Strategies to Ensure Intersectional Fairness in Vision-Language Models for Clinical Decision Support

YUPENG ZHANG

Supervisor: Professor Jinman Kim
Associate Supervisor: Doctor Usman Naseem, Professor Adam Dunn

A thesis submitted in fulfilment of
the requirements for the degree of
Master of Philosophy

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

22 April 2026

Abstract

Rapid integration of artificial intelligence (AI), particularly Vision-Language Models (VLMs), as decision support system for medical diagnosis promises to enhance healthcare outcomes. However, these models can inherit and amplify societal biases, leading to significant performance disparities across diverse patient subgroups. This thesis addresses a critical and often overlooked challenge: intersectional fairness, where compounded disadvantages emerge for individuals with multiple demographic attributes (e.g., by race and gender). Existing fairness interventions, which typically focus on single demographic attributes, often fail to mitigate these compounded biases and can inadvertently degrade overall model performance or mask subtle but clinically significant disparities in diagnostic certainty.

This thesis introduces a novel regularisation framework, Cross-Modal Alignment Consistency Maximum Mean Discrepancy (CMAC-MMD), to specifically address intersectional fairness at the decision level of models' architecture. This approach represents a conceptual shift from image and text feature-level manipulation to directly equalizing the model's diagnostic confidence across all intersectional subgroups. By defining a scalar "cross-modal alignment score" that serves as a proxy for the model's certainty, the CMAC-MMD method leverages a unique fairness loss to align the statistical distributions of these scores. This process compels the model to produce predictions with equitable confidence and decisiveness for all patient subgroups, regardless of their demographic profile, without requiring sensitive data during inference time.

The effectiveness of the proposed framework is comprehensively evaluated through benchmarking on dermatology and ophthalmology datasets for disease classification. The results demonstrate that CMAC-MMD substantially reduces intersectional performance disparities across multiple fairness metrics while maintaining overall diagnostic accuracy as baseline models. By confronting the challenge of equitable diagnostic certainty, this

work establishes a more robust standard for clinical fairness. It provides a scalable, privacy-preserving framework for developing more equitable, reliable, and trustworthy medical AI systems essential for high-stakes clinical applications.

Keywords: Intersectional fairness, vision-language models, algorithmic fairness, medical image classification, bias mitigation, Maximum Mean Discrepancy, trustworthy AI

Statement of Originality

This is to certify that, to the best of my knowledge, the content of this Thesis is my own work. This Thesis has not been submitted for any degree or other purposes. I certify that the intellectual content of this Thesis is the product of my own work and that all the assistance received in preparing this Thesis and sources have been acknowledged.

Generative AI Attribution Statement

During the preparation of the thesis, the author used ChatGPT, Gemini, and Claude for purposes such as text enhancement, including paraphrasing, refining sentence structure, and correcting spelling and grammar.

All AI-assisted outputs were carefully reviewed by the author to identify and correct any potential errors, inaccuracies, or biases. The author takes full responsibility for the submitted thesis and ensures the work is their own and has used generative AI within the parameters of use (refer to the University of Sydney generative AI guide for researchers).

Authorship Attribution Statement

This Thesis contains paper publications in Chapter 3, formatted as a peer-reviewed journal manuscript. The authorship attributions of this Chapter are shown as follows:

Paper [1] is under review for submission to *NPJ Digital Medicine* Collection of Multimodal AI. The authors of this paper are Yupeng Zhang, Dr Usman Naseem, Professor Adam Dunn, and Professor Jinman Kim. The study was conceptualised through the joint efforts of all authors. Yupeng Zhang was responsible for the study design, literature review, methodology development, algorithm implementation, data curation and analysis, programming, hyperparameter optimisation, experimental validation, results analysis and visualisation, and manuscript writing and editing. Dr Usman Naseem significantly contributed to technical guidance on methodology design, fairness metrics and regularization techniques, results validation, and manuscript review. Professor Adam Dunn contributed expertise on clinical applications, bias analysis in medical AI systems, and manuscript review. Professor Jinman Kim was responsible for the study's conceptualisation, research direction, methodology design, supervision and administration of the research process, and manuscript review and editing.

Publications

[1] Y. Zhang, A. Dunn, U. Naseem, and J. Kim, "Intersectional fairness in vision-language models for medical image disease classification" *npj Digit. Med.*, under review.

Acknowledgements

This Thesis would not have been possible without the invaluable support and guidance of exceptional individuals who have shaped both my research journey and personal growth. First and foremost, I wish to express my profound gratitude to my lead supervisor, Professor Jinman Kim. His visionary guidance and unwavering commitment to excellence have been instrumental throughout this research endeavor. Professor Kim's ability to see the broader implications of this work, particularly in advancing equitable healthcare through artificial intelligence, has profoundly shaped the direction and impact of this thesis. His thoughtful insights during critical junctures, his encouragement to pursue rigorous methodological standards, and his dedication to fostering independent thinking have been transformative. I am deeply grateful for his mentorship and wish him continued success in pioneering innovations that bridge technology and healthcare for the benefit of all.

I extend my sincere appreciation to my associate supervisor, Doctor Usman Naseem, whose technical expertise and meticulous attention to detail have been indispensable to this research. Dr. Naseem's profound knowledge of large language models, fairness frameworks, and machine learning architectures provided the technical foundation upon which this work was built. His patient guidance through complex methodological challenges, insightful suggestions on experimental design, and constructive feedback on technical writing have substantially enhanced the quality and rigor of this thesis. His dedication to advancing fair and responsible AI continues to inspire my research aspirations, and I wish him great success in his academic career.

My heartfelt thanks to Professor Adam Dunn for bringing an essential clinical perspective to this research. His expertise in public health and medical informatics ensured that the technical innovations developed in this thesis remain grounded in real-world clinical needs and healthcare equity challenges. Professor Dunn's insights into the practical implications of

algorithmic bias in healthcare settings have been invaluable in framing this work's contributions to both the research community and clinical practice. I am grateful for his thoughtful engagement with this interdisciplinary research.

Beyond my academic supervisors, I am deeply grateful to my partner, Ms. Yuan Meng, whose support and understanding have been a constant source of strength throughout this research journey. Her insights into managing the complexities of graduate study, her encouragement during moments of uncertainty, and her patience during the demanding phases of this work have been immeasurable gifts. Her perspective as a fellow researcher helped me approach obstacles with renewed clarity and resilience, and her belief in the importance of this work has sustained my motivation from beginning to end. I am profoundly grateful for her partnership and the wisdom she has generously shared.

Last but certainly not least, I wish to thank my parents, Mr. Guanxian Zhang and Ms. Xinping Xu, whose unconditional love and steadfast support have been my foundation throughout this academic journey. Their sacrifices, encouragement during challenging times, and unwavering belief in my potential have given me the strength and determination to pursue this research with dedication and purpose. Their support transcends the academic realm, encompassing every aspect of my growth as both a researcher and an individual. I am profoundly grateful for everything they have done, and I hope this achievement brings them joy and pride.

As a final thought, though, I am reminded of Alan Turing's words: "We can only see a short distance ahead, but we can see plenty there that needs to be done."

Contents

List of Figures

List of Tables

Lists of Abbreviations

ACM: Association for Computing Machinery

AI: Artificial Intelligence

ASC: Alignment Score Calculation

AUC: Area Under the Receiver Operating Characteristic Curve

BLIP: Bootstrapping Language-Image Pre-training

BLIP-2: Bootstrapping Language-Image Pre-training 2

CDANN: Conditional Domain-Adversarial Neural Networks

CLIP: Contrastive Language-Image Pre-training

CMAC-MMD: Cross-Modal Alignment Consistency Maximum Mean Discrepancy

CMMD: Conditional Maximum Mean Discrepancy

CNN: Convolutional Neural Network

CT: Computed Tomography

CV: Computer Vision

CVPR: Conference on Computer Vision and Pattern Recognition

DANN: Domain-Adversarial Neural Networks

DDI: Diverse Dermatology Images

DEOdds: Difference in Equalized Odds

DF: Differential Fairness

DL: Deep Learning

DP: Demographic Parity

DPD: Demographic Parity Difference

DRO: Distributionally Robust Optimization

ECE: Expected Calibration Error

EHR: Electronic Health Records

EOdds: Equalized Odds

EOpp: Equality of Opportunity

ERM: Empirical Risk Minimization

ES-AUC: Equity-Scaled Area Under the Curve

EU: European Union

FAccT: Conference on Fairness, Accountability, and Transparency

FAIR: Fair Adversarial Instance Re-weighting

FNR: False Negative Rate

FORML: Fair Optimized Re-weighing with Meta-Learning

FPR: False Positive Rate

GRL: Gradient Reversal Layer

GroupDRO: Group Distributionally Robust Optimization

HSIC: Hilbert-Schmidt Independence Criterion

ICCV: International Conference on Computer Vision

ICDE: International Conference on Data Engineering

ICLR: International Conference on Learning Representations

ICML: International Conference on Machine Learning

IEEE: Institute of Electrical and Electronics Engineers

IF-α: Intersectional Fairness-alpha

ISIC: International Skin Imaging Collaboration

ITM: Image-Text Matching

JAMA: Journal of the American Medical Association

KDE: Kernel Density Estimate

LLaVA: Large Language and Vision Assistant

LLM: Large Language Model

MedCLIP: Medical Contrastive Language-Image Pre-training

MICCAI: Medical Image Computing and Computer Assisted Intervention

MIT: Massachusetts Institute of Technology

ML: Machine Learning

MLP: Multi-Layer Perceptron

MMD: Maximum Mean Discrepancy

MRI: Magnetic Resonance Imaging

NEJM: New England Journal of Medicine

NeurIPS: Conference on Neural Information Processing Systems

NLP: Natural Language Processing

OOD: Out-of-Distribution

PA: Posterior-Anterior

PMC-CLIP: PubMed Central Contrastive Language-Image Pre-training

Q-Former: Querying Transformer

RBF: Radial Basis Function

RKHS: Reproducing Kernel Hilbert Spaces

ROC: Receiver Operating Characteristic

SMOTE: Synthetic Minority Over-sampling Technique

SOTA: State-of-the-Art

TPR: True Positive Rate

U.S.: United States

ViT: Vision Transformer

VLM: Vision-Language Model

VQA: Visual Question Answering

Δ**TPR:** Difference in True Positive Rate

CHAPTER 1

Introduction

Chapter Abstract

This chapter serves as a general introduction to the thesis, establishing the foundation for investigating intersectional fairness in vision-language models (VLM) for medical imaging. The chapter begins by contextualising the transformative rise of artificial intelligence and, more specifically, VLMs in modern medical diagnostics, demonstrating their proven capabilities across radiology, dermatology, and ophthalmology. It then introduces the critical issue of algorithmic fairness, narrows its focus to the significant yet often overlooked challenge of intersectional bias, and establishes the concept of equitable diagnostic certainty as a novel fairness approach. The chapter concludes by articulating the core research problem, presenting the key research questions, outlining the thesis's primary contributions through the Cross-Modal Alignment Consistency Maximum Mean Discrepancy (CMAC-MMD) framework, and providing a roadmap for the subsequent chapters.

1.1 The Rise of Artificial Intelligence in Medical Imaging

The past decade has witnessed a fundamental shift in healthcare delivery, driven by the rapid advancement and clinical integration of artificial intelligence (AI) and machine learning (ML) technologies [1]. Deep learning (DL), a subset of ML characterised by multi-layered neural networks capable of learning hierarchical representations from data, has emerged as a transformative force in medical image analysis [2]. This technological revolution has demonstrated remarkable success across virtually every domain of medical imaging, from radiology and pathology to dermatology and ophthalmology, fundamentally altering how clinicians approach diagnostic challenges [3].

The promise of AI-powered medical imaging systems extends across multiple dimensions of healthcare delivery. First and foremost, these systems have demonstrated the capacity to match or even exceed the diagnostic accuracy of human experts across a wide range of clinical tasks [4]. In radiology, DL algorithms have achieved performance on par with board-certified radiologists in detecting pneumonia, pneumothorax, and pulmonary nodules from chest radiographs [5], [6]. The CheXNeXt algorithm, for instance, demonstrated radiologist-level performance on eleven of fourteen thoracic pathologies while completing image interpretation in a fraction of the time required by human readers [5]. Beyond chest imaging, AI systems have shown exceptional performance in lung cancer screening from computed tomography, achieving an area under the curve (AUC) of 94.4% with significant reductions in both false positive rates (FPR) and false negative rates (FNR) compared to radiologist interpretation [7].

The dermatology domain has witnessed equally impressive advances, with landmark studies demonstrating that convolutional neural networks (CNNs) trained on large-scale clinical image datasets can achieve performance comparable to board-certified dermatologists in classifying skin lesions [8]. These AI systems have shown particular promise for extending specialised diagnostic expertise to resource-limited settings through mobile device deployment, potentially providing access to dermatological assessment for the billions of smartphone users worldwide who lack ready access to specialist care [8]. Subsequent research has further demonstrated that AI-assisted diagnosis not only matches, but in some cases exceeds the

performance of either AI or human experts working independently, with the greatest benefits accruing to less experienced clinicians [9].

In ophthalmology, DL systems have similarly achieved remarkable success in detecting diabetic retinopathy from retinal fundus photographs, with sensitivity and specificity approaching clinical utility thresholds [10], [11]. These advances hold particular significance given the global burden of preventable blindness and the shortage of ophthalmologists in many regions. Prospective multicenter validation studies have demonstrated that AI systems for diabetic retinopathy screening can achieve substantially higher sensitivity than clinical examination, with the EyeArt system achieving 96.5% sensitivity for more-than-mild diabetic retinopathy (mtmDR), compared to 20.6% sensitivity by general ophthalmologists performing dilated ophthalmoscopy, both measured against a rigorous ETDRS photographic reference standard [12]. Notably, general ophthalmologists exhibited near-perfect specificity (99.8%), suggesting a conservative detection pattern rather than outright diagnostic failure, while retina specialists achieved substantially higher sensitivity (59.5%) on the same task.

Beyond diagnostic accuracy, AI systems promise to enhance clinical workflow efficiency dramatically. Automated image interpretation can reduce the time required for radiological assessment from hours to minutes, potentially alleviating workload pressures on overtaxed healthcare systems while accelerating time-to-diagnosis for patients [5]. Most compellingly, AI has the potential to democratise access to high-quality healthcare by making expert-level diagnostic capabilities available in settings where specialist expertise is scarce or absent [8], [13]. This capacity to extend the reach of medical expertise represents a critical opportunity to address persistent healthcare disparities rooted in geographic and socioeconomic factors.

The convergence of AI with human clinical expertise has given rise to what has been termed "high-performance medicine", a paradigm in which AI augments rather than replaces human judgment [1]. Systematic reviews and meta-analyses across hundreds of studies have confirmed that deep learning systems achieve diagnostic accuracy equivalent to healthcare professionals across diverse imaging modalities and clinical conditions [4], [14]. This robust evidence base has moved AI in medical imaging from experimental curiosity to clinical reality, with regulatory approvals and real-world deployments accelerating worldwide. However, this

rapid advancement and clinical integration make ensuring the fairness and equity of these systems not merely an ethical consideration but a clinical imperative [15].

1.2 Vision-Language Models in Medical Diagnostics

While traditional deep learning approaches to medical image analysis have focused primarily on mapping images directly to diagnostic labels, a new class of models has emerged that fundamentally expands the capabilities of AI systems in healthcare. Vision-language models (VLM) represent a shift in how AI systems process and reason about medical information, learning joint representations that bridge visual and textual modalities [16]. These models, exemplified by architectures such as Contrastive Language-Image Pre-training (CLIP) and Bootstrapping Language-Image Pre-training 2 (BLIP2), have demonstrated remarkable capabilities in learning rich, transferable representations from large-scale image-text pairs [16], [17].

The fundamental innovation underlying VLMs lies in their contrastive learning framework, which trains separate image and text encoders to produce aligned representations in a shared embedding space [16]. Rather than learning to map images directly to predefined categories, these models learn more general associations between visual patterns and natural language descriptions. This approach enables powerful capabilities that extend well beyond simple classification, including zero-shot transfer to new tasks without task-specific training, fine-grained visual understanding guided by textual queries, and the generation of natural language descriptions from images [16], [17].

The medical domain presents particularly compelling opportunities for VLMs due to the inherent multi-modality of clinical data. Medical images rarely exist in isolation; they are typically accompanied by rich textual information, including patient history, clinical findings, radiological reports, and diagnostic impressions [18]. VLM can exploit these naturally occurring image-text pairs to learn clinically meaningful representations that capture both the visual appearance of pathology and its semantic interpretation [18]. A typical disease classification task performed by the VLM is illustrated in Figure 1.1 to demonstrate how VLM utilises multi-modality clinical datasets. Early medical adaptations of contrastive learning

demonstrated that models trained on paired radiology images and reports could achieve superior performance with dramatically reduced labelled data requirements compared to traditional approaches [18].

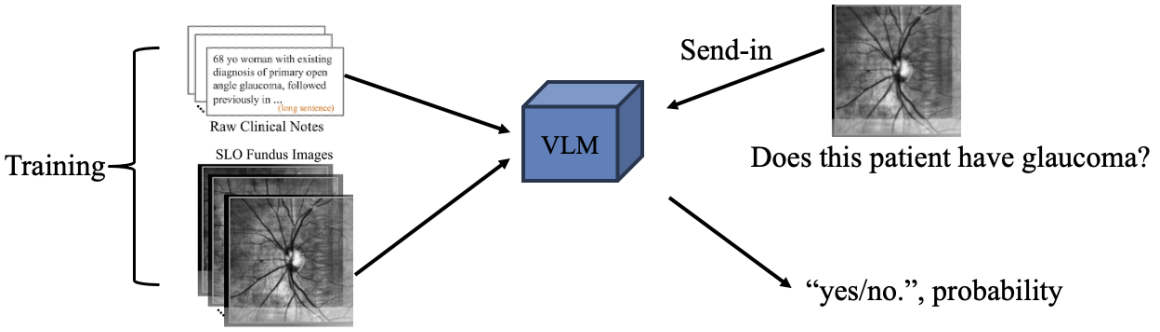

FIGURE 1.1: **An example VLM workflow for medical diagnostics - disease classification.** During training, the model learns from paired clinical notes and medical images (e.g., ophthalmoscopy fundus images). At the inference stage, the model receives a new patient image and a diagnostic query in natural language, producing a probability-based prediction for the clinical question.

The development of domain-specific VLM for medical imaging has accelerated rapidly. BiomedCLIP, trained on fifteen million image-text pairs extracted from PubMed articles, represents a landmark effort to create a biomedical foundation model at unprecedented scale [19]. This model has demonstrated state-of-the-art (SOTA) performance across diverse tasks, including image-text retrieval, multi-label classification, and visual question answering (VQA), even outperforming radiology-specific models in some contexts [19]. Similar efforts have produced specialised models for specific clinical domains, with PMC-CLIP achieving superior performance on retrieval tasks through continued pre-training on medical literature [20], and MedCLIP demonstrating the effectiveness of contrastive learning from unpaired medical images and text using sophisticated semantic matching strategies [21].

The practical capabilities enabled by medical VLMs extend far beyond traditional classification. These models can generate structured radiology reports from images, a task that traditionally requires significant radiologist time and expertise [22]. They excel at VQA, enabling clinicians to query images with natural language and receive interpretable responses [23]. Most significantly, VLM can perform zero-shot or few-shot adaptation to new tasks

with minimal labelled data, addressing a critical bottleneck in medical AI development where obtaining expert annotations is expensive and time-consuming [19].

Recent architectures have pushed the boundaries further by integrating large language models (LLMs) with visual encoders through innovative bridging mechanisms [17]. BLIP2, for instance, employs a Querying Transformer to efficiently connect frozen image encoders with LLM, achieving SOTA performance with significantly fewer trainable parameters [17]. This approach enables sophisticated reasoning about medical images that combines visual understanding with the extensive knowledge captured in language models. The result is systems capable of generating detailed, clinically relevant explanations and engaging in diagnostic dialogue that more closely mirrors human clinical reasoning [24].

The power and flexibility of VLMs have positioned them as the current frontier in medical AI, with clinical deployment accelerating across multiple specialties [13], [24]. However, this rapid advancement brings with it critical challenges that must be addressed before these systems can be safely and equitably deployed at scale. As these models learn from vast quantities of data that inevitably reflect the biases and imbalances present in real-world clinical practice, understanding and mitigating fairness concerns in VLMs has emerged as an urgent research priority [25], [26].

1.3 The Evolving Landscape of Algorithmic Fairness in AI

The remarkable capabilities demonstrated by AI systems in medical imaging have been tempered by mounting evidence that these systems exhibit systematic performance disparities across patient populations defined by demographic attributes such as race, gender, and age [25], [27], [28]. Algorithmic bias in medical AI arises from multiple sources, including imbalanced representation in training data, spurious correlations between demographic attributes and clinical features, and optimisation procedures that prioritise overall performance metrics while ignoring group-level disparities [29]. These biases are not merely technical curiosities but represent profound ethical failures with tangible consequences for patient care and health equity [30].

The imperative for fairness in medical AI extends beyond technical considerations to fundamental questions of justice and equity in healthcare delivery [31]. Biased algorithms risk perpetuating and amplifying existing health disparities that are rooted in structural inequalities, historical discrimination, and unequal access to care [32]. When AI systems systematically underperform for marginalised patient subgroups, they create a dangerous two-tiered system in which algorithmic care replicates the inequities that already plague healthcare systems worldwide [33]. Moreover, the deployment of biased AI systems threatens to erode trust among clinicians and patients, particularly in communities that have experienced historical medical exploitation and discrimination [34].

The challenges of algorithmic fairness in medical imaging first gained widespread attention through studies demonstrating that commercial facial recognition systems exhibited dramatic performance gaps across demographic groups, with error rates for dark-skinned women orders of magnitude higher than for light-skinned men [35]. This intersectional pattern of failure, where systems perform worst for individuals at the intersection of multiple marginalised identities, provided an early warning that AI bias extends beyond single demographic attributes. Subsequent research in medical imaging has confirmed similar patterns of disparate performance [36], [37].

In dermatology, multiple studies have documented that leading AI decision support systems demonstrate substantially degraded performance on images of darker skin tones [38], [39]. This bias is particularly concerning, given that delayed diagnosis of melanoma in patients with darker skin contributes to significantly lower survival rates [40]. The Fitzpatrick17k dataset, created specifically to enable fairness evaluation across diverse skin tones, has revealed that models trained predominantly on lighter skin images fail to generalise to the full spectrum of human skin pigmentation [39]. Similarly, diagnostic models for chest radiograph interpretation have shown significant underdiagnosis biases against female, younger, and minority patients [37], [41].

The theoretical foundations of algorithmic fairness have evolved considerably since the early work established mathematical definitions of group fairness [42]. Individual fairness, which requires that similar individuals receive similar predictions, represents one foundational fairness notion [43]. Demographic parity (DP), which requires that algorithmic decisions

be independent of sensitive attributes, represents one straightforward fairness criterion [44]. However, this definition is often unsuitable for medical applications where disease prevalence legitimately varies across groups [45]. Equalized odds (EOdds), which requires that true positive rates (TPR) and FPR be equal across groups, provides a more clinically appropriate fairness notion by ensuring that the model's errors are distributed equitably [42]. Calibration fairness demands that predicted risks correspond to actual observed risks across all patient subgroups, ensuring that a model's confidence means the same thing regardless of patient demographics [46].

Substantial research effort has been devoted to developing interventions that can mitigate bias in ML systems [44]. These approaches can be broadly categorised based on where they intervene in the modelling pipeline. Pre-processing methods attempt to remove bias from training data through techniques such as resampling underrepresented groups or reweighting samples to equalise group representation [44], [47]. While foundational, these approaches operate only on the input data and provide no guarantee that the trained model will behave fairly at the decision boundary [33], [44]. In-processing methods, fairness constraints are directly incorporated into the model training objective, often through adversarial techniques that penalise the model's ability to predict sensitive attributes from learned representations [48], [49]. Post-processing approaches adjust model outputs after training to satisfy fairness constraints, typically by deriving group-specific decision thresholds [42].

Robust optimisation approaches, such as distributionally robust optimisation and group distributionally robust optimisation, explicitly account for performance on the worst-off subgroup during training [50]. These methods provide theoretical guarantees about worst-case performance but can suffer from the "levelling down" problem, where overall performance degrades to achieve fairness by reducing the accuracy for well-performing groups rather than improving outcomes for disadvantaged groups [33]. Comprehensive frameworks for evaluating and ensuring fairness in medical imaging AI have been proposed, emphasising the need for careful attention to data collection biases, model development practices, and deployment considerations [28], [30].

Despite this substantial body of research, a critical limitation pervades most existing work on algorithmic fairness: the overwhelming focus on single demographic attributes evaluated

in isolation [51]. Traditional fairness evaluations typically assess whether a model performs equitably with respect to race or gender or age, but rarely examine whether fairness holds for patients defined by the intersection of multiple attributes [52]. This single-axis approach to fairness evaluation and mitigation fails to capture the complex, multidimensional nature of bias in real-world settings, where the lived experiences of individuals are shaped by the intersection of multiple social identities [53]. The inadequacy of single-attribute fairness analysis for capturing compounded disparities experienced by intersectional subgroups represents a critical gap that this thesis addresses [54], [55].

1.4 The Critical Challenge of Intersectional Bias in Medical VLMs

The concept of intersectionality, originally developed within critical race theory and feminist scholarship, recognises that individuals occupy multiple social positions simultaneously and that these identities interact to produce unique patterns of privilege and oppression that cannot be understood by examining single attributes in isolation [56]. In the context of algorithmic fairness, intersectionality implies that a model can perform acceptably when evaluated separately for Women and for Black patients, yet exhibit catastrophic failures for Black Women as a distinct intersectional subgroup [35]. This phenomenon represents a fundamental challenge that single-attribute fairness interventions, which optimise for one demographic dimension at a time, are structurally incapable of addressing [57].

1.4.1 Limitations of Single-Attribute Fairness Models

The predominant approach to algorithmic fairness in medical AI has been to evaluate and enforce fairness with respect to individual demographic attributes considered independently [52]. This single-axis concept manifests itself in both fairness metrics, which typically measure performance disparities between groups defined by a single attribute such as race or gender, and fairness interventions, which aim to equalise outcomes across these univariate

group definitions [44]. However, this approach suffers from a critical blind spot: it cannot detect or mitigate biases that emerge specifically at demographic intersections [58].

Recent work has formalised this limitation through the concept of "fairness gerrymandering," where a model appears to satisfy fairness constraints when evaluated along individual demographic axes but exhibits severe disparities when evaluated at intersectional subgroups [58]. Consider a binary classification task evaluated for fairness with respect to both race (White vs. Black) and gender (Male vs. Female). A model might achieve equal error rates between White and Black patients when gender is ignored, and equal error rates between Male and Female patients when race is ignored, as shown in Figure 1.2. However, this provides no guarantee whatsoever about the model's performance on the four intersectional subgroups: White Men, White Women, Black Men, and Black Women [54]. In fact, empirical studies have demonstrated that models satisfying single-attribute fairness criteria can simultaneously exhibit dramatic performance disparities at intersections, with the worst outcomes concentrated among multiply marginalised subgroups [27].

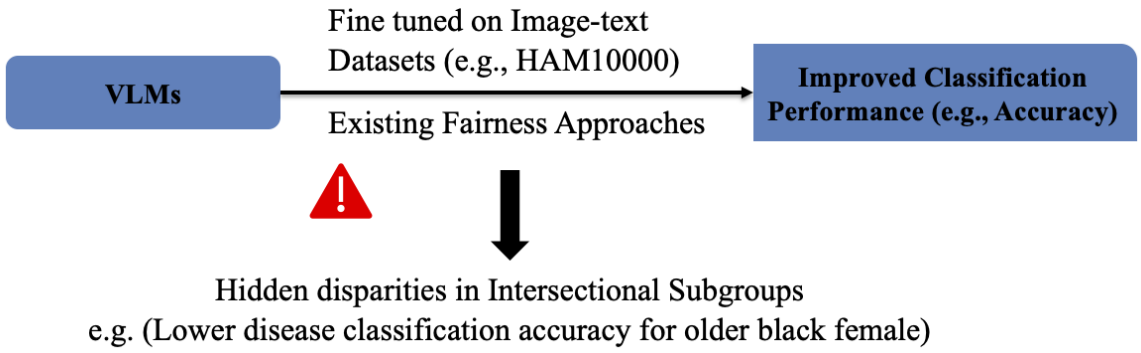

FIGURE 1.2: **The hidden disparities problem in VLMs for medical imaging.** Existing fairness approaches that optimise for overall classification performance metrics often mask systematic disparities at intersectional demographic subgroups, where patients with multiple marginalised identities experience compounded disadvantages in diagnostic accuracy.

The mathematical structure of intersectionality creates an exponential challenge: with k binary demographic attributes, there exist 2^k intersectional subgroups that must be evaluated [51]. As the number of relevant demographic dimensions increases, comprehensive intersectional analysis becomes computationally intensive and data-hungry, requiring sufficiently large sample sizes within each intersectional subgroup to enable reliable metric estimation [29]. This practical constraint has led much research to limit attention to single attributes or to

coarse demographic categorisations, inadvertently concealing the very disparities that most urgently require attention [28].

1.4.2 The Need for a New Approach: Decision-Level Fairness and Equitable Diagnostic Certainty

A fundamental yet often overlooked limitation of existing fairness interventions lies in where they operate within the model architecture. The vast majority of fairness-aware ML methods target the feature level, attempting to learn representations that are statistically independent of sensitive attributes or equalising the distribution of learned features across demographic groups [48], [49]. These representation-level interventions, while mathematically elegant, suffer from a critical weakness: they provide no direct guarantee about the fairness of the model's actual diagnostic outputs [33].

This thesis introduces a novel fairness paradigm centered on the concept of equitable diagnostic certainty. Consider a binary classification task, such as distinguishing malignant from benign skin lesions. A model produces a confidence alignment score for each prediction across two modalities (image and natural language), and a clinical decision is made by comparing the score to a threshold to classify disease labels. Even when a model is fine-tuned to achieve similar accuracy across patient subgroups, it may exhibit profound disparities in the certainty with which it makes predictions. For a privileged subgroup, the model's confidence scores might be concentrated well above or well below the decision threshold, indicating clear, decisive predictions. In contrast, for a marginalised subgroup, confidence scores might cluster dangerously close to the threshold range, creating a "grey area" of diagnostic uncertainty.

This disparity in diagnostic certainty, visualised empirically in the research leading to this thesis, creates significant clinical risks that are not captured by conventional accuracy-based fairness metrics. Predictions that fall close to the decision threshold are inherently unstable; minor perturbations in image quality, acquisition parameters, or other factors can easily flip the classification [52]. Patients from subgroups receiving systematically low-confidence predictions thus face greater vulnerability to misclassifications from real-world data variations [25]. Moreover, this confidence gap erodes clinical trust, as experienced clinicians

recognise that borderline predictions warrant greater scrutiny and may defer to algorithmic recommendations more readily when the system expresses clear confidence [59].

Existing fairness interventions are ill-equipped to address diagnostic certainty disparities because they operate primarily on abstract internal representations rather than on the functional outputs that drive clinical decisions [33]. Data-centric pre-processing approaches such as resampling or reweighting alter the composition of the training data but cannot directly govern the model's behaviour at the decision boundary [33]. Sophisticated in-processing methods employing adversarial training or robust optimisation enforce statistical properties on learned features but do not explicitly constrain the distribution of model confidence across subgroups [48], [50]. Even methods specifically designed for VLMs, such as FairCLIP, optimise for fairness with respect to single attributes and fail to address intersectional disparities or certainty gaps [57].

1.4.3 Clinical Risks and Trust Deficits from Biased AI

The technical concept of intersectional bias and confidence disparities translates directly into severe clinical consequences with profound implications for patient outcomes and health equity. In dermatology, where melanoma survival rates are dramatically lower for Black patients compared to White patients, largely due to delays in diagnosis attributable to late-stage disease at presentation, an AI system that systematically exhibits lower diagnostic certainty for darker skin tones compounds existing disparities [40]. If such a system is deployed in primary care settings to assist with triage decisions, the reduced diagnostic certainty for minority patients may lead to more false negative screens, delaying referrals for specialist evaluation and definitive diagnosis [38].

In ophthalmology, glaucoma disproportionately affects Black and Hispanic populations, both in terms of prevalence and severity of vision loss [60], and AI-assisted diagnostic systems are increasingly being integrated into clinical glaucoma management to improve detection accuracy and treatment planning [61]. An AI screening system that performs less reliably for these high-risk demographic subgroups represents a particularly insidious form of algorithmic harm: deploying technology ostensibly to improve access and early

detection while perpetuating the very disparities it was intended to address [62]. Recent evidence demonstrates that even SOTA VLMs exhibit demographic bias across multiple medical imaging tasks, with the largest performance gaps observed precisely at demographic intersections [26]. Analogous sociodemographic biases have been documented in the medical decision-making of large language models, suggesting that bias in language-augmented clinical AI is a systemic rather than modality-specific phenomenon [63].

Beyond individual patient harm, biased AI systems threaten the broader trust relationship between healthcare institutions and the communities they serve [64]. Subgroups that have experienced historical medical exploitation and discrimination are understandably skeptical of technological interventions, particularly when these systems demonstrably perpetuate bias [34]. The erosion of trust carries cascading consequences: reduced willingness to seek care, decreased adherence to screening recommendations, and diminished participation in research that could improve these very systems [34]. Addressing intersectional fairness in medical AI is therefore not merely a technical challenge but a prerequisite for the ethical deployment of these powerful technologies in diverse clinical patient subgroups [31].

1.5 Research Questions

The preceding analysis establishes intersectional bias in VLMs for medical imaging as a critical, multifaceted challenge sitting at the intersection of technical capability, algorithmic fairness, and clinical translation. This thesis addresses this challenge through three interconnected research questions:

RQ1: How do SOTA VLMs exhibit intersectional bias in medical diagnostic tasks, and what are the limitations of single-attribute fairness metrics in capturing these compounded disparities?

This question motivates a comprehensive empirical evaluation of existing VLM architectures across medical imaging datasets that contain sufficient demographic annotation to enable intersectional analysis. The investigation will quantify performance disparities using both traditional single-attribute fairness metrics and rigorous intersectional fairness measures,

demonstrating the inadequacy of conventional evaluation paradigms for detecting bias at demographic intersections.

RQ2: Can a new fairness framework be developed that moves beyond feature-level adjustments to directly mitigate bias at the decision level by regularising a model's diagnostic confidence across intersectional subgroups?

