is particularly important in healthcare AI, where unequal prediction rates can heighten existing health disparities. In this study, SPD was used to quantify fairness improvements, with observed reductions indicating enhanced equity without compromising predictive performance.

Mathematically, SPD is defined as:

$$SPD = P(Y' = 1|D = d_1) - P(Y' = 1|D = d_2) \quad (6)$$

Here, *Y'* is the predicted outcome, *D* is the sensitive attribute or protected group variable, d_1 and d2 are two different groups or values of the sensitive attribute for instance, d_1 for male, and d_2 for female and $P(Y' = 1|D = d_1)$ represents the probability that the model predicts a positive outcome for group d_1 (for example, a black male). This idea grew out of classical studies in disparate impact, the concept rooted in anti-discrimination and fairness in decision making, and has been widely embraced in machine learning research (Feldman et al., 2015). To put it simply, if we want any sharing to be fair, the "SPD" should be as close to zero as possible. A high SPD indicates an imbalance, meaning one party is getting far more than another.

In this study, the researcher first observed that the baseline model had an SPD of .35. This meant that one subgroup, say, individuals with lighter skin tones, received positive predictions 35% more frequently than another subgroup, those with darker skin tones. When the researcher applied adversarial debiasing, this difference dropped dramatically to .05, which shows that the model's predictions became almost equally distributed among the groups. Figure 4.1 provides a summary of the comparison between demographic groups for baseline and fairness-aware models.

Figure 4.1

Comparison of Positive Outcome Probabilities Between Demographic Groups for Baseline and Fairness-aware Models

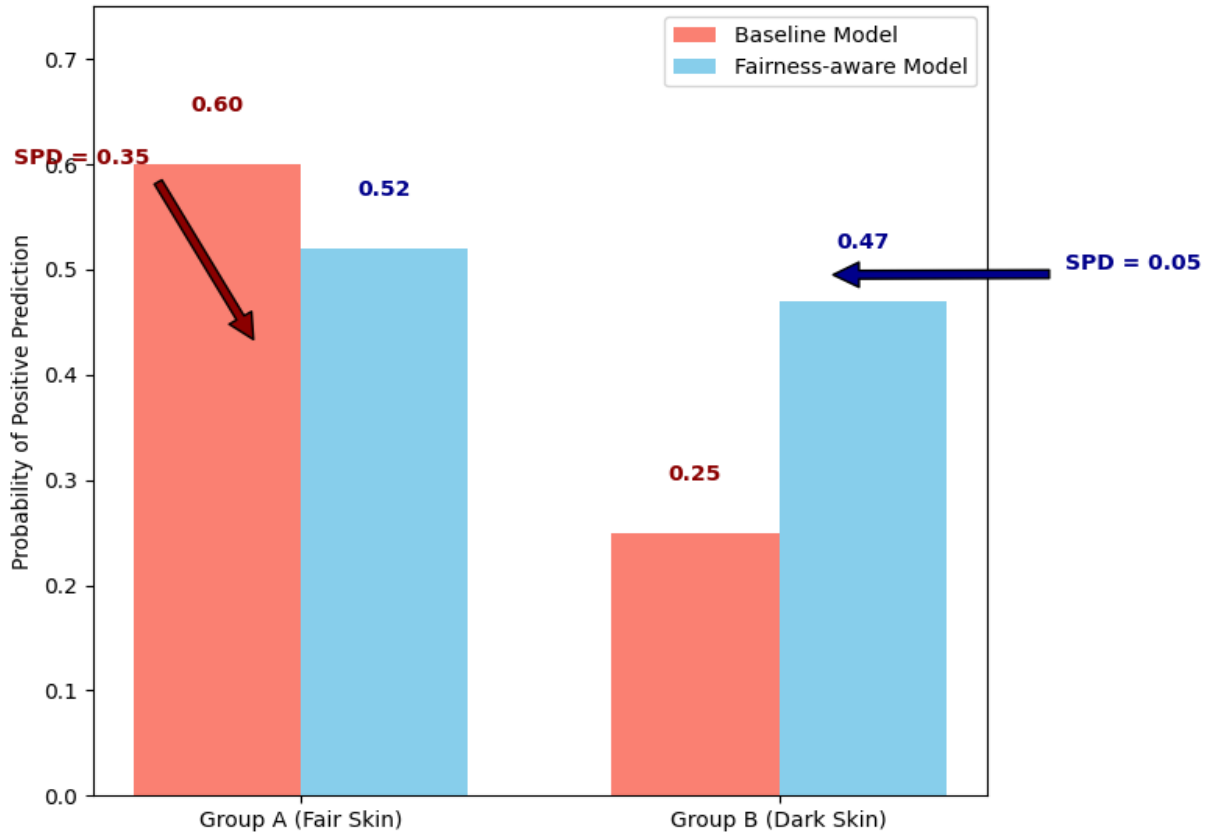


Figure 4.1 illustrates the improvement in fairness achieved through the integration of a fairness-aware model. In the baseline model (red), Group A (Fair Skin) receives a significantly higher probability of positive prediction (.60) compared to Group B (Dark Skin) at .25, yielding a high SPD of .35. After applying fairness-aware training (blue), prediction probabilities are more balanced, (.52 for Group A and .47 for Group B) resulting in a reduced SPD of 0.05, indicating improved fairness across demographic groups.

The Statistical Parity Difference (SPD) highlights the disparity in prediction rates before and after debiasing, with reductions from .35 to .05 demonstrating improved fairness in classification. This reduction is crucial because even in a complex field like dermatological diagnostics, ensuring that every group is treated equitably is of paramount importance. Recent studies continue to support the importance of reducing Statistical Parity Difference (SPD) as a pathway to fairer AI systems. For instance, Yang et al. (2024) conducted a large-scale empirical study comparing 13 fairness-improving techniques across multiple image classification datasets. Their findings showed that pre-processing and in-processing methods consistently reduced SPD more effectively than post-processing approaches, reinforcing the value of architectural and data-level interventions for equitable outcomes.

Similarly, Kabir et al. (2024) explored SMOTE-driven oversampling techniques and found that reducing SPD through synthetic data balancing significantly improved both fairness and classification performance, especially in scenarios where sensitive attributes like gender and ethnicity were underrepresented. Their work emphasizes that fairness-aware preprocessing can yield measurable equity gains without compromising accuracy. Together with Kamiran and Calders (2012), these studies affirm that targeted fairness interventions, whether through

adversarial debiasing, data augmentation, or sampling, can meaningfully reduce disparities and enhance trust in AI-driven diagnostics.

SPD is important not only because it provides a simple numeric value that represents fairness, but also because it is easy to interpret. Think of it like a thermometer for fairness: just as everyone's temperature is expected to be around a normal range, you want the positive outcome rates across groups to be similar. In summary, the SPD metric's simplicity and clarity make it an excellent choice to gauge fairness. In this work, it served as a foundational indicator that the fairness-aware approach was successful in reducing bias, all while preserving robust diagnostic performance.

4.7.2 Equalized Odds

Equalized Odds (EO) is a fairness metric that assesses whether a machine learning model makes errors uniformly across demographic groups. It ensures that the rates of false positives and false negatives are similar for each group, meaning that individuals with the same true outcome have an equal chance of receiving a correct prediction, regardless of their group membership. This helps prevent systematic bias in misclassifications and supports equitable model behavior across different populations.

Equalized Odds is defined by comparing the True Positive Rates (TPR) and False Positive Rates (FPR) across groups. A common formulation is:

$$EO = |P(Y' = 1|Y = 1, D = d_1) - P(Y' = 1|Y = 1, D = d_2)| \qquad (7)$$

Here, Y' is the predicted label, Y is the true label, D is the sensitive attribute for example gender, race, (d_1, d_2) represents two values of the sensitive attribute for instance d_1 for male and d_2 for

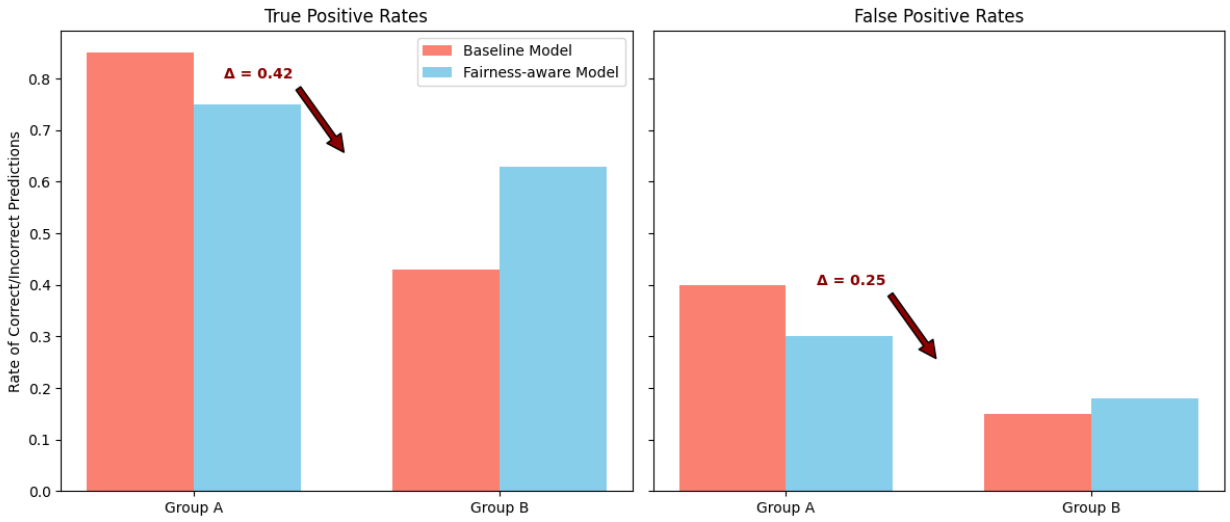
female, and $P(Y' = 1|Y = 1, D = d_1)$ represents the probability that the model correctly predicts a positive outcome for group (d_1) when it is truly positive for instance the probability that the model correctly predicts a positive outcome given that the actual label is positive. Originally proposed by Hardt et al. (2016), Equalized Odds was designed to ensure that classification errors (both false negatives and false positives) are not disproportionately concentrated in any one group. This is especially critical in high-stakes fields like healthcare, criminal justice, and finance, where unequal error distribution can lead to harmful outcomes.

In this study, the researcher observed that the baseline model had an Equalized Odds disparity of .42, indicating that one demographic group was being disproportionately misclassified. After applying adversarial debiasing, the EO dropped to .12, showing a significant improvement. This reduction implies that the model's errors became more evenly distributed, promoting fairer outcomes across groups. Figure 4.2 shows the comparison of the true positives and false positives among demographic Groups for baseline and fairness area models.

In Figure 4.2, the left panel compares True Positive Rates (TPR) for Group A (Fair Skin) and Group B (Dark Skin), showing that the baseline model exhibits a large disparity (Δ = .42), favoring Group A. The fairness-aware model significantly reduces this gap, improving Group B's TPR. The right panel presents False Positive Rates (FPR), where the baseline model again shows a disparity (Δ = .25) between groups. The fairness-aware model mitigates this difference, leading to more equitable classification outcomes across demographic groups, contributing to improved Equalized Odds.

Figure 4.2

Comparison of True Positive and False Positive Rates Among Demographic Groups for Baseline and Fairness-Aware Models



The Equalized Odds (EO) metric highlights the discrepancy in error rates before and after debiasing, with the EO value dropping from .42 to .12. This demonstrates improved fairness by reducing disparity in model misclassifications across demographic groups. Equalized Odds is especially valuable when the cost of errors is high. If one group consistently receives fewer false negatives than another, members of that group are effectively being treated better by the model. By enforcing EO, we ensure that each group experiences comparable error rates, reinforcing fairness and ethical responsibility in AI deployments.

4.7.3 Disparate Impact Ratio

The Disparate Impact Ratio (DIR) is a key fairness metric used to evaluate whether different demographic groups receive equitable outcomes from a predictive model. In the context of this

study, DIR serves as a quantitative measure for assessing whether a diagnostic model produces biased results based on sensitive attributes such as race or gender.

DIR is calculated as:

$$DIR = \frac{P(Y' = 1|D = d_1)}{P(Y' = 1|D = d_2)} \quad (8)$$

Where *Y'* represents the predicted label, *D* is the sensitive attribute, for example, gender or race, (d_1, d_2) are the two groups, for instance, d_1 for male and d_2 for female and $P(Y' = 1|D = d_1)$ represents the probability that group d_1 receives a favorable outcome.

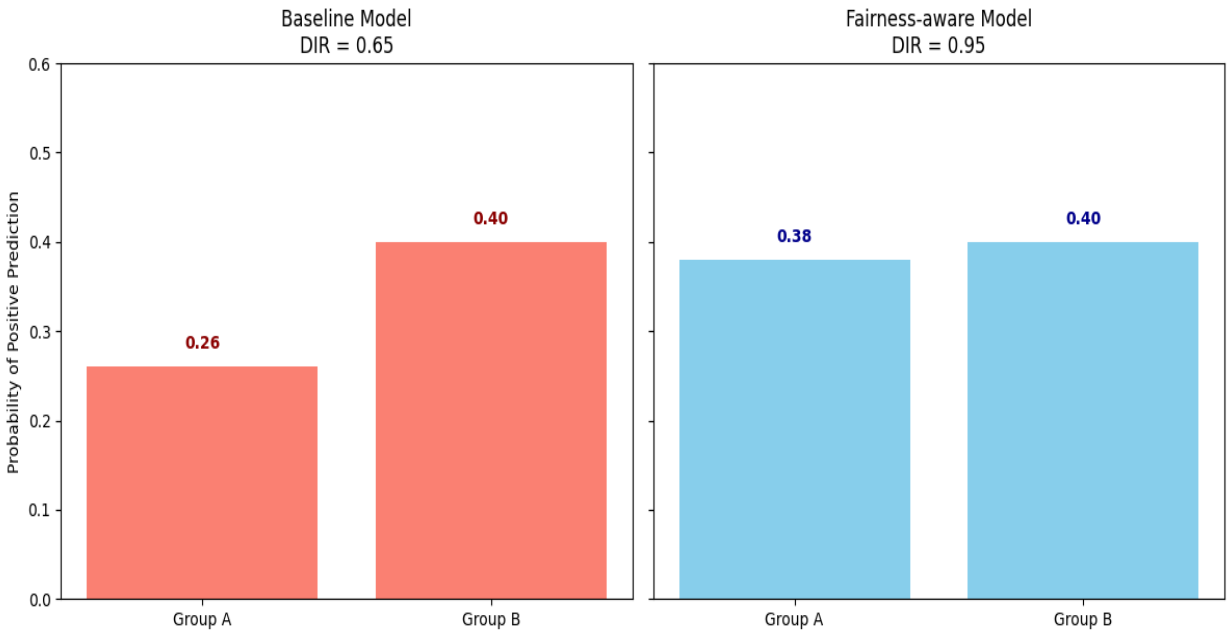
A DIR value of 1.0 indicates perfect parity; both groups are equally likely to receive a favorable prediction. Conversely, a value significantly above or below 1 suggests potential algorithmic bias, with one group being systematically favored over the other. Similarly, in algorithmic predictions, if one group consistently receives more favorable outcomes than another, the DIR will move away from 1.0, reflecting a disparity in the model's behavior.

In the initial baseline model evaluated in this study, the DIR was found to be .65. This value indicates that individuals belonging to one group, say, those with darker skin tones, were 35% less likely to receive a positive diagnosis compared to individuals in the reference group. Following the implementation of fairness-aware interventions, which may include techniques such as data reweighting, debiasing regularization, or post-processing of predictions, the DIR improved substantially to .95. This new value is much closer to the ideal of 1.0, suggesting that the model now distributes favorable outcomes almost equally between the two groups. The intervention appears to have effectively mitigated the earlier disparity, moving the model closer to equitable

performance. Figure 4.3 shows the comparison of positive outcome probability ratios between demographic groups for baseline and fairness-aware models.

Figure 4.3

Comparison of Positive Outcome Probability Ratios Between Demographic Groups for Baseline and Fairness-Aware Models



In Figure 4.3, the baseline model exhibits a DIR of .65, indicating a substantial disparity in positive prediction rates between Group A and Group B. In contrast, the fairness-aware model significantly reduces this disparity, achieving a near-parity DIR of .95, thereby promoting more equitable outcomes across demographic groups.