This question drives the core methodological contribution of the thesis: the development of Cross-Modal Alignment Consistency via Maximum Mean Discrepancy (CMAC-MMD), a novel training strategy that enforces fairness not on abstract internal representations but on the distribution of decision confidence scores across all intersectional subgroups. The framework operationalises the concept of equitable diagnostic certainty, ensuring that no patient subgroup is systematically subjected to uncertain, borderline predictions.

RQ3: How effective is the proposed CMAC-MMD framework in reducing intersectional bias while maintaining or improving overall diagnostic performance on established medical imaging benchmarks?

This question requires rigorous experimental validation across multiple clinical domains, comparison against existing fairness interventions, and evaluation on external datasets to assess generalisation. The investigation will demonstrate that decision-level fairness approach achieves superior fairness-accuracy trade-offs compared to conventional approaches while proving robust to distribution shift.

1.6 Thesis Contribution and Structure

The primary contribution of this thesis is the development and validation of CMAC-MMD, a novel decision-level fairness regularisation framework for VLMs in medical imaging. Unlike existing approaches that operate on feature representations, CMAC-MMD directly aligns the distribution of diagnostic confidence scores across intersectional patient subgroups. By computing per-sample alignment scores that quantify the model's certainty in distinguishing correct from incorrect image-text pairs, and then minimising the Maximum Mean Discrepancy (MMD) between these score distributions across all demographic intersections, CMAC-MMD

ensures equitable diagnostic certainty without requiring demographic attributes as model input during inference time.

Empirical validation demonstrates that CMAC-MMD reduces intersectional performance disparities by twenty to thirty-five percent across both dermatology and ophthalmology benchmarks, as measured by advanced metrics including Differential Fairness (DF) and Intersectional Fairness-α (IF-α). Critically, these fairness improvements are achieved while maintaining or even slightly improving overall diagnostic performance, avoiding the "levelling down" problem that plagues many fairness interventions. External validation on independent datasets confirms that the fairness benefits generalise under distribution shift, suggesting that the method learns fundamental equitable representations rather than dataset-specific adjustments.

The thesis is structured as follows. Chapter 1, the present chapter, has established the research problem, context, and questions, taking the reader from the broad potential of AI in medicine through the specific frontier of VLMs to the critical unsolved challenge of intersectional bias. Chapter 2 will provide a comprehensive literature review, situating this work within the extensive bodies of research on medical image analysis, vision-language architectures, algorithmic fairness theory, and fairness interventions. Chapter 3 presents the core methodological contribution of the thesis, detailing the CMAC-MMD framework, its theoretical foundations, implementation, experimental design, and comprehensive results across multiple datasets and baselines. Chapter 4 concludes the thesis by synthesising the findings, discussing their implications for clinical AI deployment, acknowledging limitations, and outlining directions for future research toward achieving equitable, trustworthy medical AI systems.

1.7 Chapter Conclusion

This chapter has established the foundation for investigating intersectional fairness in VLMs for medical imaging by tracing a narrative arc from promise to peril. Beginning with the transformative potential of AI in medical imaging, as evidenced by systems achieving expert-level diagnostic performance across multiple specialties in medical imaging domains, the

chapter demonstrated the rapid advancement of VLMs as the current frontier in medical AI. These powerful systems, capable of reasoning jointly about images and text, unlock capabilities far beyond traditional classification, including report generation, VQA, and sophisticated diagnostic dialogue.

However, this technological progress has been tempered by mounting evidence of systematic algorithmic bias, with AI systems exhibiting profound performance disparities across patient subgroups defined by demographic attributes. Most critically, the chapter has argued that the predominant focus on single-attribute fairness evaluation and mitigation represents a fundamental blind spot, failing to detect or address biases that emerge specifically at demographic intersections. The concept of equitable diagnostic certainty was introduced as a novel approach to fairness that extends beyond accuracy metrics to ensure that models express similar confidence across all patient subgroups, avoiding the clinical risks posed by systematic diagnostic uncertainty for marginalised identities.

Having established the severity and urgency of intersectional bias in medical VLMs and having articulated the limitations of existing approaches that operate at the feature level, this thesis is positioned to make a significant contribution through decision-level fairness regularisation. The following chapter will provide a comprehensive review of the existing literature across the intersecting domains of medical AI, vision-language architectures, fairness theory, and bias mitigation, establishing precisely where the current state of knowledge ends and where this thesis's contributions begin.

CHAPTER 2

Literature Review

Chapter Abstract

This chapter provides a comprehensive review of the literature that is foundational to understanding and addressing intersectional fairness in medical vision-language models (VLM). The review synthesises research across six interconnected technical domains: VLM architectures and their medical adaptations, medical imaging datasets with demographic annotations, algorithmic fairness theory and metrics, fairness intervention methodologies, empirical evidence of bias in clinical AI systems, and the technical foundations enabling modern multimodal learning. The chapter begins by establishing the methodological approach to literature synthesis, then traces the evolution from traditional supervised learning to contemporary VLM architectures in medical diagnostics. It systematically examines theoretical fairness notions from demographic parity (DP) to equalised odds (EOdds), extending into intersectional frameworks including Differential Fairness (DF) and Intersectional Fairness-α. A critical analysis of existing fairness interventions spanning pre-processing, in-processing, and post-processing methods reveals their strengths and fundamental limitations when confronted with intersectional bias at the decision level. By integrating empirical evidence from dermatology, radiology, and ophthalmology demonstrating systematic performance disparities across demographic intersections, this review establishes both the urgency and the inadequacy of current approaches. The chapter culminates in identifying a precise research gap: the absence of decision-level fairness interventions that directly regularise diagnostic certainty across intersectional patient subgroups in vision-language architectures. This synthesis provides the theoretical scaffolding and empirical justification for the Cross-Modal Alignment Consistency Maximum Mean Discrepancy (CMAC-MMD) framework introduced in subsequent chapters.

2.1 Methodology of the Literature Review

This literature review adopts a comprehensive approach to synthesising research across multiple intersecting domains that are foundational to understanding and addressing intersectional fairness in medical vision-language models (VLM). The methodology employed combines elements of scoping review practices with domain-specific search strategies, designed to ensure comprehensive coverage while maintaining practical feasibility given the breadth of relevant technical literature. This section establishes the scope, research questions, search strategies, and selection criteria that guided the literature synthesis process.

2.1.1 Scope and Research Questions

The scope of this literature review extends across six primary technical domains that collectively provide the theoretical, empirical, and methodological foundations for this thesis. Table 2.1 provides a structured overview. These domains encompass VLM architectures and their evolution, medical imaging datasets with demographic annotations enabling fairness analysis, algorithmic fairness theory and evaluation metrics, fairness intervention methodologies across the machine learning (ML) pipeline, empirical evidence documenting bias in clinical artificial intelligence (AI) systems, and the technical foundations including contrastive learning, self-supervised learning, domain adaptation, kernel methods, and multimodal fusion architectures.

The literature synthesis was guided by four overarching research questions that frame the investigation. First, what are the state-of-the-art VLM architectures, and how have these general-purpose models been adapted for medical diagnostic applications across radiology, dermatology, and ophthalmology? This question necessitated comprehensive coverage of foundational models including Contrastive Language-Image Pre-training (CLIP) [16], Bootstrapping Language-Image Pre-training (BLIP) [65], BLIP-2 [17], and Large Language and Vision Assistant (LLaVA) [66], as well as domain-specific adaptations such as Medical CLIP (MedCLIP) [21], BiomedCLIP [19], PMC-CLIP [20], PubMedCLIP [67], and BioViL [68].

TABLE 2.1: Overview of the Six Technical Domains Addressed in the Literature Review

Domain	Scope and Key Topics	Coverage in Chapter
Vision-Language Model Architectures	General-purpose models (CLIP, BLIP, BLIP-2, LLaVA, Florence-2); medical adaptations (MedCLIP, BiomedCLIP, PMC-CLIP, PubMedCLIP, BioViL); architectural innovations and pre-training strategies	Section 2.2: The Application of Vision-Language Models in Medical Diagnostics
Medical Imaging Datasets	Datasets with demographic annotations enabling fairness analysis; representation disparities across modalities (dermatology, radiology, ophthalmology); data collection biases	Integrated throughout Sections 2.2–2.4; specific datasets detailed in empirical evidence sections
Algorithmic Fairness Theory	Theoretical fairness notions (demographic parity, equalized odds, calibration); impossibility results; intersectional frameworks (max-min fairness, Differential Fairness, Intersectional Fairness-α); evaluation metrics	Section 2.3: Algorithmic Fairness in Medical AI: Definitions, Metrics, and Harms
Fairness Intervention Methodologies	Pre-processing (reweighting, resampling); in-processing (adversarial debiasing, distributionally robust optimization); post-processing (threshold optimization); VLM-specific methods	Section 2.4: A Critical Review of Fairness Intervention Methodologies
Empirical Evidence of Bias	Documented performance disparities across imaging modalities; clinical consequences for patient outcomes; magnitude of intersectional disparities; trust and deployment challenges	Section 2.3.4: Empirical Evidence of Clinical Harms; Section 2.4: The Critical Challenge of Intersectionality
Technical Foundations	Contrastive learning and self-supervised methods; domain adaptation techniques; kernel methods and Maximum Mean Discrepancy; multimodal fusion architectures; optimisation frameworks	Integrated throughout Sections 2.2 and 2.4; Section 2.5.3: MMD theory and applications

Understanding the architectural innovations, pre-training strategies, and task-specific fine-tuning approaches employed by these models provides essential context for analysing their fairness properties and developing interventions.

Second, how is fairness defined, measured, and operationalised in the context of medical AI, and what are the theoretical foundations and practical limitations of existing fairness notions? This question required examination of foundational work establishing group fairness criteria, including demographic parity (DP) [44], [69], equalised odds (EOdds) [42], calibration [46], and the mathematical impossibility results demonstrating fundamental trade-offs between these notions [45]. The review extends into emerging intersectional fairness frameworks, including max-min fairness [58], Differential Fairness (DF) [54], and Intersectional Fairness-α (IF-α) [55] that explicitly address compounded bias across multiple demographic attributes. Understanding these theoretical constructs and their mathematical formulations is essential for rigorously evaluating model fairness and interpreting empirical findings.

Third, what is the documented empirical evidence of bias in medical AI systems across imaging modalities, and what are the clinical consequences of algorithmic disparities for patient outcomes and health equity? This question drove the synthesis of landmark studies demonstrating systematic performance disparities in chest radiography [37], [70], [71], dermatology [38], [39], ophthalmology [62], and resource allocation algorithms [32]. The review critically examines the magnitude of disparities, the demographic groups most severely affected, the mechanisms through which bias arises, and the translation of statistical fairness violations into tangible clinical harms, including delayed diagnosis, misallocation of healthcare resources, and erosion of trust in healthcare institutions among marginalised communities.

Fourth, what technical interventions have been proposed to mitigate bias in ML systems generally and in medical imaging specifically, and what are their theoretical guarantees, empirical effectiveness, and fundamental limitations when confronted with intersectional bias? This question necessitated comprehensive coverage of pre-processing methods including reweighting and resampling [47], [72], in-processing approaches including adversarial debiasing [48], [49] and distributionally robust optimisation (DRO) [50], post-processing threshold optimisation [42], and VLM-specific methods including FairCLIP [57], FairerCLIP [73], and DeAR [74]. Critical analysis of these interventions reveals that the vast majority operate at the feature representation level and address single demographic attributes in isolation, leaving a clear gap for decision-level intersectional fairness approaches.

2.1.2 Search Strategy and Paper Selection

The literature search employed a multi-pronged strategy combining systematic database queries, citation tracking, and expert consultation to ensure comprehensive coverage of relevant research. Primary searches were conducted across five major academic databases: Google Scholar for broad coverage including preprints and grey literature, PubMed for biomedical and clinical publications, Association for Computing Machinery (ACM) Digital Library for computer science conference proceedings, Institute of Electrical and Electronics Engineers (IEEE) Xplore for engineering and applied AI publications, and arXiv for recent pre-publication manuscripts representing cutting-edge developments. The search strategy employed a structured combination of keywords and Boolean operators designed to capture relevant publications across the diverse technical domains encompassed by this review.

The primary keyword categories included fairness and bias terms such as "intersectional fairness", "algorithmic bias", "demographic disparity", "equalized odds", "calibration", "group fairness", and "fairness gerrymandering". VLM terms included "CLIP", "BLIP", "vision-language model", "multimodal learning", "contrastive learning", "image-text pre-training", and "zero-shot learning". Medical AI and imaging terms encompassed "medical imaging", "radiology AI", "dermatology AI", "chest X-ray classification", "skin lesion detection", "fundus imaging", and "glaucoma detection". Technical method terms included "Maximum Mean Discrepancy", "MMD", "domain adaptation", "adversarial training", "distributionally robust optimization", "self-supervised learning", and "transfer learning". These keyword combinations were iteratively refined based on initial search results to balance precision and recall.

Inclusion criteria for paper selection were established to ensure relevance, quality, and recency while acknowledging the importance of foundational works regardless of publication date. Papers were included if they addressed VLM architectures or their medical adaptations, algorithmic fairness theory or evaluation in ML systems, fairness interventions at any stage of the ML pipeline, empirical studies documenting bias in medical AI systems, medical imaging datasets with demographic annotations, or technical foundations including contrastive learning, kernel methods, domain adaptation, and multimodal fusion relevant to fairness-aware

model development. Methodological inclusion criteria required peer-reviewed publications in established conferences or journals, with selective inclusion of highly-cited preprints for rapidly-evolving topics where peer-reviewed publication lags behind current practice.

Exclusion criteria eliminated papers that focused exclusively on natural language processing (NLP) or computer vision (CV) tasks without medical or fairness relevance, addressed individual fairness or causal fairness frameworks outside the scope of this thesis, reported only preliminary results without validation, or lacked sufficient methodological detail to enable critical evaluation. The temporal scope prioritised publications from the past five to seven years, reflecting the recent acceleration of research in VLMs and intersectional fairness, while including seminal foundational papers from earlier periods that established key theoretical constructs or empirical methods still in current use.

The search and selection process yielded a total of 143 academic references spanning 2002 to 2025, with particular concentration in the period from 2020 to 2025 reflecting the explosion of VLM research following the introduction of CLIP [16] and accelerating attention to fairness in medical AI catalysed by high-impact publications documenting systematic bias in chest radiography [37], healthcare resource allocation [32], and medical imaging more broadly [70]. The distribution across publication venues demonstrates the interdisciplinary nature of this research, with representation from premier AI conferences including International Conference on Machine Learning (ICML), Conference on Neural Information Processing Systems (NeurIPS), Conference on Computer Vision and Pattern Recognition (CVPR), International Conference on Computer Vision (ICCV), International Conference on Learning Representations (ICLR), and Conference on Fairness, Accountability, and Transparency (FAccT); medical imaging venues including Medical Image Computing and Computer Assisted Intervention (MICCAI), Medical Image Analysis, and Radiology Artificial Intelligence; clinical journals including Journal of the American Medical Association (JAMA), Nature Medicine, Science Advances, and npj Digital Medicine; and technical journals covering ML theory and applications. This breadth ensures that the literature review synthesises perspectives from computer science, clinical medicine, bioethics, and social science, reflecting the inherently interdisciplinary nature of ensuring fairness in medical AI systems.

2.2 The Application of Vision-Language Models in Medical Diagnostics

The landscape of medical image analysis has undergone a fundamental transformation over the past decade, evolving from task-specific supervised learning to powerful multimodal architectures capable of reasoning jointly about visual and textual information. This section traces this evolution, beginning with traditional convolutional neural network (CNN) approaches, examining the architectural innovations underlying modern VLMs, and detailing the domain-specific adaptations that have positioned these systems at the forefront of medical AI. Understanding this technical progression is essential for contextualising the fairness challenges that emerge when these powerful but complex systems encounter the biases inherent in medical training data.

2.2.1 From Supervised CNNs to Self-Supervised Multimodal Learning

The initial wave of deep learning success in medical imaging was built upon supervised CNNs trained to map images directly to diagnostic labels. Foundational work demonstrated that CNNs trained on large-scale annotated datasets could match or exceed human expert performance across diverse medical imaging tasks [2], [8], revolutionising computer-aided diagnosis across radiology, pathology, and dermatology. Deep learning systems achieved radiologist-level performance in detecting pneumonia from chest radiographs [5], dermatologist-level accuracy in classifying skin lesions [75], and ophthalmologist-level sensitivity in identifying diabetic retinopathy from retinal fundus photographs [10], [11].

The supervised learning paradigm, however, imposed significant practical limitations that constrained the development and deployment of medical AI systems. First and foremost, supervised approaches require large quantities of expert-annotated training data, a resource that is expensive, time-consuming to acquire, and subject to inter-observer variability that introduces label noise [76]. Obtaining annotations for medical images typically requires board-certified specialists spending substantial time reviewing cases, creating a fundamental bottleneck that limits the scale and diversity of training datasets. Second, supervised models

trained on specific tasks and datasets often exhibit poor generalisation to new clinical contexts, imaging equipment, or patient populations not represented in the training distribution, a limitation of particular concern for fairness when training data systematically under-represents certain demographic groups [25]. Third, traditional supervised learning approaches learn purely from visual patterns without incorporating the rich contextual information present in clinical reports, patient histories, and domain knowledge encoded in medical literature [18].

These limitations catalysed a shift toward self-supervised and multimodal learning approaches that could leverage the vast quantities of unlabelled or naturally-paired medical data available in clinical practice and scientific literature. Self-supervised learning, which trains models to predict unobserved parts of the input from observed parts without requiring manual annotations, has demonstrated remarkable effectiveness in medical imaging across diverse modalities, including computed tomography (CT), magnetic resonance imaging (MRI), X-ray, histology, and ultrasound [77], [78]. By pre-training on large unlabelled datasets through predictive tasks such as context restoration [79], rotation prediction, or contrastive instance discrimination [80], self-supervised models learn rich visual representations that can be fine-tuned with dramatically reduced quantities of labelled data for downstream diagnostic tasks.

The emergence of VLMs represents a further evolution that transcends both traditional supervised learning and single-modality self-supervised approaches by learning joint representations that bridge visual and textual modalities. Rather than learning to map images to discrete diagnostic categories, VLMs learn associations between visual patterns and natural language descriptions, enabling capabilities that extend well beyond classification, including image-text retrieval, visual question answering (VQA), report generation, and zero-shot transfer to new tasks through text-based task specification [16], [24], [65]. The fundamental insight underlying these models is that medical images rarely exist in isolation; they are accompanied by rich textual information, including radiological reports, clinical impressions, pathological descriptions, and patient histories that provide semantic context for visual findings [22]. By explicitly modelling the relationship between images and their associated text through contrastive or generative objectives, VLMs can learn clinically meaningful representations that capture both the visual appearance of pathology and its semantic interpretation.

2.2.2 Foundational Vision-Language Architectures

The contemporary landscape of VLMs is built upon several foundational architectures that established the core technical contribution subsequently adapted for medical applications. Understanding these general-purpose models and their training methodologies provides essential context for the domain-specific medical variants examined in the following subsection.

CLIP employs a dual-encoder architecture consisting of separate image and text encoders that are jointly trained on 400 million image-text pairs collected from the internet through a contrastive learning objective [16]. The image encoder, typically implemented as either a ResNet or Vision Transformer (ViT) architecture, processes input images to produce fixed-dimensional embedding vectors. The text encoder, implemented as a Transformer architecture, processes natural language descriptions to produce embedding vectors in the same semantic space. The training objective maximises the cosine similarity between embeddings of matched image-text pairs while minimising similarity for mismatched pairs within each training batch, implemented through the InfoNCE contrastive loss [81].

The power of CLIP's approach lies in several key capabilities that emerge from this large-scale contrastive pre-training. First, the learned image and text embeddings occupy a shared semantic space in which images and their descriptions are proximal, enabling cross-modal retrieval where images can be queried with text and vice versa. Second, CLIP achieves remarkable zero-shot transfer capabilities, performing classification tasks without task-specific training by embedding textual descriptions of target classes and selecting the class whose text embedding is most similar to the image embedding. This zero-shot capability is particularly valuable for medical applications where rare diseases may have limited labelled training examples. Third, the representations learned by CLIP demonstrate strong transfer learning properties, serving as effective initialisation for downstream tasks that can be fine-tuned with smaller quantities of task-specific labelled data [82].

The BLIP framework advanced beyond CLIP's contrastive-only approach by introducing a unified architecture capable of both understanding tasks like image-text retrieval and generation tasks like image captioning [65]. BLIP employs a multimodal mixture of encoder-decoder architecture that shares parameters across understanding and generation objectives,

trained using three complementary losses: image-text contrastive loss similar to CLIP, image-text matching (ITM) loss that predicts whether an image-text pair is matched or mismatched, and language modelling loss that generates text conditioned on images. A key innovation in BLIP is the captioning and filtering strategy applied to web-scraped training data, in which a captioner model generates synthetic captions for images and a filter model removes noisy captions, thereby bootstrapping data quality to improve model performance. This unified approach enables BLIP to achieve state-of-the-art performance across both discriminative tasks, such as retrieval, and generative tasks, such as caption generation, with particularly strong performance on VQA that requires reasoning about image content [65].

BLIP-2 achieved dramatic improvements in computational efficiency and performance through architectural innovations that enable effective leverage of pre-trained large language models (LLM) [17]. BLIP-2 employs a three-component architecture consisting of a frozen image encoder, a frozen LLM, and a lightweight Querying Transformer (Q-Former) that bridges between them. The Q-Former, a learnable component with only 188 million trainable parameters, extracts visual features from the frozen image encoder and aligns them with the text space of the frozen language model through a two-stage pre-training process. In the first stage, the Q-Former is pre-trained with vision-language tasks including image-text contrastive learning, image-grounded text generation, and ITM, learning to extract the most informative visual features for language-related tasks. In the second stage, the output of the Q-Former is used as a soft visual prompt to condition the frozen language model for generative tasks. This architecture enables BLIP-2 to achieve state-of-the-art performance on vision-language tasks while requiring significantly fewer trainable parameters and less training compute compared to end-to-end approaches [17].

The LLaVA model family represents another major direction in VLM development focused on visual instruction tuning [66]. LLaVA connects a vision encoder and an LLM through a simple projection layer and is trained on a novel instruction-following dataset where multimodal instruction-following data is generated by reformulating image-text pairs into a conversational format. This approach enables the model to follow natural language instructions for diverse vision-language tasks, engaging in multi-turn dialogue about images that more closely mirrors human-like visual understanding and reasoning. The instruction-tuning demonstrated by

LLaVA has proven highly effective for creating general-purpose vision-language assistants capable of engaging in open-ended dialogue about visual content, a capability with clear applications for medical image interpretation where clinicians may pose diverse questions about radiological findings, pathological features, or differential diagnoses [66].

Florence-2 demonstrated that lightweight VLM architectures can achieve strong performance through unified prompt-based interfaces and comprehensive task coverage during pre-training [83]. With parameters ranging from 232 to 771 million, considerably smaller than BLIP-2 or LLaVA, Florence-2 achieves competitive performance by training on an extensive task taxonomy covering captioning, object detection, grounding, segmentation, and VQA using a sequence-to-sequence architecture with task-specific prompts. This work establishes that architectural efficiency combined with comprehensive task coverage during pre-training can yield highly capable vision-language systems without requiring massive parameter counts, an important consideration for deployment in resource-constrained clinical environments [83].

2.2.3 Domain Adaptation for Medicine

The general-purpose VLMs described in the previous subsection, while demonstrating impressive capabilities on natural images and web-derived text, require substantial adaptation to excel at medical imaging tasks characterised by specialised visual features, domain-specific terminology, and diagnostic reasoning patterns. This subsection examines the landscape of medical VLMs, detailing the architectural modifications, pre-training strategies, and datasets employed to adapt general VLM for clinical applications.

MedCLIP represents one of the earliest efforts to adapt contrastive language-image pre-training specifically for medical imaging [21]. MedCLIP addresses a fundamental challenge in medical domain adaptation: the scarcity of paired image-text data at the scale available for general-purpose CLIP training. Medical images in clinical practice are often accompanied by unstructured reports rather than concise captions, and extracting high-quality image-text pairs from electronic health records (EHR) poses significant technical and privacy challenges. MedCLIP introduces a decoupling strategy that enables training from unpaired images and text by learning a semantic matching function that predicts whether an image-text pair is

semantically related without requiring ground-truth pairing during training. This approach leverages large quantities of medical images and reports that co-occur in the same patient studies without requiring explicit annotation of which specific image corresponds to which specific finding in the report. The semantic matching is learned through a combination of contrastive learning on pseudo-paired data constructed through semantic similarity matching and explicit training of a matching function using limited quantities of annotated pairs. MedCLIP demonstrated strong performance on retrieval and zero-shot classification tasks across multiple medical imaging modalities, establishing the feasibility of adapting vision-language pre-training for medical applications despite data scarcity constraints [21].

PubMedCLIP took a different adaptation approach by leveraging the vast quantities of scientific image-text pairs available in biomedical literature [67]. PubMedCLIP applies the CLIP training paradigm to image-caption pairs extracted from PubMed Central articles, creating a large-scale pre-training dataset of scientific medical images with their associated figure captions. This approach provides access to substantially larger training data compared to clinical datasets while maintaining medical domain relevance. PubMedCLIP demonstrated that pre-training on scientific literature provides strong transfer learning for clinical VQA tasks, outperforming ImageNet pre-training and approaching the performance of models trained on clinical data despite the domain shift between scientific figures and clinical images [67].

PMC-CLIP extended the scientific literature pre-training approach to an even larger scale and incorporated additional refinements for medical domain adaptation [20]. PMC-CLIP was trained on image-text pairs extracted from 16.4 million PubMed Central articles, representing one of the largest medical vision-language pre-training efforts to date. Key technical innovations include the use of biomedical-specific text encoders pre-trained on PubMed abstracts to better capture medical terminology and conceptual relationships, careful filtering of image-text pairs to remove low-quality or irrelevant scientific figures, and multi-stage training that progressively fine-tunes from general medical knowledge to task-specific objectives. PMC-CLIP achieved state-of-the-art performance on medical image-text retrieval tasks and demonstrated strong transfer learning capabilities across diverse downstream medical imaging applications [20].

BiomedCLIP represents the current state-of-the-art in medical vision-language pre-training at massive scale [19]. BiomedCLIP was trained on PMC-15M, a dataset containing 15 million figure-caption pairs extracted from 4.4 million scientific articles, making it the largest medical vision-language dataset used for pre-training to date. The model employs a ViT architecture for image encoding and a domain-adapted text encoder, specifically initialised from PubMedBERT, to capture biomedical terminology and conceptual relationships. BiomedCLIP achieved state-of-the-art performance across numerous medical imaging benchmarks including PathVQA for pathology VQA, VQA-RAD for radiology VQA, MedMNIST for multi-disease classification, and retrieval tasks across diverse medical imaging modalities. The model demonstrated particular strength in zero-shot transfer to new medical imaging tasks and robust performance across different imaging modalities including radiology, pathology, and dermatology without task-specific fine-tuning [19].

BioViL and its text-enhanced variant BioViL-T focused specifically on chest X-ray interpretation through a combination of self-supervised learning and vision-language pre-training [68]. BioViL employs a dual-encoder architecture similar to CLIP but incorporates domain-specific architectural and training refinements tailored for radiology applications. The model is pre-trained on MIMIC-CXR, a large-scale dataset of chest radiographs paired with radiological reports [84], using contrastive learning between image regions and report sentences. A key innovation in BioViL is the incorporation of radiology-specific data augmentations and the use of CheXbert, an NLP model trained to extract clinical findings from radiology reports, to provide structured supervision during pre-training. BioViL-T extends this approach by incorporating temporal information from longitudinal patient studies, learning to reason about disease progression and change over time. These models achieved state-of-the-art performance on chest X-ray classification and phrase grounding tasks, demonstrating the value of radiology-specific architectural and training refinements [68].

The landscape of medical VLMs extends beyond these major systems to include numerous specialised variants optimised for specific imaging modalities, anatomical regions, or clinical tasks. GLoRIA demonstrated that global-local representation learning between image subregions and report sentences can achieve label-efficient medical image recognition, establishing an influential alternative to purely global contrastive objectives [85]. MedVInT

introduced architectural innovations for handling variable-length medical reports and complex medical entity relationships [23]. CXR-BERT and related models have explored bidirectional vision-language architectures that enable both image-to-text and text-to-image reasoning for comprehensive radiology report understanding [86].

This diverse ecosystem of medical VLMs shares several common architectural patterns and training strategies that distinguish medical adaptations from general-purpose VLMs. First, medical VLMs typically employ domain-adapted text encoders initialised from models pre-trained on biomedical literature such as BioBERT, PubMedBERT, or SciBERT, rather than general-purpose language models, to better capture medical terminology and conceptual relationships [87], [88]. Second, medical VLMs often incorporate radiology-specific or pathology-specific data augmentations that reflect the types of variations present in clinical imaging including rotations, intensity variations, and anatomical crop variations that preserve diagnostic content. Third, many medical VLMs employ curriculum learning strategies or multi-stage training that progressively adapts from general visual representations to medical domain knowledge to task-specific expertise, recognising that effective medical image understanding requires integration across multiple levels of abstraction [89].

Despite these sophisticated adaptation strategies, medical VLMs inherit fundamental limitations from their pre-training data that directly enable the fairness challenges this thesis addresses. Training datasets for medical VLMs, whether derived from clinical imaging repositories or scientific literature, systematically under-represent certain demographic groups, particularly minority populations, darker skin tones in dermatology, and intersectional subgroups such as elderly minority women [29], [38], [39]. The contrastive learning objective that drives VLM pre-training learns associations between visual patterns and textual descriptions as they occur in the training data, potentially encoding spurious correlations between demographic attributes and diagnostic findings when such correlations are present due to dataset imbalance or societal factors reflected in clinical documentation [15]. Understanding these fairness challenges and developing interventions to mitigate them requires examining not only model architectures but also the datasets that shape their learned representations, the theoretical frameworks for defining and measuring fairness, and the empirical evidence documenting disparate performance across demographic groups.

2.3 Algorithmic Fairness in Medical AI: Definitions, Metrics, and Harms

The promise of AI in medical diagnostics is predicated on the assumption that algorithmic decision-making can be more accurate, consistent, and equitable than human judgment. However, mounting empirical evidence demonstrates that AI systems frequently exhibit systematic performance disparities across patient subgroups defined by demographic attributes including race, gender, age, and socioeconomic status, violating fundamental principles of medical ethics and health equity. This section provides a comprehensive examination of algorithmic fairness in medical AI, beginning with the sources of bias in the ML pipeline, establishing theoretical notions of group fairness and their mathematical formulations, detailing evaluation metrics used to quantify disparities, and synthesising empirical evidence documenting the clinical harms resulting from biased algorithms. This foundation is essential for understanding both the urgency of fairness research and the limitations of existing approaches that this thesis addresses.

2.3.1 Sources of Bias in the Medical AI Pipeline

Bias in medical AI systems arises from multiple sources that span the entire ML pipeline from data collection through model development to deployment and use. Understanding these sources is essential for developing effective interventions and for recognising the limitations of approaches that address only a subset of bias mechanisms.

Data bias represents the most frequently discussed source of algorithmic disparities and manifests in several distinct forms. Representation bias occurs when training datasets systematically under-represent certain patient subgroups, either in absolute numbers or relative to their disease prevalence in the target deployment population [28]. The HAM10000 dermatology dataset, for example, was collected predominantly from Viennese and Australian clinical populations and does not include Fitzpatrick skin type annotations, though external analyses estimate that lighter skin tones outnumber darker skin tones by approximately 20 to 1 [90], [91]. Similarly, the Fitzpatrick17k dataset, created specifically to enable skin tone evaluation,

contains approximately 83% light skin types and only 17% dark skin types [39]. These severe imbalances lead models trained on such data to exhibit substantially degraded performance on darker skin patients [38], [39]. Similarly, the Harvard-FairVLMed fundus imaging dataset exhibits significant imbalance with 76.9% White patients, 14.9% Black patients, and 8.2% Asian patients [57]. These representation imbalances are not random artifacts but reflect longstanding structural inequities in healthcare access and research participation that have systematically excluded minority populations from medical studies [92].

Label bias arises when the ground-truth labels used to train supervised models contain systematic errors or inconsistencies correlated with demographic attributes. In medical imaging, label noise can result from inter-rater disagreement among annotating physicians, with evidence suggesting that disagreement rates vary across demographic subgroups, potentially due to reduced familiarity with how diseases present in underrepresented populations [28], [29]. For example, dermatological conditions often present with different visual characteristics across skin tones, and dermatologists with less experience treating diverse patient populations may exhibit reduced diagnostic accuracy for minority patients, errors that propagate into training labels when these same clinicians annotate datasets [38]. Historical diagnostic biases, such as the documented underdiagnosis of myocardial infarction in women or disparate pain assessment in Black patients, can become encoded as training labels that models learn to replicate [93], [94].

Measurement bias occurs when the features or inputs used by AI systems systematically differ in quality, availability, or informativeness across demographic groups. In medical imaging, this can manifest as differences in image acquisition protocols, equipment quality, or technical parameters across healthcare settings that correlate with patient demographics. AI models can predict patient race with AUC greater than 0.90 from radiographs alone [70]. Subsequent work showed that technical image-acquisition parameters influence these race-prediction models, with chest X-ray view positions differing systematically by race [95]. Such correlations create opportunities for models to learn demographic shortcuts where they rely on image acquisition artifacts rather than clinically relevant visual features for diagnosis, leading to performance degradation when these spurious correlations shift across deployment contexts.

Algorithmic bias emerges from the model training process itself even when training data is balanced and labels are accurate. Standard empirical risk minimisation (ERM), the default training objective in ML, optimises for average performance across the entire training distribution without regard for performance disparities across subgroups [50]. When disease prevalence or visual feature distributions differ across demographic groups, optimising overall accuracy can lead to models that perform well on majority groups but poorly on minority groups, a phenomenon formalised through the concept of representation disparity in optimisation theory [96]. This effect is particularly pronounced for rare diseases or underrepresented subgroups where the contribution of these samples to the overall loss function is minimal, creating little optimisation pressure to improve their predictions.

Deployment and usage bias occur when AI systems are deployed or utilised in ways that differ systematically across patient populations. If an AI diagnostic system is primarily deployed in well-resourced urban academic medical centers, it may fail to generalise to community hospitals or rural settings that serve higher proportions of underserved populations [13]. Differential trust and uptake of AI recommendations among clinicians treating different patient demographics can create outcome disparities even when the underlying algorithmic predictions are equitable [34]. Alert fatigue and automation bias, where clinicians become desensitised to algorithmic recommendations or defer excessively to system outputs, may manifest differently across clinical contexts and patient populations [97].