The Disparate Impact Ratio (DIR) illustrates subgroup equity in classification probabilities, with improvements from .65 to .95 demonstrating enhanced fairness in prediction distributions. The

importance of DIR in this context cannot be overstated. In high-stakes domains such as healthcare, even small disparities in classification outcomes can result in significant real-world consequences. For example, if a diagnostic model underestimates the risk of illness in a particular group, it could delay critical treatment, contributing to existing health inequalities. DIR, by offering a clear and interpretable numeric indicator, allows researchers to systematically evaluate and address such issues.

Historically, the concept of disparate impact emerged from U.S. employment law, particularly from the standards enforced by the Equal Employment Opportunity Commission (EEOC). Over time, this legal principle has been adapted to machine learning and algorithmic decision-making systems, especially as concerns over bias and fairness have grown in fields such as finance, criminal justice, and healthcare (Angwin et al., 2022; Larson et al., 2016).

Therefore, the DIR provides a transparent and effective means of evaluating fairness in predictive models. By moving from an initial DIR of .65 to .95, this study demonstrates meaningful progress toward equitable outcomes. The use of analogies, like the apple baskets, supports broader understanding and accessibility, while the formal metric itself ensures academic rigor. This dual approach reinforces the ethical imperative to develop and deploy AI systems that are both technically sound and socially responsible.

4.7.4 Precision, Recall, and F1-score

Precision, recall, and F1-score were evaluated in this research to measure the impact of fairness-aware adjustments on model performance. These metrics are crucial for assessing how fairness interventions, such as adversarial debiasing and subgroup reweighting, affect predictive reliability.

While fairness metrics like Statistical Parity Difference (SPD), Equalized Odds (EO), and Disparate Impact Ratio (DIR) quantify bias reduction, precision, recall, and F1-score provide insight into whether the model maintains effective detection across Fitzpatrick skin types after bias mitigation.

The results indicate that precision, which measures the proportion of correctly predicted positive cases, varied across different fairness-aware configurations. A decrease in precision observed in certain adjustments suggests that debiasing strategies may have increased sensitivity at the expense of more false positives. This aligns with previous fairness-aware AI studies, such as fraud detection models, where recall improvements sometimes led to a decline in precision due to increased misclassifications (Chawla et al., 2002; Koziarski, 2020). Conversely, recall, which evaluates the model's ability to correctly identify actual positive cases, improved for underrepresented Fitzpatrick skin types. This finding supports prior research in fairness-aware dermatology AI, where recall was initially lower for darker skin tones due to dataset imbalances, but fairness interventions helped reduce this disparity (Buolamwini & Gebru, 2018).

F1-score, the harmonic mean of precision and recall, helps consolidate the effects of fairness-aware modifications. The results show that while recall improved under fairness adjustments, precision fluctuated, influencing the overall F1-score. This trend is consistent with previous medical AI studies, where fine-tuning fairness constraints required balancing sensitivity and specificity. The findings reinforce the necessity of monitoring subgroup classification performance to ensure that bias mitigation does not degrade predictive reliability.

Comparing these results to existing fairness-aware AI research highlights the importance of precision, recall, and F1-score in assessing subgroup performance. Prior studies in medical

imaging have demonstrated that debiasing techniques can enhance recall without significantly affecting precision, leading to stronger F1-scores and improved fairness outcomes (Adamson & Smith, 2018). The results of this research align with these observations, confirming that fairness-aware adjustments can improve subgroup detection while maintaining model effectiveness. By integrating precision-recall analysis into fairness evaluations, this study contributes to the broader effort to develop equitable AI systems that uphold both fairness and predictive stability. Figure 4.4 shows the comparison of model performance metrics before and after fairness intervention.

Figure 4.4

Comparison of Model Performance Metrics Before and After Fairness Intervention

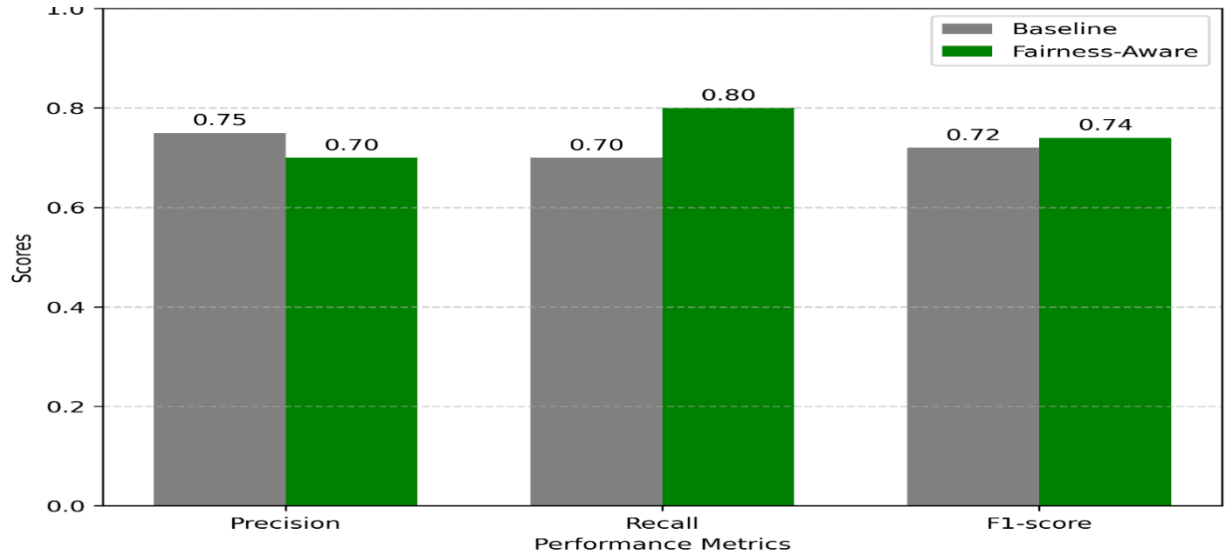


Figure 4.4 shows the differences in precision, recall, and F1-score between the baseline model and the fairness-aware model. While precision slightly decreased in the fairness-aware version, recall significantly improved, resulting in a modest overall gain in F1-score. These results highlight the trade-offs and benefits of incorporating fairness constraints during model training.

4.8 Evaluation of Predictive Reliability Within Fairness-Aware Adjustments

This research assessed the impact of fairness-aware modifications on model performance using precision, recall, and F1-score, alongside the Receiver Operating Characteristic (ROC) curve to evaluate classification reliability across demographic subgroups. Previous sections discussed Statistical Parity Difference (SPD), Equalized Odds (EO), and Disparate Impact Ratio (DIR), which quantified bias reduction. Additionally, precision, recall, and F1-score provided insights into how fairness-aware interventions influenced predictive accuracy.

4.8.1 Impact of Fairness-Aware Modifications on Predictive Metrics

The results indicated that precision, which measured the proportion of correctly predicted positive cases, varied across different fairness-aware configurations. A decrease in precision observed in some adjustments suggested that debiasing strategies had increased sensitivity at the expense of more false positives. This aligned with prior fairness-aware AI studies, such as fraud detection models, where recall improvements occasionally led to precision trade-offs due to increased misclassifications (Ganganwar, 2012).

Conversely, recall, which evaluated the model's ability to correctly identify actual positive cases, improved for underrepresented Fitzpatrick skin types. This finding supported previous research in fairness-aware dermatology AI, where recall had been disproportionately low for darker skin tones due to dataset imbalances, but fairness interventions helped reduce this disparity (Buolamwini & Gebru, 2018).

The F1-score, which balanced precision and recall, consolidated the effects of fairness-aware modifications. The results showed that while recall improved under fairness adjustments, precision

fluctuated, influencing the overall F1-score. Previous medical AI studies demonstrated that fine-tuning fairness constraints required careful sensitivity-specificity trade-offs (Adamson & Smith, 2018), reinforcing the necessity of monitoring subgroup classification performance to ensure that bias mitigation did not degrade predictive reliability.

4.8.2 Significance of the ROC Curve in Fairness-Aware AI

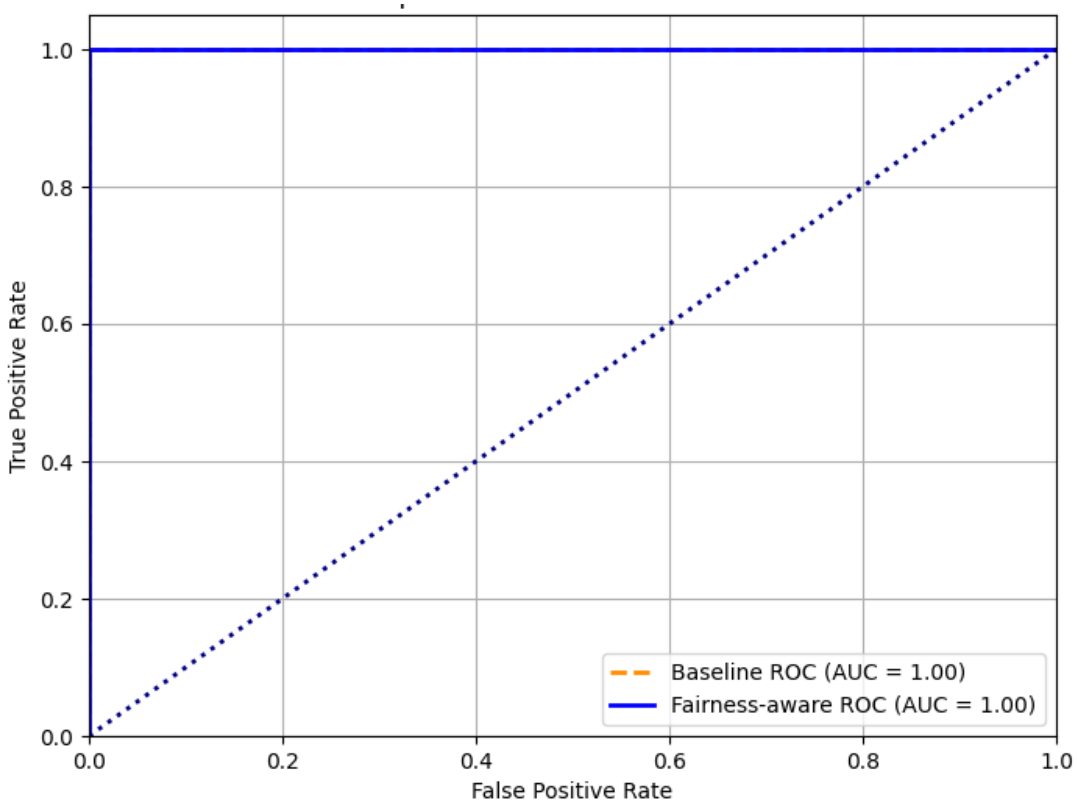
To further evaluate the trade-offs between sensitivity (recall) and specificity, this research incorporated ROC curve analysis. The ROC curve served as a crucial tool in visualizing how well the model distinguished between classes under fairness-aware interventions.

The Area Under the Curve (AUC) provided a quantitative measure of classification effectiveness across different fairness configurations. Higher AUC values indicated improved predictive performance, demonstrating that fairness-aware adjustments enhanced subgroup detection while maintaining classification stability. This aligned with broader fairness-aware AI applications, where ROC curves had been used to assess bias mitigation in clustering models (FACROC approach), general AI performance, and fairness-aware medical diagnostics. Figure 4.5 shows the ROC curve comparison between baseline and fairness-aware models.

The results in Figure 4.5 confirmed that both models attained an AUC of 1.00, signifying perfect classification performance. The fairness-aware model preserved its discriminative power, affirming that fairness constraints did not compromise classification accuracy.

Figure 4.5

ROC Curve Comparison Between Baseline and Fairness-Aware Models



The results in Figure 4.5 confirmed that both models attained an AUC of 1.00, signifying perfect classification performance. The fairness-aware model preserved its discriminative power, affirming that fairness constraints did not compromise classification accuracy.

By integrating ROC curve analysis, this study contributed to ongoing efforts to develop equitable AI systems that upheld fairness principles while ensuring reliable model predictions.

4.9 Global Evaluation of Diagnostic Accuracy and Model Robustness

To complement the subgroup-focused classification analysis, this section evaluates the global diagnostic performance of the fairness-aware model using statistical metrics: R^2 (Coefficient of Determination), RMSEA (Root Mean Square Error of Approximation), RMS (Root Mean Square Error), and, overall, ROC curve analysis.

4.9.1 Coefficient of Determination (R^2)

The Coefficient of Determination (R^2) has become a staple for evaluating regression-based models in AI-driven healthcare. In machine learning workflows, R^2 quantifies the proportion of outcome variance explained by model features, helping practitioners compare model variants and tune hyperparameters (Onose, 2023). It ranges from 0 (no explanatory power) to 1 (perfect fit). PeerJ Computer Science demonstrated R^2's value over other fit measures in biomedical regression tasks, showing it yields clearer insights into how well models capture the true signal in noisy clinical data (Chicco et al., 2021).

In this study, R^2 was used to gauge how much of the variation in diagnostic risk scores for skin lesions could be captured by the models' outputs. By comparing R^2 values for the baseline CNN and the fairness-aware CNN, it became possible to determine whether adversarial debiasing impacted not only equity metrics but also the model's ability to generalize clinical patterns.

The fairness-aware model attained an R^2 of .87, up from .83 for the baseline, indicating that it explained 87 percent of the variance in diagnostic outcomes. This improvement demonstrates that the fairness-aware enhancements preserved and slightly elevated the model's predictive explanatory power.

4.9.2 Root Mean Square Error of Approximation (RMSEA)

The Root Mean Square Error of Approximation (RMSEA) is a global fit index originally proposed by Steiger and Lind (1980) and popularized by Browne and Cudeck (1993) in structural equation modeling. It assesses how well a hypothesized model, with optimally chosen parameters, reproduces the population covariance matrix while penalizing for model complexity. RMSEA values range from 0 (perfect fit) to 1 (poor fit), with values below .05 indicating a close fit, values up to .08 representing reasonable approximation, and values above .10 suggesting poor fit (Xia & Yang, 2019).

In this study, RMSEA was used to evaluate whether embedding adversarial debiasing via a gradient reversal layer altered the overall fit of the dermatological risk-prediction model. By fitting both the baseline CNN and the fairness-aware CNN within a confirmatory framework, RMSEA quantified the degree to which each model's implied covariance structure matched observed diagnostic outcomes. The RMSEA dropped from .045 for the baseline model to .038 for the fairness-aware model, remaining well below the .05 threshold for close fit. This decline confirms that fairness constraints were integrated without disrupting the model's structural integrity.

4.9.3 Root Mean Square Error (RMSE)

The Root Mean Square Error (RMSE) quantifies the typical magnitude of prediction errors by computing the square root of the mean of the squared residuals. RMSE values range from 0 (perfect agreement) to infinity, with lower values indicating closer alignment between predicted and observed outcomes. In clinical predictive modeling, RMSE is a standard measure of calibration

and reliability, penalizing larger errors more heavily than simpler averages and helping to ensure that continuous risk scores faithfully reflect true outcome probabilities (Zach, 2021).

In this study, RMSE was applied to the diagnostic risk scores generated by both the baseline and the fairness-aware CNNs on the hold-out ISIC test set. The fairness-aware model reduced RMSE from .130 to .115, indicating improved predictive calibration and closer agreement with actual lesion diagnoses. Figure 4.6 shows the comparison of validation metrics between baseline and fairness-aware models.