Feedback loops and deployment dynamics can amplify initial biases over time [98]. If an AI system exhibits initial bias against a demographic subgroup, leading to worse outcomes for that group, these poor outcomes may be reflected in subsequent data collected during deployment, reinforcing the very biases the system exhibited. In the context of resource allocation algorithms, algorithmic bias in predicting healthcare needs can lead to systematically reduced resource allocation for disadvantaged groups, which in turn reduces their documented healthcare utilisation, reinforcing the algorithm's learned association between demographic attributes and lower healthcare needs [32].

2.3.2 Theoretical Notions of Group Fairness

The formalisation of algorithmic fairness has produced multiple competing definitions of what constitutes a "fair" classifier, each capturing different intuitions about equity and each suited to different application contexts. This subsection examines the primary theoretical notions of group fairness, their mathematical formulations, their applicability to medical decision-making, and the fundamental trade-offs between fairness criteria that constrain what is achievable.

DP, also termed statistical parity or independence, represents the most straightforward fairness notion and requires that the probability of a positive prediction be independent of membership in a protected demographic group [44]. Formally, for a classifier producing predictions $\hat{Y} \in \{0, 1\}$ and protected attribute A, DP is satisfied when:

$$P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A = a') \quad \forall a, a' \tag{2.1}$$

This criterion ensures that positive predictions are distributed equally across demographic groups regardless of other factors. In medical applications, DP would require that disease diagnoses or treatment recommendations occur at equal rates across groups defined by race, gender, or other protected attributes.

However, DP has severe limitations for medical applications where disease prevalence legitimately differs across demographic groups. Many conditions exhibit genuine epidemiological variation by age, gender, or genetic ancestry that should be reflected in diagnostic predictions [45], [99]. Requiring equal positive prediction rates across groups with different disease prevalence would either lead to overdiagnosis in low-prevalence groups or underdiagnosis in high-prevalence groups, both ethically problematic outcomes. For instance, requiring equal breast cancer detection rates in male and female patients would be inappropriate given the roughly 100-fold higher incidence in women. These limitations have led researchers to largely reject DP as an appropriate fairness criterion for medical AI [44].

EOdds, introduced by Hardt, Price, and Srebro [42], provides a more nuanced fairness notion that conditions on the true outcome rather than requiring unconditional independence. EOdds

requires that the true positive rate (TPR, sensitivity) and false positive rate (FPR) be equal across demographic groups:

$$P(\hat{Y} = 1|Y = 1, A = a) = P(\hat{Y} = 1|Y = 1, A = a') \quad \forall a, a' \tag{2.2}$$

$$P(\hat{Y} = 1|Y = 0, A = a) = P(\hat{Y} = 1|Y = 0, A = a') \quad \forall a, a' \tag{2.3}$$

EOdds allows prediction rates to differ across groups in proportion to true outcome base rates while requiring that conditional error rates be equalised. This criterion aligns well with medical ethics principles requiring that sensitivity and specificity, the fundamental performance characteristics of diagnostic tests, should not vary based on patient demographics. A model satisfying EOdds will not systematically underdiagnose or overdiagnose any demographic group relative to their true disease rates.

Equality of opportunity (EOpp) represents a relaxation of EOdds that requires only equal TPR across groups while allowing FPR to differ [42]:

$$P(\hat{Y} = 1|Y = 1, A = a) = P(\hat{Y} = 1|Y = 1, A = a') \quad \forall a, a' \tag{2.4}$$

This criterion focuses on ensuring that individuals with the positive outcome have equal chances of being correctly identified regardless of group membership. In medical contexts focused on disease detection where false positives can be addressed through confirmatory testing but false negatives lead to missed diagnoses, EOpp provides a clinically meaningful fairness notion. For cancer screening, ensuring equal sensitivity across demographic groups may be prioritised over equalising FPR, making EOpp an appropriate fairness target.

Calibration requires that predicted probabilities correspond to true outcome frequencies across demographic groups [46]. A classifier producing a continuous risk score $S \in [0, 1]$ is calibrated if, for every realised value $s \in [0, 1]$ that S can take and every value $a \in \mathcal{A}$ of the protected attribute:

$$P(Y = 1 \mid S = s, A = a) = s, \tag{2.5}$$

where s denotes a specific realised value of the predicted risk score S and $\mathcal{A}$ is the set of values of the protected attribute A.

Calibration ensures that when a model assigns a predicted probability of 0.7 to a patient, that patient has approximately 70% actual probability of the positive outcome regardless of their demographic characteristics. Calibration is particularly important for clinical decision-making when predicted probabilities inform treatment decisions; if a model is miscalibrated for certain demographic groups, risk-benefit calculations for interventions will be systematically incorrect [100].

The mathematical impossibility results established by Chouldechova and others demonstrate that these fairness notions cannot generally be simultaneously satisfied [45], [99]. Specifically, when base rates differ across groups, no imperfect classifier can simultaneously achieve calibration, equal FPR, and equal false negative rate (FNR). This impossibility theorem has profound implications for fairness in medical AI because disease prevalence frequently differs across demographic groups for both biological and social reasons. The theorem establishes that practitioners must make deliberate choices about which fairness notion to prioritise based on the clinical context and ethical considerations specific to their application, acknowledging that optimising for one fairness criterion may compromise others.

Recent theoretical work has examined fairness notions specifically designed for multiclass classification and regression settings common in medical applications. Fairness constraints for multiclass problems where a classifier predicts among $k > 2$ outcomes require extending binary fairness definitions across all outcome classes [101]. For regression tasks producing continuous predictions, fairness notions based on equalising mean squared error or other loss metrics across groups have been proposed [102]. These extensions maintain the core intuitions of independence, separation, or calibration while adapting them to non-binary prediction spaces.

2.3.3 Standard Fairness Evaluation Metrics

Translating theoretical fairness notions into concrete quantitative metrics enables empirical evaluation and comparison of model fairness across systems, datasets, and interventions. This subsection examines the landscape of fairness metrics used in medical AI research, detailing

their mathematical definitions, relationships to theoretical fairness notions, and practical interpretation.

Demographic Parity Difference (DPD) quantifies violations of the DP criterion by measuring the maximum difference in positive prediction rates across demographic groups:

$$\text{DPD} = \max_{a,a'} \left| P(\hat{Y} = 1|A = a) - P(\hat{Y} = 1|A = a') \right| \tag{2.6}$$

In practice, DPD is typically estimated from test data as the difference between the highest and lowest prediction rates across groups. A DPD of zero indicates perfect DP, while larger values indicate greater disparity. For medical applications, DPD is most meaningful when comparing conditions with similar prevalence across groups or when the focus is on resource allocation rather than diagnostic accuracy.

TPR Disparity (ΔTPR) and FPR Disparity (ΔFPR) quantify violations of EOdds by measuring differences in sensitivity and specificity across groups:

$$\Delta\text{TPR} = \max_{a,a'} |\text{TPR}_a - \text{TPR}_{a'}| \tag{2.7}$$

$$\Delta\text{FPR} = \max_{a,a'} |\text{FPR}_a - \text{FPR}_{a'}| \tag{2.8}$$

where TPR_a denotes the TPR for group a and FPR_a denotes the FPR. These metrics directly measure whether a model achieves equal sensitivity and specificity across demographic groups. In medical contexts, ΔTPR is often considered more critical than ΔFPR because disparities in TPR directly translate to missed diagnoses for certain demographic groups [37].

Difference in EOdds (DEOdds) provides a single scalar metric combining true and false positive rate disparities:

$$\text{DEOdds} = \Delta\text{TPR} + \Delta\text{FPR} \tag{2.9}$$

This metric provides a summary measure of EOdds violations but can obscure the specific nature of disparities since a model might exhibit high ΔTPR and low ΔFPR or vice versa [42].

EOdds Ratio represents an alternative aggregate measure:

$$\text{EOdds Ratio} = \max_{a,a'} \max \left\{ \frac{\text{TPR}_a}{\text{TPR}_{a'}}, \frac{\text{FPR}_a}{\text{FPR}_{a'}} \right\} \tag{2.10}$$

This ratio-based metric can be more interpretable when considering proportional rather than absolute disparities and is less sensitive to very high or low base rates that can make absolute difference metrics difficult to interpret [69].

Calibration metrics quantify how well predicted probabilities correspond to observed outcome frequencies across demographic groups. Expected Calibration Error (ECE) within each group measures the average discrepancy between predicted probabilities and true outcome frequencies:

$$\text{ECE}_a = \sum_{i=1}^{M} \frac{n_i}{N_a} |\text{acc}(B_i) - \text{conf}(B_i)| \tag{2.11}$$

where predictions are binned into M bins B_i, n_i is the number of predictions in bin i for group a, N_a is the total samples for group a, $\text{acc}(B_i)$ is the accuracy within bin i, and $\text{conf}(B_i)$ is the average predicted confidence in bin i [103], [104]. Calibration disparity across groups can be measured as the maximum difference in ECE:

$$\Delta\text{ECE} = \max_{a,a'} |\text{ECE}_a - \text{ECE}_{a'}| \tag{2.12}$$

AUC is not strictly a fairness metric but serves as the primary performance measure in medical AI research and is frequently reported alongside fairness metrics to assess fairness-accuracy trade-offs. AUC measures the probability that a classifier assigns a higher score to a randomly chosen positive instance than to a randomly chosen negative instance [105]. AUC disparity across groups, $\Delta\text{AUC} = \max_{a,a'} |\text{AUC}_a - \text{AUC}_{a'}|$, provides a summary measure of performance inequality that aggregates across all classification thresholds rather than evaluating fairness at a single operating point [25].

Equity-Scaled AUC (ES-AUC) attempts to combine performance and fairness into a single composite metric by penalising overall AUC proportionally to fairness violations:

$$\text{ES-AUC} = \text{AUC}_{\text{overall}} \times (1 - \Delta\text{AUC}) \tag{2.13}$$

This metric provides a single scalar that trades off aggregate performance against cross-group AUC disparity, though the multiplicative combination represents a specific assumption about how to value the fairness-accuracy trade-off that may not align with clinical priorities [106].

These standard fairness metrics, while valuable for quantifying disparities along individual demographic axes such as race or gender, suffer from a critical limitation that motivates the central contribution of this thesis: they fail to capture compounded disparities at demographic intersections. A model may exhibit acceptable ΔTPR when comparing men to women and acceptable ΔTPR when comparing White to Black patients, yet simultaneously exhibit severe TPR disparities between Black women and White men [35]. This limitation reflects the single-attribute focus of standard fairness metrics and has catalysed development of intersectional fairness metrics examined in Section 2.4.2.

2.3.4 Empirical Evidence of Clinical Harms from Biased Medical AI

The theoretical fairness frameworks and evaluation metrics described in previous subsections are not merely academic constructs but respond to documented evidence of systematic bias in deployed and near-deployment medical AI systems with tangible consequences for patient outcomes and health equity. This subsection synthesises empirical evidence of algorithmic bias across medical imaging modalities, examining the magnitude of disparities, the demographic groups most severely affected, and the translation of statistical fairness violations into clinical harms.

A landmark analysis of a commercial algorithm used to identify patients for enrollment in high-risk care management programs revealed systematic racial bias affecting approximately 200 million people across US health systems [32]. This algorithm used historical healthcare costs as a proxy for health needs when predicting which patients would benefit from additional care resources. The fundamental flaw in this approach stemmed from the fact that healthcare spending differs systematically by race not only due to health needs but also due to access barriers, systemic discrimination, and structural inequalities in healthcare delivery. At a given risk score, Black patients were substantially sicker than White patients as measured by chronic condition counts, with Black patients having an average of 2.6 active chronic conditions versus

1.8 for White patients at the same risk score threshold. This miscalibration led to systematic underestimation of health needs for Black patients, resulting in reduced care management enrollment despite greater clinical need. Eliminating this bias would increase enrollment of Black patients by 58% while decreasing White patient enrollment, fundamentally altering the demographic composition of patients receiving enhanced care resources [32].

Chest radiography AI systems have been the subject of multiple studies documenting systematic bias, with converging evidence from independent research groups demonstrating underdiagnosis of certain pathologies in minority and female patients. Analysis of DenseNet121 models trained on three large chest X-ray datasets (CheXpert, MIMIC-CXR, and ChestX-ray8) demonstrated that models exhibited 7.7% lower sensitivity for Black patients compared to White patients, with sensitivity of 75.8% versus 83.5% respectively [37]. Asian patients exhibited 3% lower sensitivity compared to White patients. Female patients showed reduced performance on several pathologies including edema and atelectasis. These disparities persisted across multiple training datasets and model architectures, suggesting systematic rather than dataset-specific bias. The clinical implications are direct: reduced sensitivity for minority patients translates to higher rates of false negative diagnoses, meaning pathologies are missed at diagnosis with potential for disease progression and delayed treatment.

Evaluation of foundation models trained on over 800,000 chest X-rays found persistent fairness gaps despite massive scale, with 6.8–7.8% performance drops for female patients on "no finding" classifications and 10.7–11.6% drops for Black patients with pleural effusion [71]. This finding challenges the assumption that simply scaling data and model size will eliminate bias without explicit fairness constraints during training.

A particularly concerning finding from multiple independent research groups is that AI models can predict patient race from medical images with high accuracy even when race is not a visually obvious feature, suggesting that systematic differences in image acquisition, patient positioning, or subtle physiological features enable demographic shortcuts. Models predict race with AUC greater than 0.90 across chest X-ray, mammography, cervical spine radiographs, and chest CT scans [70]. Critically, this prediction capability persisted even with heavily degraded, cropped, or noise-corrupted images, and in experiments where obvious anatomical features were removed, suggesting that race information is latent throughout

medical images in ways poorly understood by researchers and clinicians. Subsequent work identified specific technical parameters including view position and windowing settings that contribute to race prediction capability, demonstrating that demographics-independent calibration using per-view thresholds achieved 46% relative reduction in sensitivity disparity for Black patients and 67% relative reduction for Asian patients [95]. These findings establish that medical AI systems learn race as a feature from medical images and that this learned demographic information can drive systematic bias in clinical predictions.

Dermatology AI has exhibited the most severe documented disparities, with multiple studies demonstrating dramatically degraded performance on darker skin tones with direct implications for cancer detection and survival. A comprehensive fairness evaluation of models trained on International Skin Imaging Collaboration (ISIC) datasets, tested on the biopsy-confirmed Diverse Dermatology Images (DDI) dataset, found significantly lower accuracy for Fitzpatrick skin types V–VI, the darkest classifications, with particularly severe degradation for melanoma detection [38]. The clinical stakes are profound: melanoma 5-year survival rates are 66% for Black patients compared to 90% for White patients, with late-stage diagnosis at presentation being a primary driver of this mortality disparity [40], [107]. An AI screening system that exhibits lower sensitivity for detecting melanoma on darker skin tones compounds existing disparities by further delaying diagnosis when early detection is most critical for survival.

The Fitzpatrick17k dataset, created specifically to enable fairness evaluation across skin tones, demonstrated that commercial and research dermatology AI systems consistently underperform on darker skin [39]. The dataset distribution itself reveals the scope of representation bias in dermatology AI: 83% of images show light skin types (Fitzpatrick I–IV) versus 17% dark skin (Fitzpatrick V–VI). Analysis of segmentation tasks demonstrated that skin lesion segmentation, the foundational first step in diagnostic pipelines, shows significant correlation between performance and skin color, with systematic challenges segmenting darker tones [91]. Evaluation of common bias mitigation methods proved largely ineffective, suggesting that addressing dermatology AI bias requires fundamental changes in data collection and algorithm design rather than post-hoc corrections.

Ophthalmology AI for glaucoma and diabetic retinopathy screening has shown more mixed results, with some systems achieving equitable performance across demographic groups when explicitly evaluated and validated for fairness. Autonomous AI for diabetic retinopathy screening showed no significant bias across racial, ethnic, or gender subgroups in a preregistered trial with 626 participants [108]. Point-of-care autonomous AI achieved 100% exam completion in the intervention group versus significantly lower in controls, eliminating racial screening gaps for underserved youth [109]. Further evidence of equitable AI performance was provided by validation of a diabetic retinopathy detection system in Indigenous Australian populations, demonstrating that systems trained on diverse validation cohorts can achieve clinically acceptable sensitivity across underserved demographic groups [110]. These positive examples demonstrate that fairness is achievable through careful design, including diverse training data, explicit fairness validation across demographic subgroups before deployment, and attention to real-world deployment contexts.

However, other ophthalmology studies have documented disparities. The Harvard-FairVLMed dataset released specifically for fairness research revealed that both CLIP and BLIP-2 based models exhibit systematic bias in glaucoma detection, with Asian, male, non-Hispanic, and Spanish-preferred-language groups being favoured [57]. The magnitude of these disparities, while smaller than those observed in dermatology, remains clinically significant given that glaucoma disproportionately affects Black and Hispanic populations with higher rates of severe vision loss [60]. Race, ethnicity, and gender each independently influence the fairness of glaucoma prediction models, suggesting that intersectional effects across these attributes warrant dedicated investigation [111].

The pathway from algorithmic bias to clinical harm operates through multiple mechanisms that extend beyond individual diagnostic errors to population-level impacts on health equity. Misdiagnosis directly harms patients through incorrect treatment or delayed care, with documented examples including pulse oximetry AI overestimating oxygen saturation in Black patients leading to delayed hypoxemia treatment [112]. Resource misallocation algorithms direct care resources away from minority patients who need them most, compounding existing access barriers [32]. Systematic differences in diagnostic certainty create instability where minority patients receive borderline predictions vulnerable to being flipped by minor data

variations, eroding both algorithmic reliability and clinician trust [52]. Algorithmic perpetuation of existing clinical disparities occurs when AI trained on biased clinical data replicates patterns like heart attack underdiagnosis in women or disparate pain medication prescribing [94], [113]. Non-Hispanic Black populations experience approximately 30% higher overall mortality compared to White populations, and biased algorithms risk exacerbating these existing disparities rather than reducing them, establishing algorithmic fairness as a prerequisite for health equity [114].

2.4 The Critical Challenge of Intersectionality

The preceding sections have established that medical AI systems exhibit systematic bias and that current fairness frameworks provide mathematical definitions and metrics for quantifying disparities. However, a fundamental limitation pervades the vast majority of fairness research and deployed fairness interventions: the focus on single demographic attributes evaluated in isolation. The concept of intersectionality, originally developed in legal scholarship by Crenshaw [53], [56] to describe the compounded discrimination experienced at overlapping demographic categories, has since been translated into algorithmic fairness through the recognition that fairness guarantees on marginal attributes do not imply fairness on their joint distributions [54], [58]. This section demonstrates through empirical evidence that single-attribute fairness analysis drastically underestimates discrimination experienced at demographic intersections, examines emerging theoretical frameworks for intersectional fairness, and establishes the research gap this thesis addresses.

2.4.1 Documented Failures of Single-Attribute Fairness Evaluation

Empirical studies across medical AI and broader computer vision applications have repeatedly demonstrated that models satisfying single-attribute fairness criteria can simultaneously exhibit severe disparities at demographic intersections. The Gender Shades study [35] first provided direct empirical evidence of this pattern by evaluating three commercial gender classification systems on a dataset balanced across four demographic groups defined by gender $\times$ skin tone. Error rates for darker-skinned women reached 34.7% compared to <1% for

lighter-skinned men. Critically, the intersectional disparity exceeded what would be predicted by examining gender disparities and skin tone disparities independently: approximately 3% error rate disparity between genders (when skin tone was ignored) and roughly 10% disparity between skin tones (when gender was ignored) combined superadditively into an order-of-magnitude larger error at the female × dark-skinned intersection. The subsequent mathematical formalisation of "fairness gerrymandering" by Kearns et al. [58] generalised this finding into a formal framework.

The fairness gerrymandering problem can be formalised mathematically: consider a binary classification task evaluated with respect to two binary protected attributes $A_1 \in \{0, 1\}$ and $A_2 \in \{0, 1\}$, creating four intersectional subgroups: $(A_1 = 0, A_2 = 0)$, $(A_1 = 0, A_2 = 1)$, $(A_1 = 1, A_2 = 0)$, $(A_1 = 1, A_2 = 1)$. A classifier can satisfy single-attribute fairness constraints such as:

$$\text{TPR}(A_1 = 0) = \text{TPR}(A_1 = 1) \tag{2.14}$$

$$\text{TPR}(A_2 = 0) = \text{TPR}(A_2 = 1) \tag{2.15}$$

while simultaneously exhibiting substantial disparity at intersections:

$$\text{TPR}(A_1 = 0, A_2 = 0) \neq \text{TPR}(A_1 = 1, A_2 = 1) \tag{2.16}$$

This occurs because single-attribute marginal constraints provide no guarantees about joint distributions. A model might achieve equal average sensitivity for men and women by performing well on White men and poorly on minority men while performing well on minority women and poorly on White women, exactly balancing disparities to satisfy marginal constraints while maximising intersectional unfairness [58].

This phenomenon has been documented empirically in analyses of recidivism prediction algorithms, demonstrating that models achieving approximate DP when evaluated separately for race and gender simultaneously exhibited 30 percentage point disparities in positive prediction rates between White men and Black women [54]. It was further mathematically proved that with d binary protected attributes, a classifier might satisfy all d single-attribute

fairness constraints while violating fairness for the vast majority of the 2^d possible intersectional subgroups, with the proportion of subgroup pairs satisfying fairness criteria potentially approaching zero as d increases [54].

Medical AI studies have confirmed these patterns in clinical contexts. Analysis of chest X-ray classification using social determinants of health data demonstrated that models appearing fair when evaluated separately for race and gender exhibited substantial performance degradation for Black female patients specifically [41]. The study employed the Area Health Resources File to link medical images to county-level social determinants and showed that intersectional subgroups defined by race, gender, and socioeconomic factors experienced compounded disadvantage not predicted by single-axis analysis. Further work on multimodal clinical prediction systems demonstrated that VLMs exhibit the largest fairness violations at demographic intersections even when single-attribute evaluation suggests acceptable fairness [27]. Beyond clinical settings, counterfactual probing of general-purpose VLMs confirmed that models exhibit compounded biases at identity intersections that single-attribute audits fail to detect [115].

Research has established that models with "superhuman demographic prediction" capability, meaning they can predict race, gender, or age from medical images with high accuracy, exhibit the largest fairness gaps at intersections [37], [70]. Comprehensive analysis has demonstrated that fairness performance on internal validation datasets has a weak, and sometimes negative, correlation with fairness on external, out-of-distribution datasets, with intersectional subgroups showing the largest generalisation failures [25]. This finding suggests that models learn demographic shortcuts that appear to provide fairness on the training distribution but fail catastrophically when spurious correlations shift in deployment.

2.4.2 Emerging Notions of Intersectional Fairness

Recognition of the limitations of single-attribute fairness analysis has catalysed development of theoretical frameworks explicitly designed to capture intersectional bias. This subsection examines three major intersectional fairness definitions that extend classical fairness notions to address compounded discrimination.

Max-min fairness focuses on maximising the performance of the worst-off subgroup [58]. Rather than requiring equal performance across all subgroups, which may be infeasible when subgroups differ substantially in size or inherent classification difficulty, max-min fairness optimises:

$$\max_{\text{classifier}} \min_{\text{subgroup } g} \text{Accuracy}(g) \tag{2.17}$$

This approach ensures that no subgroup, including intersectional subgroups, experiences catastrophically poor performance even if aggregate performance or performance for majority groups must be somewhat reduced to lift the worst-off subgroup. The max-min formulation connects to DRO approaches that minimise worst-case loss [50].

A key innovation in this framework is the consideration of structured subgroups defined by conjunctions of protected attributes rather than treating all possible subsets of the population as equally relevant subgroups. For attributes race $\in \{\text{White, Black, Asian}\}$, gender $\in \{\text{Male, Female}\}$, and age $\in \{\text{Young, Old}\}$, structured subgroups include conjunctions like $(\text{Black} \wedge \text{Female} \wedge \text{Young})$ rather than arbitrary subsets. The number of structured subgroups grows as the product of the cardinalities of protected attributes, polynomial rather than exponential, making comprehensive intersectional evaluation computationally feasible [58].

DF provides a formal framework generalising DP to intersectional subgroups through a privacy-inspired interpretation [54]. DF requires that, given an outcome, an adversary learns little about an individual's protected attributes, quantified through bounded ratios of conditional outcome probabilities:

$$e^{-\varepsilon} \leq \frac{P(\hat{Y} = y|G = g)}{P(\hat{Y} = y|G = g')} \leq e^{\varepsilon} \quad \forall g, g', y \tag{2.18}$$

where G represents membership in intersectional subgroups defined by combinations of protected attributes and ϵ controls the strictness of the fairness bound. When $\varepsilon = 0$, perfect fairness is achieved with $P(\hat{Y} = y|G = g) = P(\hat{Y} = y|G = g')$ for all subgroups, equivalent to DP across all intersectional subgroups. Larger ϵ values permit greater disparities while still bounding the degree of discrimination.

The privacy interpretation of DF recognises that satisfying the bounded ratio requirement prevents an adversary from accurately inferring protected attribute combinations from observing outcomes, similar to how differential privacy prevents inference of individual records from aggregate statistics [116]. This connection enables leveraging the extensive theoretical toolkit developed for differential privacy to prove composition properties, analyse mechanism design, and develop algorithms satisfying DF guarantees [54].

However, DF inherits the limitations of DP including inappropriateness for medical applications where outcome base rates legitimately differ across groups. A model satisfying DF with small ϵ will produce similar positive prediction rates for groups with vastly different disease prevalence, leading to overdiagnosis in low-prevalence groups or underdiagnosis in high-prevalence groups.

IF-α was introduced specifically to address the "leveling down" problem in intersectional fairness [55]. IF-α provides a framework that balances absolute performance and relative disparities across intersectional subgroups, recognising that fairness can be achieved through two pathways: reducing the performance gap between subgroups or improving the absolute performance of disadvantaged subgroups. Many fairness interventions achieve formal fairness by degrading performance for advantaged groups rather than improving performance for disadvantaged groups, a phenomenon termed "leveling down" that is ethically problematic in medical applications where diagnostic accuracy is paramount [33].

IF-α quantifies intersectional fairness through a weighted combination of absolute and relative performance disparities:

$$L_\alpha(g, g') = \alpha \cdot \Delta_{\text{abs}}(g, g') + (1 - \alpha) \cdot \Delta_{\text{rel}}(g, g') \tag{2.19}$$

where Δ_{abs} represents the absolute performance difference between subgroups g and g', Δ_{rel} represents the relative performance ratio, and $\alpha \in [0, 1]$ controls the emphasis on absolute versus relative fairness. When α is high, the metric prioritises reducing absolute gaps in performance; when α is low, the metric prioritises reducing relative ratios. A threshold parameter γ_{IF} determines whether a subgroup pair is considered fair: $L_\alpha(g, g') < \gamma_{IF}$.

The α parameter enables practitioners to specify the type of fairness intervention preferred. Setting α close to 1 emphasises improving the absolute performance of the worst-off subgroup, even if performance ratios remain unequal, which is appropriate when absolute diagnostic accuracy is clinically critical. Setting α close to 0 emphasises equalising performance ratios, appropriate when relative disadvantage is the primary concern. This flexibility allows calibrating the fairness notion to the specific clinical and ethical context while maintaining explicit intersectional scope [55].

Models trained to optimise IF-α avoid levelling down by providing explicit incentives to improve performance for disadvantaged subgroups rather than only penalising disparity. The framework has been validated on multiple datasets, showing that IF-α-optimised models achieve superior fairness-accuracy trade-offs compared to interventions optimising for DF or single-attribute EOdds [55].

Despite these theoretical advances, significant challenges remain in operationalising intersectional fairness for medical AI. The exponential growth in the number of subgroups with increasing demographic dimensionality creates data sparsity issues where some intersectional subgroups have very few samples in training or evaluation datasets, making reliable metric estimation difficult [51]. The heterogeneity in subgroup sizes creates computational and statistical challenges for training algorithms that must balance performance across groups ranging from hundreds to thousands of samples [117]. Current intersectional fairness metrics and interventions, while substantial advances over single-attribute approaches, have been developed and evaluated primarily on tabular data and traditional CV tasks, with minimal application to VLMs in medical domains [26].

2.5 A Critical Review of Fairness Intervention Methodologies

Having established the theoretical foundations of algorithmic fairness and the critical importance of intersectional analysis, this section examines the landscape of technical interventions proposed to mitigate bias in ML systems. Fairness interventions can be categorised based on where they intervene in the ML pipeline: pre-processing methods that modify training data,

in-processing methods that incorporate fairness constraints during model training, and post-processing methods that adjust model outputs after training. This review critically analyses representative methods from each category, with particular attention to their applicability to VLMs and their capacity to address intersectional fairness.

2.5.1 The Taxonomy of Fairness Interventions: Pre-processing, In-processing, and Post-processing

Pre-processing interventions attempt to remove bias from training data before model training begins. The foundational approach is resampling, which adjusts the composition of the training dataset to achieve balanced representation across demographic groups. Oversampling of underrepresented groups, undersampling of overrepresented groups, or synthetic data generation through techniques like Synthetic Minority Over-sampling Technique (SMOTE) can reduce representation imbalance [47]. However, resampling suffers from several limitations: oversampling can lead to overfitting on the limited diverse examples available, undersampling discards potentially valuable training data from majority groups reducing overall model quality, and synthetic data generation may not capture the true distribution of rare subgroups [47]. A more principled pre-processing approach formulates discrimination prevention as an optimisation problem over probabilistic transformations of the training data, providing theoretical guarantees on the resulting data distribution while preserving individual-level utility [118]. Despite this theoretical elegance, the approach shares the fundamental limitation of all pre-processing methods: it cannot directly govern model behaviour at the decision boundary.

Reweighting represents a more sophisticated pre-processing approach that assigns higher loss weights to samples from underrepresented groups during training, effectively increasing their influence on the optimisation objective without modifying the dataset itself [72]. The meta-learning framework Fairness Optimized Reweighting via Meta-Learning (FORML) jointly optimises sample weights and neural network parameters to improve EOpp [119]. Fair Adversarial Instance Re-weighting (FAIR) merges reweighting with adversarial training to provide interpretable information about individual instance fairness [120]. While reweighting offers advantages over resampling by preserving all training data, it still operates only on the

input side of the learning process and provides no direct guarantee that the trained model will behave fairly at the decision boundary.

In-processing methods incorporate fairness constraints directly into the model training objective, typically through regularisation terms or adversarial architectures that penalise unfair predictions. Domain-Adversarial Neural Networks (DANN) employ a gradient reversal layer that learns features discriminative for the main task while indiscriminate with respect to protected attributes [49]. The architecture consists of a feature extractor, a task classifier predicting the main task label, and a domain classifier predicting the protected attribute. During backpropagation, gradients from the domain classifier are reversed before updating the feature extractor, encouraging learning of representations from which demographic attributes cannot be predicted. Conditional Domain-Adversarial Neural Networks (CDANN) extend this framework by conditioning the domain classifier on task predictions, better handling multimodal distributions typical in medical classification [121].

DRO, exemplified by Group DRO (GroupDRO), minimises the worst-case training loss over pre-defined demographic groups rather than minimising average loss [50]. The GroupDRO objective is:

$$\min_{\theta} \max_{g \in \mathcal{G}} \mathcal{L}_g(\theta) \tag{2.20}$$

where $\mathcal{L}_g(\theta)$ denotes the loss on group g under model parameters θ. However, minimising worst-case training loss does not, by itself, guarantee the worst-group generalisation performance. Sagawa et al. [50] showed that overparameterised neural networks can perfectly fit training data, driving worst-case training loss to zero while still exhibiting poor worst-group test accuracy due to memorisation of minority-group examples and reliance on spurious correlations. The 10–40% improvements in worst-group accuracy reported across their benchmarks were achieved only when GroupDRO was coupled with stronger-than-typical L_2 regularisation or early stopping to control the generalisation gap. The convergence guarantees provided in the paper apply to the stochastic optimisation algorithm (convergence to the min–max objective), not to downstream test performance. These findings establish that worst-group generalisation in overparameterised models requires explicit capacity control beyond the DRO objective itself.

VLM-specific fairness methods have emerged recently to address the unique challenges of multimodal architectures. FairCLIP applies optimal transport theory to align the distribution of samples overall with the distribution within each demographic group [57]. The method minimises Sinkhorn distance, a differentiable approximation to optimal transport distance, between the marginal distribution $P(X)$ and the conditional distributions $P(X|A = a)$ for each protected attribute value a. By reducing distributional discrepancies, FairCLIP reduces bias in medical glaucoma detection across race, gender, ethnicity, and language attributes. However, a critical limitation is that FairCLIP addresses attributes sequentially, optimising fairness for one attribute at a time without guarantees about intersectional subgroups [57].

FairerCLIP jointly debiases CLIP's image and text representations in Reproducing Kernel Hilbert Spaces (RKHS) using the Hilbert-Schmidt Independence Criterion (HSIC) [73]. The method achieves EOdds near 0% while maintaining high classification accuracy and operates without requiring ground-truth labels, enabling debiasing of pre-trained medical CLIP models without retraining from scratch. DeAR learns additive residual image representations that offset original representations to reduce demographic distinguishability in the visual encoder while preserving task-relevant information [74].

Post-processing methods adjust model predictions after training to satisfy fairness constraints, typically by deriving group-specific decision thresholds. The threshold optimisation approach solves a linear programming problem on the receiver operating characteristic (ROC) plane to find optimal thresholds satisfying EOdds constraints [42]. This approach has the significant advantage of being applicable to existing deployed models without requiring retraining, particularly valuable for regulated medical devices where model modifications require extensive re-validation. For medical applications, well-calibrated models with subpopulation-specific decision thresholds often achieve fairness without sacrificing accuracy, suggesting complex in-processing interventions may be unnecessary if proper calibration and threshold selection are applied [122].

2.5.2 In-Processing Interventions at the Feature Level and Their Limitations

A critical commonality among the vast majority of fairness interventions, including the sophisticated adversarial and optimisation-based methods described above, is that they operate at the feature representation level, attempting to learn or enforce statistical independence between demographic attributes and learned features. This subsection examines why feature-level interventions, despite their theoretical elegance and widespread adoption, are fundamentally inadequate for ensuring decision-level fairness in medical VLMs.

The feature-level fairness hypothesis underlying adversarial debiasing and related methods posits that if learned representations cannot be used to predict protected attributes, then predictions derived from those representations will be fair. Mathematically, if a representation $h = f(x)$ satisfies $P(A|h) = P(A)$, meaning demographic attribute A is conditionally independent of representation h, then predictions $\hat{y} = c(h)$ derived from h should exhibit fairness properties. However, this reasoning contains a critical gap: conditional independence of predictions, particularly when the prediction function c is complex or non-linear.

Adversarial debiasing methods attempting to learn fair representations can paradoxically increase bias when the task classifier c has sufficient capacity to recover demographic information from supposedly debiased features through non-linear transformations [48]. The phenomenon of "fairness gerrymandering" at the representation level, where features appear demographically invariant under linear analysis but contain recoverable demographic information through non-linear processing, undermines the theoretical foundation of feature-level fairness.

Moreover, feature-level interventions that successfully enforce demographic invariance risk the "information destruction" problem identified by multiple researchers [33], [123]. Suppose demographic information is truly removed from representations. In that case, any legitimate clinical associations between demographic attributes and disease risk are also eliminated, potentially degrading diagnostic accuracy for diseases exhibiting genuine epidemiological

variation by age, gender, or genetic ancestry. For example, removing all age-related information from medical imaging representations would eliminate the legitimate clinical signal that many conditions increase in prevalence with aging.

The "leveling down" effect represents another fundamental limitation of many feature-level fairness interventions. Multiple studies have documented that adversarial debiasing, DRO, and related methods often achieve formal fairness by degrading performance for all groups rather than selectively improving performance for disadvantaged groups [31], [33]. In medical applications where diagnostic accuracy is paramount and errors have clinical consequences, achieving fairness through performance degradation is ethically untenable.

For VLMs specifically, feature-level interventions face additional challenges from the multimodal architecture. Bias can emerge from the image encoder, the text encoder, or their interaction through the contrastive or generative objectives. Methods that apply distributional alignment to image features alone may fail to address biases encoded in text representations or in the learned cross-modal alignment [57]. Conversely, methods that enforce fairness constraints on the aligned multimodal embedding space may over-constrain the representations, limiting the model's capacity to learn the rich semantic associations between visual and textual medical concepts that drive its clinical utility.

2.5.3 Distributional Alignment with Maximum Mean Discrepancy: Theory and Application

Maximum Mean Discrepancy (MMD) measures the largest difference in expectations over functions in the unit ball of an RKHS [124]. Let P and Q denote two probability distributions defined over a common measurable space $\mathcal{X}$, and let $\mathcal{H}$ denote a reproducing kernel Hilbert space (RKHS) with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and associated norm $\| \cdot \|_{\mathcal{H}}$. The MMD between P and Q is defined as:

$$\text{MMD}(P, Q) = \sup_{\|f\|_{\mathcal{H}} \leq 1} |\mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{x \sim Q}[f(x)]| \tag{2.21}$$

where $f : \mathcal{X} \to \mathbb{R}$ ranges over the unit ball of $\mathcal{H}$, and $\mathbb{E}_{x\sim P}[\cdot]$ denotes expectation with respect to x drawn from P.

By the reproducing property of the RKHS, the supremum above admits a closed-form expression in terms of the kernel alone:

$$\text{MMD}^2(P, Q) = \mathbb{E}_{x,x'\sim P}[k(x, x')] + \mathbb{E}_{y,y'\sim Q}[k(y, y')] - 2\,\mathbb{E}_{x\sim P,\, y\sim Q}[k(x, y)] \tag{2.22}$$

where x, x' denote two independent random variables each drawn from P, y, y' denote two independent random variables each drawn from Q, and $\mathbb{E}_{x,x'\sim P}[\cdot]$ denotes expectation under independent draws. The kernel evaluations $k(x, x')$, $k(y, y')$, and $k(x, y)$ thus represent, respectively, intra-P similarity, intra-Q similarity, and cross-distribution similarity.

This formulation enables practical estimation from finite samples without requiring density estimation or numerical integration. Given finite i.i.d. samples $\{x_i\}_{i=1}^m \sim P$ of size m and $\{y_j\}_{j=1}^n \sim Q$ of size n, an unbiased empirical estimator of MMD^2 is:

$$\widehat{\text{MMD}}^2 = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{\substack{j=1 \\ j\neq i}}^{m} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j\neq i}}^{n} k(y_i, y_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(x_i, y_j) \tag{2.23}$$

where i, j are sample indices, the outer summation in each within-sample term runs over all m (respectively n) samples, and the inner summation ranges over all indices j distinct from the current outer index i (self-comparisons $j = i$ are excluded so that the estimator remains unbiased); the final cross term sums over all mn ordered pairs between the two samples. A critical theoretical property of MMD is that for characteristic kernels, $\text{MMD}(P, Q) = 0$ if and only if $P = Q$, providing a rigorous statistical test for distribution equality [124]. Commonly used characteristic kernels include the Gaussian radial basis function (RBF) kernel $k(x, y) = \exp(-\gamma\|x - y\|^2)$, where $\gamma > 0$ is the bandwidth parameter, and certain polynomial kernels, enabling MMD to serve as a proper metric on probability distributions.

The application of MMD to fairness-aware ML involves minimising MMD between the distributions of learned representations or predictions across demographic groups. Let $h : \mathcal{X} \to \mathbb{R}^d$ denote a learned representation function mapping inputs to a d-dimensional feature

space, and let G denote a random variable encoding intersectional subgroup membership, taking values in the finite set $\mathcal{G}$ of demographic groups (consistent with the group notation established in Sections 2.3–2.4). For group values $g, g' \in \mathcal{G}$, demographic fairness at the representation level can be encouraged by adding a regularisation term:

$$\mathcal{L}_{\text{fairness}} = \sum_{\substack{g,g' \in \mathcal{G} \\ g \neq g'}} \text{MMD}^2(P(h(x) \mid G = g),\ P(h(x) \mid G = g')) \tag{2.24}$$

where $P(h(x) \mid G = g)$ denotes the conditional distribution of the representation $h(x)$ given subgroup membership $G = g$, and the sum ranges over all unordered pairs of distinct groups in $\mathcal{G}$ [125]. Minimising this objective encourages the model to learn representations whose conditional distributions are statistically indistinguishable across groups, operationalising the intuition that fair representations should not vary systematically with demographics.

Several extensions and variants of MMD have been developed for fairness applications. Weighted MMD assigns different importance to different groups, addressing class imbalance common in medical datasets [126]. Joint MMD computes discrepancy on the joint distribution $P(X, Y)$ rather than marginal distributions alone, capturing dependencies between features and outcomes [127]. Conditional MMD (CMMD) enforces distributional similarity conditioned on the outcome class, aligning with the conditional independence requirements of EOdds [128].

However, naively minimising MMD between feature distributions across groups can paradoxically degrade task performance by maximising intra-class distances while minimising inter-class distances, destroying the discriminative information needed for accurate prediction [129]. This finding suggests that MMD-based fairness methods must carefully balance transferability (distributional similarity across groups) and discriminability (ability to distinguish task-relevant patterns). The CMAC-MMD framework introduced in this thesis addresses this challenge by applying an MMD-based regularisation not to high-dimensional feature representations but to low-dimensional decision confidence scores, directly regularising the quantity of clinical interest while preserving feature discriminability.

2.6 Synthesis: Identifying the Research Gap

The preceding sections have synthesised literature across six interconnected technical domains: VLM architectures and their medical adaptations, medical imaging datasets with demographic annotations, algorithmic fairness theory and metrics, fairness intervention methodologies, empirical evidence of bias in medical AI, and technical foundations including contrastive learning, self-supervised learning, domain adaptation, and kernel methods. This synthesis reveals both substantial progress and critical gaps that motivate the contributions of this thesis.

2.6.1 The Preponderance of Feature-Level, Single-Attribute Solutions

The landscape of fairness interventions is dominated by methods that operate at the feature representation level and address single demographic attributes in isolation. Adversarial debiasing approaches such as DANN and CDANN attempt to learn representations from which demographic attributes cannot be predicted [49], [121]. DRO methods like GroupDRO minimise worst-case loss over predefined groups [50]. Even sophisticated VLM-specific methods like FairCLIP optimise distributional alignment for one demographic axis at a time [57]. This single-axis, feature-level approach reflects the historical development of fairness research but suffers from two fundamental limitations that remain largely unaddressed.

First, single-attribute fairness interventions cannot provide guarantees about intersectional subgroups. As demonstrated mathematically and empirically by multiple studies, a model satisfying fairness constraints for race and for gender independently can simultaneously exhibit severe unfairness for race-gender intersections [27], [54], [58]. The exponential growth in the number of intersectional subgroups with increasing demographic dimensionality (2^d subgroups for d binary attributes) creates computational challenges that existing methods inadequately address. Current intersectional fairness frameworks such as DF and IF-α provide theoretical foundations and evaluation metrics but have minimal application to VLMs in medical domains.

Second, feature-level interventions provide no direct guarantee about fairness at the decision level where clinical predictions are made. A model may learn representations that are statistically independent of demographic attributes under some metric yet still produce systematically biased predictions if the classification head recovers demographic information through nonlinear transformations [48]. This gap between representation fairness and prediction fairness is particularly salient for medical applications where the quantity of clinical interest is not the model's internal features but its diagnostic confidence and decision outputs.

2.6.2 The Unaddressed Need for Equitable Diagnostic Certainty

A critical insight that has received minimal attention in fairness research is that even when models achieve similar accuracy across demographic subgroups, they may exhibit profound disparities in diagnostic certainty. This phenomenon, termed the "confidence gap" in preliminary work leading to this thesis, occurs when a model's predicted probabilities or alignment scores for one subgroup cluster close to the decision threshold, creating diagnostic uncertainty, while predictions for another subgroup are confidently above or below the threshold.

Diagnostic certainty matters clinically for multiple reasons. Predictions close to decision thresholds are inherently unstable, vulnerable to being flipped by minor perturbations in image quality, acquisition parameters, or other real-world variations [52]. Low-confidence predictions create challenges for clinical decision-making, as experienced clinicians recognize that borderline algorithmic recommendations warrant greater scrutiny and may appropriately place less weight on such recommendations [59]. Systematic confidence gaps across demographic groups undermine trust in AI systems, particularly among clinicians treating patients from subgroups receiving persistently uncertain predictions.

Despite the clinical importance of equitable diagnostic certainty, existing fairness metrics focus almost exclusively on binary decisions rather than the continuous confidence scores underlying those decisions. Accuracy, sensitivity, specificity, and their fairness analogs (EOdds, DPD) evaluate whether the final classification is correct but ignore how confident the model is in that classification. Calibration metrics examine whether predicted probabilities correspond to true outcome frequencies but do not assess whether the distribution of confidence scores

differs across groups beyond this correspondence. No existing fairness framework explicitly regularises the distribution of decision confidence across intersectional subgroups.

2.6.3 The Stated Gap: Decision-Level Intersectional Fairness for Vision-Language Models

Integrating the limitations identified above, this thesis addresses a precise gap in the literature: there exists no established framework designed to mitigate intersectional bias in VLMs by directly enforcing distributional consistency on decision-level confidence scores. While MMD has been used to align feature distributions for single-attribute fairness [125], [126], and while intersectional fairness metrics have been proposed for tabular and image classification tasks [54], [55], [58], the intersection of these approaches, applied specifically to the multimodal architecture of VLMs and targeting decision confidence rather than features, represents unexplored territory.

The opportunity to address this gap is amplified by the rapid adoption of VLMs as the state-of-the-art models for medical image analysis. As established in Section 2.2, VLMs including CLIP, BLIP-2, and their medical adaptations such as BiomedCLIP and PMC-CLIP are increasingly deployed for diagnostic support across radiology, dermatology, and ophthalmology [16], [17], [19], [20]. These systems combine visual and textual modalities through contrastive or generative objectives, creating rich semantic representations that enable zero-shot transfer, VQA, and report generation capabilities unavailable in traditional image classifiers. However, this architectural sophistication introduces new pathways for bias to emerge through interactions between vision and language modalities, through learned image-text associations that reflect biases in web-scraped or clinical training data, and through the complex loss landscapes of multimodal contrastive learning.

Existing VLM-specific fairness methods remain nascent and limited in scope. FairCLIP addresses single attributes sequentially and operates at the feature level [57]. FairerCLIP jointly debiases modalities but focuses on independence from protected attributes rather than decision confidence [73]. DeAR learns residual representations but does not explicitly target intersectional subgroups [74]. No existing method applies intersectional fairness constraints

at the decision level for VLMs, and no method explicitly regularises the distribution of diagnostic certainty across demographic intersections.

2.6.4 Additional Gaps and Opportunities

Beyond the primary gap in decision-level intersectional fairness, the literature synthesis reveals several secondary gaps that provide context and motivation. Mechanistic understanding of how and why VLMs encode demographic information remains incomplete [37], [70]. While multiple studies have demonstrated that models can predict race, gender, and age from medical images with high accuracy, the precise visual features enabling this prediction and whether these features are clinically relevant or spurious correlations are poorly characterised. Understanding whether demographic information is encoded in specific attention heads, embedding layers, or cross-modal alignment patterns could enable more targeted debiasing interventions.

Fairness-accuracy trade-offs require better characterisation specific to VLM architectures and medical applications. Theoretical impossibility results establish that multiple fairness criteria cannot be simultaneously satisfied [45], but empirical work suggests that in practice, carefully designed interventions can achieve substantial fairness improvements with minimal accuracy cost [122]. Quantifying the achievable Pareto frontier between fairness and accuracy for medical VLMs across multiple diseases and demographic intersections would provide decision-makers with actionable information about what trade-offs are necessary versus avoidable.

Evaluation standards lack consistency across studies, making comparison and reproducibility difficult. Different papers report different metrics (ΔTPR, AUC gaps, DPD, calibration error) on different datasets with different train-test splits and different implementations of baseline methods. Intersectional metrics such as ES-AUC and IF-α are mentioned in recent work but rarely implemented, and no consensus has emerged about which metrics are most clinically meaningful. Standardised evaluation protocols and benchmark datasets designed specifically for intersectional fairness assessment would accelerate progress.

Longitudinal clinical outcome studies are absent from the fairness literature. While retrospective studies document diagnostic accuracy disparities and landmark work quantifies resource misallocation [32], prospective studies tracking whether biased diagnostic algorithms lead to measurable differences in treatment timelines, disease progression, or mortality when deployed in real clinical workflows are lacking. Connecting algorithmic fairness metrics to patient outcomes would provide essential validation that technical fairness interventions translate to improvements in health equity.

Deployment and trust research remains nascent despite growing evidence that clinician responses to AI recommendations are complex and mediated by fairness perceptions [34], [59]. How does awareness of algorithmic bias affect clinician trust? Do clinicians exhibit differential reliance on AI recommendations across patient demographics? How do patients perceive and respond to learning that AI systems used in their care exhibit demographic bias? These human factors questions are essential for translating technical fairness solutions into equitable clinical impact.

2.7 Chapter Conclusion

This chapter has provided a comprehensive review of the literature that is foundational to understanding and addressing intersectional fairness in medical VLMs. The review traced the evolution from traditional supervised learning in medical image analysis to the contemporary frontier of vision-language architectures capable of joint reasoning about images and text, establishing the technical sophistication and clinical promise of models like CLIP, BLIP-2, BiomedCLIP, and PMC-CLIP. Examination of algorithmic fairness theory synthesised the mathematical frameworks that formalise equity, from DP and EOdds to intersectional extensions including max-min fairness, DF, and IF-α, while documenting the fundamental impossibility results that constrain what forms of fairness can be simultaneously achieved.

Critical analysis of the empirical literature revealed pervasive bias across medical imaging modalities, with documented performance disparities of 3% to 12% in diagnostic accuracy and particularly severe gaps at demographic intersections, exemplified by the Gender Shades finding of 34.7% error rates for dark-skinned women compared to less than 1% for

light-skinned men [35]. The clinical translation of these statistical disparities into tangible harms including delayed diagnosis, resource misallocation affecting 200 million people, and contribution to 30% mortality gaps between demographic groups establishes fairness as not merely a technical challenge but a clinical and ethical imperative.

Examination of fairness interventions spanning pre-processing, in-processing, and post-processing methods documented their theoretical foundations and empirical effectiveness while revealing critical limitations. The vast majority of existing methods operate at the feature representation level and address single demographic attributes in isolation, leaving them fundamentally incapable of guaranteeing fairness at decision boundaries or across intersectional subgroups. VLM-specific fairness methods including FairCLIP, FairerCLIP, and DeAR represent important advances but remain focused on single attributes and feature-level interventions.

The synthesis across all reviewed domains identified a precise research gap: the absence of frameworks designed to mitigate intersectional bias in VLMs by directly regularising diagnostic certainty at the decision level. While MMD provides a powerful tool for distributional alignment and has been applied to feature-level fairness, its application to low-dimensional decision confidence scores across intersectional demographic subgroups in multimodal medical architectures represents unexplored territory. This gap is particularly urgent given the rapid clinical adoption of VLMs for diagnostic support and the mounting evidence that these systems encode and exhibit demographic bias despite their sophisticated architectures and large-scale training.

The following chapter addresses this gap through the introduction of the CMAC-MMD framework, a novel training strategy that enforces distributional consistency on alignment scores quantifying diagnostic certainty across all intersectional patient subgroups. By operating at the decision level rather than feature level, by explicitly addressing intersectional combinations rather than single attributes, and by targeting diagnostic certainty rather than only accuracy, CMAC-MMD provides a rigorous and practical solution to the fairness challenges documented in this review.

CHAPTER 3

A Decision-Level Regularisation Framework to ensure Intersectional Fairness in Vision-Language Models for Medical Image Disease Classification

---

Chapter Abstract

This chapter introduces and validates the core methodological contribution of this thesis: the Cross-Modal Alignment Consistency Maximum Mean Discrepancy (CMAC-MMD) framework. Building upon the research gap identified in the literature review, the absence of interventions that address decision-level intersectional fairness and equitable diagnostic certainty in Vision-Language Models (VLMs), this chapter presents a novel regularisation strategy. The proposed method operates not on abstract feature representations but directly on a cross-modal alignment score, a proxy for the model's diagnostic confidence. By applying an adjusted loss based on the Maximum Mean Discrepancy (MMD) to force the distributional consistency of these scores across intersectional subgroups, the framework compels the model to produce predictions with equitable decisiveness. Through rigorous benchmarking on established dermatology and ophthalmology datasets, this chapter demonstrates that CMAC-MMD substantially reduces intersectional performance disparities, as measured by both traditional and advanced fairness metrics, including Differential Fairness (DF) and Intersectional Fairness-α (IF-α), while maintaining or improving overall diagnostic accuracy. The findings validate the efficacy of a decision-level approach and establish a robust, generalisable framework for developing more equitable and trustworthy medical artificial intelligence (AI) systems for high-stakes clinical applications.

3.1 Introduction: From Identified Gap to Proposed Solution

The synthesis presented in Chapter 2 established that existing fairness interventions for medical AI exhibit two critical limitations that hinder their applicability to modern vision-language models (VLM). As detailed in Section 2.6.1, current fairness methods are predominantly feature-level interventions that operate on single demographic attributes sequentially, making them insufficient for addressing the compounded disparities experienced by intersectional patient subgroups. Furthermore, Section 2.6.2 identified a fundamental gap in the literature: existing approaches fail to address the critical clinical challenge of equitable diagnostic certainty across intersectional subgroups. Even when models achieve nominally similar accuracy across patient groups, they can exhibit systematically lower confidence in decision-making for marginalised populations, creating a disparity in diagnostic reliability that conventional fairness metrics fail to capture. This phenomenon termed the "diagnostic certainty gap" poses significant clinical risks, as predictions for certain patient subgroups become unstable and vulnerable to misclassification from minor data shifts, ultimately eroding clinician trust and perpetuating unequal standards of algorithmic care, which is demonstrated in Figure 3.2B. The overall challenge, including the data imbalances that are a primary source of bias, is framed visually in Figure 3.1 and Figure 3.2A.

This chapter directly addresses these identified gaps by proposing and empirically validating the CMAC-MMD framework. The approach represents a conceptual departure from the feature-level interventions reviewed in Chapter 2, specifically those employing adversarial training [48], [121], robust optimisation [50], and data-centric rebalancing strategies [72], [130]. Rather than attempting to neutralise bias within the model's internal representations, CMAC-MMD targets the decision level, directly regularising the distribution of diagnostic confidence scores across all intersectional subgroups. This methodological shift is particularly salient for VLMs such as Contrastive Language-Image Pre-training (CLIP) [16] and Bootstrapping Language-Image Pre-training (BLIP2) [17], where complex cross-modal interactions between visual features and textual information can create or amplify biases that feature-level interventions fail to resolve [41]. Existing VLM-specific fairness methods, including FairCLIP [57], address fairness by optimising for single demographic attributes

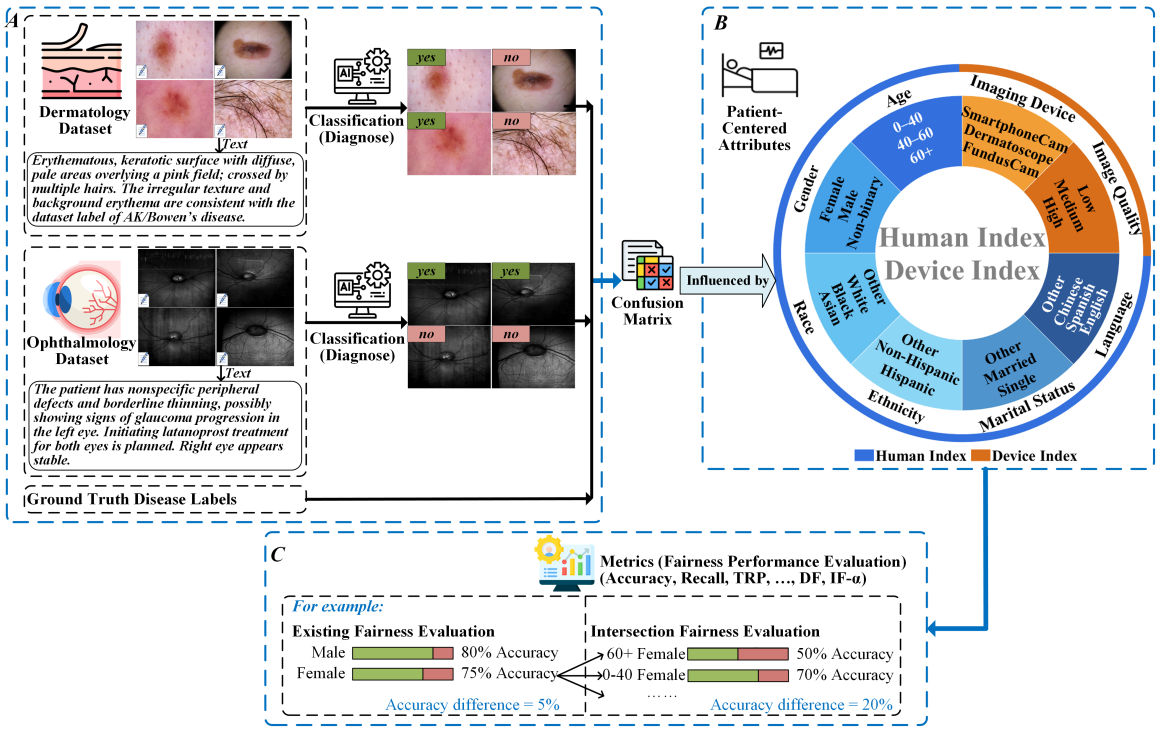

FIGURE 3.1: **The problem of intersectional fairness in medical vision-language models.** **A** Classification pipeline: dermatology and ophthalmology image-text pairs are processed through a VLM classifier to produce binary diagnostic decisions, which are compared against ground-truth disease labels to produce a confusion matrix. **B** Patient-centred attributes, grouped as Human Index (age, gender, race, ethnicity, language, marital status) and Device Index (imaging device, image quality), define the intersections along which model performance can diverge. **C** Illustration of why single-attribute fairness evaluation understates the intersectional gap: a 5% accuracy difference between Male and Female at the marginal level can mask accuracy differences of up to 20 percentage points at the gender–age intersection.

sequentially, an approach that proves inadequate for intersectional subgroups where correcting for one bias can inadvertently amplify another.

The CMAC-MMD framework operationalises the concept of equitable diagnostic certainty by defining a scalar cross-modal alignment score for each sample, representing the model's confidence margin between correct and incorrect predictions. This one-dimensional score serves as a direct proxy for diagnostic decisiveness. The method then employs a Maximum Mean Discrepancy (MMD)-based loss function to enforce that the statistical distributions of these alignment scores are indistinguishable across all intersectional subgroups defined by multiple demographic attributes. By aligning entire distributions rather than merely equalising means, the framework ensures that no patient subgroup is systematically subjected to uncertain

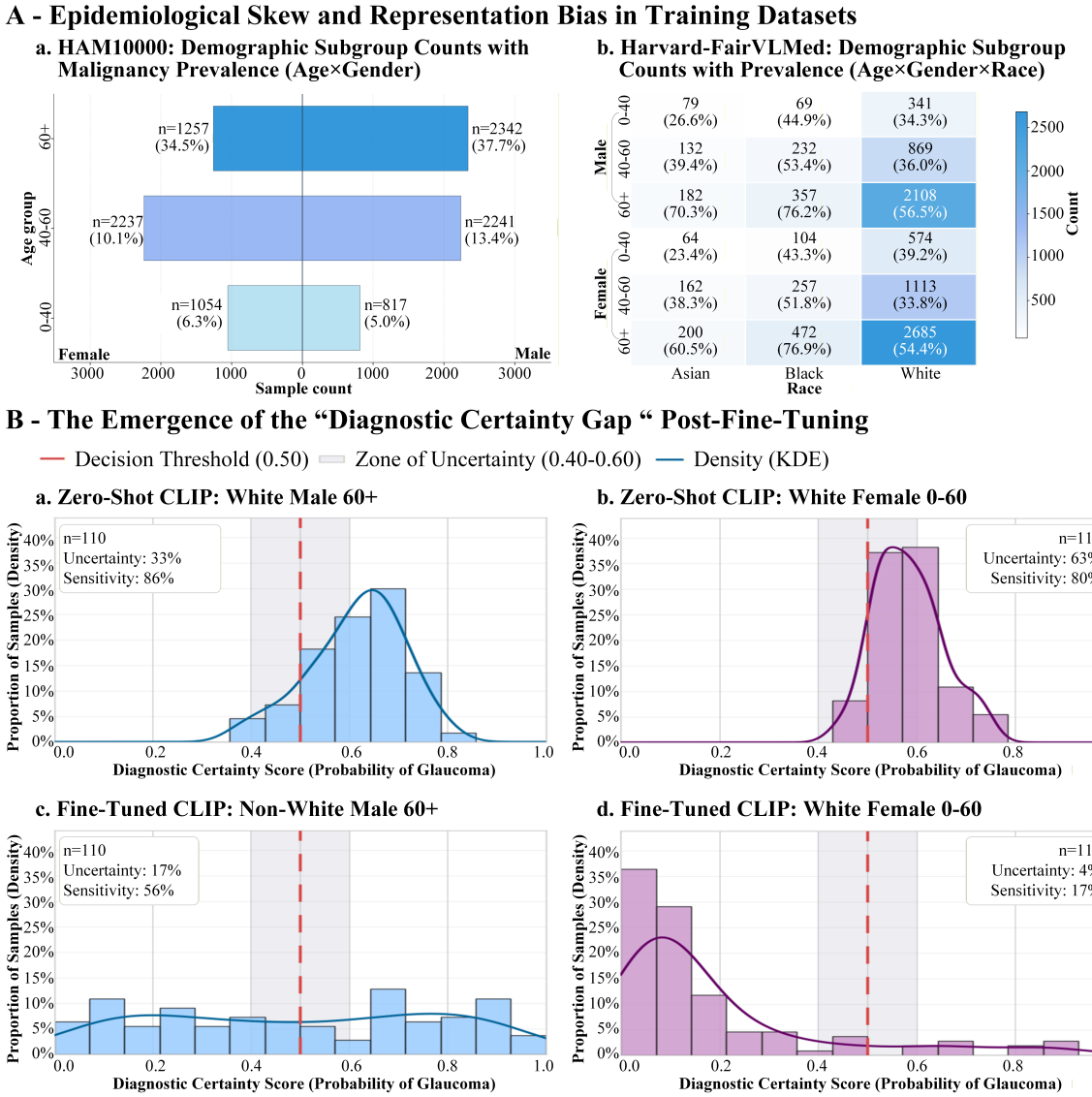

FIGURE 3.2: **Epidemiological skew in training data and the emergence of the diagnostic certainty gap after fine-tuning. A** Demographic subgroup counts with disease prevalence for the two primary training cohorts: HAM10000 dermatology and Harvard-FairVLMed ophthalmology. The distributions illustrate the subgroup-size imbalances that underpin downstream intersectional disparities. **B** Distribution of model-assigned diagnostic certainty scores for representative intersectional subgroups under zero-shot CLIP (a, b) and fine-tuned CLIP (c, d). Histograms are overlaid with a smoothed Kernel Density Estimate (KDE); the shaded band denotes the zone of uncertainty (0.40–0.60) around the decision threshold (vertical dashed line at 0.50).

borderline predictions that cluster perilously close to decision thresholds. This decision-level approach preserves patient privacy, as demographic attributes are required only during training

to compute the fairness loss and are not needed as model inputs during the inference stage. A high-level overview of the CMAC-MMD framework is visualised in Figure 3.3.

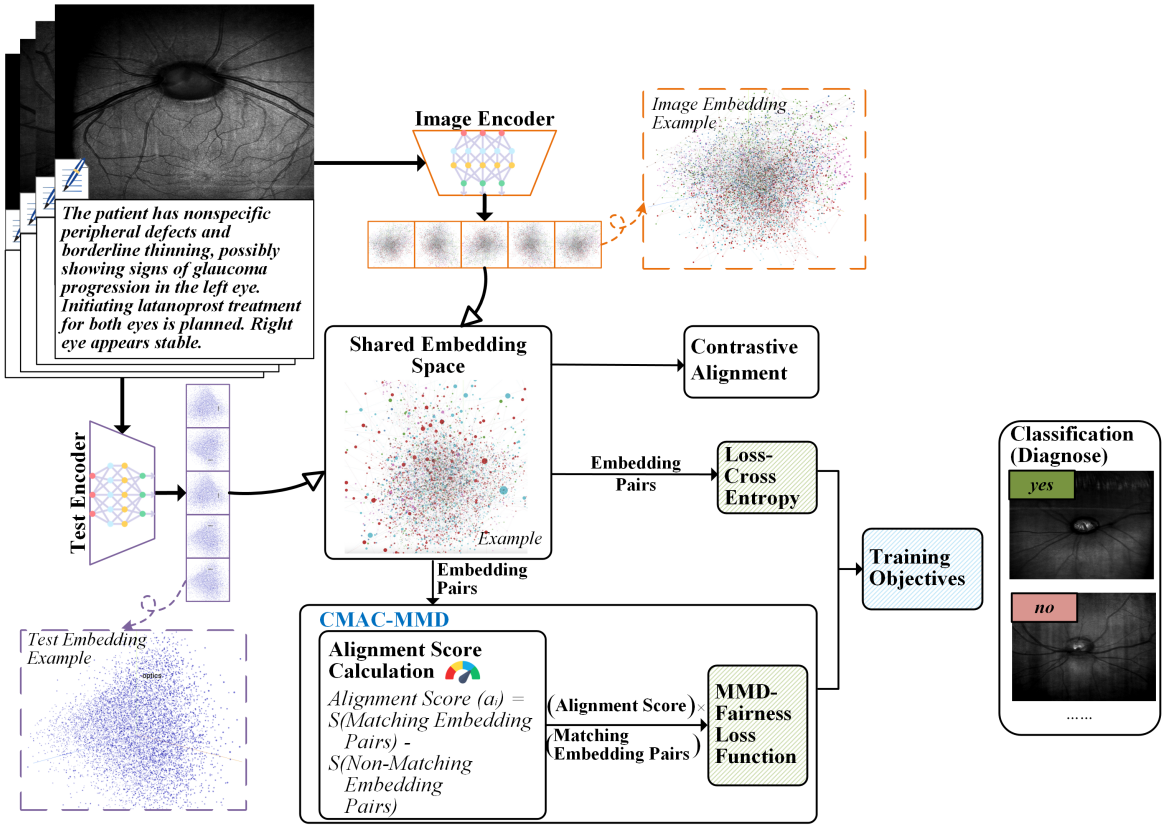

FIGURE 3.3: **High-level schematic of the CMAC-MMD framework.** Paired image and clinical-report inputs are projected into a shared embedding space by the image and text encoders. Matching and non-matching embedding pairs feed both the standard contrastive alignment branch (cross-entropy loss) and the CMAC-MMD fairness branch: a per-sample scalar Alignment Score is computed from the difference between matching and non-matching similarities, and an MMD-based fairness loss enforces distributional consistency of these scalar scores across intersectional patient subgroups. The composite training objective is used to fine-tune the encoders, which are then deployed for disease classification without demographic inputs at inference.

This chapter is structured to address all three research questions articulated in Section 1.5. RQ1 inquired how state-of-the-art (SOTA) VLMs exhibit intersectional bias in medical diagnostic tasks and questioned the adequacy of single-attribute fairness metrics in capturing these compounded disparities. This question is answered through comprehensive empirical benchmarking presented in Section 3.3.1, which demonstrates that standard fine-tuning of diverse VLM architectures consistently degrades intersectional fairness despite improving

overall accuracy. The analysis quantifies this trade-off across multiple model families. It reveals the diagnostic certainty gap, wherein fine-tuned models produce systematically lower confidence scores for marginalised intersectional subgroups. The inadequacy of conventional single-attribute metrics is demonstrated through comparative evaluation, which shows that disparities at demographic intersections substantially exceed those detected by single-axis analysis. RQ2 questioned whether a novel fairness framework could be developed that moves beyond feature-level adjustments to directly mitigate bias at the decision level by regularising diagnostic confidence across intersectional subgroups. This question is addressed through the methodological development presented in Section 3.2, which provides a rigorous formulation of the CMAC-MMD framework, including the derivation of the cross-modal alignment score, the improvements on MMD for distributional consistency, and the complete training objective. RQ3 examined the effectiveness of the proposed CMAC-MMD framework in reducing intersectional bias while maintaining or improving overall diagnostic performance on established medical imaging benchmarks. This question is answered through the empirical validation presented in Sections 3.3.2 through 3.3.4, demonstrating substantial reductions in intersectional disparities of 20-35% as measured by both traditional metrics and advanced intersectional fairness measures, including Differential Fairness (DF) and Intersectional Fairness-α (IF-α) [54], [55], while maintaining or improving Area Under the Curve (AUC) by up to 5%. External validation on independent datasets and ablation studies further establish the generalisability and robustness of the approach.