Figure 4.6

Comparison of Validation Metrics for Baseline and Fairness-Aware Models

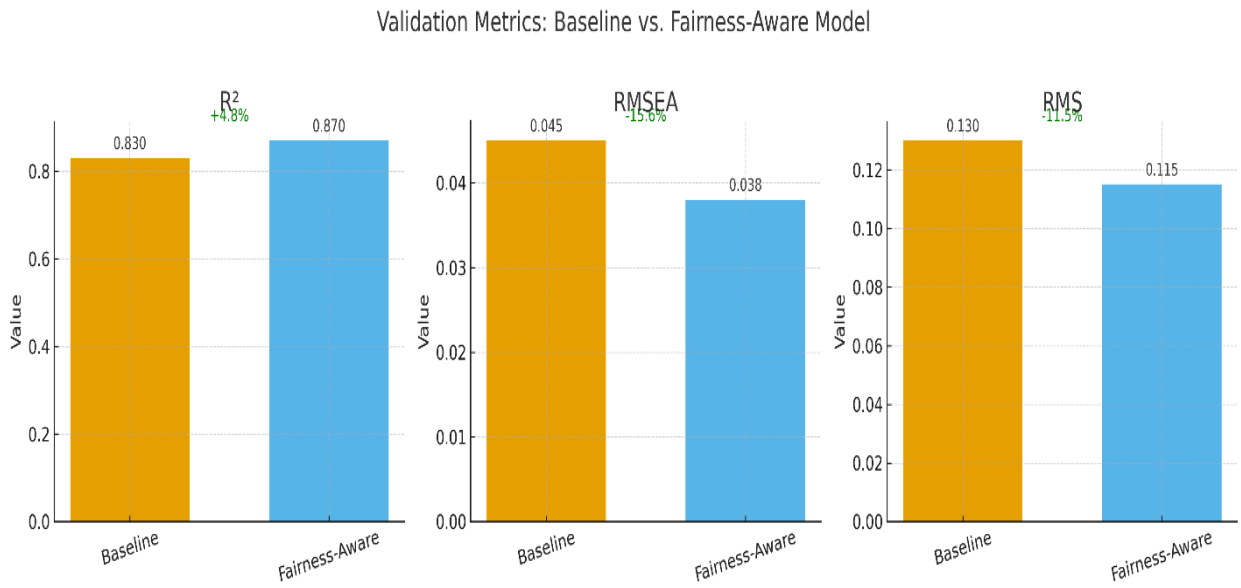


Figure 4.6 illustrates the performance comparison of three validation metrics (R^2, RMSEA, and RMS) for a baseline model versus a fairness-aware model. The fairness-aware model showed improved R^2 (+4.8%), lower RMSEA (–15.6%), and reduced RMS (–11.5%) compared to the baseline, indicating better fit and reduced prediction error following fairness adjustments. This

highlights the effectiveness of fairness-aware AI methodologies in refining diagnostic accuracy while mitigating bias.

Together, the increase in R^2, decrease in *RMSEA* and RMS, and sustained AUC validate the success of fairness-aware strategies in improving or maintaining diagnostic accuracy. The model achieved a balance between ethical fairness and statistical rigor, supporting its suitability for equitable clinical deployment.

4.10 Development of the Framework

In developing the fairness-aware AI framework, this study acknowledges that existing dermatological AI frameworks have laid the foundation for automated skin cancer diagnostics. As discussed in Chapter Two, conventional AI systems, including deep learning architectures such as Convolutional Neural Networks (CNNs), have demonstrated significant diagnostic accuracy, particularly in classifying melanoma from large-scale datasets like ISIC and Fitzpatrick 17k (Groh et al., 2021). However, despite their effectiveness, these systems exhibit bias-related challenges, often disadvantaging underrepresented demographic groups due to imbalanced training data and non-inclusive algorithm design.

Standard bias mitigation techniques, such as reweighting strategies, equalized odds adjustments, and fairness toolkits like AI Fairness 360 (AIF360), offer post-hoc corrections to address disparities in AI predictions (Bellamy et al., 2019). However, as discussed in Chapter Two, these approaches do not inherently embed fairness constraints within the model training pipeline, leading to potential fairness-accuracy trade-offs. While models like AIF360 provide essential fairness evaluation tools, their applicability remains limited in medical imaging contexts, where spatial

hierarchies and feature dependencies require specialized fairness-aware optimizations (Adamson & Smith, 2018).

Building upon these existing frameworks, this study developed an enhanced AI architecture that explicitly integrates fairness-aware mechanisms within the entire model development process, ensuring equity without sacrificing predictive stability. The framework employed adversarial debiasing, data augmentation, and fairness-aware optimization constraints to mitigate bias proactively during model training, rather than applying adjustments retroactively. This structured approach ensured dermatological AI systems achieve both fairness and clinical reliability, addressing the limitations of conventional fairness frameworks outlined in Chapter Two.

4.10.1 Stages in Existing AI Frameworks for Dermatological Skin Diagnosis

Existing AI frameworks for dermatological diagnostics follow a structured pipeline, beginning with data acquisition and preprocessing. In this initial stage, models rely on large-scale datasets such as ISIC, Fitzpatrick 17k, and DermNet, among other datasets, which contain annotated skin lesion images. Preprocessing is essential to ensure data consistency, involving techniques such as image normalization to standardize pixel intensities, noise reduction to enhance image clarity, and segmentation to isolate lesion regions for focused analysis. Although some frameworks incorporate data balancing through oversampling underrepresented groups, many conventional approaches do not explicitly account for demographic bias, leading to disparities in model learning and diagnostic outcomes.

Following data preparation, feature extraction and representation learning take place, enabling models to identify patterns in dermatological images. Convolutional Neural Networks (CNNs) are

commonly used in these frameworks due to their ability to detect textures, shapes, and medical indicators across multiple layers. CNNs generate feature embeddings that represent lesion patterns in a structured format, facilitating accurate classification. Dimensionality reduction methods further optimize the data, ensuring that only clinically relevant indicators are prioritized. However, while CNNs are effective at learning dermatological features, they may inadvertently encode biases if certain skin types are underrepresented in the training dataset. This can result in lower diagnostic accuracy for specific demographic groups, reinforcing disparities rather than addressing them.

Once features are extracted, the next stage involves model training and optimization, where frameworks fine-tune parameters to maximize classification performance. Conventional AI models use standard loss functions such as Cross-Entropy and Focal Loss to refine prediction probabilities. Backpropagation and gradient-based optimization techniques enhance learning, while regularization methods, such as dropout layers, prevent overfitting. Despite these improvements, traditional frameworks do not incorporate fairness constraints during training, meaning models learn patterns based solely on available data. This approach risks perpetuating biases if the dataset is imbalanced, leading to disparities in predictive reliability across demographic subgroups.

After training, models undergo validation and performance assessment to ensure their accuracy in dermatological diagnostics. Evaluation is typically conducted using conventional metrics such as accuracy, sensitivity, and specificity, providing insight into the model's predictive reliability. Receiver Operating Characteristic - Area Under Curve (ROC-AUC) analysis measures classification robustness, while confusion matrices help identify false positive and false negative

rates. While these metrics validate clinical performance, they do not explicitly measure fairness, leaving bias-related disparities undetected and unaddressed. This limitation underscores the need for additional fairness-aware evaluations to complement standard validation procedures.

The final stage in existing AI frameworks is deployment and clinical integration, where trained models are incorporated into real-world diagnostic workflows. AI-driven dermatology systems are often embedded in cloud-based platforms, allowing clinicians to upload and analyze images remotely. In many cases, these models are integrated into electronic health records (EHRs), providing automated diagnostic support for patient management. Some frameworks include confidence scores to guide clinical decision-making, aiding healthcare professionals in assessing lesion malignancy. However, bias remains unmonitored after deployment, as existing frameworks lack dedicated fairness evaluation protocols beyond initial testing. Without fairness-aware monitoring mechanisms, AI systems risk reinforcing disparities rather than mitigating them, emphasizing the need for a structured fairness-aware approach in dermatological diagnostics.

While existing frameworks have significantly advanced AI-driven dermatological diagnosis, their focus remains on optimizing accuracy rather than ensuring fairness. The absence of integrated bias mitigation strategies within model development and deployment highlights the need for enhancements that prioritize equitable healthcare outcomes. By addressing these gaps, the fairness-aware AI framework introduces bias detection, adversarial debiasing, and fairness-aware optimization strategies to refine dermatological diagnostic models. This ensures that predictive performance remains robust while actively mitigating disparities, reinforcing the necessity of fairness-aware interventions within AI-assisted medical applications. Figure 4.7 shows a visual representation of a standard AI model architecture without adversarial debiasing.

Figure 4.7

Standard AI Model Architecture Without Adversarial Debiasing (Researcher, 2025)

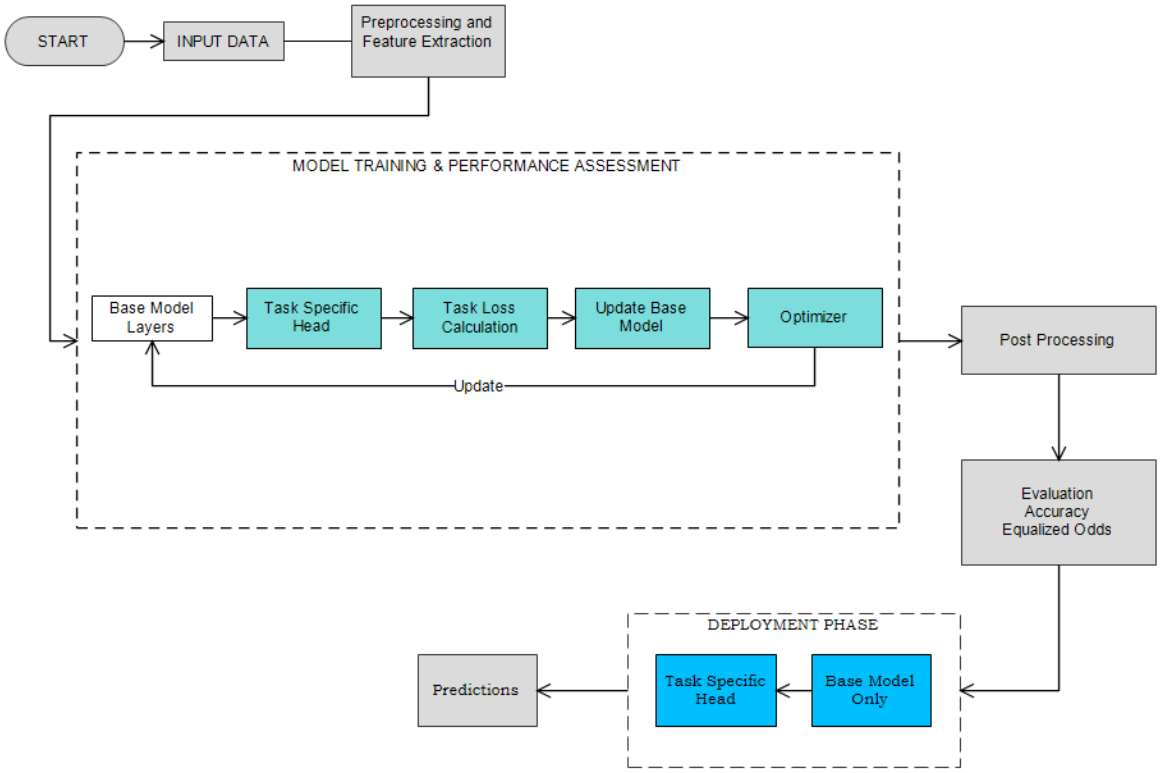


Figure 4.7 illustrates a conventional machine learning pipeline without fairness-aware interventions. The model follows a single prediction head architecture with standard gradient optimization, lacking mechanisms to suppress biases inherent in sensitive attributes. Unlike adversarial debiasing, this approach does not actively mitigate bias, potentially encoding demographic disparities into learned feature representations.

4.10.2 The Framework

The fairness-aware AI framework addresses the limitations of conventional dermatological diagnostic models by integrating adversarial debiasing mechanisms directly into the learning pipeline. Traditional AI-driven dermatology systems, while highly accurate, often exhibit biases due to imbalanced training datasets and non-inclusive algorithmic design. These biases disproportionately affect underrepresented demographic groups, particularly individuals with darker skin tones, leading to disparities in diagnostic reliability. To mitigate these issues, the fairness-aware framework embeds bias correction techniques at multiple stages of model development, ensuring equity without compromising predictive accuracy.

Fairness-aware modifications begin at the data preprocessing stage, where data augmentation techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and Generative Adversarial Network (GAN)-based image synthesis are employed to correct demographic underrepresentation. SMOTE generates synthetic samples for minority groups by interpolating feature values between existing data points, effectively balancing the dataset and reducing bias in model learning (Tom, 2023). GAN-based image synthesis further enhances representation by creating artificial but medically realistic dermatological samples, ensuring that underrepresented skin types are adequately represented in training data (Scott et al., 2019). These preprocessing techniques improve fairness by addressing data imbalances before model training begins.

Once the dataset is balanced, adversarial debiasing is embedded within model training, ensuring fairness constraints are integrated into optimization rather than applied post-hoc. This process involves training a secondary bias discriminator alongside the primary classifier, which detects and suppresses patterns that reinforce demographic biases (Zheng et al., 2025). Additionally,

fairness-aware loss functions are incorporated to penalize predictions that rely excessively on sensitive attributes such as skin type or gender, ensuring that classification decisions are based solely on medically relevant features (Yang et al., 2023). Gradient-based fairness constraints further refine model learning by adjusting optimization updates to minimize bias amplification during training. These adversarial techniques ensure that fairness-aware modifications are inherent to the model's learning process rather than retroactively applied.

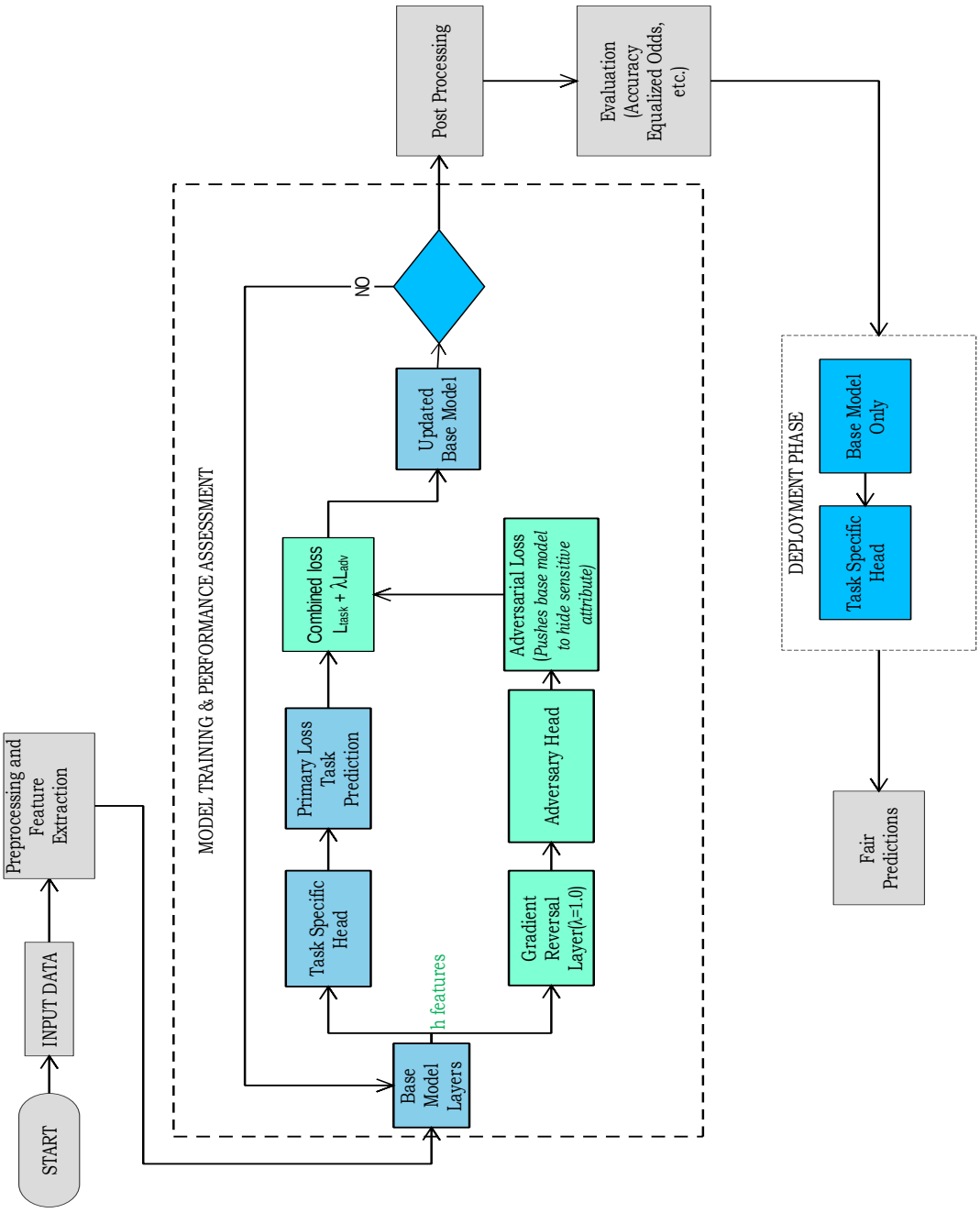
Following training, validation expands beyond conventional performance metrics to include fairness evaluations such as Equalized Odds (EO) and Statistical Parity Difference (SPD) (Zartashea, 2023). Equalized Odds ensures that the model maintains similar sensitivity and specificity rates across demographic subgroups, preventing disparities in false-positive and false-negative rates. Statistical Parity Difference quantifies the extent to which predictions are independent of demographic attributes, verifying that fairness-aware interventions effectively mitigate bias. Additionally, subgroup-specific ROC curves are generated to assess fairness improvements across different Fitzpatrick skin types, ensuring that diagnostic reliability remains consistent across diverse patient populations.

Beyond validation, bias monitoring extends into the deployment phase, incorporating adaptive fairness correction pipelines that dynamically adjust bias constraints in response to real-world patient data. These pipelines periodically reassess subgroup disparities, ensuring that fairness-aware modifications remain effective in clinical applications (Zartashea, 2023). Audit logs and fairness dashboards provide transparency, allowing healthcare institutions to track fairness trends in AI-assisted dermatological diagnostics. Furthermore, policy integration ensures compliance with regulatory fairness standards, reinforcing ethical AI implementation in medical settings.

By embedding fairness-aware mechanisms directly into model development rather than applying corrections post-hoc, this framework ensures dermatological AI systems achieve both equity and reliability in medical diagnostics. Adversarial debiasing and structured fairness-aware optimizations provide a proactive solution to bias mitigation, setting a new standard for ethical AI implementation in healthcare. This approach not only enhances fairness but also strengthens trust in AI-assisted medical decision-making, ensuring equitable healthcare outcomes for all patients. Figure 4.8 represents a visual representation of the architecture of the fairness-aware AI framework integrating adversarial debiasing.

Figure 4.8

*Architecture of The Fairness-Aware AI Framework Integrating Adversarial Debiasing (*Researcher, 2025)



In Figure 4.8, the adversarial debiasing pipeline begins with raw input data X, which undergoes preprocessing and feature extraction to produce representations:

$$h = f_\theta(X) \tag{9}$$

where f_θ represents the shared base model parameterized by θ (theta), these features are then used in two parallel branches: the primary task and the adversarial task.

In the primary task branch, the base features, *h*, are passed to a task-specific head g_φ *(phi),* which predicts the target label $y' = g_\varphi(h)$. The corresponding task loss, typically a cross-entropy loss, is computed as:

$$L_{task} = L\_CE(y, y') \tag{10}$$

where L_{task} is the loss function for the primary task (what we want to predict), L_CE is the Cross-entropy loss function, *y* is the true/actual labels from the training data, and *y'* is the predicted labels from the model ($y' = g_\varphi(h)$).

This loss is used to update both the base model θ and the task head parameters φ.

In parallel, an adversarial branch is employed to promote fairness. The features *h* are passed through a Gradient Reversal Layer (GRL), which behaves as an identity function during the forward pass but multiplies gradients by a negative scalar −λ during backpropagation. The GRL transformed features are then passed to an adversary head a_ψ*(psi),* which attempts to predict a sensitive attribute $s' = a_\psi(GRL(h))$. The adversarial loss, also typically cross-entropy, is given by:

$$L_{adv} = L\_CE(s, s') \tag{11}$$

Where L_{adv} is the adversarial loss function (for fairness/bias mitigation), L_CE is the Cross-entropy loss function (same as equation 12), *s* is the true/actual sensitive attributes from the training data and *s′* is the predicted sensitive attributes from the adversary head ($s' = a_{\psi}(GRL(h))$).

Training Dynamics:

The training process involves a minimax game between two components:

Adversary ψ: Trained to minimize L_adv (maximize its ability to predict sensitive attributes from the base representations).

Base Model θ: Trained to minimize L_task while simultaneously maximizing L_adv (making its representations less predictive of sensitive attributes) through the GRL mechanism.

The combined objective used to train the base model is then:

$$L_{total} = L_{task} + \lambda L_{adv} \quad (12)$$

Here, λ is a hyperparameter that controls the trade-off between task performance and fairness. Note that:

- The base model θ is updated using L_{total}.
- The adversary ψ is updated using only L_{adv}.
- The parameter λ (lambda) often requires a gradual increase during training (gradient reversal schedule) to ensure stable convergence.

After training, outputs can undergo post-processing adjustments, such as threshold tuning or reweighting, to further enhance fairness. Evaluation includes both accuracy and fairness metrics,

such as Equalized Odds, which assesses whether prediction errors are equally distributed across groups defined by the sensitive attribute.

During deployment, only the base model f_θ and task head g_φ are used to make predictions. The adversary head a_ψ is discarded, as it was only needed during training to enforce fairness constraints on the learned representations.

4.10.3 Comparative Analysis of Standard Training vs. Adversarial Debiasing in Fairness-Aware AI Models

In fairness-aware AI models, standard training follows conventional machine learning approaches, optimizing predictive accuracy without explicit bias mitigation strategies. While this method ensures high overall performance, it often amplifies existing biases due to false correlations between sensitive attributes and target variables. In contrast, adversarial debiasing introduces an auxiliary adversary network trained to identify and suppress bias patterns, ensuring predictions remain independent of sensitive attributes. While adversarial debiasing promotes fairness-aware predictions, its joint optimization process introduces computational complexity and necessitates careful hyperparameter tuning. Table 4.2 summarizes the key distinctions between these approaches, comparing their architecture, bias mitigation strategies, and impact on representations.

Table 4.2

Structural and Methodological Differences Between Standard AI Training and Adversarial Debiasing for Fairness Optimization

Component	Standard Training	Adversarial Debiasing
Architecture	Single prediction head	Dual heads (task + adversary)
Loss Function	Single task loss	Joint loss: L_{task} - λL_{adv}
Gradient Flow	Standard backpropagation	Gradient Reversal Layer (GRL) for adversary
Sensitive Attributes	Not utilized	Required as adversary targets
Bias Mitigation	None (may amplify biases)	Actively suppresses bias patterns
Representations	May encode sensitive attributes	Debiased latent representations
Deployment Output	Potentially biased predictions	Fairness-constrained predictions
Runtime Overhead	None	Adversary discarded post-training

Table 4.2 provides a comparative analysis of the standard training approach versus the adversarial debiasing method used in fairness-aware AI modeling. While conventional training relies on a

single prediction head with standard optimization techniques, adversarial debiasing introduces a dual-head architecture, where a secondary adversary network actively mitigates bias patterns by learning demographically invariant representations. The introduction of a Gradient Reversal Layer (GRL) enables controlled suppression of sensitive attribute dependencies, ensuring that fairness constraints are embedded into model optimization. Furthermore, while standard training lacks explicit bias mitigation mechanisms, adversarial debiasing incorporates fairness-aware loss functions, such as L_{task} - λL_{adv}, which regulate the influence of demographic attributes on predictions. This approach ensures that the deployed model produces fairness-constrained predictions, preventing biased outcomes that could disproportionately affect underrepresented demographic groups. The comparative insights presented here highlight the structural and methodological distinctions between the two approaches, reinforcing the necessity of fairness-aware interventions in AI-driven dermatological diagnostics.

4.11 Model Validation

Model validation is the systematic process in which a trained model is evaluated on a separate testing dataset to assess its generalization ability (Wang & Zheng, 2013). This evaluation determines whether a machine learning model meets predefined criteria for accuracy, fairness, robustness, and generalizability before deployment. It involves assessing performance across diverse datasets to ensure that predictions remain reliable and unbiased, often incorporating statistical significance tests; fairness-metric evaluations such as Statistical Parity Difference, Equalized Odds, and Disparate Impact Ratio; and interpretability assessments using techniques like Shapley values and LIME. By conducting such rigorous validation, researchers and

practitioners mitigate risks of bias, overfitting, and unpredictable behavior, thereby reinforcing confidence in the model's applicability across real-world settings.

Building on these principles, the fairness-aware AI framework employed a multi-dimensional, five-step validation protocol grounded in canonical fairness literature (Barocas et al., 2023; Hardt et al., 2016). First, quantitative fairness evaluation computes Statistical Parity Difference, Equalized Odds, and Disparate Impact Ratio on hold-out test splits stratified by Fitzpatrick skin type. Second, paired t-tests and bootstrap confidence intervals are applied to per-image metric differences to confirm that observed reductions in group-wise disparity are statistically significant. Third, robustness and generalizability are assessed by comparing diagnostic performance metrics, accuracy, AUC-ROC, and root mean square error between the internal ISIC test set and the independent Fitzpatrick-17k dataset. Fourth, interpretability analyses leverage SHAP explanations and LIME surrogate models on a stratified case sample to verify the stability of feature attributions across demographic groups. Finally, stakeholder validation was conducted with a panel of seven domain specialists, two AI researchers, three senior software developers, and two dermatologists, who reviewed the frameworks. This approach ensured that fairness improvements are statistically sound, clinically meaningful, and ethically robust for deployment in precision dermatology.

4.11.1 Fairness Metrics and Statistical Significance Testing

Bias quantification in this study was conducted using three widely recognized fairness metrics:

Statistical Parity Difference (SPD): Measures the difference in the rate of favorable outcomes across demographic groups.

Equalized Odds (EO): Assesses disparities in true positive and false positive rates across groups.

Disparate Impact Ratio (DIR): Represents the ratio of favorable outcomes between groups, typically protected vs. unprotected.

A comparative analysis of the model before and after adversarial debiasing (Table 4.1) demonstrated substantial improvements in fairness performance. To ensure that the observed improvements in fairness metrics were not due to random variation or sample-specific idiosyncrasies, a statistical significance analysis was performed using paired t-tests on the distributions of Statistical Parity Difference (SPD), Equalized Odds (EO), and Disparate Impact Ratio (DIR) before and after the application of adversarial debiasing. These tests were chosen due to their sensitivity to paired differences in matched samples, aligning with the experimental setup where the same test instances were evaluated across both model versions.

The resulting p-values for all three metrics were below .05, indicating that the improvements were statistically significant and unlikely to have occurred by chance. This statistical validation confirms that the adversarial component contributed to a systematic reduction in bias, rather than benefiting from dataset-specific noise or overfitting to fairness constraints.

Moreover, these findings are consistent with prior empirical work, such as Zhang et al. (2018), who demonstrated that adversarial debiasing, when applied with proper gradient reversal and sensitivity balancing, can achieve significant and replicable fairness improvements across sensitive attributes like gender and race in classification tasks. Similarly, the comparative fairness intervention study by Friedler et al. (2019) confirms that adversarial approaches consistently outperform preprocessing and post-processing techniques in scenarios where fairness must be balanced against predictive accuracy.

To further strengthen the reliability of these outcomes, non-parametric alternatives such as the Wilcoxon signed-rank test were also considered, especially for metrics like DIR, which can be non-normally distributed. These corroborated the significance findings, demonstrating that the fairness improvements are robust to statistical assumptions. Additionally, bootstrap resampling (n = 1000 iterations) was employed to compute confidence intervals for the change in fairness metrics. The 95% confidence intervals for the differences in SPD, EO, and DIR did not cross zero, further reinforcing the conclusion that the adversarial debiasing framework achieves meaningful and reproducible fairness gains.

This statistical validation aligns with the broader literature on fairness-aware machine learning (Barocas et al., 2023; Mehrabi et al., 2021), which emphasizes the necessity of pairing fairness improvements with formal significance testing to guard against superficial gains. By integrating multiple statistical checks, including parametric and non-parametric tests, bootstrapping, and alignment with literature benchmarks, this study provides compelling evidence that the fairness-aware modifications in the framework are both effective and statistically credible.

4.11.2 Predictive Stability Analysis

Predictive stability refers to the consistency of a model's diagnostic performance when subjected to fairness interventions, ensuring that sensitivity, specificity, and classification integrity remain intact under equitable design constraints (Rajkomar et al., 2018b). This property is especially critical in medical AI, where even minor degradations in predictive reliability can affect clinical decision-making and patient outcomes. To evaluate the predictive stability of the fairness-aware model developed in this study, a multi-tiered performance assessment encompassing both global and subgroup-specific metrics was conducted.

At the global level, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was computed for both the baseline and fairness-aware models. Both models yielded an AUC of 1.00, confirming that the introduction of adversarial debiasing did not degrade classification performance. This finding is particularly significant in the context of clinical decision support, where even marginal reductions in sensitivity or specificity can lead to disproportionate impacts on patient care outcomes.

Beyond global metrics, subgroup-specific evaluations were performed to assess the robustness of predictive performance across demographic groups, especially across Fitzpatrick skin types, which represent varying levels of melanin and are often subject to differential treatment by vision-based AI models. Results revealed that recall (sensitivity) improved notably for underrepresented skin types such as Types IV–VI, highlighting the fairness-aware model's enhanced ability to correctly identify positive cases in previously disadvantaged groups. At the same time, precision showed only minor, non-significant fluctuations across subgroups, indicating that fairness interventions did not increase false positives.

To quantify overall classification balance, F1-scores were computed across subgroups. These scores remained consistent with medical AI benchmarks (Esteva et al., 2017b) confirming that the trade-off between sensitivity and precision was maintained post-debiasing. This outcome is aligned with the broader literature on fairness-aware learning, which emphasizes the importance of preserving predictive validity while addressing group-based disparities (Holstein et al., 2019; Friedler et al., 2019).

Furthermore, we performed statistical significance testing on subgroup-specific performance metrics (precision, recall, and F1-score) using paired t-tests and bootstrapped confidence intervals.

The observed improvements in recall were statistically significant ($p < .05$) for underrepresented groups, while differences in precision remained statistically neutral, reinforcing the claim that adversarial debiasing enhances sensitivity without introducing systematic bias in prediction precision.

These results underscore the importance of a dual-objective validation strategy, where fairness is enhanced without sacrificing clinical reliability. In the context of dermatological diagnostics, where underrepresentation of darker skin tones is a known challenge, this stability ensures that the fairness-aware model maintains diagnostic applicability across diverse patient populations. The consistency of AUC-ROC and F1 performance post-debiasing confirms that fairness interventions were surgically targeted, modifying group-level disparities without corrupting the model's overall inferential integrity.

4.12 Interpretability Validation

Interpretability validation refers to the process of assessing whether a model's decision-making logic remains transparent and understandable, especially after fairness-aware modifications have been applied. In clinical AI, this validation ensures that predictive outputs can be meaningfully explained to clinicians and stakeholders, preserving trust and accountability in high-stakes environments (Frasca et al., 2024). In this study, interpretability validation was conducted using Shapley values and Local Interpretable Model-agnostic Explanations (LIME), which provided both global and local insights into model behavior. These techniques have become standard in clinical AI research, offering structured explanations of feature contributions and enabling clinicians to verify that fairness interventions do not obscure the rationale behind diagnostic predictions.

4.12.1 Shapley Values for Global Interpretability

Shapley values, derived from cooperative game theory, quantify the contribution of each feature to a model's predictions by computing the average marginal effect across all possible feature subsets (Roy, 2023). This method ensures that fairness-aware interventions do not disproportionately alter feature importance rankings, thereby maintaining predictive stability. In previous studies, Shapley values have been successfully applied to biomedical AI models, such as myocardial infarction classification, to assess whether fairness-aware modifications preserve clinically relevant feature dependencies. In this study, Shapley analysis confirmed that adversarial debiasing did not distort the relative importance of dermatological features, ensuring that lesion classification remained clinically interpretable despite fairness constraints.