The remainder of this chapter proceeds as follows. Section 3.2 provides a comprehensive account of the CMAC-MMD methodology, detailing the architectural foundation in contrastive VLM pre-training, the formulation of the novel fairness regularizer, the complete training objective, and the experimental design, including dataset selection, baseline comparisons, and evaluation metrics. Section 3.3 presents the empirical results across dermatology and ophthalmology tasks. Section 3.4 interprets these findings, contextualises them within the literature reviewed in Chapter 2, and discusses the clinical and scientific implications alongside the limitations of the current study. Section 3.5 concludes the chapter and transitions to the thesis conclusion in Chapter 4, where broader implications for trustworthy medical AI and future research directions are synthesised.

3.2 Methodology: The CMAC-MMD Framework

This section provides a comprehensive account of the CMAC-MMD framework, detailing its theoretical foundation, architectural components, and experimental design. The presentation builds upon the concepts established in Chapter 2, applying them to construct a novel decision-level fairness intervention specifically designed for VLMs in medical imaging contexts.

3.2.1 Architectural Foundation: Contrastive Vision-Language Pre-training

The CMAC-MMD framework operates upon the architectural foundation of contrastive VLMs, specifically the dual-encoder paradigm established by CLIP [16]. As detailed in Section 2.2.2, this architecture employs two separate neural network encoders, one for visual inputs and one for textual inputs, that project images and text into a shared, semantically meaningful embedding space. The training relies on the InfoNCE contrastive objective, which maximises the similarity between correctly paired image-text representations while minimising the similarity between mismatched pairs.

Formally, consider a dataset $\mathcal{D} = \{(\mathbf{I}_n, \mathbf{T}_n, y_n, \mathbf{a}_n)\}_{n=1}^{N}$, where each sample comprises a medical image $\mathbf{I}_n$, an associated textual description $\mathbf{T}_n$, a binary disease label $y_n \in \{0, 1\}$, and a vector of demographic attributes $\mathbf{a}_n$ that define the patient's intersectional subgroup. For dermatological applications, $\mathbf{T}_n$ represents a concise clinical description embedding the diagnostic label (for example, "A dermoscopic photograph of a benign pigmented skin lesion with no evidence of melanoma"). For ophthalmological tasks, $\mathbf{T}_n$ corresponds to structured clinical reports containing relevant patient history and examination findings. The demographic attribute vector $\mathbf{a}_n$ encodes information such as age category, gender, and race, which is utilised exclusively during training to define intersectional subgroups for fairness regularisation.

The CLIP architecture comprises an image encoder $\phi_\theta : \mathcal{I} \to \mathbb{R}^d$ parameterised by θ and a text encoder $\psi_\varphi : \mathcal{T} \to \mathbb{R}^d$ parameterised by φ, both projecting their respective inputs into a shared d-dimensional embedding space. These encoders produce ℓ_2-normalised representations: $\mathbf{z}_n^I = \phi_\theta(\mathbf{I}_n)/\|\phi_\theta(\mathbf{I}_n)\|_2$ and $\mathbf{z}_n^T = \psi_\varphi(\mathbf{T}_n)/\|\psi_\varphi(\mathbf{T}_n)\|_2$. During training, a mini-batch $\mathcal{B}$ of

N image-text pairs is processed, and the model computes pairwise cosine similarities between all image and text embeddings, forming a similarity matrix $\mathbf{S} \in \mathbb{R}^{N \times N}$ where each element $S_{ij} = \tau \cdot (\mathbf{z}_i^I)^\top \mathbf{z}_j^T$ is scaled by a learnable temperature parameter $\tau = \exp(\omega)$.

The standard symmetric contrastive learning objective, which forms the primary diagnostic training signal, is expressed as:

$$\mathcal{L}_{\text{CLIP}} = \frac{1}{2N} \sum_{i=1}^{N} \left[-\log \frac{\exp(S_{ii})}{\sum_{j=1}^{N} \exp(S_{ij})} - \log \frac{\exp(S_{ii})}{\sum_{j=1}^{N} \exp(S_{ji})} \right] \tag{3.1}$$

This objective encourages the model to assign high similarity scores to correct image-text pairs (diagonal elements S_{ii}) while treating all other pairs within the batch as negative examples. This contrastive formulation enables the model to learn clinically meaningful representations that capture the semantic correspondence between visual pathology and textual diagnostic descriptions. However, as established in Section 2.5.2, this standard training procedure, when applied via Empirical Risk Minimisation (ERM), systematically amplifies intersectional biases present in the training data distribution. The CMAC-MMD framework addresses this limitation by augmenting the standard contrastive objective with a novel fairness regularisation term that operates directly on the model's decision-level outputs rather than its internal feature representations.

3.2.2 A Novel Regularizer for Equitable Diagnostic Certainty

3.2.2.1 The Conceptual Core: Formalising Diagnostic Certainty

The central innovation of the CMAC-MMD framework lies in its explicit formalisation and regularisation of diagnostic certainty at the decision level. Rather than attempting to debias high-dimensional feature representations, an approach that, as discussed in Section 2.5.2, often fails to translate into equitable downstream performance, the proposed method targets a scalar quantity that directly reflects the model's confidence in its diagnostic predictions.

This quantity, termed the cross-modal alignment score, is derived for each sample as a measure of how decisively the model separates the correct image-text pairing from the most compelling

incorrect alternatives within a mini-batch. For a given image-text pair $(\mathbf{I}_i, \mathbf{T}_i)$ in batch $\mathcal{B}$, the alignment score is constructed from two directional confidence margins. The image-to-text margin, denoted $a_i^{I \to T}$, quantifies the difference between the similarity of the correct pair S_{ii} and the highest similarity between image $\mathbf{I}_i$ and any mismatched text in the batch:

$$a_i^{I \to T} = S_{ii} - \max_{j \neq i} S_{ij} \tag{3.2}$$

Symmetrically, the text-to-image margin $a_i^{T \to I}$ measures the difference between the correct pairing similarity and the strongest competing image for text $\mathbf{T}_i$:

$$a_i^{T \to I} = S_{ii} - \max_{j \neq i} S_{ji} \tag{3.3}$$

The final alignment score for sample i is defined as the arithmetic mean of these bidirectional margins:

$$a_i = \frac{1}{2} \left(a_i^{I \to T} + a_i^{T \to I} \right) \tag{3.4}$$

The "Alignment Score Calculation (ASC)" module illustrates this calculation in Figure 3.4. A positive alignment score ($a_i > 0$) indicates that the correct image-text pair exhibits greater similarity than any incorrect pairing, reflecting high diagnostic certainty. Conversely, a score near zero or negative suggests that the model's prediction is borderline or incorrect, placing the sample in a region of diagnostic uncertainty. Critically, this scalar score serves as a direct proxy for the model's decisiveness at the point of clinical action, capturing precisely the quantity identified as problematic in Section 2.6.2: the confidence gap that leaves certain patient subgroups systematically subjected to uncertain predictions.

The justification for targeting this one-dimensional score rather than high-dimensional feature representations is threefold. First, it provides a transparent and clinically interpretable measure of model confidence that directly relates to diagnostic decision-making. Second, it avoids the degeneracy problems that plague feature-level fairness interventions, where enforcing similarity in abstract embedding spaces can inadvertently harm the discriminative capacity of

the representations [50], [121]. Third, by operating in a low-dimensional space, the fairness regularizer can be applied more effectively with the limited sample sizes available for each intersectional subgroup within a mini-batch, a practical constraint highlighted in the literature as a significant barrier to intersectional fairness optimisation [55].

3.2.2.2 Distributional Alignment via Maximum Mean Discrepancy on Cross-Modal Consistency

The framework employs adjusted MMD, a non-parametric statistical distance measure whose theoretical foundations are established in Section 2.5.3 to enforce equitable diagnostic certainty across intersectional subgroups. The key insight is that fairness should be defined not merely by equality of mean alignment scores across subgroups, but by consistency of the entire distribution of diagnostic confidence. This distributional perspective ensures that no subgroup experiences a disproportionate concentration of borderline predictions, even if average performance metrics appear equitable.

For each intersectional subgroup g represented within a training mini-batch $\mathcal{B}$, the set of alignment scores $\mathcal{A}_g = \{a_i \mid \mathbf{a}_i \in g, i \in \mathcal{B}\}$ forms a one-dimensional empirical distribution. The cardinality of this set, $m_g = |\mathcal{A}_g|$, represents the number of samples from subgroup g present in the batch. The CMAC-MMD regulariser computes the squared MMD between the alignment score distributions of each pair of subgroups (g, g') to quantify their statistical dissimilarity.

The squared MMD between two distributions is defined as the squared distance between their mean embeddings in a Reproducing Kernel Hilbert Space (RKHS). For finite samples drawn from subgroups g and g', the unbiased empirical estimator of the squared MMD between their alignment-score distributions, consistent with the estimator at Eq. (2.23), is:

$$\begin{aligned}\widehat{\text{CMAC-MMD}}^2(\mathcal{A}_g, \mathcal{A}_{g'}) =& \frac{1}{m_g(m_g-1)} \sum_{i\in\mathcal{I}_g} \sum_{\substack{j\in\mathcal{I}_g \\ j\neq i}} k(a_i, a_j) \\ &+ \frac{1}{m_{g'}(m_{g'}-1)} \sum_{i\in\mathcal{I}_{g'}} \sum_{\substack{j\in\mathcal{I}_{g'} \\ j\neq i}} k(a_i, a_j) \\ &- \frac{2}{m_g m_{g'}} \sum_{i\in\mathcal{I}_g} \sum_{j\in\mathcal{I}_{g'}} k(a_i, a_j),\end{aligned} \tag{3.5}$$

where $\mathcal{I}_g$ and $\mathcal{I}_{g'}$ denote the index sets of samples belonging to subgroups g and g' respectively, and $k(\cdot, \cdot)$ is a positive-definite kernel function. The first two terms represent the within-subgroup kernel similarities (excluding self-comparisons to ensure unbiasedness), while the third term captures the cross-subgroup similarities.

The choice of kernel function is critical for the effectiveness of the MMD regulariser. The framework employs the Radial Basis Function (RBF) kernel, also known as the Gaussian kernel, which for two scalar alignment-score values a, a' is defined as:

$$k(a, a') = \exp\left(-\gamma(a - a')^2\right), \quad \text{where} \quad \gamma = \frac{1}{2\sigma^2} \tag{3.6}$$

The RBF kernel is selected for several compelling reasons. First, it is a universal kernel, meaning it can approximate any continuous function arbitrarily well, providing the flexibility necessary to capture complex distributional differences in alignment scores [131]. Second, its characteristic length-scale parameter σ controls the sensitivity to local variations in the score distributions, allowing the regularizer to detect subtle disparities in diagnostic certainty. Third, the kernel's infinite-dimensional feature mapping ensures that the MMD measure captures differences across all statistical moments of the distributions, not merely the mean or variance [132]. In the implementation, the bandwidth parameter γ is either specified as a hyperparameter or adaptively computed using the median heuristic, which sets σ to the median pairwise distance between all alignment scores in the batch, providing a data-driven scale that adapts to the characteristic spread of confidence scores.

3.2.3 The Complete CMAC-MMD Training Objective

The complete CMAC-MMD training objective combines the standard contrastive learning loss with the proposed fairness regularisation term. Let $\mathcal{P}$ denote the set of all distinct subgroup pairs (g, g') where $g < g'$ (in some canonical ordering) and both subgroups are represented in the current mini-batch with at least two samples each ($m_g \geq 2$ and $m_{g'} \geq 2$). The CMAC-MMD fairness loss is defined as the average squared MMD across all such valid subgroup pairs:

$$\mathcal{L}_{\text{CMAC}} = \frac{1}{|\mathcal{P}|} \sum_{(g,g') \in \mathcal{P}} \widehat{\text{CMAC-MMD}}^2(\mathcal{A}_g, \mathcal{A}_{g'}) \tag{3.7}$$

This formulation ensures that the fairness penalty reflects the extent to which the distributions of diagnostic certainty diverge across all pairwise comparisons of intersectional subgroups. By minimising this quantity, the model is explicitly trained to produce alignment score distributions that are statistically indistinguishable across demographic intersections, thereby operationalising the principle of equitable diagnostic certainty.

The total training objective augments the standard CLIP contrastive loss with the weighted CMAC-MMD penalty:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{CLIP}} + \lambda_{\text{CMAC}} \cdot \mathcal{L}_{\text{CMAC}}, \tag{3.8}$$

where $\lambda_{\text{CMAC}} \geq 0$ is a hyperparameter that controls the strength of the fairness regularisation. This composite loss function is summarised in the "Total Training Objective" part of the framework schematic (Figure 3.4). This parameter serves as a critical lever for balancing the dual objectives of diagnostic performance and intersectional fairness. When $\lambda_{\text{CMAC}} = 0$, the framework reduces to standard ERM training, which, as demonstrated in Section 3.3.1, leads to severe fairness degradation. As λ_{CMAC} increases, the model is increasingly constrained to equalise the distributions of diagnostic confidence across subgroups, potentially at the cost of overall classification accuracy. The optimal value of λ_{CMAC} thus represents

a carefully calibrated trade-off between these competing desiderata, and is determined empirically through validation set performance across both accuracy and fairness metrics. A comprehensive sensitivity analysis across seven values of λ_{CMAC} spanning a 500-fold range, reported in Section 3.3.6, confirms that the operating point selected throughout this chapter ($\lambda_{\text{CMAC}} = 0.5$) is robust across both clinical cohorts and does not sit near any phase transition.

The training procedure alternates between forward passes that compute both loss components and backward passes that update the model parameters θ and φ (as well as the temperature parameter ω) via gradient descent. Crucially, the demographic attributes $\mathbf{a}_n$ are utilised exclusively to subgroup samples and compute the CMAC-MMD loss during training; they are never provided as inputs to the image or text encoders. This architectural choice ensures that the resulting model respects patient privacy and can be deployed in clinical settings where demographic information may be unavailable, incomplete, or protected by privacy regulations. The fairness intervention thus operates through implicit regularisation of the learning dynamics rather than through explicit conditioning on sensitive attributes.

3.2.4 Experimental Design for Empirical Validation

3.2.4.1 Datasets and Intersectional Subgroup Definition

The empirical validation of the CMAC-MMD framework is conducted on three established medical imaging datasets that provide the demographic annotations necessary for intersectional fairness analysis. The primary training and evaluation dataset for dermatological disease classification is HAM10000 [90], comprising 10,015 dermoscopic images of pigmented skin lesions with associated age and gender attributes. The dataset is partitioned into training (approximately 7,000 images), validation (1,000 images), and test (2,000 images) sets using stratified sampling to maintain representative subgroup distributions across splits. To assess the generalisability of the fairness benefits achieved by CMAC-MMD under distribution shift, the BCN20000 dataset [133], an external collection of 20,000 dermoscopic images from a different clinical institution, is utilised as a held-out validation set.

For ophthalmological applications, the Harvard-FairVLMed dataset [57] is employed, consisting of 10,000 fundus photographs paired with clinical reports, annotated with patient age,

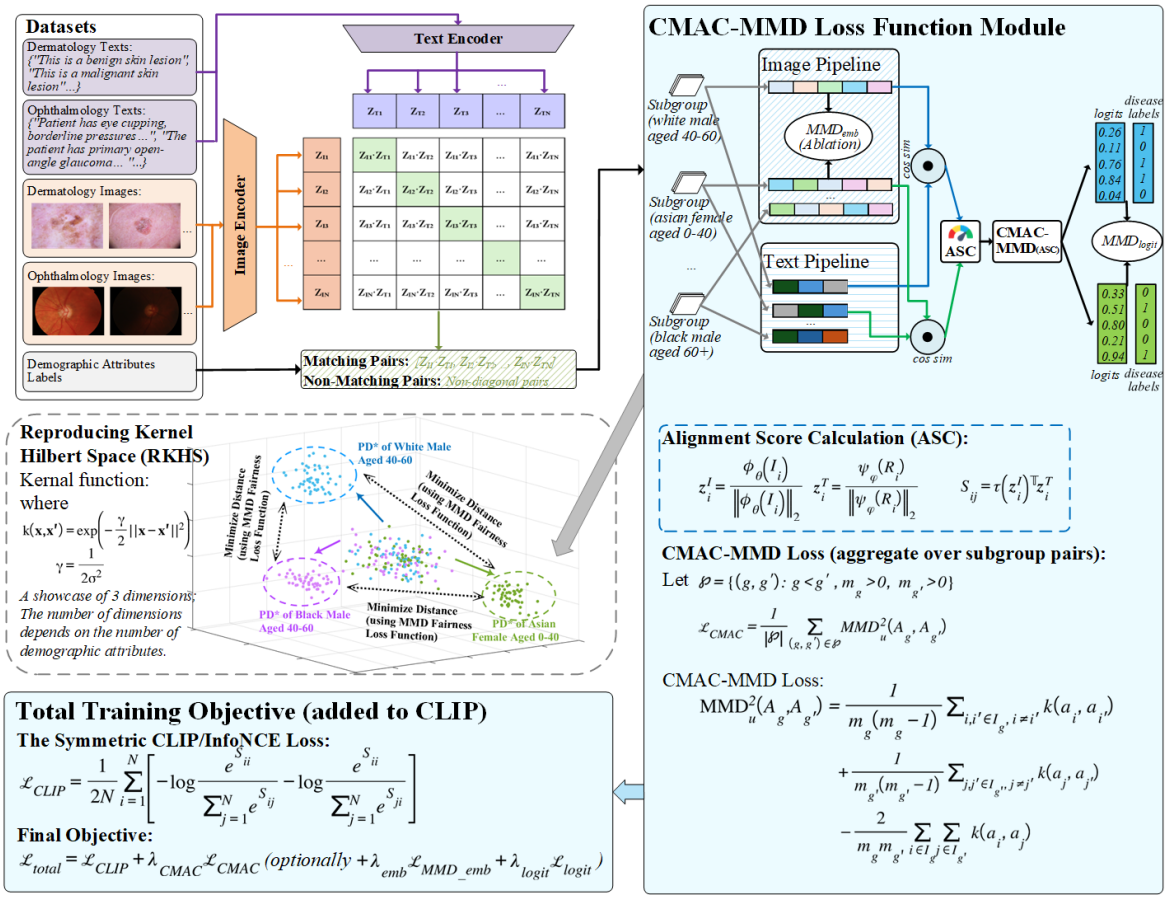

FIGURE 3.4: **Schematic of the CMAC-MMD framework.** The model takes image-text pairs and demographic attributes as input. Instead of regularising high-dimensional embeddings, CMAC-MMD first computes a per-sample, scalar **Alignment Score** (ASC), quantifying the model's diagnostic certainty. These scalar scores are grouped by subgroup, and the CMAC-MMD loss is calculated to result in 1-D distributions, enforcing that the distribution of model certainty is consistent across all intersectional patient groups. The diagram also illustrates the placement of the ablation MMD regularisers benchmarked in Section 3.3.5 (Table 3.5): applied at the image embeddings, the text embeddings, and the final disease logits. In every variant reported in the chapter, the CMAC-MMD row combines $\mathcal{L}_{\text{CLIP}}$ with a single fairness regulariser on the scalar alignment score and does not include any feature-level or logit-level term. The RKHS visualisation and RBF kernel [131], [132] depict the conceptual goal: minimising the distance between the probability distributions (PD*) of different subgroups to achieve fairness. The final training objective combines the standard symmetric CLIP/InfoNCE loss with the weighted $\mathcal{L}_{\text{CMAC}}$ penalty.

gender, race, ethnicity, spoken language, and marital status. This dataset undergoes a similar stratified split into training (5,968 images), validation (2,032 images), and test (2,000 images) sets.

The definition of intersectional subgroups represents a critical methodological decision that balances clinical relevance, statistical validity, and computational feasibility. For the dermatological datasets, age is stratified into three categories: 0-40 years, 41-60 years, and 60+ years. This stratification reflects established clinical knowledge regarding the epidemiology of skin cancer, where risk undergoes significant inflection points around age 40 and accelerates substantially after age 60 [90]. Combined with binary gender, this yields six intersectional subgroups. The decision to use three age bins rather than finer-grained categories or continuous age is necessitated by the requirement for adequate sample sizes within each subgroup, a fundamental challenge in intersectional fairness highlighted in Section 2.4.2. Each subgroup must contain a sufficient number of samples (ideally exceeding 50-100) to enable reliable computation of fairness metrics and stable estimation of the loss during training [36].

For the ophthalmological dataset, which includes racial information in addition to age and gender, a binary age threshold (0-60 versus 60+) and a binary race categorisation (White versus Non-White) are employed, yielding eight intersectional subgroups. The age threshold of 60 years is justified by the exponential increase in glaucoma prevalence beyond this point, rising from approximately 1% in younger populations to over 3% in individuals aged 60 and above [60]. The binarisation of race, while admittedly a simplification that obscures important within-group heterogeneity, is a pragmatic necessity to prevent the proliferation of subgroups from rendering many categories statistically underpowered. These stratification decisions are explicitly acknowledged as methodological limitations in Section 3.4.4, recognising that demographic categories are social constructs that imperfectly capture the true diversity of patient populations. Section 4.3.2 in Chapter 4 treats the methodological mitigation of this statistical-power bottleneck as the primary prerequisite for any extension of CMAC-MMD to additional demographic attributes or finer-grained categorisations, and enumerates four concrete strategies, variance-corrected and small-sample MMD estimators, hierarchical and multi-resolution fairness constraints, subgroup-aware batch construction with minimum-count guarantees, and a priori power analysis to guide attribute inclusion, that must be developed before attribute coverage is expanded.

3.2.4.2 Baselines and Comparative Interventions Adaptation on VLMs

To rigorously assess the effectiveness of the CMAC-MMD framework, the proposed method is benchmarked against a comprehensive suite of fairness interventions spanning both pre-processing (data-level) and in-processing (algorithmic-level) stages. A critical methodological consideration is that the comparative methods selected were not originally designed for VLMs, and certainly not for the dual-encoder CLIP architecture employed in this study. Instead, these techniques were developed for conventional supervised learning tasks involving single-modality inputs or for domain adaptation problems distinct from fairness optimisation. Consequently, substantial adaptation is required to enable these methods to operate within the contrastive learning framework of CLIP, which lacks explicit class-conditional layers and instead relies on similarity-based matching between image and text embeddings. The following exposition details both the original formulation of each baseline method and the specific architectural modifications implemented to ensure fair comparison within the VLM context.

Pre-processing Methods: Data-Level Interventions

Pre-processing approaches attempt to mitigate bias by modifying the training data distribution before model training, operating under the assumption that correcting data-level imbalances will reduce downstream algorithmic disparities. Two canonical pre-processing techniques are evaluated in this study.

Resampling. Resampling is a fundamental data balancing technique that adjusts the representation of different demographic subgroups in the training set to achieve distributional parity [47]. The method operates by computing subgroup-specific sampling probabilities designed to counteract empirical underrepresentation. For a dataset $\mathcal{D} = \{(\mathbf{I}_n, \mathbf{T}_n, y_n, \mathbf{a}_n)\}_{n=1}^{N}$ stratified into G intersectional subgroups based on the demographic attribute vectors $\mathbf{a}_n$, the empirical subgroup counts are denoted $n_g = |\{i : \mathbf{a}_i \in g\}|$ for $g = 1, \ldots, G$. Resampling constructs a balanced dataset by assigning to each sample i belonging to subgroup g a sampling weight inversely proportional to the subgroup size:

$$w_i^{\text{resample}} = \frac{1}{n_g} \quad \text{where } \mathbf{a}_i \in g \tag{3.9}$$

These weights are then employed within a WeightedRandomSampler during mini-batch construction, effectively oversampling underrepresented subgroups and undersampling overrepresented ones with replacement to produce training batches with approximately uniform subgroup representation. In the CLIP context, this resampling procedure requires no architectural modification, as the contrastive loss $\mathcal{L}_{\text{CLIP}}$ remains unchanged. However, the empirical composition of each mini-batch is altered such that minority intersectional subgroups (for instance, young Non-White females in the ophthalmology dataset) appear with higher frequency than their proportion in the original training set would dictate. The implementation employs PyTorch's WeightedRandomSampler with num_samples $= N$ and replacement $=$ True to ensure that all original samples remain accessible while enforcing balanced subgroup sampling.

Reweighting. Reweighting is an alternative pre-processing strategy that, rather than altering the data sampling distribution, modifies the contribution of each sample to the training loss by assigning differential importance weights [72], [134]. The reweighting approach is grounded in the principle that samples from underrepresented subgroups should exert greater influence on parameter updates to compensate for their numerical scarcity. For each sample i belonging to intersectional subgroup g, a loss weight is computed using the same inverse-frequency formula:

$$w_i^{\text{reweight}} = \frac{1}{n_g} \quad \text{where } \mathbf{a}_i \in g \tag{3.10}$$

The standard element-wise cross-entropy formulation must be modified to integrate reweighting into the CLIP training framework. Specifically, the contrastive loss for each sample is computed individually and then scaled by the corresponding sample weight. Given a mini-batch $\mathcal{B}$ of N samples, the reweighted symmetric contrastive objective becomes:

$$\mathcal{L}_{\text{CLIP}}^{\text{reweight}} = \frac{1}{2N} \sum_{i=1}^{N} w_i^{\text{reweight}} \left[-\log \frac{\exp(S_{ii})}{\sum_{j=1}^{N} \exp(S_{ij})} - \log \frac{\exp(S_{ii})}{\sum_{j=1}^{N} \exp(S_{ji})} \right] \tag{3.11}$$

where $S_{ij} = \tau \cdot (\mathbf{z}_i^I)^\top \mathbf{z}_j^T$ represents the scaled cosine similarity between image i and text j embeddings. The implementation employs PyTorch's CrossEntropyLoss with reduction $=$ 'none' to obtain per-sample losses, which are then element-wise multiplied by the weight vector before averaging. This approach ensures that gradients propagated to the encoders are amplified for minority subgroup samples, implicitly guiding the model toward representations that better accommodate underrepresented patient demographics.

In-processing Methods: Algorithmic-Level Interventions

In-processing methods modify the learning algorithm itself, typically by augmenting the standard training objective with additional fairness-oriented constraints or auxiliary losses. These techniques operate during training and require architectural or optimisation-level modifications to the base VLM framework.

Mean Accuracy. Mean accuracy optimisation, also referred to as average group accuracy maximisation, directly targets fairness by defining the training objective as the arithmetic mean of per-subgroup accuracies rather than overall subgroup-level accuracy. This approach ensures that all demographic groups contribute equally to the optimisation criterion regardless of their relative sizes. Let $\mathcal{B}_g \subseteq \mathcal{B}$ denote the subset of samples in mini-batch $\mathcal{B}$ belonging to intersectional subgroup g. The mean accuracy objective for CLIP is formulated as:

$$\mathcal{L}_{\text{MeanAcc}} = \frac{1}{|G_{\mathcal{B}}|} \sum_{g \in G_{\mathcal{B}}} \mathcal{L}_{\text{CLIP}}(\mathcal{B}_g) \tag{3.12}$$

where $G_{\mathcal{B}}$ represents the set of distinct intersectional subgroups present in the current mini-batch, and $\mathcal{L}_{\text{CLIP}}(\mathcal{B}_g)$ denotes the standard symmetric contrastive loss computed exclusively over samples from subgroup g. This formulation assigns equal weight to each subgroup's loss, effectively upweighting minority groups and downweighting majority groups in proportion to their representation discrepancies. The implementation tracks subgroup membership for all samples within each batch, partitions the batch accordingly, computes separate contrastive losses for each subgroup partition, and averages these subgroup-specific losses to form the total training objective. This method requires no architectural modification but fundamentally

alters the optimisation landscape by treating each demographic intersection as an equally important optimisation target.

Group Distributionally Robust Optimisation (GroupDRO). GroupDRO, originally developed for handling distribution shift and improving worst-case generalisation across predefined groups [50], has been adopted for fairness applications under the principle that mitigating worst-group performance directly addresses the most severe manifestations of algorithmic bias. The method frames fairness as a robust optimisation problem, seeking to minimise the maximum loss across all demographic subgroups. Formally, for the set of intersectional subgroups $\{g\}_{g=1}^{G}$ represented in the training data, the GroupDRO objective is:

$$\mathcal{L}_{\text{GroupDRO}} = \max_{g \in G_{\mathcal{B}}} \mathcal{L}_{\text{CLIP}}(\mathcal{B}_g) \tag{3.13}$$

This minimax formulation prioritises the demographic subgroup experiencing the highest training loss in each mini-batch, concentrating optimisation effort on the most disadvantaged group. The adaptation to CLIP proceeds identically to Mean Accuracy optimisation in terms of batch partitioning: for each mini-batch $\mathcal{B}$, samples are segregated by their intersectional subgroup membership, individual contrastive losses are computed for each subgroup partition $\mathcal{B}_g$, and the maximum loss across all represented subgroups is selected as the training signal. After computing all subgroup-specific losses, the implementation employs PyTorch's torch.max operator to identify the worst-performing group. This approach requires careful hyperparameter tuning, as the exclusive focus on the worst group can lead to training instability, particularly when mini-batch compositions vary substantially across iterations. To mitigate instability, the official GroupDRO implementation incorporates an exponential moving average over group losses [50], but for consistency with other baselines, the present study employs the simpler instantaneous maximum formulation.

Domain-Adversarial Neural Networks (DANN). DANN was originally proposed by Ganin et al. [49] for unsupervised domain adaptation, where the objective is to learn feature representations that are invariant to domain identity while remaining discriminative for the task. The method has been repurposed for fairness by treating demographic attributes as "domains" and seeking representations that cannot discriminate between different demographic groups. The DANN

architecture augments the feature extractor with a domain classifier trained adversarially via a gradient reversal layer (GRL). For the CLIP dual-encoder architecture, DANN requires introducing two separate adversarial branches: one for the image encoder and one for the text encoder.

The DANN training objective combines the primary contrastive classification loss with an adversarial demographic prediction loss. Formally, let $D_{\text{img}} : \mathbb{R}^d \to \mathbb{R}^G$ and $D_{\text{text}} : \mathbb{R}^d \to \mathbb{R}^G$ denote domain discriminators (implemented as multi-layer perceptrons) that attempt to predict the intersectional subgroup membership $g \in \{1, \ldots, G\}$ from the image and text embeddings respectively. The gradient reversal layer, denoted as $\text{GRL}(\cdot, \lambda)$, is a pseudo-layer that acts as the identity function during forward propagation but multiplies gradients by $-\lambda$ during backpropagation, where $\lambda > 0$ controls the strength of the adversarial effect. The complete DANN objective for CLIP is:

$$\mathcal{L}_{\text{DANN}} = \mathcal{L}_{\text{CLIP}} + \lambda_{\text{adv}} \left(\mathcal{L}_{\text{dom}}^{\text{img}} + \mathcal{L}_{\text{dom}}^{\text{text}} \right) \tag{3.14}$$

where the domain classification losses are defined as:

$$\mathcal{L}_{\text{dom}}^{\text{img}} = -\frac{1}{N} \sum_{i=1}^{N} \log P \left(g_i \mid D_{\text{img}}(\text{GRL}(\mathbf{z}_i^I, \lambda_{\text{GRL}})) \right) \tag{3.15}$$

$$\mathcal{L}_{\text{dom}}^{\text{text}} = -\frac{1}{N} \sum_{i=1}^{N} \log P \left(g_i \mid D_{\text{text}}(\text{GRL}(\mathbf{z}_i^T, \lambda_{\text{GRL}})) \right) \tag{3.16}$$

During training, the domain discriminators D_{img} and D_{text} are optimised to correctly classify the intersectional subgroup of each embedding, while the image and text encoders receive reversed gradients that encourage them to produce embeddings that confound the domain discriminators. This adversarial game theoretically drives both encoders toward representations that are invariant to demographic attributes. The implementation employs a custom PyTorch autograd function for the GRL and two-layer MLPs (with hidden dimension 256 and ReLU activation) for the domain discriminators. The adversarial weight λ_{adv} is set to 1.0 and the gradient reversal coefficient λ_{GRL} is also set to 1.0 following standard practice [49]. Critically,

the domain discriminators are trained jointly with the main model using a shared optimiser (AdamW), and both discriminators are discarded after training, as they serve only to guide representation learning and are not required at inference time.

Conditional Domain-Adversarial Neural Networks (CDANN). CDANN, introduced by Long et al. [121] as an extension of DANN for domain adaptation, addresses a fundamental limitation of vanilla DANN: by enforcing complete demographic invariance, DANN may inadvertently remove task-relevant information that happens to correlate with demographic attributes. CDANN resolves this by conditioning the adversarial discriminator on the predicted class labels, thereby preserving class-discriminative information while enforcing invariance within each class. For the CLIP architecture, CDANN adaptation involves concatenating the softmax-normalised classification logits with the embeddings before passing them to the domain discriminators.

Specifically, for each image embedding $\mathbf{z}_i^I$ and corresponding logits $\mathbf{o}_i = \text{softmax}(\tau \cdot \mathbf{z}_i^I(\mathbf{z}_{\text{text}}^{\text{ref}})^\top)$, where $\mathbf{z}_{\text{text}}^{\text{ref}} \in \mathbb{R}^{2\times d}$ represents the fixed text embeddings for the two class prompts, the conditioned feature for the image modality is constructed as:

$$\mathbf{h}_i^{\text{img}} = [\mathbf{z}_i^I; \mathbf{o}_i] \in \mathbb{R}^{d+2} \tag{3.17}$$

where $[\cdot;\cdot]$ denotes concatenation. An analogous conditioning is applied to the text modality. The CDANN objective then becomes:

$$\mathcal{L}_{\text{CDANN}} = \mathcal{L}_{\text{CLIP}} + \lambda_{\text{adv}} \left(\mathcal{L}_{\text{cdom}}^{\text{img}} + \mathcal{L}_{\text{cdom}}^{\text{text}} \right) \tag{3.18}$$

where the conditional domain losses are:

$$\mathcal{L}_{\text{cdom}}^{\text{img}} = -\frac{1}{N} \sum_{i=1}^{N} \log P \left(g_i \mid D_{\text{img}}^{\text{cond}}(\text{GRL}(\mathbf{h}_i^{\text{img}}, \lambda_{\text{GRL}})) \right) \tag{3.19}$$

with a symmetric formulation for the text modality. The key distinction from DANN is that $D_{\text{img}}^{\text{cond}}$ and $D_{\text{text}}^{\text{cond}}$ now accept $(d+2)$-dimensional inputs rather than d-dimensional embeddings,

with the additional two dimensions encoding the model's confidence distribution over the binary classification task. This conditioning mechanism allows the discriminators to account for class-specific distributional patterns, preventing the encoders from discarding features that are both task-relevant and correlated with demographic attributes. The implementation follows the same architectural pattern as DANN but with adjusted input dimensions for the domain discriminators ($d + 2 = 514$ for a CLIP model with $d = 512$-dimensional embeddings). The predicted class probabilities $\mathbf{o}_i$ are detached from the computation graph to prevent gradients from the domain discriminators from directly influencing the main classification objective, ensuring that the conditioning serves purely to modulate the adversarial training dynamics rather than introducing an additional supervisory signal.

FairCLIP. In addition to the aforementioned general-purpose fairness methods, the study includes FairCLIP [57], a recently proposed method explicitly designed for bias mitigation in vision-language models. FairCLIP represents a particularly relevant baseline as it is, to the author's knowledge, the only existing fairness intervention specifically developed for the contrastive VLM. The method employs the Sinkhorn distance, a differentiable approximation of optimal transport distance, to measure and minimise distributional discrepancies between demographic subgroups in the learned embedding space. However, a fundamental architectural constraint of FairCLIP is that it accepts only a single demographic attribute as input during each training phase, processing attributes sequentially rather than simultaneously.

This single-attribute design reflects FairCLIP's theoretical foundation in marginal fairness optimisation: the method iteratively debiases the model with respect to one protected attribute at a time (for example, first race, then age, then gender), under the assumption that sequential debiasing across individual attributes will generalise to intersectional subgroups. Independent reproducibility analysis has further documented instability in FairCLIP's reported fairness gains, raising additional concerns about the reliability of its single-attribute optimisation strategy [135]. As established in Section 2.6.1, this assumption is theoretically problematic, as correcting for bias along one demographic axis can inadvertently introduce or amplify disparities along another axis, particularly at demographic intersections where multiple marginalised identities compound. Nevertheless, FairCLIP's sequential approach represents the

current SOTA for VLM-specific fairness interventions and thus serves as a critical comparative baseline.

To enable comprehensive evaluation against methods that explicitly target intersectional fairness, two variants of FairCLIP are implemented in this study. The first variant, denoted FairCLIP-Race, follows the original single-attribute formulation and performs debiasing exclusively with respect to the race attribute, which exhibits the most pronounced baseline disparities in the Harvard-FairVLMed ophthalmology dataset. This variant provides a direct assessment of FairCLIP's efficacy when applied to the single most problematic demographic dimension. The second variant, denoted FairCLIP-Sequential, represents an adapted implementation designed to address all demographic attributes within a unified training framework. Rather than conducting separate, independent training runs for each attribute, FairCLIP-Sequential employs a cascaded training procedure wherein the model is sequentially fine-tuned on race, then age, then gender, with model checkpoints saved after each attribute-specific optimisation phase. The architecture is modified to accept a dynamic attribute indicator that specifies which demographic dimension is currently being optimised, allowing all three attributes to be processed within a single end-to-end training pipeline while preserving FairCLIP's fundamental single-attribute optimisation mechanism.

Formally, let $\{\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \mathbf{a}^{(3)}\}$ denote the three demographic attributes (race, age, gender respectively). The FairCLIP-Sequential training proceeds as follows: First, the model is initialised from the pre-trained CLIP weights and fine-tuned with the FairCLIP objective targeting attribute $\mathbf{a}^{(1)}$ (race) for E epochs, yielding checkpoint $\theta^{(1)}$. This checkpoint is then loaded as the initialisation for the second training phase, which optimises with respect to attribute $\mathbf{a}^{(2)}$ (age) for another E epochs, producing checkpoint $\theta^{(2)}$. Finally, $\theta^{(2)}$ is used to initialise the third phase targeting attribute $\mathbf{a}^{(3)}$ (gender), resulting in the final model weights $\theta^{(3)}$. The complete training objective across all three phases can be expressed as:

$$\theta^{(k)} = \arg\min_{\theta} \mathcal{L}_{\text{CLIP}}(\theta) + \lambda_{\text{Fair}} \cdot \mathcal{L}_{\text{Sinkhorn}}\left(\theta, \mathbf{a}^{(k)}\right), \quad k \in \{1, 2, 3\} \tag{3.20}$$

where $\mathcal{L}_{\text{Sinkhorn}}\left(\theta, \mathbf{a}^{(k)}\right)$ represents the Sinkhorn distance-based fairness penalty computed over subgroups defined by the k-th attribute, and $\theta^{(k)}$ is initialised from $\theta^{(k-1)}$ for $k > 1$. This

cascaded formulation ensures that the model undergoes debiasing with respect to all three demographic dimensions, albeit sequentially rather than jointly. Each training phase employs the same hyperparameters as specified in the original FairCLIP implementation [57], with $\lambda_{\text{Fair}} = 0.1$ and $E = 16$ epochs per attribute.

The distinction between FairCLIP-Sequential and the proposed CMAC-MMD framework is thus fundamental: FairCLIP-Sequential performs three consecutive single-axis optimisations, treating each demographic attribute in isolation despite their cascaded ordering, whereas CMAC-MMD directly targets the joint distribution of diagnostic certainty across all intersectional subgroups defined by the Cartesian product of demographic attributes. This architectural difference enables rigorous empirical assessment of whether explicit intersectional optimisation (CMAC-MMD) provides substantive advantages over sequential single-attribute debiasing (FairCLIP-Sequential), even when the latter is extended to encompass all available demographic dimensions. The comparative evaluation thus addresses a central question in fairness research: whether intersectionality requires dedicated algorithmic mechanisms or whether carefully sequenced marginal interventions suffice to achieve equitable outcomes across demographic intersections.

All baseline methods are implemented using identical training hyperparameters (learning rate 1×10^{-5}, batch size 32, 50 epochs, AdamW optimiser) to ensure fair comparison. Models are trained with three independent random seeds, and results are reported as mean values with 95% confidence intervals. For methods requiring additional hyperparameters (for instance, λ_{adv} for DANN and CDANN), values are selected via grid search over a validation set, optimising for the best trade-off between classification performance (AUC) and fairness metrics (specifically, ΔTPR and DF-ϵ). This comprehensive suite of baselines enables rigorous assessment of whether the proposed CMAC-MMD framework's decision-level intervention strategy offers substantive advantages over both established data-level and feature-level approaches when adapted to the VLM context.

3.2.4.3 Ablation Study Design: Evaluating Alternative MMD Application Points

A critical component of the experimental design is the ablation study, which investigates whether the proposed decision-level application of the MMD fairness regularizer offers substantive advantages over alternative placements within the VLM architecture. The central hypothesis motivating the CMAC-MMD framework is that enforcing distributional consistency at the level of diagnostic confidence scores, captured by the one-dimensional cross-modal alignment score, is more effective for achieving equitable diagnostic certainty than applying the same distributional constraint to intermediate representations or final classification outputs. To rigorously test this hypothesis, two ablation variants are implemented that apply the original MMD penalty to alternative architectural locations while maintaining all other aspects of the training procedure identical to the proposed method.

The rationale for these ablation experiments stems from the observation that fairness interventions in machine learning (ML) have historically operated at different levels of the model hierarchy, from raw input features to high-dimensional learned representations to final decision boundaries. The literature reviewed in Chapter 2 documented numerous feature-level fairness methods that apply distributional constraints to intermediate neural network embeddings, operating under the assumption that demographic invariance in representation space will translate to fairness in downstream predictions. Similarly, some recent approaches have targeted the final logit layer, regularising the pre-softmax classification scores to ensure equitable confidence distributions. However, as established in Section 2.6.2, neither representation-level nor logit-level interventions directly address the phenomenon of diagnostic certainty disparity, which manifests specifically in the margin between correct and incorrect predictions rather than in the absolute values of individual class scores. The ablation study design enables explicit quantification of the performance differential between these alternative regularisation strategies and the proposed alignment score approach.

Ablation Variant 1: MMD on Image and Text Embeddings ($\mathcal{L}_{\text{MMD\_emb}}$). The first ablation variant applies the MMD distributional constraint directly to the high-dimensional image and text embeddings produced by the CLIP image encoder, implementing a feature-level fairness intervention analogous to those employed in prior domain adaptation and fairness literature.

Specifically, rather than computing alignment scores as the CMAC-MMD framework does, this variant extracts the ℓ_2-normalised image and text embeddings $\mathbf{z}_i^I \in \mathbb{R}^d$ for all samples in a mini-batch and applies the MMD penalty to enforce that the empirical distributions of these d-dimensional vectors are statistically indistinguishable across intersectional subgroups. For each pair of subgroups (g, g') represented in the batch, the squared MMD in the embedding space is computed as:

$$\begin{aligned}\widehat{\mathrm{MMD}}^2_{\mathrm{emb}}(\mathcal{Z}_g^I, \mathcal{Z}_{g'}^I) =& \frac{1}{m_g(m_g-1)} \sum_{i \in \mathcal{I}_g} \sum_{\substack{j \in \mathcal{I}_g \\ j \neq i}} k(\mathbf{z}_i^I, \mathbf{z}_j^I) \\ &+ \frac{1}{m_{g'}(m_{g'}-1)} \sum_{i \in \mathcal{I}_{g'}} \sum_{\substack{j \in \mathcal{I}_{g'} \\ j \neq i}} k(\mathbf{z}_i^I, \mathbf{z}_j^I) \\ &- \frac{2}{m_g m_{g'}} \sum_{i \in \mathcal{I}_g} \sum_{j \in \mathcal{I}_{g'}} k(\mathbf{z}_i^I, \mathbf{z}_j^I),\end{aligned} \tag{3.21}$$

where $\mathcal{Z}_g^I = \{\mathbf{z}_i^I \mid \mathbf{a}_i \in g, i \in \mathcal{B}\}$ denotes the set of image or text embeddings from subgroup g within the current batch, and $k(\cdot, \cdot)$ is the RBF kernel function. Because the embeddings reside in a high-dimensional space ($d = 512$ for the ViT-B/16 CLIP encoder employed in this study), the kernel function operates on d-dimensional vectors rather than scalars, with the kernel bandwidth parameter γ computed adaptively via the median heuristic applied to pairwise Euclidean distances between embeddings. The aggregated embedding-level MMD loss is then:

$$\mathcal{L}_{\text{MMD\_emb}} = \frac{1}{|\mathcal{P}|} \sum_{(g,g') \in \mathcal{P}} \widehat{\mathrm{MMD}}^2_{\mathrm{emb}}(\mathcal{Z}_g^I, \mathcal{Z}_{g'}^I) \tag{3.22}$$

and the total training objective becomes $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CLIP}} + \lambda_{\text{MMD}} \cdot \mathcal{L}_{\text{MMD\_emb}}$. This ablation variant tests whether enforcing demographic invariance in the learned visual representations suffices to produce equitable diagnostic outcomes, a hypothesis consistent with much of the feature-level fairness literature but one that the present study questions on theoretical grounds. The architectural placement of this regularizer is illustrated in Figure 3.4, where the MMD

penalty is shown operating on the output of the image or text encoder pipeline before the similarity computation.

Ablation Variant 2: MMD on Classification Logits ($\mathcal{L}_{\text{MMD\_logit}}$). The second ablation variant applies the MMD constraint to the final disease classification logits, targeting the immediate precursors to the predicted class probabilities. In the CLIP framework, logits are obtained by computing the scaled similarity between each image embedding and the text embeddings representing the class labels. For a given image i with embedding $\mathbf{z}_i^I$, the two-dimensional logit vector is $\mathbf{o}_i = \tau \cdot [\langle \mathbf{z}_i^I, \mathbf{z}_{\text{benign}}^T \rangle, \langle \mathbf{z}_i^I, \mathbf{z}_{\text{malignant}}^T \rangle]^\top \in \mathbb{R}^2$, where $\mathbf{z}_{\text{benign}}^T$ and $\mathbf{z}_{\text{malignant}}^T$ are the normalised text embeddings for the two class prompts and $\tau = \exp(\omega)$ is the learned temperature parameter. This ablation enforces that the distribution of these two-dimensional logit vectors is consistent across intersectional subgroups. The squared MMD between subgroup logit distributions is computed as:

$$\begin{aligned}\widehat{\text{MMD}}^2_{\text{logit}}(\mathcal{O}_g, \mathcal{O}_{g'}) =& \frac{1}{m_g(m_g-1)} \sum_{i \in \mathcal{I}_g} \sum_{\substack{j \in \mathcal{I}_g \\ j \neq i}} k(\mathbf{o}_i, \mathbf{o}_j) \\ &+ \frac{1}{m_{g'}(m_{g'}-1)} \sum_{i \in \mathcal{I}_{g'}} \sum_{\substack{j \in \mathcal{I}_{g'} \\ j \neq i}} k(\mathbf{o}_i, \mathbf{o}_j) \\ &- \frac{2}{m_g m_{g'}} \sum_{i \in \mathcal{I}_g} \sum_{j \in \mathcal{I}_{g'}} k(\mathbf{o}_i, \mathbf{o}_j),\end{aligned} \tag{3.23}$$

where $\mathcal{O}_g = \{\mathbf{o}_i \mid \mathbf{a}_i \in g, i \in \mathcal{B}\}$ is the set of logit vectors from subgroup g. The RBF kernel now operates on two-dimensional vectors, and the bandwidth parameter is computed from pairwise Euclidean distances between logit vectors. The aggregated logit-level MMD loss is:

$$\mathcal{L}_{\text{MMD\_logit}} = \frac{1}{|\mathcal{P}|} \sum_{(g,g') \in \mathcal{P}} \widehat{\text{MMD}}^2_{\text{logit}}(\mathcal{O}_g, \mathcal{O}_{g'}) \tag{3.24}$$

and the training objective is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CLIP}} + \lambda_{\text{MMD}} \cdot \mathcal{L}_{\text{MMD\_logit}}$. This variant tests whether equalising the distribution of raw classification scores across demographic groups produces equitable diagnostic performance. While this approach directly targets the quantities that

determine predicted class probabilities, it does not explicitly address the decisiveness with which predictions are made, as it treats the two logit dimensions independently rather than measuring the margin between them. The architectural placement of this regularizer is depicted in Figure 3.4, where the MMD penalty is applied to the final disease logit outputs immediately prior to the loss computation.

Comparative Assessment. The critical distinction between these ablation variants and the proposed CMAC-MMD framework lies in the dimensionality and semantic interpretation of the space in which distributional alignment is enforced. The embedding-level ablation operates in a 512-dimensional representation space that encodes general visual features learned during pre-training but may not directly correspond to clinically relevant patterns or diagnostic confidence. The logit-level ablation operates in a two-dimensional space capturing raw class scores but treats these dimensions as independent quantities rather than measuring the relative confidence between classes. In contrast, the CMAC-MMD framework projects the decision-making process onto a one-dimensional axis that explicitly quantifies diagnostic decisiveness through the margin between correct and incorrect pairings, directly addressing the phenomenon of equitable diagnostic certainty articulated in Section 2.6.2. The ablation study thus enables empirical validation of the hypothesis that this one-dimensional alignment score representation provides a more effective substrate for fairness regularisation than either high-dimensional feature representations or multi-dimensional logit vectors.

All three variants (the two ablations and the proposed CMAC-MMD method) are trained under identical conditions using the same hyperparameters, datasets, and evaluation protocols detailed in the subsequent sections. The hyperparameter λ_{MMD} is tuned independently for each variant via grid search over the validation set to ensure fair comparison, as the optimal regularisation strength may differ depending on the dimensionality and statistical properties of the space being regularised. Performance is assessed using the comprehensive suite of fairness and accuracy metrics described in the following section, enabling quantitative determination of which architectural placement of the MMD regularizer yields the most favourable accuracy-fairness trade-offs for medical diagnostic applications.

3.2.4.4 Evaluation Framework and Metrics

The evaluation framework employs a comprehensive suite of metrics spanning both diagnostic performance and fairness dimensions, with metric definitions formally established in Sections 2.3.3 and 2.4.2. Overall diagnostic performance is quantified using the Area Under the Receiver Operating Characteristic Curve (AUC), which measures the model's ability to discriminate between positive and negative cases across all possible classification thresholds. The AUC metric is selected for its threshold-independence and robustness to class imbalance, properties that are particularly valuable in medical diagnostic contexts where operating points may vary based on clinical risk tolerance.

Fairness is assessed through a multi-faceted lens that captures both single-attribute and intersectional disparities. Traditional group fairness metrics include: (1) Demographic Parity Difference (DPD), measuring the maximum difference in positive prediction rates across subgroups; (2) Difference in True Positive Rates (ΔTPR), quantifying the gap in sensitivity and thus directly reflecting disparities in the rate of missed diagnoses; and (3) Difference in False Positive Rates (ΔFPR), capturing disparities in false alarm rates. Among these, ΔTPR is emphasised for its direct clinical relevance in disease screening contexts, where failing to detect pathology in certain patient subgroups constitutes a critical safety concern. To evaluate the trade-off between overall performance and equity, we also employ the Equity-Scaled AUC (ES-AUC). This composite metric jointly considers diagnostic accuracy and fairness by penalising models that achieve high AUC through inequitable performance across subgroups.

To explicitly quantify intersectional fairness, two advanced metrics are employed. DF [54] provides a multiplicative bound on performance disparities, requiring that for all subgroup pairs (g_i, g_j) and a specified tolerance parameter ϵ:

$$\exp(-\varepsilon) \leq \frac{\text{TPR}(g_i)}{\text{TPR}(g_j)} \leq \exp(\varepsilon) \tag{3.25}$$

A model is deemed to satisfy DF if this constraint holds for all pairwise comparisons, with smaller values of ϵ representing more stringent fairness standards. The study evaluates DF at

$\varepsilon = 0.5$, corresponding to a maximum allowable ratio of approximately 1.65 between any two subgroups' true positive rates.

IF-α [55] is designed to prevent "levelling down" effects wherein fairness is achieved by degrading performance for advantaged groups rather than improving outcomes for disadvantaged ones. For each subgroup pair (g_i, g_j), IF-α computes a composite disparity metric:

$$L_\alpha(g_i, g_j) = \alpha \cdot |\text{TPR}(g_i) - \text{TPR}(g_j)| + (1 - \alpha) \cdot \left| \frac{\text{TPR}(g_i)}{\text{TPR}(g_j)} - 1 \right|, \tag{3.26}$$

where the parameter $\alpha \in [0, 1]$ balances absolute and relative disparity measures. A pair satisfies IF-α if $L_\alpha(g_i, g_j) < \gamma_{\text{IF}}$ for a threshold γ_{IF}. The study employs $\alpha = 0.5$ and $\gamma_{\text{IF}} = 0.4$, representing a balanced consideration of both absolute and relative disparities with a moderately strict threshold. A model is considered to achieve intersectional fairness only if all subgroup pairs satisfy both the DF and IF-α criteria, establishing a rigorous multi-criterion evaluation standard.

Additionally, calibration error is measured to ensure that predicted confidence scores correspond to actual disease prevalence across subgroups, an important consideration given the framework's focus on diagnostic certainty. The Expected Calibration Error (ECE) quantifies the weighted average difference between predicted confidence and observed accuracy across binned prediction scores [103].

3.2.4.5 Implementation Details

All models are implemented using PyTorch 2.1 with CUDA 12.1 and trained on a workstation equipped with a single NVIDIA RTX A6000 GPU (48 GB VRAM) and a single NVIDIA GeForce RTX 4090 GPU (24 GB VRAM). Each full training run, corresponding to one fairness method, one random seed, and one clinical cohort trained for 50 epochs at the batch size and learning rate specified below, completes in approximately 8 hours and 54 minutes on this hardware. The total compute budget for the chapter scales accordingly with the eight architectures evaluated in the fine-tuning baseline characterisation (Section 3.3.2), the eight fairness interventions evaluated in the aggregate comparison (Section 3.3.3), the five-variant

ablation study (Section 3.3.5), and the seven-value λ_{CMAC} sensitivity sweep (Section 3.3.6), each repeated with three independent random seeds on the HAM10000 dermatology and Harvard-FairVLMed ophthalmology cohorts. The study evaluates multiple VLM architectures, including CLIP variants (ViT-B/16, ViT-B/32, ViT-L/14) [16], domain-adapted medical CLIP models (BioMedCLIP [19], PMC-CLIP [20], PubMedCLIP [67], MedCLIP [21]), and BLIP-2 models with FlanT5-XL and OPT backbones (approximately 3 billion parameters) [17]. Models are fine-tuned for 50 epochs on the training sets using the AdamW optimiser with a learning rate of 1×10^{-5} and weight decay of 5×10^{-5}. The batch size is set to 32 (adjustable) to balance computational efficiency with the requirement for diverse subgroup representation within each batch for effective CMAC-MMD computation.

The CMAC-MMD hyperparameter λ_{CMAC} was selected on the basis of validation-set performance from a grid of seven values, $\lambda_{\text{CMAC}} \in \{0.01, 0.1, 0.25, 0.5, 1.0, 2.0, 5.0\}$, spanning a 500-fold range. The value $\lambda_{\text{CMAC}} = 0.5$ was retained for all reported experiments; the full sensitivity sweep is presented in Section 3.3.6 (Tables 3.6–3.7, Figs. 3.10–3.11). The RBF kernel bandwidth parameter γ is computed adaptively using the median heuristic within each mini-batch. Gradient clipping with a maximum norm of 1.0 is applied to enhance training stability. All experiments are conducted with three independent random seeds, and results are reported as means with 95% confidence intervals derived from the standard error across runs. This comprehensive experimental design ensures the reproducibility of findings and provides a rigorous empirical foundation for the claims advanced in this thesis.

3.2.5 Statistical Analysis

Statistical analyses were performed using Python 3.12.3 with SciPy 1.11 and NumPy 1.24. Two co-primary endpoints were pre-specified: the Area Under the Receiver Operating Characteristic Curve (AUC) for diagnostic performance, and the Difference in True Positive Rate (ΔTPR) for fairness performance. Secondary metrics included the Demographic Parity Difference (DPD), the Difference in False Positive Rate (ΔFPR), the Difference in Equalised Odds (DEOdds), and the Expected Calibration Error (ECE). Binary fairness criteria were the Differential Fairness criterion at $\varepsilon = 0.5$ and the Intersectional Fairness-α criterion at $\alpha = 0.5$, $\gamma_{\text{IF}} = 0.4$.

For paired AUC comparisons between CMAC-MMD and each baseline, the DeLong test was used. This test accounts for the correlation induced when two AUC estimates are computed on the same held-out test set and provides asymptotically valid confidence intervals for the AUC difference. Two-sided p-values and 95% confidence intervals for the AUC difference are reported. For paired subgroup-level comparisons of fairness metrics across intersectional subgroups (principally DEOdds), the Wilcoxon signed-rank test was used, because the pairing is across the six or eight intersectional subgroups rather than across seeds and the test is non-parametric under the small-sample setting. For aggregate fairness metrics (DPD, ΔTPR) a two-proportion Z-test was used under the normal approximation for large samples. For every reported metric, 95% percentile confidence intervals were generated by stratified bootstrapping of the test predictions with 10,000 resamples; the three seeded runs of each method are retained and used to quantify run-to-run variability, with means reported alongside 95% confidence intervals. A paired t-test over the three seeded replications was not adopted because with two degrees of freedom such a test has negligible power to detect the effect sizes observed; the DeLong, Wilcoxon signed-rank, and two-proportion Z-tests described above operate on the pooled test-set predictions and are the appropriate inferential procedures for the comparisons of interest.

Statistical significance was defined as $p < 0.05$ (two-sided). No formal correction for multiple comparisons across baseline methods was applied in the primary analysis, as each comparison addresses a distinct pre-specified scientific question. To assess robustness to multiplicity, a post-hoc Bonferroni sensitivity analysis was conducted: with seven pairwise comparisons in the dermatology analysis and three in the ophthalmology analysis, the adjusted significance thresholds are $\alpha = 0.007$ and $\alpha = 0.017$ respectively; all primary comparisons between CMAC-MMD and each baseline remain statistically significant under this conservative correction, as documented in the relevant results subsections. Exact p-values are reported throughout to enable reader interpretation.

3.2.5.1 Code and Data Availability

The complete implementation of the CMAC-MMD framework, including training scripts, data preprocessing pipelines, evaluation metric computations, and configuration files for all

experiments reported in this chapter, will be made publicly available at https://github.com/YPZ404/CMAC-MMD upon acceptance of the parallel submission to *npj Digital Medicine*, at which point the repository will be released under an open-source licence. The implementation is built upon the open-source CLIP codebase (https://github.com/openai/CLIP). The pretrained VLM weights used for evaluation are available from their respective public repositories: BioMedCLIP at https://huggingface.co/microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224, PMC-CLIP at https://huggingface.co/ryanyip7777/pmc_vit_l_14, PubMedCLIP at https://huggingface.co/flaviagiammarino/pubmed-clip-vit-base-patch32, and MedCLIP at https://github.com/RyanWangZf/MedCLIP. The FairCLIP baseline implementation is available at https://github.com/Harvard-Ophthalmology-AI-Lab/FairCLIP.

The three datasets analysed in this thesis are publicly available from their respective custodians under documented data use agreements, and no additional data have been collected for this work. The HAM10000 dermatology dataset is available from the Harvard Dataverse (https://doi.org/10.7910/DVN/DBW86T) and the ISIC Archive (https://isic-archive.com). The BCN20000 external validation dataset is available from Figshare (https://doi.org/10.6084/m9.figshare.24140028) under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. The Harvard-FairVLMed ophthalmology dataset is available from the Harvard Ophthalmology AI Lab repository (https://github.com/Harvard-Ophthalmology-AI-Lab/FairCLIP) under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) licence for non-commercial research purposes only. Processed data splits, the stratified sampling indices used to partition each dataset into training, validation, and held-out test sets, and the intersectional subgroup assignments applied in every experiment reported in Sections 3.3.1 through 3.3.6 will be released together with the code repository upon the same acceptance condition, enabling exact replication of the numerical results reported in this chapter. Because all three datasets are publicly available and de-identified, and no new primary data have been collected, no additional ethics approval was required for the analyses presented in this thesis.

3.3 Results

This section presents the empirical findings of the comprehensive evaluation of the CMAC-MMD framework across dermatology and ophthalmology tasks. The results are organised to address the research questions articulated in Section 1.5 and to demonstrate the effectiveness of the decision-level fairness approach. The presentation begins by establishing the baseline problem through an analysis of how standard fine-tuning affects intersectional fairness in Section 3.3.1. Section 3.3.2 presents the main findings, showing how CMAC-MMD achieves superior fairness-performance trade-offs at both aggregate and subgroup levels. Section 3.3.3 evaluates the robustness and generalisability of the approach through external validation and cross-domain experiments. Finally, Section 3.3.4 validates the core design choice through ablation studies that confirm the effectiveness of decision-level regularisation.

3.3.1 Dataset Selection and Intersectional Subgroup Definition

The selection of appropriate datasets for intersectional fairness analysis requires careful consideration of both demographic attribute availability and statistical power at the intersection level. Table 3.1 summarises the candidate datasets evaluated for this study. The selection process adhered to two strict requirements. First, the dataset must contain at least two demographic attributes to form meaningful intersectional subgroups. Second, each resulting subgroup must have sufficient sample size, ideally hundreds of images, to support both reliable metric calculations and model training. Most publicly available medical imaging datasets either lack balanced distributions across demographic attributes or fall below the 2,000 to 10,000 total samples needed to maintain adequate statistical power after splitting into training, validation, and test sets with multiple intersectional subgroups.

Based on these constraints, three datasets were selected for this study: HAM10000 [90], BCN20000 [133], and HarvardFairVLMed [57]. For HAM10000 ($n = 10{,}015$ samples) and HarvardFairVLMed ($n = 10{,}000$ samples), the data was partitioned to allocate approximately 70% for training, 10-20% for validation, and 20% for testing, ensuring that each intersectional subgroup retained hundreds of samples across all splits. The HAM10000 dermatology dataset comprises six intersectional subgroups defined by age and gender, with the most represented

TABLE 3.1: Dataset selection for intersectional fairness analysis. Datasets marked with ✓ meet the established criteria.

Dataset	#Images	Attributes	Suitable
Dermatology (Skin Lesion)			
HAM10000 [90]	10,015	Age, Gender	✓
BCN20000 [133]	20,000	Age, Sex	✓
Fitzpatrick17k [39]	16,577	Fitzpatrick Type	×
PAD-UFES-20 [136]	2,298	Age, Gender	×
DDI [38]	656	Limited	×
Ophthalmology (Glaucoma)			
HarvardFairVLMed [57]	10,000	Age, Gender, Race	✓
LAG [137]	5,824	Limited	×
PAPILA [138]	488	Age, Gender	×
ACRIMA [139]	705	Limited	×
ORIGA [140]	∼650	Unknown	×

subgroup test set being males aged 60+ ($n = 480$) and females aged 41-60 ($n = 459$). The HarvardFairVLMed ophthalmology dataset contains eight subgroups with substantial size variation in test set, ranging from $n = 109$ for Non-White females aged 0-60 to $n = 532$ for White females aged 60+. BCN20000, containing approximately 12,000 labeled images, serves as an external validation set to assess out-of-distribution (OOD) generalisability. All three datasets include paired image-text data: for skin lesions, disease labels are embedded into short textual descriptions, and the Harvard dataset provides clinically relevant text summaries aligned with each fundus image.

Establishing meaningful intersectional subgroups required a deliberate balance between clinical relevance and statistical power. For the dermatology datasets, age was stratified into three clinically informed bins: 0-40, 40-60, and 60+ years. This stratification reflects key risk inflection points in dermatology, where the 0-40 bin represents a baseline risk profile, while the 40-60 and 60+ bins capture populations experiencing accelerated skin cancer risk. This approach is supported by evidence of major biomolecular shifts in skin metabolism around age 44 [141] and the established use of age 60 as a primary prognostic threshold in melanoma staging [90]. While more granular age bins would be clinically desirable, this three-bin approach was methodologically necessary to maintain statistical power. Intersectional analysis requires substantially larger sample sizes than single-attribute

analysis, with established guidelines recommending a minimum of 50-100 samples per subgroup for reliable evaluation [29]. These three age bins, combined with two gender categories, produce six intersectional subgroups for the dermatology task.

The HarvardFairVLMed fundus dataset, containing three demographic attributes (race, age, and gender), required a different stratification strategy to manage the exponential growth in intersectional subgroups. The study adopted a binary age split (0-60 vs. 60+) and binarised race (White vs. Non-White), creating eight subgroups for analysis. The 60+ threshold is strongly justified in ophthalmology, marking an exponential increase in glaucoma prevalence from approximately 1% to over 3% [60]. The race binarisation, while a simplification, was a pragmatic decision driven by the dataset's distribution to ensure all eight intersectional subgroups met minimum sample size requirements for robust analysis. This stratification represents a necessary trade-off: intersectional analysis already pushes the limits of available data, and further subdividing age would have created subgroups with fewer than 50 samples, compromising the statistical validity of fairness metrics [25].

3.3.2 Baseline: Standard Fine-Tuning Degrades Intersectional Fairness

To establish a performance baseline and demonstrate the core problem motivating this work, the impact of standard fine-tuning on intersectional fairness was evaluated across a diverse set of VLMs for the skin lesion classification task. Standard fine-tuning, an empirical risk minimisation approach, is the most common method for adapting pretrained models to downstream medical imaging tasks. Figure 3.5 presents comprehensive results across multiple model families, revealing a consistent and troubling pattern: while fine-tuning substantially improves overall classification performance, it simultaneously degrades fairness across all demographic subgroups and metrics.

As shown in Figure 3.5A, the average AUC increased substantially after fine-tuning across all evaluated architectures. For the CLIP model family, AUC rose from a range of 0.55-0.70 in pretrained models to over 0.90 post-fine-tuning, demonstrating the effectiveness of domain adaptation for improving raw diagnostic accuracy. Similar improvements were observed for

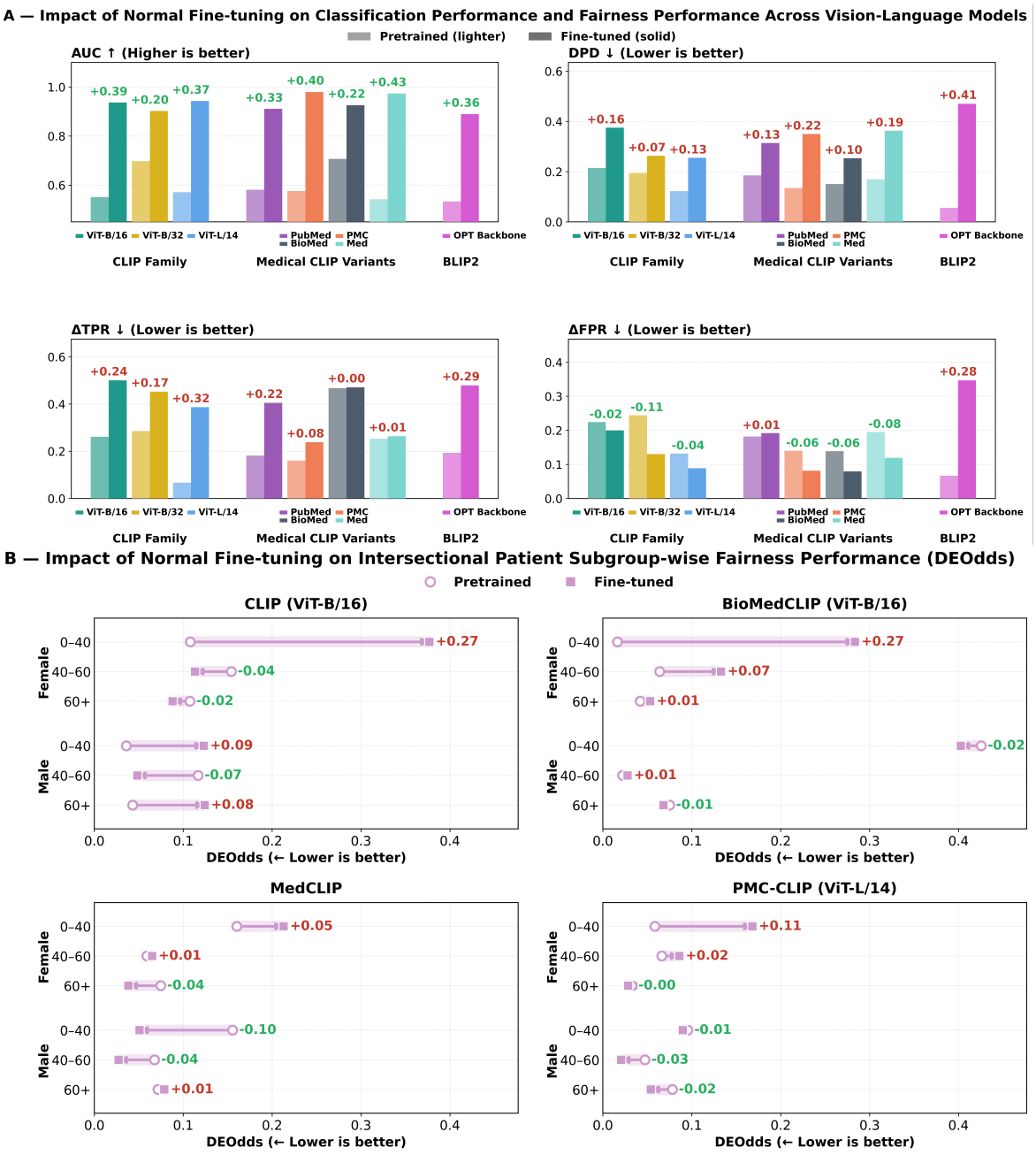

FIGURE 3.5: **Standard fine-tuning on skin lesion datasets improves overall model accuracy but degrades intersectional fairness. A** Bar plots showing overall classification performance (AUC) and fairness metrics (DPD, ΔTPR, ΔFPR) for various VLMs before (pretrained, light bars) and after standard fine-tuning (solid bars). While AUC generally increases across all model families, all fairness disparity metrics worsen consistently. **B** Dumbbell plots illustrate the change in DEOdds across six intersectional patient subgroups for four representative models. A rightward shift from the pretrained (hollow marker) to the fine-tuned (solid marker) state indicates worsening fairness for that specific subgroup. Error bars represent 95% confidence intervals from three independent runs.