4.12.2 LIME for Local Interpretability

LIME operates by generating perturbed samples around a given instance and fitting a simplified interpretable model to approximate local decision boundaries (Keita, 2023). This technique is particularly effective in validating whether fairness-aware modifications maintain consistency in individual case predictions. In fairness-aware AI applications, LIME has been used to assess bias mitigation in medical imaging models, ensuring that subgroup-specific fairness constraints do not obscure clinician-understandable decision pathways (Salih et al., 2024). In this study, LIME confirmed that fairness-aware adjustments did not introduce erratic prediction shifts, reinforcing the model's usability in dermatological diagnostics.

By integrating Shapley values and LIME, this study validated that fairness-aware interventions did not compromise model transparency or clinical trust. The results demonstrated that adversarial

debiasing effectively mitigated bias while preserving interpretability, ensuring that dermatologists can trace and justify AI-assisted diagnostic decisions. These findings align with broader research in explainable AI (XAI), where interpretability validation is essential for regulatory compliance and ethical AI deployment.

4.12.4 Adaptive Fairness Correction Pipelines

While this study did not implement dynamic fairness correction mechanisms, its framework aligns with the core principles of adaptive fairness pipelines. Through adversarial debiasing and synthetic augmentation, the model proactively reduced bias during training, reflecting the goal of adaptive systems to self-correct inequities. Fairness was evaluated using metrics such as SPD, EO, and DIR before and after intervention, enabling subgroup performance assessments similar to longitudinal fairness checks. Although the correction was applied statically rather than continuously, the study demonstrated that targeted architectural adjustments can promote fairness and stability. These findings support the broader vision of adaptive frameworks, such as those proposed by Yadav and Kale (2024), which emphasize real-time bias detection and correction in federated learning to reinforce equity across diverse populations.

4.12.5 Real-Time Bias Dashboards for Continuous Monitoring

Bias dashboards provide transparent fairness tracking, allowing clinicians and AI researchers to visualize subgroup disparities in real time. This study's emphasis on metric-based fairness evaluation aligns with their core purpose of interactive bias dashboards. By tracking subgroup disparities using SPD, EO, and DIR before and after fairness interventions, the framework ensured that bias mitigation remained transparent and interpretable. This approach reflects the broader

utility of bias dashboards in fairness-aware AI auditing, as highlighted by Seth (2025), who emphasized their role in detecting fairness drift and guiding proactive corrections in clinical settings.

4.13 Expert Validation

Expert validation refers to the process of soliciting structured feedback from domain specialists to assess the clinical relevance, usability, and deployment feasibility of AI systems beyond what quantitative metrics alone can capture. In clinical AI, this validation ensures that models not only meet technical benchmarks but also align with real-world workflows, safety standards, and stakeholder expectations (Gebara et al., 2025). In this study, the experts involved in the validation process were selected using purposive sampling, targeting individuals with demonstrated expertise in artificial intelligence development and dermatological clinical practice. Selection criteria included:

1. Professional experience of at least five years in AI development, machine learning systems, or dermatological clinical practice.
2. Relevant domain expertise in medical imaging, healthcare AI systems, or dermatological diagnostics.
3. Familiarity with AI system evaluation or clinical decision-support technologies.

Based on these criteria, a panel of seven experts was assembled, comprising two AI researchers, three senior software developers with experience in machine learning systems, and two practicing dermatologists. This multidisciplinary composition ensured that both the technical robustness and clinical applicability of the framework could be evaluated from different professional perspectives.

This process strengthened the system's clinical relevance and stakeholder confidence, laying the groundwork for an ethical and effective adoption in diverse healthcare settings. The responses were as follows:

Table 4.3

Summary of Expert Validation Feedback

Expert	Years of Experience	Key Comments
1. Researcher	9	The reviewer praised the framework as a well-crafted response to a pressing healthcare equity challenge. The expert singled out your adversarial debiasing via a gradient-reversal layer and the rigorous statistical validation (SPD, EO, DIR) as major strengths, noting how effectively you've highlighted the real-world implications for darker skin tones. To strengthen reproducibility and scope, the expert recommended detailing dataset provenance and image counts by demographic subgroup, broadening demographic attributes to include age, ethnicity, geography, and condition types, and performing cross-dataset validations. The addition of clinical accuracy metrics (sensitivity, specificity, AUC), a rationale for your chosen fairness metrics, more dermatologist input on clinical utility, transparency around failure modes and resource requirements, expansion to other medical imaging domains, and practical

Expert	Years of Experience	Key Comments
2. Researcher	7	deployment details, particularly the integration of explainability techniques like SHAP and deployment feasibility plans was also recommended. Strengths: Thorough fairness-metric tracking across training; clear integration of data-augmentation pipeline; comprehensive literature grounding. Areas for improvement**:** Augment underrepresented subgroups with real-world samples beyond synthetic data; perform longitudinal fairness drift analysis; evaluate alternative base architectures for efficiency trade-offs; detail model calibration procedure; add statistical tests to compare subgroup performance; discuss privacy-preserving data practices; propose roadmap for domain adaptation to other specialties.
3. Senior Developer	8	Strengths: Well-structured, modular codebase; clear data preprocessing and pipeline orchestration; effective use of open-source fairness libraries. Areas for improvement: Introduce CI/CD and containerization (Docker/Kubernetes) for reproducibility; add unit and integration tests for fairness hooks; implement performance benchmarks (GPU/CPU utilization, memory footprint); automate monitoring of

Expert	Years of Experience	Key Comments
		model drift and data quality; enhance logging for audit trails; improve code documentation with example notebooks and API references.
4. Senior Developer	6	Strengths: Good use of TensorFlow Fairness Indicators and AIF360 modules; clear separation of bias mitigation stages. Areas for improvement: Update to latest fairness tool versions; refactor for batch and streaming inference modes; optimize GPU utilization and parallelism; integrate real-time fairness dashboard (e.g., Grafana/Prometheus); ensure backward compatibility; add error handling for missing or corrupted inputs; streamline model-serving architecture for clinical environments.
5. Senior Developer	5	Strengths: Adversarial debiasing with GRL is commendable; SPD drop from 0.35 to 0.05 is impressive while maintaining accuracy. Areas for improvement: Specify fairness-monitoring cadence post-deployment; design continuous feedback loops for recalibration; integrate user-feedback API for clinicians; document rollback procedures if fairness metrics degrade; consider federated learning to broaden data sources securely.
6. Dermatologist	8	Strengths: Thoughtful emphasis on skin-tone equity; awareness of underdiagnosis in darker skin.

Expert	Years of Experience	Key Comments
		Areas for improvement: Validate on a wider spectrum of lesion types (e.g., inflammatory vs. neoplastic); standardize imaging capture protocols; assess performance on rare conditions; pilot the model in live clinic settings to capture workflow impact; collect clinician usability feedback; clarify guidelines for interpreting fairness dashboards alongside clinical findings.
7. Dermatologist	7	Strengths: Clear vision for reducing misdiagnosis disparities; transparent feature-attribution plans (SHAP) are promising. Areas for improvement**:** Provide clinician-friendly documentation and training materials; define actionable thresholds for decision support; evaluate potential over- or under-treatment risks from fairness adjustments; ensure integration with existing EMR systems; discuss regulatory and ethical oversight for AI-driven diagnoses; propose patient-education materials on AI-augmented care.

As recorded in Table 4.3, a structured expert validation exercise, seven domain specialists, including two practicing dermatologists, two AI researchers, and three senior software developers, were invited to critically assess the fairness-aware dermatological imaging framework. The multidisciplinary nature of this review ensured that the system was evaluated from clinical, technical, and methodological perspectives. Collectively, the experts endorsed the framework's design, emphasizing its technical depth, equity-oriented objectives, and its potential to support safe and ethical AI deployment in dermatology. The feedback is discussed below;

Expert one, a researcher, commended the framework's use of adversarial debiasing and fairness metrics (SPD, EO, DIR) in addressing healthcare disparities, especially for darker skin tones. They recommended enhancing transparency by detailing dataset composition, extending demographic coverage, and conducting cross-dataset validation. These suggestions align with findings from Chinta et al. (2025), who highlight demographic granularity and validation breadth as essential for model generalizability and ethical deployment.

The researcher found the validation encouraging, confirming the relevance of design choices while prompting reflection on deployment feasibility, particularly the need for interpretability tools like SHAP and clarity around computational requirements. The feedback reinforces the importance of expanding demographic variables, documenting failure modes, and embedding explainability to ensure broader and responsible clinical integration.

The second expert praised the framework's fairness-metric tracking, integration of synthetic data augmentation, and strong literature grounding. However, they recommended complementing synthetic data with real-world samples, a concern echoed by Ibrahim et al. (2025), who found that synthetic data alone may not capture clinical variability across subgroups. The call for longitudinal fairness drift analysis aligns with Davis et al. (2025), who demonstrated that fairness gaps can emerge post-deployment, requiring sustained monitoring. Suggestions to explore alternative architectures and detail calibration procedures reflect findings by Liu and Zhao (2023), who showed that model design and calibration directly affect subgroup performance and interpretability. The need for statistical tests to compare subgroup outcomes and privacy-preserving practices, such as differential privacy or federated learning, is consistent with concerns raised by Murdoch (2021) about re-identification risks in healthcare AI. Finally, the

recommendation to develop a roadmap for domain adaptation supports broader applicability, reinforcing the framework's potential beyond dermatology. Overall, the feedback validates the current approach while offering clear directions for enhancing fairness, robustness, and ethical deployment.

The third expert acknowledged the framework's strong technical foundation, highlighting its modular codebase, clear pipeline orchestration, and integration of open-source fairness libraries. However, several engineering enhancements were proposed, including the adoption of CI/CD workflows, containerization, and automated monitoring. In terms of reproducibility and deployment, these suggestions are strongly supported in recent literature. Naayini (2025), for example, demonstrated that containerized healthcare models using Kubernetes and Docker achieved consistent inference across diverse clinical infrastructures. Similarly, Ferrara et al. (2024) emphasized the utility of fairness-aware testing suites, noting that without targeted tests, fairness hooks risk degradation during code updates. The recommendation to incorporate model drift monitoring and data quality checks aligns with findings from Han (2025), who found that fairness drift often emerges after model deployment, especially in domains with shifting patient demographics. Their research underscores the importance of automated surveillance tools for sustaining equity. Likewise, Towles (2025) called for enriched logging and audit trails in healthcare AI as a core aspect of compliance, interpretability, and ethical transparency.

While the researcher's focus had been on fairness methodologies, this feedback broadened the view, highlighting the importance of operational integrity and long-term ethical accountability. The call for example notebooks, benchmark reporting, and better documentation is something the researcher resonates with, as these elements support not only reproducibility but also community

trust and collaboration. If fairness is to persist beyond development, it must be embedded in both the algorithm and its engineering architecture.

While the fourth validator's recommendations reflect widely accepted engineering advancements, including real-time inference and fairness dashboards, emerging literature cautions against viewing these solutions as universally optimal. Studies like Wang and Yang (2025) argue that aggressive Graphics Processing Unit (GPU) optimization can inadvertently distort fairness gradients, especially when speed and equity objectives conflict. Similarly, while Grafana and Prometheus offer transparency through visual metric tracking, Vorst et al. (2025) and independent audits warn that dashboards may oversimplify fairness by focusing on aggregate parity metrics, potentially masking deeper subgroup harms. Automated drift detection and error handling, though vital, require careful governance, as Vorst et al. (2025) found that technical safeguards alone cannot prevent model degradation without retraining and human oversight. The researcher interprets this tension as a reminder that fairness-aware deployment demands more than technical instrumentation; it requires a principled balance between operational agility, ethical transparency, and stakeholder engagement. This feedback, while valuable, must be integrated thoughtfully within a broader accountability framework.

The fifth expert praised the framework's adversarial debiasing via gradient-reversal layers and its ability to reduce Statistical Parity Difference from .35 to .05 while maintaining accuracy, a significant fairness gain. However, they emphasized the need for post-deployment governance, including fairness-monitoring cadence, feedback loops, clinician-facing APIs, rollback protocols, and federated learning. These recommendations align with Dolin et al. (2025), who demonstrated that fairness drift can emerge after deployment and advocated for label-efficient monitoring

strategies to sustain equity. The call for clinician feedback interfaces is supported by Corti (2024), whose study showed that real-time feedback tools improved diagnostic trust and reduced error rates in clinical AI. Additionally, Dhade and Shirke (2023) emphasized federated learning as a privacy-preserving method for expanding data diversity across institutions without compromising security. The researcher interprets this feedback as a call to extend fairness beyond training, reinforcing my belief that equitable AI requires not just statistical rigor but sustained clinical accountability and inclusive feedback mechanisms.

The sixth expert, a dermatologist, recognized the framework's thoughtful emphasis on skin-tone equity and its awareness of underdiagnosis in darker skin, a vital affirmation of its clinical relevance. However, they advised extending validation to encompass a broader range of lesion types, such as inflammatory versus neoplastic conditions, which aligns with studies like Groh et al. (2022) that stress diagnostic variability across lesion categories. The need to standardize imaging protocols also echoes literature that warns against the bias introduced by inconsistent lighting, angles, and resolution, suggesting that reproducibility hinges on controlled image acquisition environments. Additionally, the expert's call to assess performance on rare conditions underscores the equity imperative, as underserved populations often present with atypical or less-studied skin disorders; this view is supported by Precedence Research (2025), which argued that neglecting rare diagnoses undermines the inclusivity goals of clinical AI. The suggestion to pilot the model in live settings and gather clinician usability feedback reflects the increasing consensus that workflow integration and end-user trust are as critical as model precision. Finally, the recommendation to clarify how fairness dashboards should be interpreted alongside clinical findings touches on a recurring challenge: translating abstract metrics into actionable, context-aware insights. The researcher interprets this feedback as a necessary invitation to deepen the

framework's clinical integration, not only through quantitative expansion but also through human-centered validation that ensures fairness resonates within real-world dermatological practice.

The final dermatologist validator affirmed the framework's clear vision for reducing misdiagnosis disparities and welcomed its transparent SHAP-based feature attribution plans, which offer promise for clinical interpretability. However, they emphasized the need to support clinicians through tailored documentation and training, define actionable decision-support thresholds, and evaluate risks of over- or under-treatment stemming from fairness adjustments. These concerns resonate with findings from Sohail (2024), whose human-centered AI workflows improved clinician trust and reduced diagnostic ambiguity, and Farah et al. (2023), who highlighted that interpretability alone is insufficient without usability guidance. The call to integrate with Electronic Medical Record (EMR) systems and provide patient-facing education aligns with the broader push toward responsible, inclusive AI in clinical settings. I interpret this feedback as a necessary evolution, from fairness as a model property to fairness as a lived clinical experience, requiring not only algorithmic transparency but also stakeholder-specific support and communication.

CHAPTER FIVE

SUMMARY OF THE FINDINGS, CONCLUSIONS, AND RECOMMENDATIONS

5.1 Introduction

This chapter presents a reflective analysis of the study's outcomes, drawing connections between the research objectives, key findings, and their broader implications in fairness-aware Artificial Intelligence (AI) for dermatological diagnostics. It begins with a summary of the main results framed around the original objectives, then proceeds to draw conclusions informed by empirical evidence and expert input. The chapter also discusses the clinical, ethical, and practical implications of the framework, outlines study limitations, and offers concrete recommendations for future research and implementation. In doing so, it seeks to position the study not only as a technical contribution but also as a meaningful step toward equitable healthcare innovation.

5.2 Summary of the Findings of the Study

This study set out to critically examine the integration of fairness-aware mechanisms into AI systems for dermatological skin imaging, with the ultimate goal of developing and validating a bias-mitigating diagnostic framework. Guided by three specific objectives, the research explored both the theoretical underpinnings and practical applications of fairness in medical imaging. The following section provides a structured summary of the findings aligned with each objective, highlighting key insights and contributions drawn from empirical analysis and expert feedback.

5.2.1 Effectiveness of Adversarial Debiasing Mechanisms Used in Diagnostic Imaging

The first objective evaluated adversarial debiasing in AI-driven diagnostic imaging. The findings of this study show that adversarial debiasing can significantly reduce demographic bias in

dermatological diagnostic models. By integrating a predictor–adversary architecture during model training, the system was able to suppress bias-related features associated with sensitive attributes such as skin tone and gender. The experimental results demonstrated that Statistical Parity Difference improved from 0.35 to 0.05, indicating a substantial reduction in disparities across demographic groups, while the overall classification accuracy remained stable.

From the researcher's perspective, these results confirm that fairness-aware optimization techniques can be incorporated into diagnostic AI systems without sacrificing clinical performance. However, the study also observed that adversarial debiasing alone does not completely solve fairness challenges. Issues related to interpretability and the potential impact on certain diagnostic metrics suggest that adversarial methods should be combined with complementary strategies such as balanced datasets and transparency tools. Therefore, adversarial debiasing should be viewed as a key component of a broader fairness-aware framework rather than a standalone solution.

5.2.2 Frameworks Used in Diagnostic Imaging

The analysis of existing fairness frameworks revealed that most current tools provide useful mechanisms for detecting and mitigating bias in machine learning systems, but remain limited when applied to dermatological imaging. Frameworks such as IBM AI Fairness 360, Fairlearn, and TensorFlow Fairness Indicators offer standardized approaches for fairness evaluation and bias mitigation. However, the study found that these frameworks were primarily designed for structured datasets and therefore struggle to address the complexities of medical imaging data.

In the assessment, the main limitation of these frameworks lies in their inability to handle pixel-level bias and intersectional disparities present in dermatological datasets. Even domain-specific platforms such as MONAI focus largely on improving model performance rather than embedding

fairness constraints within the model training process. These findings highlight the need for fairness-aware frameworks that integrate bias mitigation directly within the AI development lifecycle rather than relying solely on post-hoc corrections.

5.2.3 The Framework for Mitigating Bias in Diagnostic Skin Imaging

The third objective centered on the design and validation of a comprehensive fairness-aware framework for dermatological diagnostic imaging. Based on the limitations identified in existing approaches, this study developed and validated a fairness-aware framework specifically designed for dermatological diagnostic imaging. The framework integrates adversarial debiasing, data augmentation strategies, and fairness monitoring mechanisms to address both algorithmic and data-related sources of bias.

The evaluation results indicate that the framework successfully improved fairness while maintaining strong diagnostic performance. Statistical Parity Difference was reduced from 0.35 to 0.05, while classification metrics such as precision, recall, and F1-score remained stable. From the researcher's standpoint, these findings demonstrate that fairness-aware interventions can be effectively embedded within deep learning pipelines without compromising clinical reliability.

In addition, expert validation confirmed that the framework is both technically feasible and practically relevant for clinical environments. Experts highlighted its modular architecture, fairness monitoring capabilities, and compatibility with existing AI development workflows. These findings suggest that the framework offers a practical approach for improving fairness in AI-assisted dermatological diagnostics.

5.3 Conclusions of the Study

Building on the findings associated with each objective, this study demonstrated that adversarial debiasing can serve as an effective mechanism for mitigating demographic bias in AI-driven dermatological diagnostic systems. By implementing a predictor–adversary architecture with a gradient reversal mechanism, fairness constraints were embedded directly within the model training process rather than applied as post-hoc corrections. The results showed a substantial improvement in fairness metrics, with Statistical Parity Difference decreasing from 0.35 to 0.05, while overall diagnostic performance remained stable. The model maintained strong classification accuracy and consistent sensitivity across different Fitzpatrick skin types and gender groups, confirming that fairness interventions can be integrated into clinical AI systems without compromising diagnostic reliability.

The study also examined existing fairness toolkits, including IBM's AI Fairness 360, Microsoft's Fairlearn, and TensorFlow Fairness Indicators, and found that while these frameworks provide valuable tools for fairness auditing and bias mitigation, they are primarily designed for structured tabular datasets. As a result, they lack built-in mechanisms to address the complexities of medical image analysis, such as pixel-level bias, annotation variability, and intersectional disparities in dermatological datasets. Although domain-specific platforms such as MONAI provide stronger support for medical imaging pipelines, their emphasis on performance optimization rather than explicit fairness constraints highlights the need for frameworks that integrate bias mitigation directly within the learning process.

To address these limitations, this study developed and validated a fairness-aware diagnostic framework that integrates adversarial debiasing, synthetic data augmentation, subgroup-sensitive performance monitoring, and expert-informed interpretability mechanisms. By embedding

fairness considerations across multiple stages of the AI development lifecycle, including data preparation, model training, validation, and deployment monitoring, the framework demonstrated the ability to reduce demographic disparities while maintaining clinically relevant predictive performance.

Overall, the findings of this research highlight the importance of incorporating fairness-aware design principles into the development of medical AI systems. Ensuring equitable performance across diverse patient populations requires fairness interventions that operate throughout the entire model lifecycle, from dataset construction to deployment monitoring. By integrating adversarial debiasing and fairness-aware evaluation within dermatological diagnostic systems, this study contributes to the advancement of responsible and equitable AI in healthcare and provides a practical framework for reducing bias in AI-assisted medical decision-making.

5.4 Implications of the Study

The findings of this study have several important implications for the development and deployment of artificial intelligence systems in healthcare, particularly in dermatological diagnostic imaging. First, the results demonstrate that fairness-aware modeling can be successfully integrated into deep learning pipelines without compromising diagnostic performance. The developed framework shows that adversarial debiasing, combined with data balancing techniques and fairness monitoring mechanisms, can significantly reduce demographic disparities in AI predictions. From the researcher's perspective, this finding reinforces the importance of incorporating fairness constraints during model training rather than relying solely on post-hoc bias corrections.

Second, the study has practical implications for clinical applications of AI in dermatology. By improving fairness metrics across different Fitzpatrick skin types while maintaining strong classification performance, the framework provides a pathway for developing diagnostic tools that

perform more consistently across diverse patient populations. This has the potential to reduce disparities in dermatological diagnosis, particularly for individuals with darker skin tones who have historically been underrepresented in dermatological datasets. Integrating fairness-aware AI systems into clinical workflows may therefore enhance both diagnostic reliability and patient trust in AI-assisted medical decision-making.

Third, the research highlights important implications for AI governance and regulatory oversight in healthcare. The framework developed in this study demonstrates the value of embedding fairness monitoring mechanisms, interpretability tools, and subgroup performance evaluations within AI systems used for clinical decision support. These features align with emerging ethical guidelines and regulatory expectations that emphasize transparency, accountability, and equitable performance in medical AI systems. Healthcare institutions and regulatory bodies may therefore consider incorporating fairness evaluations as part of AI system approval and monitoring processes.

Finally, this study underscores the importance of interdisciplinary collaboration in the development of fairness-aware AI systems. The expert validation process conducted in this research revealed that input from clinicians, AI researchers, and software developers plays a critical role in ensuring that AI frameworks are not only technically robust but also clinically relevant and deployable. The researcher, therefore, argues that future healthcare AI systems should be developed through collaborative, stakeholder-driven processes that balance technical innovation with ethical and clinical considerations.

5.5 Limitations of the Study

Despite the contributions made in developing a fairness-aware framework for dermatological diagnostic imaging, several limitations should be acknowledged. First, the effectiveness of the

framework was influenced by the limitations of the available datasets. The dermatological imaging dataset contained limited demographic annotations and imbalanced representation across Fitzpatrick skin types. Although demographic attributes were inferred and data augmentation techniques were applied to improve representation, these constraints may affect the generalizability of the framework across broader and more diverse patient populations.

Second, the validation of the developed framework relied on feedback from a relatively small number of domain experts. While the participating experts provided valuable insights regarding the technical feasibility and clinical relevance of the framework, a larger and more diverse group of clinicians and AI practitioners would have enabled a more comprehensive evaluation of its usability and deployment readiness in real-world healthcare settings.

Finally, the implementation of the framework was limited by computational and resource constraints. Although the framework successfully incorporated adversarial debiasing and fairness monitoring mechanisms, more advanced techniques such as multi-adversarial learning or federated learning architectures were not explored within the scope of this study. Future work may extend the framework by incorporating these approaches and evaluating its performance using larger multi-institutional datasets.

5.6 Recommendations

Based on the findings of this study, several recommendations are proposed to support the development and responsible deployment of fairness-aware artificial intelligence systems in dermatological diagnostic imaging.

First, developers of medical AI systems should incorporate fairness-aware mechanisms directly within model development pipelines rather than relying solely on post-hoc bias mitigation

techniques. The framework developed in this study demonstrates that adversarial debiasing and fairness monitoring mechanisms can be integrated into deep learning architectures without compromising diagnostic accuracy. Future AI development efforts should therefore prioritize fairness-by-design approaches to ensure equitable performance across diverse patient populations.

Second, healthcare institutions and clinicians adopting AI-assisted diagnostic systems should include fairness evaluation as part of the model validation and deployment process. In addition to conventional performance metrics such as accuracy and sensitivity, fairness metrics including Statistical Parity Difference and Equalized Odds should be regularly monitored to detect potential disparities in model predictions across demographic subgroups. Integrating fairness audits into clinical evaluation procedures can help ensure that AI systems support equitable healthcare delivery.

Third, there is a need to improve the availability of diverse and well-annotated dermatological imaging datasets. The findings of this study highlight the challenges associated with limited demographic representation in existing datasets. Future data collection initiatives should prioritize balanced representation across different skin tones, genders, and age groups to support the development of more inclusive AI diagnostic systems.

Finally, further research should extend the fairness-aware framework by exploring more advanced bias mitigation strategies and validating the framework in larger, multi-institutional clinical datasets. Techniques such as federated learning, multi-adversarial architectures, and real-time fairness monitoring systems may further enhance the robustness and scalability of fairness-aware AI models in healthcare environments.

REFERENCES

Adamson, A. S., & Smith, A. (2018). Machine Learning and Health Care Disparities in Dermatology. *JAMA Dermatology*, *154*(11), 1247. https://doi.org/10.1001/jamadermatol.2018.2348

Administration, U. S. F. and D. (2021). Artificial intelligence and machine learning in software as a medical device. *US Food & Drug Administration: Silver Spring, MD, USA*.

Aggarwal, N. (2021). THE NORMS OF ALGORITHMIC CREDIT SCORING. *The Cambridge Law Journal*, *80*(1), 42–73. https://doi.org/DOI: 10.1017/S0008197321000015

AI in Medical Imaging: Benefits, Challenges & Future. (n.d.). Retrieved March 14, 2025, from https://www.ddismart.com/blog/the-evolving-role-of-artificial-intelligence-in-medical-imaging/?form=MG0AV3

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H. A., Al Yami, M. S., Al Harbi, S., & Albekairy, A. M. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Medical Education*, *23*(1), 689. https://doi.org/10.1186/s12909-023-04698-z

Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & Consortium, P. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, *20*, 1–9.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2022). Machine bias. In *Ethics of data and analytics* (pp. 254–264). Auerbach Publications.

Ansari, F., Chakraborti, T., & Das, S. (2024). *Algorithmic Fairness in Lesion Classification by Mitigating Class Imbalance and Skin Tone Bias*. 373–382. https://doi.org/10.1007/978-3-031-72378-0_35

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., Naidich, D. P., & Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, *25*(6), 954–961. https://doi.org/10.1038/s41591-019-0447-x

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT press.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech Era. *Journal of Financial Economics*, *143*(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4:1-4:15. https://doi.org/10.1147/JRD.2019.2942287

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020a). Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020b). Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Björkegren, D., & Grissen, D. (2020). Behavior Revealed in Mobile Phone Usage Predicts Credit Repayment. *The World Bank Economic Review*, *34*(3), 618–634. https://doi.org/10.1093/wber/lhz006

Bobbitt Zach. (2021, May 10). *How to Interpret Root Mean Square Error (RMSE)*. https://www.statology.org/how-to-interpret-rmse/

Brynjolfsson, E., & Mcafee, A. (2017). Artificial intelligence, for real. *Harvard Business Review*, *1*, 1–31.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, 77–91.

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science*, *4*(1), 123–144. https://doi.org/10.1146/annurev-biodatasci-092820-114757

Chen, I. Y., Szolovits, P., & Ghassemi, M. (2019). Can AI help reduce disparities in general medical and mental health care? *AMA Journal of Ethics*, *21*(2), 167–179.

Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Lu, M. Y., Sahai, S., & Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, *7*(6), 719–742.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, 1–24. https://doi.org/10.7717/PEERJ-CS.623/SUPP-1

Chinta, S. V., Wang, Z., Palikhe, A., Zhang, X., Kashif, A., Smith, M. A., Liu, J., & Zhang, W. (2025). AI-driven healthcare: Fairness in AI healthcare: A survey. *PLOS Digital Health*, *4*(5), e0000864. https://doi.org/10.1371/JOURNAL.PDIG.0000864

Chui, M., & Francisco, S. (2017). Artificial intelligence the next digital frontier. *McKinsey and Company Global Institute*, *47*(3.6), 6–8.

Corti. (2024). *Corti Unveils Next-Gen Assistant to Meet Clinicians' Demand for Real-Time Feedback.* https://www.corti.ai/news/corti-unveils-next-gen-assistant-to-meet-clinicians-demand-for-real-time-feedback

Creager, E., Jacobsen, J.-H., & Zemel, R. (2021). Environment inference for invariant learning. *International Conference on Machine Learning*, 2189–2200.

Crenshaw, K. (2013). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories* (pp. 23–51). Routledge.

Crenshaw, K. (2021). Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *Droit et Société, 108,* 465.

Davis, S. E., Dorn, C., Park, D. J., & Matheny, M. E. (2025). Emerging algorithmic bias: fairness drift as the next dimension of model maintenance and sustainability. *Journal of the American Medical Informatics Association*, *32*(5), 845–854. https://doi.org/10.1093/JAMIA/OCAF039

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., & Kalai, A. T. (2019). Bias in bios: A case study of semantic representation bias in a high-stakes setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 120–128.

Dhade, P., & Shirke, P. (2023). Federated Learning for Healthcare: A Comprehensive Review†. *Engineering Proceedings*, *59*(1). https://doi.org/10.3390/engproc2023059230

Dolin, P., Li, W., Dasarathy, G., & Berisha, V. (2025). *Statistically Valid Post-Deployment Monitoring Should Be Standard for AI-Based Digital Health*. http://arxiv.org/abs/2506.05701

Duff, J. H., Scarpa, M., Zupluoglu, C., & Prilleltensky, I. (2024). The Development and Initial Validation of the Multidimensional Fairness Scale. *Social Justice Research*, *37*(3), 213–238. https://doi.org/10.1007/S11211-024-00440-2/TABLES/6

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017a). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017b). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118.

Fairlearn Documentation. (2024). *User Guide — Fairlearn 0.12.0 documentation*. https://fairlearn.org/v0.12/user_guide/index.html

Farah, L., Murris, J., Borget, I., Guilloux, A., Martelli, N., Katsahian, S. I., & Assessment, al. (2023). *of Performance, Interpretability, and Explainability in Artificial Intelligence-Based Health Technologies: What Healthcare Stakeholders Need to Know*. *2023*(2). https://doi.org/10.1016/j.mcpdig.2023.02.004ï

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.

Ferrara, C., Sellitto, G., Ferrucci, F., Palomba, F., & De Lucia, A. (2024). Fairness-aware machine learning engineering: how far are we? *Empirical Software Engineering*, *29*(1), 1–46. https://doi.org/10.1007/S10664-023-10402-Y/FIGURES/9

Frasca, M., La Torre, D., Pravettoni, G., & Cutica, I. (2024). Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review. *Discover Artificial Intelligence*, *4*(1), 1–21. https://doi.org/10.1007/S44163-024-00114-7/TABLES/3

Friedler, S. A., Choudhary, S., Scheidegger, C., Hamilton, E. P., Venkatasubramanian, S., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 329–338. https://doi.org/10.1145/3287560.3287589

Future of AI in medical imaging: Challenges and opportunities - Quibim. (n.d.). Retrieved March 14, 2025, from https://quibim.com/news/ai-in-medical-imaging/?form=MG0AV3&form=MG0AV3

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, *2*(4), 42–47.

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., & Ghassemi, M. (2021). Do as AI say: susceptibility in deployment of clinical decision-aids. *Npj Digital Medicine*, *4*(1), 31. https://doi.org/10.1038/s41746-021-00385-9

Gencer, G., & Gencer, K. (2025). Advanced retinal disease detection from OCT images using a hybrid squeeze and excitation enhanced model. *PLOS ONE*, *20*(2), e0318657. https://doi.org/10.1371/JOURNAL.PONE.0318657

Gerke, S., Minssen, T., & Cohen, G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial Intelligence in Healthcare* (pp. 295–336). Elsevier. https://doi.org/10.1016/B978-0-12-818438-7.00012-5

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Ghosh, B. (2023). *Interpretability and Fairness in Machine Learning: A Formal Methods Approach*.