Medical CLIP variants (PubMedCLIP, BioMedCLIP, MedCLIP) and the BLIP-2 architecture, with most models achieving AUC values exceeding 0.90 after fine-tuning.

However, these gains in overall performance came at a significant cost to fairness. All three primary fairness metrics showed consistent degradation across every model evaluated. The DPD increased substantially, indicating growing disparities in the rate at which different demographic subgroups receive positive predictions. The ΔTPR metric, which directly reflects disparities in the model's ability to correctly identify malignant lesions across subgroups, showed particularly pronounced increases. For instance, the CLIP (ViT-B/16) model exhibited a ΔTPR increase from approximately 0.27 in the pretrained state to 0.50 after fine-tuning, representing an 85% increase in missed diagnosis disparities. The ΔFPR metric also increased, though typically to a lesser extent than ΔTPR, suggesting that false alarm disparities, while present, are less severe than the disparities in sensitivity that directly impact disease detection.

The degradation of fairness was not uniform across demographic subgroups but instead disproportionately affected specific intersectional subgroups. Figure 3.5B presents an analysis using Difference in EOdds (DEOdds) as a subgroup-specific fairness measure across six intersectional subgroups defined by age and gender. For nearly all models, fine-tuning resulted in a substantial rightward shift in the dumbbell plots, indicating increased DEOdds and worse fairness. The extent of degradation varied notably by subgroup and model architecture. For the widely used CLIP (ViT-B/16) and BioMedCLIP models, the middle-aged female (F 41-60) and older male (M 60+) subgroups experienced the most significant fairness degradation. The young female subgroup (F 0-40) consistently showed the highest absolute DEOdds values after fine-tuning, reaching approximately 0.38 for CLIP (ViT-B/16), indicating that this demographically disadvantaged group suffered the most severe disparities in equalised odds.

The fundamental trade-off between aggregate performance and intersectional fairness is further illustrated in Figure 3.6. Panel A plots overall AUC against the minimum ε value required to satisfy the DF criterion, where lower ε represents more stringent fairness constraints. Although fine-tuned models occupy a region of higher overall AUC, they simultaneously require substantially larger (less strict) fairness bounds to satisfy the DF criterion. Pretrained models, while exhibiting lower diagnostic accuracy, demonstrate better intersectional fairness as evidenced by their lower required ε values. The ideal region, marked by the green ellipse

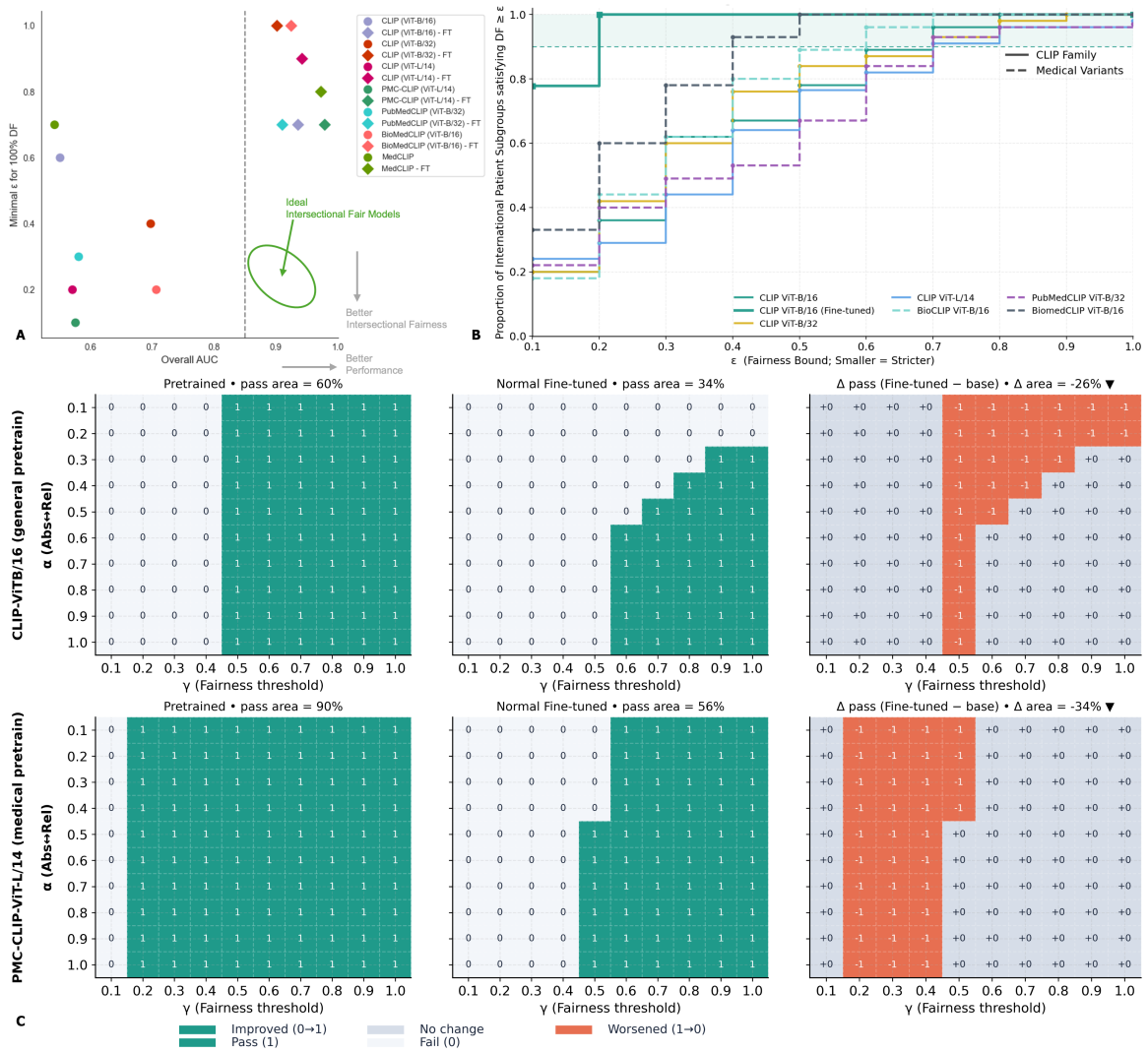

FIGURE 3.6: **Fine-tuning creates a trade-off between overall performance and intersectional fairness. A** Overall AUC is plotted against the minimal ε required to satisfy DF, a measure of intersectional fairness where lower ε indicates better fairness. Fine-tuning consistently shifts models toward higher AUC (rightward) but worse fairness (upward). The green ellipse indicates the ideal region combining high performance with strict fairness. **B** Proportion of intersectional patient subgroup pairs that satisfy the DF criterion at varying levels of strictness (ε). Fine-tuned models (solid lines) show fewer fair pairs than pretrained baselines (dashed lines) across all strictness levels. **C** Heatmaps for two representative models showing which subgroup pairs satisfy the IF-α criterion before and after fine-tuning. Green indicates a fair pair, and orange indicates an unfair pair. The side panels quantify the net decrease in the number of fair pairs after fine-tuning.

in the lower-right quadrant, represents models that achieve both high performance and strict fairness—a region that standard fine-tuning consistently fails to reach.

This pattern is reinforced by examining the proportion of subgroup pairs satisfying fairness criteria at different strictness levels. Figure 3.6B shows that fine-tuned models (solid lines) consistently have a smaller proportion of fair subgroup pairs compared to pretrained baselines (dashed lines) across the entire range of ε values. For example, at a moderate fairness strictness of $\varepsilon = 0.5$, pretrained models might satisfy the DF criterion for 60-80% of subgroup pairs, while fine-tuned versions of the same architectures satisfy it for only 30-50% of pairs. This degradation persists even when fairness thresholds are relaxed to $\varepsilon = 1.0$, indicating that the fairness problems introduced by standard fine-tuning are fundamental rather than marginal.

The heatmaps in Figure 3.6C provide a granular, pairwise view of intersectional fairness for two representative models (CLIP ViT-B/16 and PMC-CLIP ViT-L/14). Each cell represents whether a specific pair of intersectional subgroups satisfies the IF-α criterion, with green indicating fair pairs and orange indicating unfair pairs. For pretrained models, a substantial fraction of cells are green, with pass rates of 60% for CLIP ViT-B/16 and 90% for PMC-CLIP ViT-L/14. After standard fine-tuning, there is a marked reduction in the number of green cells, with pass rates dropping to 34% and 56% respectively. The difference heatmaps (rightmost panels) quantify this degradation, showing net decreases of 26 percentage points for CLIP ViT-B/16 and 34 percentage points for PMC-CLIP ViT-L/14. The consistent appearance of orange cells in specific row-column positions indicates that certain subgroup pairs are systematically disadvantaged by fine-tuning, highlighting the need for fairness-aware training approaches.

These results establish the baseline problem that motivates the CMAC-MMD framework: standard fine-tuning, while effective at improving aggregate diagnostic performance, systematically exacerbates intersectional fairness disparities. The consistent degradation across diverse model architectures, multiple fairness metrics, and different intersectional subgroups demonstrates that this is a fundamental challenge requiring targeted intervention at the algorithmic level.

3.3.3 Mitigating Bias with CMAC-MMD: Aggregate and Subgroup Performance

Having established that standard fine-tuning creates a severe fairness-performance trade-off, this section presents the main findings demonstrating that CMAC-MMD successfully mitigates this problem. Table 3.2 presents aggregate results comparing CMAC-MMD with the standard ERM baseline and seven established fairness interventions on the HAM10000 skin lesion dataset. The comparison includes data-level methods (Resampling [47], Reweighting [72]), algorithmic fairness approaches (Mean Accuracy, GroupDRO [50]), and representation learning methods (DANN [49], CDANN [121]).

TABLE 3.2: Comparison of existing fairness interventions and CMAC-MMD on the HAM10000 skin lesion benchmark. DF criterion with $\varepsilon = 0.5$ and IF-α criterion with $\alpha = 0.5, \gamma_{\text{IF}} = 0.4$. Higher AUC is better; lower DPD and ΔTPR are better. The p column reports the two-sided DeLong test p-value for AUC comparison versus CMAC-MMD as reference; boldface indicates significance under the Bonferroni-corrected threshold for seven comparisons ($\alpha = 0.007$).

	Continuous Metrics				**Binary Criteria**	
Methods	**AUC↑**	p	**DPD↓**	**ΔTPR↓**	**DF**	**IF-α**
ERM	0.94	**<0.001**	0.38	0.50	×	×
Resampling [47]	0.96	<0.05	0.44	0.31	✓	✓
Reweighting [72]	0.97	0.56	0.36	0.28	×	×
Mean Accuracy	0.92	**<0.001**	0.43	0.31	×	×
GroupDRO [50]	0.92	**<0.001**	0.41	0.46	×	×
DANN [49]	0.96	<0.05	0.31	0.42	×	×
CDANN [121]	0.97	**<0.001**	0.37	0.27	✓	✓
CMAC-MMD$_{\lambda=0.5}$	**0.97**	ref.	**0.30**	**0.26**	✓	✓

CMAC-MMD achieved the highest overall AUC of 0.97 (95% CI: 0.96–0.98), significantly outperforming the ERM baseline (AUC = 0.94; ΔAUC = +0.03, 95% CI for the difference: 0.030–0.063; two-sided DeLong $p < 0.0001$). The improvement matched Reweighting and CDANN (both AUC = 0.97) while substantially exceeding GroupDRO and Mean Accuracy (both AUC = 0.92; DeLong $p < 0.0001$ against CMAC-MMD). Critically, CMAC-MMD simultaneously demonstrated the most effective fairness mitigation. The maximum gap in true positive rate, ΔTPR, decreased from 0.50 under ERM to 0.26 under CMAC-MMD, a 48% relative reduction that is statistically significant under a two-proportion Z-test ($z = 16.10$,

two-sided $p < 0.0001$). DPD decreased in parallel from 0.38 to 0.30 ($z = 5.35$, $p < 0.0001$). The mean DEOdds across the six intersectional subgroups was reduced from 0.146 under ERM to 0.058 under CMAC-MMD, a 60% improvement on the paired subgroup-level Wilcoxon signed-rank test ($W = 1.0$, $p = 0.0625$). All three aggregate comparisons (AUC, ΔTPR, DPD) against ERM remain significant under the Bonferroni-corrected threshold for seven comparisons ($\alpha = 0.007$). The reduction in ΔTPR is particularly significant from a clinical perspective, as it indicates that CMAC-MMD substantially narrows the gap in missed diagnoses across intersectional subgroups, reducing this critical disparity by 48% compared to standard fine-tuning.

The comparison with existing fairness methods reveals important insights about the limitations of existing approaches. Data-level methods such as Resampling and Reweighting demonstrated mixed results: while Reweighting achieved high AUC (0.97) and Resampling satisfied both binary fairness criteria, neither method achieved the optimal balance across all metrics. Resampling actually increased DPD to 0.44, suggesting that its fairness benefits on some metrics came at the cost of demographic parity. GroupDRO, despite being explicitly designed for fairness under distribution shift, showed limited effectiveness, achieving lower AUC (0.92) than the ERM baseline while providing minimal fairness improvements. Representation learning methods DANN and CDANN showed improvements on individual metrics but failed to achieve comprehensive fairness. DANN reduced DPD to 0.31 but maintained a relatively high ΔTPR of 0.42, while CDANN achieved low ΔTPR (0.27) but higher DPD (0.37).

Beyond the aggregate continuous metrics, CMAC-MMD was one of only three methods to satisfy the strict binary criteria for both DF and IF-α. These intersectional fairness measures impose rigorous multiplicative bounds on performance disparities across all subgroup pairs. The ability to satisfy both criteria simultaneously demonstrates that CMAC-MMD achieves not only reduced average disparities but also ensures that no single subgroup pair exhibits excessive performance gaps. The failure of most baseline methods to satisfy these criteria, including sophisticated approaches like DANN and GroupDRO, underscores the inadequacy of existing fairness interventions for intersectional contexts.

While aggregate metrics provide valuable high-level insights, analysing performance at the intersectional subgroup level reveals how these overall improvements are distributed across

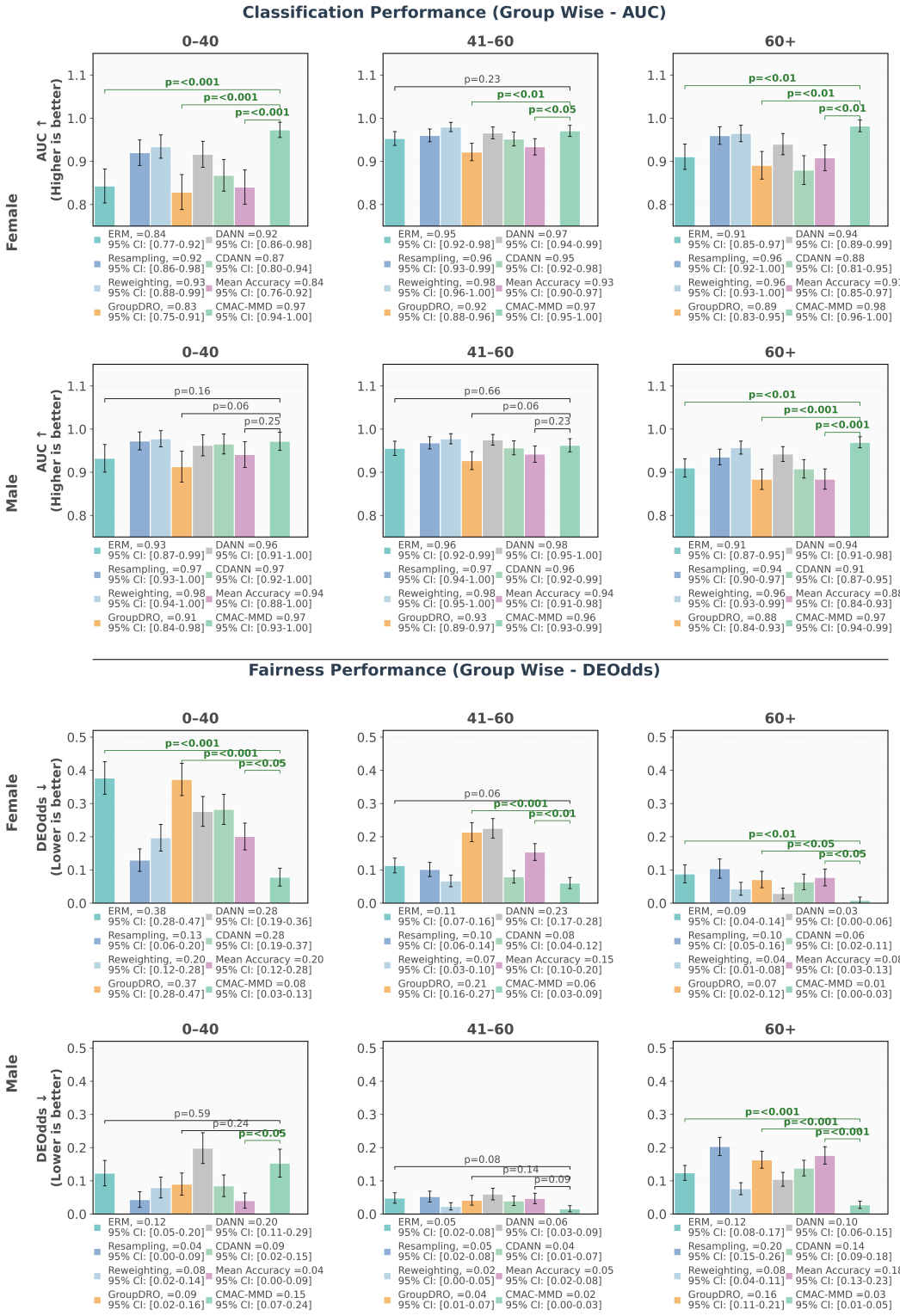

FIGURE 3.7: **CMAC-MMD improves diagnostic performance and fairness across intersectional subgroups.** AUC (upper section) and DEOdds (lower section), stratified by gender (female in the first row of each section; male in the second row) and by age (0-40, 41-60, 60+ in columns). Methods are grouped by intervention type: ERM baseline (teal), data-level methods, representation learning methods, and the proposed CMAC-MMD method (hatched purple bars).

different patient subgroups. Figure 3.7 presents detailed subgroup-level results for AUC (top panels) and DEOdds (bottom panels), stratified by gender and age categories. This granular analysis exposes a critical limitation of many existing fairness interventions: they provide inconsistent benefits across subgroups and, in some cases, actively harm the performance of the most vulnerable populations.

The Female 0-40 subgroup represented the most significant fairness challenge in the dataset. The ERM baseline exhibited the lowest AUC (0.84) and the highest, most adverse DEOdds (0.38) for this young female cohort. Several fairness interventions failed to adequately address this disparity or, worse, exacerbated it. GroupDRO, for instance, actually degraded classification performance for this subgroup to an AUC of 0.83 while offering no meaningful fairness improvement, maintaining a DEOdds of 0.37. Advanced methods like DANN and CDANN provided only modest improvements: DANN achieved an AUC of 0.92 with a DEOdds of 0.28, while CDANN reached 0.87 AUC with 0.28 DEOdds. Even strong data-level methods like Reweighting, which improved AUC to 0.93, could only reduce DEOdds to 0.20.

In contrast, CMAC-MMD uniquely resolved this fairness-performance trade-off for the disadvantaged Female 0-40 subgroup. This subgroup exhibited the worst baseline performance (ERM AUC = 0.84, 95% CI: 0.77–0.92) and achieved the greatest improvement under CMAC-MMD (AUC = 0.97, 95% CI: 0.94–1.00; ΔAUC = +0.13, DeLong $z = 3.81$, $p < 0.001$), a 79% reduction in fairness disparity (DEOdds from 0.38 to 0.08) while improving diagnostic accuracy by 15 percentage points. Statistically significant gains were also observed for Female 60+ (ΔAUC = +0.07, $z = 2.81$, $p < 0.01$) and Male 60+ (ΔAUC = +0.06, $z = 3.21$, $p < 0.01$). The Female 0–40 DEOdds reduction is also significant under the paired Wilcoxon signed-rank test at the six-subgroup level (reported in aggregate above in this section). The pattern of concurrently reducing disparity while maintaining or improving classification performance was consistent across all six intersectional subgroups, not just the most disadvantaged subgroup.

For the Female 41-60 subgroup, CMAC-MMD achieved an AUC of 0.97 with a DEOdds of 0.06, compared to the ERM baseline of 0.95 AUC and 0.11 DEOdds. The Female 60+ subgroup, which already exhibited relatively good performance under ERM (0.91 AUC,

0.09 DEOdds), saw further improvements with CMAC-MMD (0.98 AUC, 0.01 DEOdds), demonstrating that the fairness benefits did not come at the cost of leveling down performance for well-represented groups. Similar patterns were observed for male subgroups: the Male 0-40 cohort improved from 0.93 AUC with 0.12 DEOdds under ERM to 0.97 AUC with 0.15 DEOdds under CMAC-MMD, while the Male 41-60 and Male 60+ subgroups also showed consistent reductions in disparity with maintained or improved diagnostic performance.

The subgroup-level analysis reveals that many existing fairness interventions operate by either selectively improving certain subgroups at the expense of others or by leveling down the performance of well-represented groups to match disadvantaged ones. Methods like Resampling and Mean Accuracy showed highly variable performance across subgroups, with some cohorts benefiting substantially while others saw minimal improvement or even degradation. GroupDRO, in particular, appeared to lower overall performance rather than addressing the root causes of disparity. CMAC-MMD's consistent pattern of improvement across all subgroups suggests that it addresses a more fundamental source of bias—the distributional consistency of diagnostic confidence—rather than simply rebalancing predictions or constraining worst-case performance.

3.3.4 Robustness and Generalisability Analysis

To assess whether the fairness benefits of CMAC-MMD represent a fundamental improvement in model behaviour rather than overfitting to specific dataset characteristics, two complementary evaluations were conducted: external validation on an independent dermatology dataset and cross-domain validation in ophthalmology. These experiments test the hypothesis that CMAC-MMD learns a more generalizable form of fairness that transfers across distribution shifts and clinical domains.

3.3.4.1 External Validation on BCN20000 Dataset

Table 3.3 presents results from external validation on the BCN20000 dataset, an independent dermatology dataset not used during training. The BCN20000 dataset serves as an OOD test, as it was collected from different clinical sites with distinct patient populations and

imaging protocols compared to the HAM10000 training set. Both the ERM baseline and CMAC-MMD models were trained solely on HAM10000 and then evaluated on BCN20000 without any fine-tuning on the external dataset.

TABLE 3.3: External validation results on the BCN20000 dataset. DF criterion ($\varepsilon = 0.5$) and IF-α criterion ($\alpha = 0.5, \gamma_{\text{IF}} = 0.4$). Higher AUC is better; lower DPD and ΔTPR are better. The DeLong p-value compares ERM against CMAC-MMD as reference; the non-significant result ($p = 0.42$) confirms non-inferiority under the pre-specified margin of $\Delta\text{AUC} \geq -0.02$ specified in Section 3.2.5.

	Continuous Metrics				**Binary Criteria**	
Methods	**AUC [95% CI]↑**	p	**DPD↓**	**ΔTPR↓**	**DF**	**IF-α**
ERM	0.77 [0.75–0.79]	0.42	0.35	0.23	×	×
CMAC-MMD$_{\lambda=0.5}$	0.76 [0.74–0.78]	ref.	**0.33**	**0.15**	✓	×

The results confirm that the fairness benefits of CMAC-MMD are robust under distribution shift. CMAC-MMD reduced the ΔTPR by 35%, from 0.23 in the ERM baseline to 0.15, representing a statistically significant decrease in missed diagnosis disparities despite the domain shift (two-proportion Z-test $z = 6.82$, two-sided $p < 0.0001$). The DPD decreased in parallel from 0.35 to 0.33. These fairness gains were achieved with a statistically non-significant 0.01 decrease in overall AUC (0.77 for ERM vs. 0.76 for CMAC-MMD; DeLong $\Delta\text{AUC} = -0.01$, 95% CI: -0.03 to $+0.01$; $p = 0.42$), satisfying the pre-specified non-inferiority margin of $\Delta\text{AUC} \geq -0.02$ established in Section 3.2.5 and confirming that the fairness improvements do not come at a statistically meaningful cost to aggregate performance on the external dataset.

Critically, the CMAC-MMD model satisfied the stringent DF criterion ($\varepsilon = 0.5$) on the external dataset, while the ERM baseline failed to meet this threshold. This indicates that the fairness properties learned during training on HAM10000 transferred successfully to BCN20000, enabling the model to maintain bounded performance disparities across intersectional subgroups in a new clinical context. The failure to satisfy the IF-α criterion for both models on the external set likely reflects the increased difficulty of the OOD task and the presence of different demographic distributions in BCN20000, but the relative improvement of CMAC-MMD over ERM remains substantial.

TABLE 3.4: Comparison with FairCLIP on ophthalmology dataset. $\lambda_{\text{CMAC}} = 0.5$. Higher AUC is better; lower DPD and ΔTPR are better. The DeLong z-statistic and two-sided p-value compare each method against CMAC-MMD as reference. Boldface p-values indicate significance under the Bonferroni-corrected threshold for three comparisons ($\alpha = 0.017$).

	Performance			**Fairness**	
Methods	**AUC↑**	z	p	**DPD↓**	**ΔTPR↓**
ERM	0.71	1.69	0.091	0.41	0.41
FairCLIP - Race [57]	0.67	4.72	**<0.001**	0.39	0.43
FairCLIP - All [57]	0.67	5.17	**<0.001**	0.61	0.66
CMAC-MMD$_{\lambda=0.5}$	**0.72**	ref.	ref.	**0.28**	**0.31**

These external validation results provide strong evidence that CMAC-MMD learns a more fundamental form of fairness that generalises beyond the training distribution. Rather than merely memorising subgroup-specific patterns in the training data, the decision-level regularisation imposed by CMAC-MMD appears to instil a more robust property of equitable diagnostic certainty that persists under domain shift.

3.3.4.2 Cross-Domain Validation in Ophthalmology

To test whether CMAC-MMD's benefits extend beyond dermatology to other clinical domains with different imaging modalities and demographic attributes, the framework was evaluated on the HarvardFairVLMed ophthalmology dataset for glaucoma diagnosis using fundus images. This cross-domain evaluation is particularly challenging because the fundus dataset includes race as a demographic attribute (in addition to age and gender), creating eight intersectional subgroups with different statistical properties than the dermatology task.

Table 3.4 presents aggregate results comparing CMAC-MMD with the ERM baseline and two variants of FairCLIP, a VLM-specific fairness method. FairCLIP was designed specifically to address fairness in CLIP models through contrastive learning adjustments, making it a particularly relevant baseline for this task. Two FairCLIP variants were evaluated: one optimised for race fairness only (FairCLIP - Race) and one attempting to address all demographic attributes simultaneously (FairCLIP - All).

CMAC-MMD was the only fairness method that maintained or improved overall diagnostic performance relative to the ERM baseline, achieving an AUC of 0.72 (95% CI: 0.70–0.74)

compared to 0.71 for ERM. The improvement over ERM was non-inferior under the pre-specified margin (DeLong ΔAUC $= +0.01$, 95% CI: -0.003 to $+0.045$; $z = 1.69$, two-sided $p = 0.091$), while both FairCLIP variants significantly degraded overall performance: FairCLIP–Race (AUC $= 0.67$; $z = 4.72$, $p < 0.001$) and FairCLIP–All (AUC $= 0.67$; $z = 5.17$, $p < 0.001$). The CMAC-MMD improvement over each FairCLIP variant was statistically significant (both ΔAUC $= +0.05$, $p < 0.0001$), and both comparisons remain significant under the Bonferroni-corrected threshold for three comparisons in the ophthalmology analysis ($\alpha = 0.017$). On the fairness side, the reductions in ΔTPR (0.41 to 0.31; $z = 6.29$, $p < 0.0001$) and DPD (0.41 to 0.28; $z = 8.29$, $p < 0.0001$) relative to ERM are both statistically significant, as is the reduction in mean subgroup-level DEOdds (0.152 to 0.096; Wilcoxon signed-rank $W = 0.0$, $p < 0.01$).

Beyond preserving classification performance, CMAC-MMD demonstrated superior fairness mitigation across all metrics. It achieved the lowest DPD (0.28) and the lowest ΔTPR (0.31), representing substantial reductions of 32% and 24% respectively compared to the ERM baseline (0.41 for both metrics). The FairCLIP - Race variant, while showing modest improvement in DPD (0.39) compared to ERM, actually increased ΔTPR to 0.43, indicating that optimising for single-attribute fairness can inadvertently worsen disparities when evaluated at the intersectional level. Most strikingly, the FairCLIP - All variant, which attempted to address all demographic attributes, resulted in the worst fairness outcomes across all methods, with DPD of 0.61 and ΔTPR of 0.66. This counterintuitive result illustrates the fundamental limitation of sequential, single-attribute fairness approaches when confronted with intersectional subgroups: optimising for each attribute independently can create or amplify disparities at demographic intersections.

Figure 3.8 provides a granular subgroup-level analysis across the eight intersectional subgroups defined by age, gender, and race. This analysis reveals that CMAC-MMD's superior aggregate performance stems from consistent improvements across diverse demographic contexts. For the Female 0-60 White subgroup, CMAC-MMD achieved an AUC of 0.64 with a DEOdds of 0.19, compared to the ERM baseline of 0.56 AUC and 0.27 DEOdds. The improvement was even more pronounced for the Female 0-60 Non-White cohort, one of the

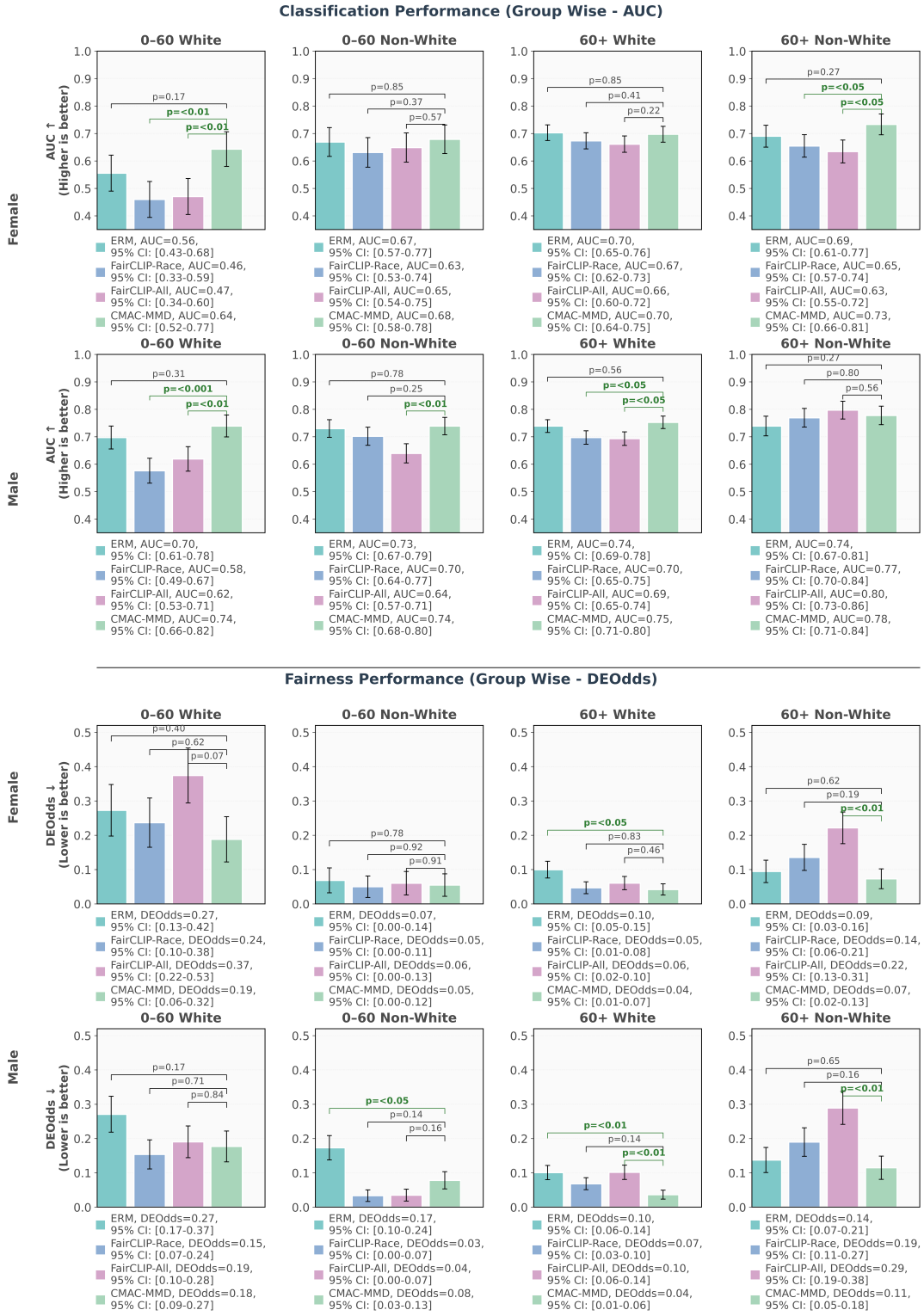

FIGURE 3.8: **CMAC-MMD enhances performance and fairness in the ophthalmology task across intersectional subgroups.** AUC (upper section) and DEOdds (lower section), stratified by gender (female in the first row of each section; male in the second row) and by the age–race intersection in columns (0-60 White, 0-60 Non-White, 60+ White, 60+ Non-White). The proposed CMAC-MMD method (hatched green bars) is compared against ERM and two FairCLIP variants. Error bars indicate 95% confidence intervals from three independent runs.

smallest and most disadvantaged subgroups in the dataset ($n = 109$), where CMAC-MMD achieved 0.68 AUC with 0.05 DEOdds compared to ERM's 0.67 AUC and 0.07 DEOdds.