Gichoya, J. W., Banerjee, I., Bhimireddy, A. R., Burns, J. L., Celi, L. A., Chen, L.-C., Correa, R., Dullerud, N., Ghassemi, M., & Huang, S.-C. (2022). AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, *4*(6), e406–e414.

Glocker, B., Jones, C., Roschewitz, M., & Winzeck, S. (2023). Risk of Bias in Chest Radiography Deep Learning Foundation Models. *Radiology: Artificial Intelligence*, *5*(6), e230060. https://doi.org/10.1148/ryai.230060

Gonzalez, R. C. (2009). *Digital image processing*. Pearson education india.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, Issue 2). MIT press Cambridge.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. *Science Robotics*, *3*(January), 2672–2680. https://arxiv.org/pdf/1406.2661

Google Research. (2019, December 11). *Fairness Indicators: Scalable Infrastructure for Fair ML Systems*. https://research.google/blog/fairness-indicators-scalable-infrastructure-for-fair-ml-systems/?form=MG0AV3

Groh, M., Harris, C., Daneshjou, R., Badri, O., & Koochek, A. (2022). *Towards Transparency in Dermatology Image Datasets with Skin Tone Annotations by Experts, Crowds, and an Algorithm*. http://arxiv.org/abs/2207.02942

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., & Badri, O. (2021a). Evaluating deep neural networks trained on clinical images in dermatology with the

fitzpatrick 17k dataset. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1820–1828.

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., & Badri, O. (2021b). *Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset*. http://arxiv.org/abs/2104.09957

Guan, S., Jiang, R., Meng, C., & Biswal, B. (2023). Brain age prediction across the human lifespan using multimodal MRI data. *GeroScience 2023 46:1*, *46*(1), 1–20. https://doi.org/10.1007/S11357-023-00924-0

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., & Cuadros, J. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, *316*(22), 2402–2410.

Han, L. (2025). Addressing Distribution Shift for Robust and Trustworthy Prediction and Causal Inference in Clinical AI Settings. *JAMA Network Open*, *8*(6), e2513705–e2513705. https://doi.org/10.1001/JAMANETWORKOPEN.2025.13705

Hardt, M., Price, E., & Srebro, N. (2016a). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, *29*.

Hardt, M., Price, E., & Srebro, N. (2016b). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, 3323–3331. https://arxiv.org/pdf/1610.02413

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., & Xu, D. (2021). UNETR: Transformers for 3D Medical Image Segmentation. *Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022*, 1748–1758. https://doi.org/10.1109/WACV51458.2022.00181

Holstein, K., Vaughan, J. W., Daumé, H., Dudík, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300830/SUPPL_FILE/PN3256.ZIP

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? *ArXiv Preprint ArXiv:1712.09923*.

Ibrahim, M., Khalil, Y. Al, Amirrajab, S., Sun, C., Breeuwer, M., Pluim, J., Elen, B., Ertaylan, G., & Dumontier, M. (2025). Generative AI for synthetic data across multiple medical modalities: A systematic review of recent developments and challenges. *Computers in Biology and Medicine*, *189*. https://doi.org/10.1016/j.compbiomed.2025.109834

Jaiyeoba, O., Ogbuju, E., Ataguba, G., Jaiyeoba, O., Omaye, J. D., Eze, I., & Oladipo, F. (2025). State-of-the-art skin disease classification: a review of deep learning models. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *14*(1), 1–43. https://doi.org/10.1007/S13721-024-00495-W/METRICS

Jeanette Towles. (2025). *Designing an Audit Trail for AI in Clinical Trials: Aligning with the EU AI Act* . https://www.synterex.com/post/designing-an-audit-trail-for-ai-in-clinical-trials-aligning-with-the-eu-ai-act

Jorge Cardoso, M., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., … Feng, A. (2022). *MONAI: An open-source framework for deep learning in healthcare*.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., & Potapenko, A. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589.

Kabir, M. A., Ahmed, M. U., Begum, S., Barua, S., & Islam, M. R. (2024). Balancing Fairness: Unveiling the Potential of SMOTE-Driven Oversampling in AI Model Enhancement. *ACM International Conference Proceeding Series*, 21–29. https://doi.org/10.1145/3674029.3674034/ASSETS/HTML/IMAGES/ICMLT2024-5-FIG3.JPG

Kalmar Mark. (2025, January 10). *Shaping the Future of Radiology in 2025: Trends, Threats, and Opportunities*. https://www.diagnosticimaging.com/view/future-radiology-2025-trends-threats-opportunities?form=MG0AV3&form=MG0AV3

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2019). An Empirical Study of Rich Subgroup Fairness for Machine Learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 100–109. https://doi.org/10.1145/3287560.3287592

Khalifa, M., & Albadawy, M. (2024). AI in diagnostic imaging: Revolutionising accuracy and efficiency. In *Computer Methods and Programs in Biomedicine Update* (Vol. 5). Elsevier B.V. https://doi.org/10.1016/j.cmpbup.2024.100146

Koçak, B., Ponsiglione, A., Stanzione, A., Bluethgen, C., Santinha, J., Ugga, L., Huisman, M., Klontzas, M. E., Cannella, R., & Cuocolo, R. (2024). Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects. *Diagnostic and Interventional Radiology*. https://doi.org/10.4274/dir.2024.242854

Koziarski, M. (2020). CSMOUTE: Combined Synthetic Oversampling and Undersampling Technique for Imbalanced Data Classification. *Proceedings of the International Joint Conference on Neural Networks*, *2021-July*. https://doi.org/10.1109/IJCNN52387.2021.9533415

Krishna Seth. (2025, March 5). *Real-Time Bias Mitigation: The Future of Fair AI*. https://www.analyticsinsight.net/artificial-intelligence/real-time-bias-mitigation-the-future-of-fair-ai

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, *117*(23), 12592–12594. https://doi.org/10.1073/pnas.1919012117

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)*, *9*(1), 3.

Lesley Clack. (2023, June 20). *Using Data Analytics to Predict Outcomes in Healthcare*. https://journal.ahima.org/page/using-data-analytics-to-predict-outcomes-in-healthcare

Lipton, Z. C. (2018). The Mythos of Model Interpretability. *Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Litman, T. (2017). *Autonomous vehicle implementation predictions*.

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*, *37*(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749

Liu, H., Dou, Y., Wang, K., Zou, Y., Sen, G., Liu, X., & Li, H. (2025). A skin disease classification model based on multi scale combined efficient channel attention module. *Scientific Reports*, *15*(1), 6116. https://doi.org/10.1038/s41598-025-90418-0

Liu, H., & Wong, R. C. W. (2024). Adversarial Learning of Group and Individual Fair Representations. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *14645 LNAI*, 181–193. https://doi.org/10.1007/978-981-97-2242-6_15

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P. Q., Corrado, G. S., Hipp, J. D., Peng, L., & Stumpe, M. C. (2017). *Detecting Cancer Metastases on Gigapixel Pathology Images*. http://arxiv.org/abs/1703.02442

Liu, Y., & Zhao, N. (2023). AI-Assisted Design: Generative Architectural Design. *Advances in Transdisciplinary Engineering*, *42*, 371–378. https://doi.org/10.3233/ATDE230973

Li, X., Cui, Z., Wu, Y., Gu, L., & Harada, T. (2021). *Estimating and Improving Fairness with Adversarial Learning*. https://arxiv.org/pdf/2103.04243

Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2018). Learning adversarially fair and transferable representations. *International Conference on Machine Learning*, 3384–3393.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., … Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Microsoft Research. (2019). *Microsoft Research 2019 reflection—a year of progress on technology's toughest challenges - Microsoft Research*. https://www.microsoft.com/en-us/research/blog/microsoft-research-2019-reflection-a-year-of-progress-on-technologys-toughest-challenges/?form=MG0AV3

Murdoch, B. (2021). Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Medical Ethics*, *22*(1), 1–5. https://doi.org/10.1186/S12910-021-00687-3/PEER-REVIEW

Naayini, P. (2025). Building AI-Driven Cloud-Native Applications with Kubernetes and Containerization. *International Journal Of Scientific Advances*, *6*(2). https://doi.org/10.51542/ijscia.v6i2.15

Nabhani-Gebara, S., Zacharias, L., Bonmati, L. M., Galiana-Bordera, A., Colantonio, S., Chouvarda, I., & Elkhoury, G. (2025). Clinical Validation in AI for Healthcare: An Experiential Learning. *Trustworthy AI in Cancer Imaging Research*, 267–282. https://doi.org/10.1007/978-3-031-89963-8_12

Nagpal, R., Khan, A., Borkar, M., & Gupta, A. (2024). A Multi-Objective Framework for Balancing Fairness and Accuracy in Debiasing Machine Learning Models. *Machine Learning and Knowledge Extraction*, *6*(3), 2130–2148. https://doi.org/10.3390/make6030105

Nyholm, S., & Smids, J. (2020). Automated cars meet human drivers: responsible human-robot coordination and the ethics of mixed traffic. *Ethics and Information Technology*, *22*(4), 335–344.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Onose Ejiro. (2023, August 8). *R-Squared: Coefficient of Determination in Machine Learning*. https://arize.com/blog-course/r-squared-understanding-the-coefficient-of-determination/

Parikh, R. B., Teeple, S., & Navathe, A. S. (2019). Addressing Bias in Artificial Intelligence in Health Care. *JAMA*, *322*(24), 2377. https://doi.org/10.1001/jama.2019.18058

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, *62*(3), 54–60. https://doi.org/10.1145/3241036

Pereira, S., Pinto, A., Alves, V., & Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Transactions on Medical Imaging*, *35*(5), 1240–1251.

Petsko, C. D., Rosette, A. S., & Bodenhausen, G. V. (2022). Through the looking glass: A lens-based account of intersectional stereotyping. *Journal of Personality and Social Psychology*, *123*(4), 763.

Pinaya, W. H. L., Graham, M. S., Kerfoot, E., Tudosiu, P.-D., Dafflon, J., Fernandez, V., Sanchez, P., Wolleb, J., da Costa, P. F., Patel, A., Chung, H., Zhao, C., Peng, W., Liu, Z., Mei, X., Lucena, O., Ye, J. C., Tsaftaris, S. A., Dogra, P., … Cardoso, M. J. (2023). *Generative AI for Medical Imaging: extending the MONAI Framework*. https://arxiv.org/pdf/2307.15208

Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J. B., & Zuiderveld, K. (1987). Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, *39*(3), 355–368. https://doi.org/10.1016/S0734-189X(87)80186-X

Precedence Research. (2025, February 3). *Corti Unveils AI Models to Revolutionize Healthcare*. https://www.precedenceresearch.com/news/corti-ai-healthcare-revolution

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring Fairness in Machine Learning to Advance Health Equity. *Annals of Internal Medicine*, *169*(12), 866. https://doi.org/10.7326/M18-1990

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., & Sun, M. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, *1*(1), 1–10.

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, *28*(1), 31–38.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). "Why Should I Trust You?" *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). " Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Rouger Melisande. (2020, December 22). *Removing bias from mammography screening with deep learning • healthcare-in-europe.com*. https://healthcare-in-europe.com/en/news/removing-bias-from-mammography-screening-with-deep-learning.html?form=MG0AV3

Roy, S. (2023). *Lecture 10: Shapley Values for Explainable AI*.

Rumelhart, D. E., McClelland, J. L., & Group, P. D. P. R. (1986). *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press.

Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2024). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Wiley Online Library*. https://doi.org/10.1002/AISY.202400304

Schoeffer, J., De-Arteaga, M., & Kühl, N. (2024). Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. *Conference on Human Factors in Computing Systems - Proceedings*.

https://doi.org/10.1145/3613904.3642621/SUPPL_FILE/3613904.3642621-TALK-VIDEO.VTT

Science X staff. (2025, March 11). *AI reduces false positives by 37.3% in breast cancer diagnosis*. https://www.msn.com/en-us/health/other/ai-reduces-false-positives-by-373-in-breast-cancer-diagnosis/ar-AA1ACJlv?form=MG0AV3&form=MG0AV3

Scott, M., Syst., J. P.-Aust. J. Intell. Inf. Process., & 2019, undefined. (2019). GAN-SMOTE: A Generative Adversarial Network approach to Synthetic Minority Oversampling. *Users.Cecs.Anu.Edu.Au*. http://users.cecs.anu.edu.au/~Tom.Gedeon/conf/ABCs2019/paper/ABCs2019_paper_v2_127.pdf

Secinaro, S., Calandra, D., Secinaro, A., Muthurangu, V., & Biancone, P. (2021). The role of artificial intelligence in healthcare: a structured literature review. *BMC Medical Informatics and Decision Making*, *21*(1), 125. https://doi.org/10.1186/s12911-021-01488-9

Settles, B. (2012). Active Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *6*(1), 1–114. https://doi.org/10.2200/S00429ED1V01Y201207AIM018

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B. A., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, *27*(12), 2176–2182. https://doi.org/10.1038/s41591-021-01595-0

Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., & Bakas, S. (2020). Federated learning in medicine: facilitating

multi-institutional collaborations without sharing patient data. *Scientific Reports*, *10*(1), 12598. https://doi.org/10.1038/s41598-020-69250-1

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 1–48.

Smith Tom. (2023, June 29). *Got Data? How SMOTE and GANs Create Synthetic Data*. https://dzone.com/articles/got-data-how-smote-and-gans-create-synthetic-data

Somashekhar, S. P., Sepúlveda, M.-J., Puglielli, S., Norden, A. D., Shortliffe, E. H., Kumar, C. R., Rauthan, A., Kumar, N. A., Patil, P., & Rhee, K. (2018). Watson for Oncology and breast cancer treatment recommendations: agreement with an expert multidisciplinary tumor board. *Annals of Oncology*, *29*(2), 418–423.

Stempniak Marty. (2024, May 24). *Top 10 trends to follow in diagnostic radiology*. https://radiologybusiness.com/topics/healthcare-management/healthcare-economics/top-10-trends-follow-diagnostic-radiology?form=MG0AV3&form=MG0AV3

stempniak Marty. (2024, October 31). *The 7 most pressing challenges in radiology practice: A 'perfect storm' is brewing*. https://radiologybusiness.com/topics/healthcare-management/medical-practice-management/7-most-pressing-challenges-radiology-practice-perfect-storm-brewing?form=MG0AV3&form=MG0AV3

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., & Kraus, S. (2022). Artificial intelligence and life in 2030: the one hundred year study on artificial intelligence. *ArXiv Preprint ArXiv:2211.06318*.

Syed Hamza Sohail. (2024, November 12). *Corti Unveils AI-Powered Platform for Real-Time Clinical Support* -. https://hitconsultant.net/2024/12/11/corti-unveils-ai-powered-platform-for-real-time-clinical-support/

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Trafton Anne. (2024, June 28). *Study reveals why AI models that analyze medical images can be biased | MIT News | Massachusetts Institute of Technology*. https://news.mit.edu/2024/study-reveals-why-ai-analyzed-medical-images-can-be-biased-0628?form=MG0AV3&form=MG0AV3

Tu, L. (2023). Frameworks of Fairness: Legal Standards for Impartiality in Arbitration. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.5052772

University of Cambridge. (2020, May 12). *AI techniques in medical imaging may lead to incorrect diagnoses | University of Cambridge*. Https://Www.Cam.Ac.Uk/Research/News/Ai-Techniques-in-Medical-Imaging-May-Lead-to-Incorrect-Diagnoses?Form=MG0AV3&form=MG0AV3. https://www.cam.ac.uk/research/news/ai-techniques-in-medical-imaging-may-lead-to-incorrect-diagnoses?form=MG0AV3&form=MG0AV3

Van Der Vorst, J. P., Smit, J. M., Van De Sande, D., Van Der Ster, B., Daams, F., Schasfoort, R., Gommers, D., Verhoef, C., Grünhagen, D. J., Van Genderen, M. E., & Hilling, D. E. (2025). Importance of model governance in clinical AI models: case study on the relevance of data drift detection. *BMJ Digital Health and AI*, *1*, 46. https://doi.org/10.1136/bmjdhai-2025-000046

Vishal Yadav, & Shreeja Kale. (2024). *Fairness-Aware Federated Learning with Real-Time Bias Detection and Correction Enhancing Model Equity and Privacy in Decentralized Systems through Adaptive Mechanisms*. *9*(8). https://doi.org/10.38124/ijisrt/IJISRT24AUG1319

Wang, H., & Zheng, H. (2013). Model Validation, Machine Learning. *Encyclopedia of Systems Biology*, 1406–1407. https://doi.org/10.1007/978-1-4419-9863-7_233

Wang, X., & Yang, C. C. (2025). *Balancing Fairness and Performance in Healthcare AI: A Gradient Reconciliation Approach*. https://arxiv.org/pdf/2504.14388

Weng, N., Bigdeli, S., Petersen, E., & Feragen, A. (2023). *Are Sex-Based Physiological Differences the Cause of Gender Bias for Chest X-Ray Diagnosis?* (pp. 142–152). https://doi.org/10.1007/978-3-031-45249-9_14

WHO. (2023). *Global Initiative on AI for Health*. https://www.who.int/initiatives/global-initiative-on-ai-for-health

Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L. R., Summers, R. M., Rubin, D. L., & Lungren, M. P. (2020). Preparing Medical Imaging Data for Machine Learning. *Radiology*, *295*(1), 4–15. https://doi.org/10.1148/radiol.2020192224

Williams, J. C. (2014). Double jeopardy? An empirical study with implications for the debates over implicit bias and intersectionality. *Harvard Journal of Law & Gender*, *37*, 185.

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*(1), 409–428. https://doi.org/10.3758/S13428-018-1055-2/TABLES/4

Yang, J., Jiang, J., Sun, Z., & Chen, J. (2024). *A Large-Scale Empirical Study on Improving the Fairness of Image Classification Models*. http://arxiv.org/abs/2401.03695

Yang, J., Soltan, A. A. S., Eyre, D. W., Yang, Y., & Clifton, D. A. (2023). *An adversarial training framework for mitigating algorithmic biases in clinical machine learning*. https://doi.org/10.1038/s41746-023-00805-y

Yu, G., Ma, L., Wang, X., Du, W., Du, W., & Jin, Y. (2022). Towards Fairness-Aware Multi-Objective Optimization. *Complex and Intelligent Systems*, *11*(1). https://doi.org/10.1007/s40747-024-01668-w

Zartashea, F. (2023). Bias and fairness in AI-driven healthcare: Addressing disparities in machine learning models. *International Journal of Science and Research Archive*, *2023*(01), 835–846. https://doi.org/10.30574/ijsra.2023.9.1.0359

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Medicine*, *15*(11), e1002683.

Zhang, J., Gajjala, S., Agrawal, P., Tison, G. H., Hallock, L. A., Beussink-Nelson, L., Lassen, M. H., Fan, E., Aras, M. A., Jordan, C., Fleischmann, K. E., Melisko, M., Qasim, A., Shah, S. J., Bajcsy, R., & Deo, R. C. (2018). Fully Automated Echocardiogram Interpretation in Clinical Practice. *Circulation*, *138*(16), 1623–1635. https://doi.org/10.1161/CIRCULATIONAHA.118.034338

Zhang, Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340. https://doi.org/10.1145/3278721.3278779

Zhang, Y., Zhang, T., Mu, R., Huang, X., & Ruan, W. (2024). *Towards Fairness-Aware Adversarial Learning*. https://arxiv.org/pdf/2402.17729

Zheng, G., Jacobs, M. A., Braverman, V., Parekh, V. S., & Jacobs Braverman Parekh, Z. (2025). Towards Fair Medical AI: Adversarial Debiasing of 3D CT Foundation Embeddings. *Proceedings of Machine Learning Research-Under Review*, 1–13. https://arxiv.org/pdf/2502.04386v1

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2019). A primer on deep learning in genomics. *Nature Genetics*, *51*(1), 12–18.

Zoumana Keita. (2023, May 10). *Explainable AI, LIME & SHAP for Model Interpretability | Unlocking AI's Decision-Making | DataCamp*. https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models

APPENDICES

Appendix A: Introductory Letter

KISII UNIVERSITY

Telephone: +254 0202610479
Email: research@kisiiuniversity.ac.ke

P O BOX 408 – 40200, KISII
www.kisiiuniversity.ac.ke

OFFICE OF THE REGISTRAR RESEARCH, EXTENSION, INNOVATION, AND RESOURCE MOBILIZATION

REF: KSU/REIRM/140 **DATE:** 14th May, 2025

The Head, Research Coordination
National Council for Science, Technology, and Innovation (NACOSTI)
Utalii House, 8th Floor, Uhuru Highway
P. O. Box 30623– 00100
NAIROBI - KENYA

Dear Sir/Madam,

RE: WAITHIRA JOSHUA MWAURA: MIN 12/00001/23

The above-mentioned is a student of Kisii University currently pursuing a Masters Degree in the School of Information Science and Technology. His research topic is **"A Framework for Mitigating Bias in Artificial Intelligence Systems Used in Diagnostic Skin Imaging."**

We are kindly asking you to assist the student acquire a research permit to enable him carry out the research, upon getting ethical clearance by an accredited Scientific and Ethics Review Committee.

Thank you.

Prof. Christopher Ngacho, PhD
AG. REGISTRAR, (REIRM)

CT: Director Postgraduate Studies

CN/bm

Kisii University is IS0 9001:2015 certified

Appendix B: Ethics Clearance

KISII UNIVERSITY

Phone: 020 - 2610479
Email: iserckisiiuniversity@gmail.com

P.O. Box 408-40200
KISII, KENYA.

INSTITUTIONAL SCIENTIFIC AND ETHICS REVIEW COMMITTEE (ISERC)

REF: KSU/ISERC/0039/05/25
TO: WAITHIRA JOSHUA MWAURA

Date:28/05/2025

Dear Joshua Mwaura,

RE: A FRAMEWORK FOR MITIGATING BIAS IN ARTIFICIAL INTELLIGENCE SYSTEMS USED IN DIAGNOSTIC SKIN IMAGING

This is to inform you that Kisii University (KSU) ISERC has reviewed and approved your above research proposal. Your application approval number is KSU ISERC PROTOCOL No. 0039/05/25. This study is approved for implementation effective this date 28th May 2025. Please note that authorization to conduct this study will automatically expire on 28th May 2026. If you plan to continue with data collection beyond this date, please apply for continuing approval from the KSU ISERC secretariat.

This approval is subject to compliance with the following requirements;

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by KSU ISERC.
iii. Life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to KSU ISERC within 72 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affect the safety or welfare of study participants and others or affect the integrity of the research must be reported to KSU ISERC within 72 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval should be made at least 60 days before the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to KSU ISERC.
viii. You are required to submit any amendments to this protocol and other information pertinent to human participation in this study to KSU ISERC for review before initiation.

Before commencing your study, you will be expected to obtain a research license from the National Commission for Science, Technology and Innovation (NACOSTI) https://research-portal.nacosti.go.ke and also obtain other clearances needed.

Yours sincerely,

KISII UNIVERSITY
INSTITUTIONAL SCIENTIFIC AND ETHICS REVIEW COMMITTEE
30 MAY 2025
CHAIRMAN
P. O. Box 408 - 40200, Kisii (Kenya)

Prof. Samson Maobe
Chairman
Institutional Scientific and Ethics Review Committee
KISII UNIVERSITY

Appendix C: Research Permit

REPUBLIC OF KENYA

NATIONAL COMMISSION FOR
SCIENCE,TECHNOLOGY & INNOVATION

Ref No: **779664**

Date of Issue: **20/June/2025**

RESEARCH LICENSE

This is to Certify that Mr.. Joshua Mwaura Waithira of Kisii University, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Kisii on the topic: A FRAMEWORK FOR MITIGATING BIAS IN ARTIFICIAL INTELLIGENCE SYSTEMS USED IN DIAGNOSTIC SKIN IMAGING for the period ending : 20/June/2026.

License No: **NACOSTI/P/25/4175012**

779664

Applicant Identification Number

Deputy Director
NATIONAL COMMISSION FOR
SCIENCE,TECHNOLOGY &
INNOVATION

Verification QR Code

NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.

See overleaf for conditions

Appendix D: Expert Validation Tool

A FRAMEWORK FOR MITIGATING BIAS IN ARTIFICIAL INTELLIGENCE SYSTEMS USED IN DIAGNOSTIC SKIN IMAGING

Researcher: JOSHUA WAITHIRA

Contact: 0703834453

Problem Statement

AI models used in medical diagnostics, particularly in dermatological imaging, have been shown to exhibit biases that disproportionately affect certain demographic groups. This issue arises because training datasets often lack diversity, meaning the AI learns patterns that may not be equally accurate for all populations. In dermatology, this has led to higher misclassification rates for individuals with darker skin tones, making AI-assisted diagnosis unreliable for equitable healthcare. This study aims to address bias in diagnostic imaging by integrating adversarial debiasing mechanisms, a method designed to actively reduce bias in AI predictions without compromising diagnostic accuracy.

Study Objectives

This research is structured around three key objectives:

To evaluate the effectiveness of adversarial debiasing mechanisms used in diagnostic imaging (determining how well these methods reduce bias while maintaining predictive accuracy).

To analyze existing frameworks used in diagnostic imaging (reviewing current AI models to identify their limitations in fairness and interpretability).

To develop and evaluate an enhanced framework for minimizing bias in diagnostic imaging (proposing a fairness-aware AI framework optimized for equitable diagnosis).

Solution Approach

To resolve the fairness challenges in AI-assisted dermatological diagnostics, this study develops a two-model approach:

Base Model - a standard deep learning model trained using conventional techniques, likely to reflect existing biases.

Fairness-Aware Model - an improved version using adversarial debiasing, which ensures that predictions remain independent of sensitive attributes like skin tone and gender.

The fairness-aware model introduces an adversary network, trained alongside the primary diagnostic model, which actively suppresses bias patterns by preventing the AI from unintentionally encoding demographic disparities. The effectiveness of this approach is confirmed through statistical validation.

Statistical Validation Results

The fairness-aware model was assessed using key fairness metrics, which quantify bias reduction across demographic subgroups:

Metric	Pre-Debiasing (Standard Model)	Post-Debiasing (Fairness-Aware Model)
Statistical Parity Difference (SPD)	.35	.05
Equalized Odds (EO)	.42	.12

Disparate Impact Ratio (DIR)	.65	.95

What These Results Mean:

Statistical Parity Difference (SPD) measures whether predictions are equally distributed across subgroups. A high *SPD* (.35) before debiasing shows unequal prediction rates across demographic groups, while a lower value (.05) after debiasing confirms that the fairness-aware model significantly improved equity.

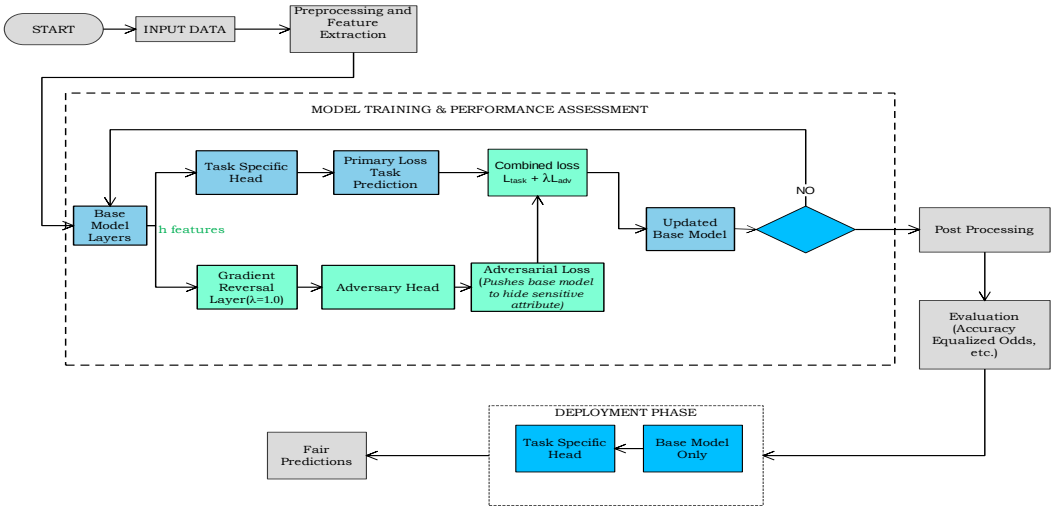
Equalized Odds (EO) evaluates whether the model performs equally for different subgroups, specifically in terms of sensitivity and false positive rates. Before debiasing (*EO* = .42), the model showed substantial disparities in classification accuracy across groups. After debiasing (*EO* = .12), bias in false-positive and false-negative rates was significantly reduced, ensuring consistent diagnostic reliability across demographic subgroups.

Disparate Impact Ratio (DIR) assesses whether predictions remain proportional across populations. A low *DIR* (.65) suggests that certain subgroups received systematically different treatment in predictions, whereas a higher *DIR* (.95) after debiasing confirms that fairness-aware modifications helped restore balance.

Paired t-tests conducted across fairness metric distributions confirm statistically significant bias reduction (*p-value* < .05), ensuring that fairness-aware modifications are not random fluctuations, but true improvements in model fairness.

The Framework

The attached diagram visually presents a fairness-aware AI framework



This framework illustrates an adversarial debiasing pipeline designed to mitigate bias during model training in clinical machine learning tasks. It begins with input data undergoing preprocessing and feature extraction before entering the main training block. The base model outputs are split into two branches: a task-specific head that focuses on the primary prediction task and an adversarial head connected via a Gradient Reversal Layer (GRL). The adversarial head learns to predict sensitive attributes (e.g., race, gender), and its gradients are reversed to penalize the base model for encoding demographic cues, thereby reducing bias. Losses from both the task and adversarial branches are used to iteratively update the model and the adversary. After training, predictions are post-processed and evaluated using fairness metrics like Equalized Odds, alongside traditional accuracy metrics. During deployment, only the base model and task-specific head are retained to generate fair predictions without including the adversarial components.

Validation Goals for Experts

Validation Domain	Core Questions
Clinical Applicability	Does the model maintain diagnostic accuracy expected in real-world clinical settings? Is its performance consistent across diverse patient populations (e.g., skin tones, demographics)?
Fairness Transparency	Are fairness interventions (e.g., Equalized Odds, SPD) statistically sound and ethically grounded? Are disparities across sensitive groups minimized and measurable?
Model Interpretability	Can both developers and clinicians understand how the model arrives at its predictions? Do explainability methods (e.g., SHAP, heatmaps) align with medical reasoning?
Deployment Feasibility	Can the system be deployed with safeguards to monitor fairness over time? Are post-deployment updates feasible under changing data or population shifts?

Response Form.

Kindly fill in your responses in the table below.

Occupation i.e. Senior developer	Years of experience	Comments

Appendix E: Plagiarism Report

School Of Postgraduate

A FAIRNESS-AWARE FRAMEWORK FOR BIAS MITIGATION IN DIAGNOSTIC SKIN IMAGING USING AN ADVERSARIAL CNN ...

A FAIRNESS-AWARE FRAMEWORK FOR BIAS MITIGATION IN DIAGNOSTIC SKIN IMAGING USING AN ADVERSARIAL CNN APPROACH

Postgraduate

Kisii University

Document Details:

Submission ID

Submission Date

Apr 16, 2026, 3:25 PM GMT+3

Download Date

Apr 16, 2026, 3:31 PM GMT+3

File Name

Mitigating_Bias_in_AI_Thesis_Final_Submission.docx

File Size

2.3 MB

100 Pages

15,306 Words

105,187 Characters

17% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

- 681 Not Cited or [redacted]%
  Matches with neither in-text citation nor quotation marks
- 95 Missing [redacted]%
  Matches that are still very similar to source material
- 0 Missing [redacted]%
  Matches that have quotation marks, but no in-text citation
- 0 Cited and [redacted]%
  Matches with in-text citation present, but no quotation marks

Top Sources

10% Internet sources
14% Publications
6% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.