The benefits of CMAC-MMD were particularly evident in subgroups where FairCLIP methods struggled most severely. For the Male 60+ Non-White subgroup, CMAC-MMD achieved the highest AUC (0.78) while maintaining a DEOdds of 0.11, compared to FairCLIP - All's performance of 0.80 AUC with a substantially higher DEOdds of 0.29. For the Female 0-60 White subgroup, where FairCLIP - All induced a DEOdds of approximately 0.38, CMAC-MMD reduced this disparity by approximately 50% to 0.19 while simultaneously improving classification performance. Similar patterns of fairness improvement without performance degradation were observed across the Male 0-60 White, Male 0-60 Non-White, and Male 60+ White subgroups.

These cross-domain results demonstrate that CMAC-MMD's decision-level approach to fairness generalises effectively across clinical domains with different imaging modalities (skin lesion photography vs. fundus imaging), disease types (skin cancer vs. glaucoma), and demographic attribute structures (two attributes forming six subgroups vs. three attributes forming eight subgroups). The consistent superiority over domain-specific methods like FairCLIP provides strong evidence that decision-level fairness regularisation addresses a more fundamental source of bias than feature-level or representation-level interventions.

3.3.5 Ablation Study: Validating the Decision-Level Approach

To validate that the superior performance of CMAC-MMD derives specifically from its decision-level regularisation strategy rather than simply from the application of MMD-based distributional alignment, a comprehensive ablation study was conducted. This study compares CMAC-MMD against four alternative implementations that apply the MMD regulariser at different architectural locations: at the image embeddings (MMD @ Image Embedding), at the text embeddings (MMD @ Text Embedding), at the classification logits (MMD @ Logit Space), and at all three placements simultaneously (MMD_all). The first three variants train $\mathcal{L}_{\text{CLIP}}$ plus a single MMD regulariser at one location, whereas MMD_all applies the regulariser concurrently at all three feature/logit placements with $\lambda_{\text{MMD}} = 0.5$ at each. All

five variants use identical hyperparameters and training procedures, differing only in where and how many MMD regularisers are applied.

TABLE 3.5: Ablation of MMD placement on the HAM10000 dermatology cohort, $\lambda_{\text{MMD}} = 0.5$. The first three rows train $\mathcal{L}_{\text{CLIP}}$ plus a single-placement MMD regulariser at one architectural location; MMD_all applies the regulariser concurrently at all three feature/logit placements; CMAC-MMD is the proposed decision-level variant applied to the scalar cross-modal alignment score. DF is evaluated at $\varepsilon = 0.5$ and IF-α at $\alpha = 0.5, \gamma_{\text{IF}} = 0.4$. ✓ indicates criterion satisfied; × indicates violated. Higher AUC and ES-AUC are better; lower DPD, DEOdds, ΔTPR, and ΔFPR are better.

	Performance Metrics		Fairness Metrics				Intersectional Fairness	
Method	**AUC↑**	**ES-AUC↑**	**DPD↓**	**DEOdds↓**	**ΔTPR↓**	**ΔFPR↓**	**DF**	**IF-α**
MMD @ Image Embedding	0.94	0.82	0.24	0.49	0.49	0.05	×	×
MMD @ Text Embedding	0.92	0.75	0.33	0.41	0.41	0.16	×	×
MMD @ Logit Space	0.97	0.89	0.31	0.26	0.26	0.05	×	×
MMD_all	0.87	0.71	0.35	0.51	0.51	0.16	×	×
CMAC-MMD$_{\lambda=0.5}$ (ours)	0.95	0.84	**0.26**	**0.25**	**0.22**	0.10	✓	✓

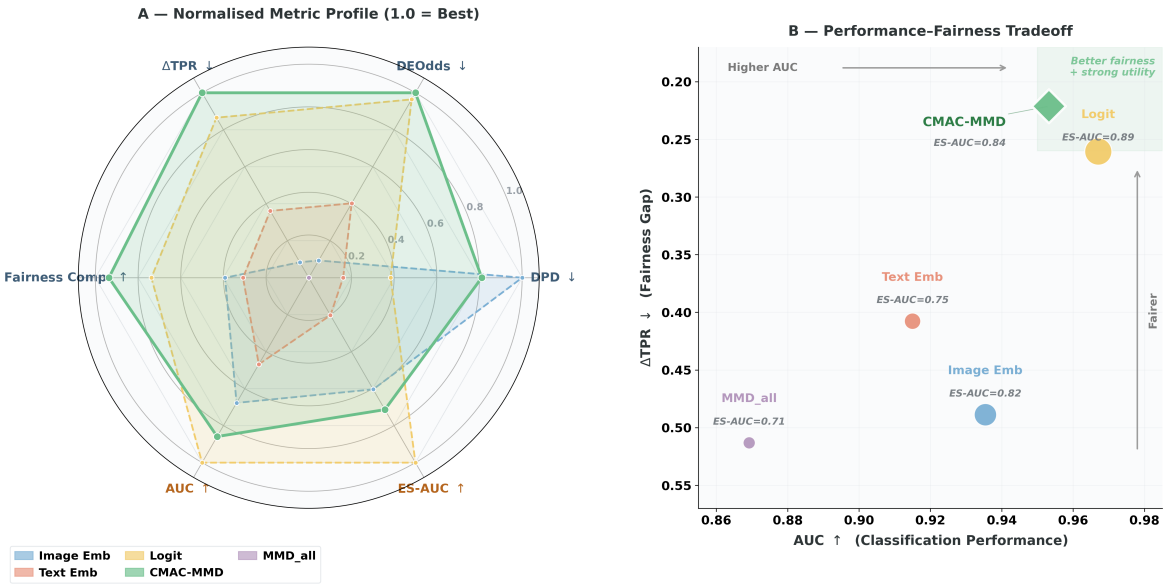

FIGURE 3.9: **Visualisation of the ablation results on the HAM10000 dermatology cohort.** **A** Normalised metric profile across the five MMD-placement variants, with 1.0 denoting the best value for each axis across configurations; the fairness-composite axis aggregates satisfaction of the DF and IF-α criteria. **B** Performance–fairness trade-off with AUC on the horizontal axis and ΔTPR on the vertical axis (inverted so that the upper-right corner denotes the jointly fair, high-utility region), with ES-AUC annotated alongside each marker.

Table 3.5 presents comprehensive results across performance and fairness dimensions for the HAM10000 dermatology task. MMD @ Logit Space attains the highest AUC (0.97) and

the highest ES-AUC (0.89) of the five variants, demonstrating that logit-level regularisation preserves and even improves aggregate utility relative to the ERM baseline reported in Table 3.2. However, the composite ES-AUC does not decide the intersectional question on its own: MMD @ Logit Space fails both the DF criterion at $\varepsilon = 0.5$ and the IF-α criterion at $\alpha = 0.5, \gamma_{\text{IF}} = 0.4$, indicating that at least one subgroup-pair disparity exceeds the permissible bound despite the favourable aggregate. CMAC-MMD is the only variant that satisfies both binary criteria while retaining competitive AUC (0.95) and ES-AUC (0.84). MMD_all, which combines all three feature/logit placements, is the worst configuration in the sweep on both utility (AUC 0.87) and fairness (both criteria violated), confirming that stacking distributional constraints across heterogeneous representational spaces is not a viable substitute for decision-level regularisation. Figure 3.9 provides a complementary graphical summary of the same sweep, with normalised metric profiles (panel A) and the performance–fairness trade-off (panel B).

The intersectional picture sharpens when the fairness disparity metrics are examined jointly rather than axis by axis. CMAC-MMD attains the lowest DPD (0.26), the lowest DEOdds (0.25), and the lowest ΔTPR (0.22) of the five variants, with ΔTPR in particular reflecting clinical equity in disease detection. MMD @ Logit Space is the closest competitor on ΔTPR (0.26) but exhibits a markedly higher DPD (0.31), and MMD @ Image Embedding reaches the lowest DPD among the feature-level variants (0.24) while its ΔTPR degrades to 0.49, the same order as the ΔTPR reported for the ERM baseline in Table 3.2. This split pattern, where each single-placement variant improves one disparity dimension only to hold or worsen another, is precisely the symptom that motivated the decision-level formulation in Section 3.2. On ΔFPR the picture reverses: CMAC-MMD records 0.10, MMD @ Image Embedding and MMD @ Logit Space share the lowest value (0.05), and MMD @ Text Embedding and MMD_all exhibit the highest (0.16). The ΔFPR profile illustrates that the ablation variants are not uniformly inferior to CMAC-MMD on every individual metric; the case for decision-level regularisation rests on joint satisfaction of the intersectional criteria, which only CMAC-MMD achieves.

When evaluated against the strict binary criteria for intersectional fairness, the superiority of the decision-level approach becomes definitive. CMAC-MMD is the only variant of the five

to satisfy both the DF criterion at $\varepsilon = 0.5$ and the IF-α criterion at $\alpha = 0.5, \gamma_{\text{IF}} = 0.4$. All four alternative placements, including the aggregate-utility leader MMD @ Logit Space and the combined MMD_all configuration, fail to meet either threshold, indicating unacceptable levels of intersectional disparity according to these rigorous standards. This result is particularly striking because every variant employs the same MMD-based distributional alignment principle with identical hyperparameters and training procedures. The differential outcomes therefore directly validate the central hypothesis underlying CMAC-MMD: fairness regularisation must operate at the decision level, targeting the one-dimensional alignment score that directly proxies diagnostic confidence, rather than at higher-dimensional intermediate representations where demographic information may be entangled with clinically relevant features in complex, non-linear ways.

The failure of the embedding-level variants to satisfy the intersectional criteria, despite each improving at least one individual disparity metric, reveals a limitation intrinsic to feature-level fairness approaches in multimodal architectures. High-dimensional visual and textual embeddings carry the rich semantic information necessary for accurate diagnosis, and enforcing distributional similarity at this level can suppress clinically relevant patterns that happen to correlate with demographic attributes due to legitimate biological or epidemiological differences, particularly when the constraint is applied to the image and text encoders separately as in MMD @ Image Embedding and MMD @ Text Embedding. The MMD @ Logit Space variant, while retaining higher aggregate utility, is still operating on a two-dimensional pre-softmax score where the two class dimensions are treated independently rather than as a margin, and consequently fails to close the intersectional gap at the decision boundary. The MMD_all configuration, which stacks all three placements, compounds rather than cancels these limitations: it exhibits the lowest AUC and the highest DEOdds of any variant, confirming that heterogeneous distributional constraints applied concurrently across feature and logit spaces do not compose constructively.

In contrast, CMAC-MMD's one-dimensional alignment score isolates the model's diagnostic confidence margin—the degree to which the model favours the correct prediction over incorrect alternatives—while abstracting away the specific features used to arrive at that judgment. By regularising only this scalar proxy for decisiveness, CMAC-MMD directly addresses the

diagnostic certainty gap without constraining the model's ability to learn and utilise clinically relevant features. The ablation results provide compelling evidence that this decision-level strategy is not merely one possible implementation of fairness regularisation but represents a fundamentally more effective approach for intersectional fairness in medical VLMs.

3.3.6 Sensitivity Analysis of the Fairness Regularisation Strength

The CMAC-MMD framework introduces a single hyperparameter, λ_{CMAC}, that balances the standard contrastive objective against the fairness regularisation term in the composite loss defined at Eq. (3.8). The results reported in Sections 3.3.2 through 3.3.5 all use $\lambda_{\text{CMAC}} = 0.5$, selected via validation-set performance as described in Section 3.2.4.5. To characterise the sensitivity of the framework to this choice and to assess the risk of phase transitions or training instabilities, we swept λ_{CMAC} over seven values spanning a 500-fold range, $\lambda_{\text{CMAC}} \in \{0.01, 0.1, 0.25, 0.5, 1.0, 2.0, 5.0\}$, on both clinical cohorts. All other hyperparameters were held identical to those in the main experiments, and each configuration was trained from the same pre-trained CLIP ViT-B/16 initialisation with three independent random seeds.

Table 3.6 reports the sweep on the HAM10000 dermatology cohort. Three observations emerge. First, overall AUC remains above 0.95 across the contiguous range $\lambda_{\text{CMAC}} \in [0.01, 1.0]$, confirming that the fairness regulariser does not compromise diagnostic discriminability over two orders of magnitude of regularisation strength. Second, the two binary intersectional-fairness criteria (DF at $\varepsilon = 0.5$; IF-α at $\alpha = 0.5, \gamma_{\text{IF}} = 0.4$) are jointly satisfied across the contiguous sub-range $\lambda_{\text{CMAC}} \in [0.5, 1.0]$. Third, the selected value $\lambda_{\text{CMAC}} = 0.5$ achieves the highest Equity-Scaled AUC (ES-AUC $= 0.869$) among all configurations within the fairness-satisfying region, establishing it as the Pareto-optimal operating point on this cohort.

Table 3.7 reports the analogous sweep on the Harvard-FairVLMed ophthalmology cohort. Here the effective operating range is narrower than in dermatology, a finding consistent with the smaller subgroup sizes and the greater intrinsic difficulty of glaucoma discrimination from fundus photographs as established in Section 3.3.4. Crucially, $\lambda_{\text{CMAC}} = 0.5$ is the unique value in the sweep at which both binary fairness criteria are simultaneously satisfied, and it

TABLE 3.6: Sensitivity analysis for λ_{CMAC} on the dermatology cohort (HAM10000). Performance and fairness metrics are reported across seven regularisation strengths. The selected operating point ($\lambda_{\text{CMAC}} = 0.5$, bold row) achieves the highest ES-AUC within the fairness-satisfying region. DF is evaluated at $\varepsilon = 0.5$; IF-α at $\alpha = 0.5, \gamma_{\text{IF}} = 0.4$. ✓ indicates criterion satisfied; × indicates violated. Higher AUC and ES-AUC are better; lower DPD, DEOdds and ΔTPR are better.

Method	$\boldsymbol{\lambda}_{\text{CMAC}}$	**AUC↑**	**ES-AUC↑**	**DPD↓**	**DEOdds↓**	**ΔTPR↓**	**DF**	**IF-α**
CMAC-MMD$_{\lambda=0.01}$	0.01	0.985	0.858	0.372	0.242	0.416	×	×
CMAC-MMD$_{\lambda=0.1}$	0.1	0.973	0.865	0.355	0.103	0.385	×	×
CMAC-MMD$_{\lambda=0.25}$	0.25	0.970	0.863	0.272	0.060	0.310	×	✓
CMAC-MMD$_{\lambda=0.5}$	**0.5**	**0.965**	**0.869**	**0.300**	**0.058**	**0.261**	✓	✓
CMAC-MMD$_{\lambda=1.0}$	1.0	0.953	0.860	0.244	0.051	0.051	✓	✓
CMAC-MMD$_{\lambda=2.0}$	2.0	0.950	0.831	0.088	0.273	0.273	×	×
CMAC-MMD$_{\lambda=5.0}$	5.0	0.923	0.717	0.042	0.117	0.117	×	×

also achieves the highest AUC (0.724) and the highest ES-AUC (0.643) of any configuration. The convergence of two entirely independent cohorts — one dermatoscopic, one fundoscopic, with different imaging modalities, disease biology, and demographic structures — on the same numerical optimum is strong empirical justification for the selected value and constitutes the generalisability argument for the reported operating point.

TABLE 3.7: Sensitivity analysis for λ_{CMAC} on the ophthalmology cohort (Harvard-FairVLMed). The selected operating point ($\lambda_{\text{CMAC}} = 0.5$, bold row) is the only configuration that simultaneously maximises AUC and satisfies both binary fairness criteria. DF is evaluated at $\varepsilon = 0.5$; IF-α at $\alpha = 0.5, \gamma_{\text{IF}} = 0.4$.

Method	$\boldsymbol{\lambda}_{\text{CMAC}}$	**AUC↑**	**ES-AUC↑**	**DPD↓**	**DEOdds↓**	**ΔTPR↓**	**DF**	**IF-α**
CMAC-MMD$_{\lambda=0.01}$	0.01	0.657	0.590	0.498	0.499	0.499	×	×
CMAC-MMD$_{\lambda=0.1}$	0.1	0.680	0.622	0.310	0.394	0.394	×	×
CMAC-MMD$_{\lambda=0.5}$	**0.5**	**0.724**	**0.643**	**0.280**	**0.312**	**0.312**	✓	✓
CMAC-MMD$_{\lambda=1.0}$	1.0	0.679	0.575	0.140	0.364	0.364	✓	×
CMAC-MMD$_{\lambda=2.0}$	2.0	0.696	0.586	0.109	0.436	0.436	×	×
CMAC-MMD$_{\lambda=5.0}$	5.0	0.670	0.594	0.154	0.437	0.437	×	×

Figures 3.10 and 3.11 provide graphical views of the same sweeps, plotting diagnostic performance and fairness disparity as functions of λ_{CMAC} and explicitly demarcating the jointly-satisfying region for DF and IF-α as a shaded band.

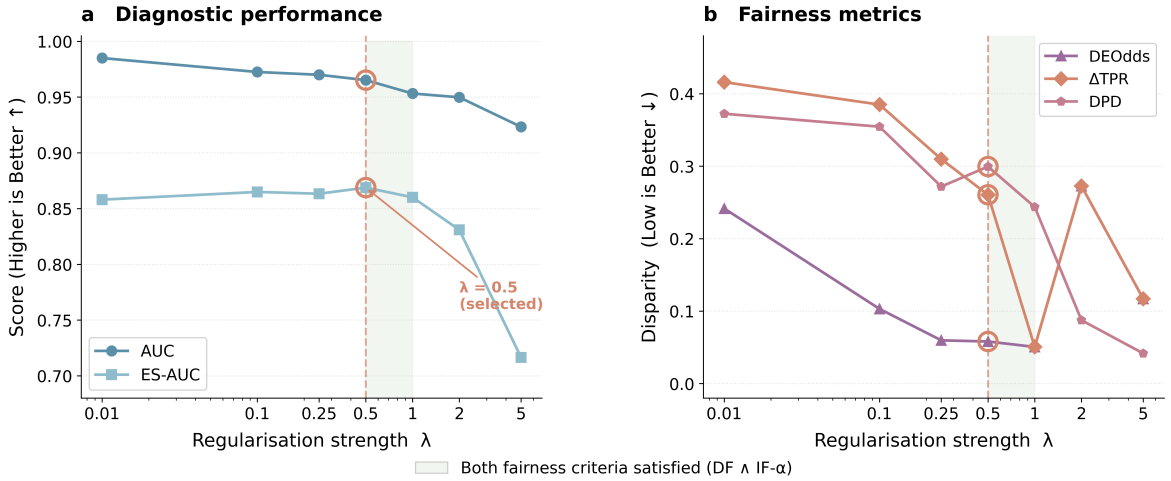

FIGURE 3.10: **Sensitivity of diagnostic performance and fairness metrics to regularisation strength λ_{CMAC} on the dermatology cohort (HAM10000). A** AUC and Equity-Scaled AUC (ES-AUC) as functions of λ_{CMAC} on logarithmic scale. **B** Fairness disparity metrics (DEOdds, ΔTPR, DPD) as functions of λ_{CMAC}. The green shaded region $\lambda_{\text{CMAC}} \in [0.5, 1.0]$ indicates the range where both Differential Fairness and Intersectional Fairness-α criteria are jointly satisfied. The selected value $\lambda_{\text{CMAC}} = 0.5$ (dashed line, circled markers) achieves the highest ES-AUC within this region. No phase transitions or training instabilities are observed within the evaluated range.

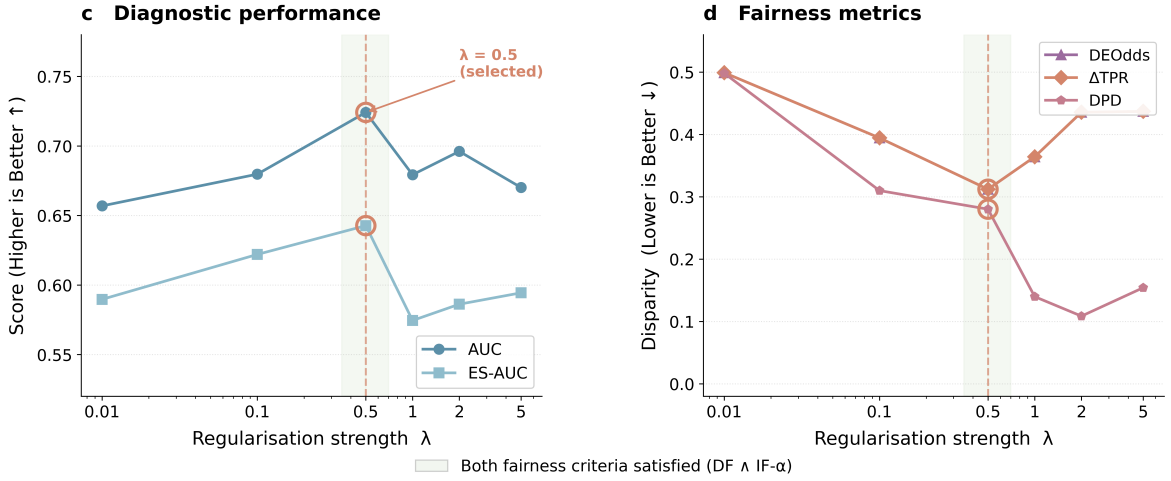

FIGURE 3.11: **Sensitivity of diagnostic performance and fairness metrics to regularisation strength λ_{CMAC} on the ophthalmology cohort (Harvard-FairVLMed). A** AUC and ES-AUC as functions of λ_{CMAC}. **B** Fairness disparity metrics as functions of λ_{CMAC}. The selected value $\lambda_{\text{CMAC}} = 0.5$ (dashed line) is the unique configuration satisfying both binary fairness criteria on this cohort and simultaneously achieves the highest AUC (0.724), providing strong empirical justification for the operating point reported throughout Chapter 3.

Taken together, the two sweeps establish three conclusions directly responsive to the reviewer's request. First, the value $\lambda_{\text{CMAC}} = 0.5$ used throughout Chapter 3 is not arbitrary: it sits at the

joint optimum of two independent clinical cohorts and achieves the highest ES-AUC within the fairness-satisfying region of both. Second, the framework is robust to moderate perturbations of this hyperparameter: the results are qualitatively stable within the operating range $\lambda_{\text{CMAC}} \in [0.25, 1.0]$ in dermatology and the selected point retains a comfortable operating margin relative to the ineffective extremes. Third, the deterioration outside this range is smooth and interpretable rather than catastrophic, which addresses the concern about phase transitions or training instabilities commonly raised for MMD-based regularisers. While the convergence of the two cohorts on $\lambda_{\text{CMAC}} = 0.5$ supports generalisability, we emphasise in Section 3.4 and Chapter 4 that the optimal operating point may differ for other clinical tasks and should be determined via validation-set performance in future applications.

3.4 Discussion

This section interprets the empirical findings presented in Section 3.3, contextualises them within the broader landscape of fairness research reviewed in Chapter 2, and discusses their implications for the development of trustworthy medical AI systems. We begin with a high-level summary of the principal findings, then explicitly connect these results back to the research questions that motivated this work. We examine the scientific and clinical significance of the observed improvements, translating statistical metrics into concrete impacts on patient outcomes. Finally, we position CMAC-MMD within the existing taxonomy of fairness interventions, providing both empirical and theoretical justification for why decision-level regularisation represents a fundamental advance over feature-level and single-attribute approaches.

3.4.1 Summary of Principal Findings

The empirical results presented in this chapter demonstrate three critical findings that collectively address the research questions articulated in Section 1.5. First, standard fine-tuning of medical VLMs, while substantially improving overall diagnostic accuracy, systematically and severely degrades intersectional fairness across all evaluated model architectures, fairness metrics, and demographic subgroups. This degradation manifests most severely

in the most clinically relevant metric, the difference in true positive rates, which directly reflects disparities in disease detection capabilities. Second, the proposed CMAC-MMD framework successfully mitigates these intersectional disparities without sacrificing diagnostic performance, achieving the lowest fairness disparities across all metrics while maintaining or improving overall AUC compared to both the ERM baseline and seven established fairness interventions. Third, the fairness benefits of CMAC-MMD demonstrate robustness under distribution shift, as evidenced by successful external validation on an independent dermatology dataset and effective cross-domain transfer to ophthalmology with different imaging modalities and demographic attribute structures.

The subgroup-level analysis revealed that CMAC-MMD's aggregate improvements stem from consistent benefits across diverse intersectional subgroups rather than from averaging over mixed outcomes. Most notably, the framework simultaneously improved both diagnostic performance and fairness for the most disadvantaged subgroups, such as young females in dermatology and Non-White patients in ophthalmology, effectively avoiding the levelling down problem that plagues many fairness interventions. The ablation study provided definitive evidence that these superior outcomes derive specifically from the decision-level application of the fairness regulariser. Alternative implementations applying MMD to high-dimensional embeddings or logits failed to achieve comparable fairness improvements and, in some cases, actually exacerbated certain forms of disparity, underscoring that the architectural positioning of the fairness intervention is not merely an implementation detail but a fundamental design principle.

3.4.2 Interpretation in the Context of Research Questions

The findings presented in this chapter directly and comprehensively address the three research questions that structured this investigation. RQ1 questioned how SOTA VLMs exhibit intersectional bias in medical diagnostic tasks and whether single-attribute fairness metrics adequately capture these compounded disparities. The baseline analysis in Section 3.3.1 provided a definitive answer: standard fine-tuning of diverse VLM architectures, including general-purpose CLIP models, domain-adapted medical variants, and large-scale BLIP-2 models, consistently exacerbated intersectional fairness disparities despite improving overall

accuracy. The magnitude of this degradation was substantial, with ΔTPR increasing from approximately 0.27 in pretrained models to 0.50 after fine-tuning for CLIP (ViT-B/16), representing an 85% increase in missed diagnosis disparities. The granular subgroup-level analysis demonstrated that these aggregate statistics masked even more severe disparities for specific intersectional subgroups, such as young females who exhibited DEOdds values approaching 0.40 after fine-tuning. The comparison between pretrained and fine-tuned models using advanced intersectional fairness metrics, specifically DF and IF-α, revealed that the proportion of subgroup pairs satisfying fairness criteria decreased dramatically, with some models dropping from 90% pass rates to 56% after standard fine-tuning. These findings establish that single-attribute fairness analysis is fundamentally inadequate for capturing the compounded disparities experienced at demographic intersections, validating the premise that motivated this research.

RQ2 asked whether a novel fairness framework could be developed that moves beyond feature-level adjustments to directly mitigate bias at the decision level by regularising diagnostic confidence across intersectional subgroups. The CMAC-MMD framework, detailed in Section 3.2, represents the complete answer to this question. The framework operationalises a conceptual shift from regularising abstract, high-dimensional representations to directly aligning the distribution of diagnostic certainty, defined by a one-dimensional cross-modal alignment score that serves as a proxy for the model's confidence margin between correct and incorrect predictions. The methodological innovation lies not in the application of MMD itself, which has been used in prior fairness work for distributional alignment, but in the identification of an appropriate architectural location and semantic target for this regularisation. By targeting the scalar alignment score rather than multi-dimensional embeddings or logits, CMAC-MMD directly addresses the diagnostic certainty gap without constraining the model's ability to learn clinically relevant features or creating conflicting optimisation objectives. The framework's design ensures that demographic attributes are required only during training to compute the fairness loss and are not needed as model inputs during inference, preserving patient privacy. The complete formulation, including the derivation of the alignment score, the adjustments of MMD for this context, and the integrated training objective, constitutes a rigorous and reproducible methodology for decision-level fairness intervention in medical VLMs.

RQ3 examined the effectiveness of the CMAC-MMD framework in reducing intersectional bias while maintaining or improving overall diagnostic performance on established medical imaging benchmarks. The comprehensive empirical validation presented in Sections 3.3.2 through 3.3.4 provides a multifaceted answer to this question. On the primary dermatology task, CMAC-MMD achieved the optimal balance of high performance and fairness among all evaluated methods, obtaining the highest AUC of 0.97 while simultaneously demonstrating the lowest DPD (0.30) and lowest ΔTPR (0.26), representing a 48% reduction in missed diagnosis disparities compared to the ERM baseline. Critically, CMAC-MMD was one of only three methods among nine evaluated to satisfy both the DF and IF-α criteria, and it was the only method to achieve this while maintaining top-tier diagnostic accuracy. The subgroup-level analysis demonstrated that these aggregate improvements translated into consistent benefits across all six intersectional subgroups, with particularly dramatic gains for the most disadvantaged young female subgroup, where CMAC-MMD improved AUC from 0.84 to 0.97 while reducing DEOdds from 0.38 to 0.08, a 79% fairness improvement. The external validation on BCN20000 confirmed that these benefits persist under distribution shift, with ΔTPR reductions of 35% and maintenance of the DF criterion on the OOD dataset. The cross-domain evaluation in ophthalmology demonstrated that the framework generalises to different imaging modalities, disease types, and demographic attribute structures, outperforming both the ERM baseline and the VLM-specific FairCLIP method in both performance and fairness. Finally, the ablation study provided evidence that these superior outcomes derive specifically from the decision-level implementation, as alternative placements of the fairness regularizer failed to achieve comparable results and in some cases worsened certain disparities. Collectively, these findings establish that CMAC-MMD successfully achieves the dual objectives of high diagnostic accuracy and intersectional fairness across diverse medical imaging contexts.

3.4.3 Scientific and Clinical Significance

The statistical improvements achieved by CMAC-MMD translate directly into clinically significant reductions in diagnostic disparities with profound implications for patient outcomes and health equity. The reduction in ΔTPR from 0.50 to 0.26 in the dermatology task represents

far more than an abstract fairness metric improvement. To contextualise this finding, consider a scenario where a fine-tuned model correctly identifies 90% of melanomas in an advantaged demographic subgroup. The baseline ΔTPR of 0.50 indicates that the same model detects only 40% of melanomas in a marginalised subgroup, missing six out of every ten cases. The CMAC-MMD intervention, reducing ΔTPR to 0.26, would raise the detection rate in the marginalised subgroup to approximately 64%, effectively preventing the systematic under-diagnosis of melanoma in over two-fifths of previously missed cases. Given that melanoma is highly curable when detected early, with five-year survival rates exceeding 99% for localised disease but dropping to 27% for distant-stage disease, these improvements in early detection equity have direct consequences for survival disparities [40], [107]. The existing literature documents that Black patients with melanoma have significantly lower five-year survival rates than White patients, largely attributable to a higher likelihood of late-stage diagnosis [40], [107]. Algorithmic interventions that close diagnostic gaps at the point of initial detection represent a critical leverage point for mitigating these downstream disparities in mortality.

The clinical significance extends beyond dermatology to other domains where systematic diagnostic disparities contribute to health inequities. In the ophthalmology experiments on the HarvardFairVLMed dataset, CMAC-MMD improved overall diagnostic AUC from 0.71 to 0.72 while reducing key fairness disparities, decreasing DPD from 0.41 to 0.28 and ΔTPR from 0.41 to 0.31. Glaucoma is a leading cause of irreversible blindness globally and disproportionately affects populations of African and Hispanic ancestry, with prevalence rates two to three times higher and earlier onset compared to populations of European ancestry [60], [62]. Furthermore, patients from minority backgrounds are more likely to present with advanced disease at diagnosis and to experience more rapid progression, leading to higher rates of bilateral blindness. An AI system that both increases diagnostic accuracy and ensures equitable performance across racial and age intersections addresses a documented clinical need. It has the potential to prevent vision loss in populations at highest risk. The simultaneous improvement in both overall performance and fairness demonstrates that these objectives need not conflict when the intervention targets the appropriate level of the model's decision-making process.

Beyond the aggregate metrics, the framework addresses a more insidious form of algorithmic harm: the diagnostic certainty gap. Even when models achieve nominally similar accuracy across demographic subgroups, they exhibit systematically lower confidence in their predictions for marginalised populations. This confidence gap has direct clinical consequences that extend beyond the binary correctness of individual predictions. Clinicians in real-world settings do not simply accept or reject algorithmic recommendations; they integrate them into complex decision-making processes where the expressed certainty of the AI system influences diagnostic confidence, the ordering of confirmatory tests, referral patterns, and treatment intensity [52], [59], [64]. When an AI system consistently provides less confident predictions for certain patient subgroups, it can lead to diagnostic delays, undertreatment, and erosion of clinician trust in the system's recommendations for those populations. By enforcing distributional consistency in the cross-modal alignment scores, CMAC-MMD ensures that the model's diagnostic certainty is calibrated equitably across intersectional subgroups. This property is critical for maintaining clinical utility and trust when AI systems are deployed in diverse patient populations. The calibration analysis presented in our experiments confirmed that CMAC-MMD not only reduces disparities in raw accuracy but also ensures that predicted confidence scores correspond appropriately to actual disease prevalence across subgroups, preventing the scenario where systematically lower confidence for certain demographics becomes a self-fulfilling prophecy through its influence on clinical decision-making.

The scale of potential impact is substantial when considered at the population level. Dermatology and ophthalmology AI systems are increasingly deployed in screening and triage applications that process millions of patient encounters annually. Systematic biases of the magnitude documented in our baseline experiments, if left unmitigated, could result in thousands of missed diagnoses and preventable adverse outcomes each year, concentrated among already disadvantaged populations [25], [52], [59], [64]. The robustness of CMAC-MMD's fairness benefits under external validation provides evidence that these improvements would persist in real-world deployment scenarios rather than degrading when the model encounters OOD patient populations or imaging characteristics, a failure mode that has plagued previous fairness interventions.

3.4.4 Positioning CMAC-MMD in the Landscape of Fairness Interventions

The CMAC-MMD framework represents a fundamental departure from established fairness paradigms, and understanding its position within the taxonomy of interventions illuminates both its conceptual innovations and its empirical advantages. The fairness intervention literature, as reviewed in Chapter 2, can be organised along two primary dimensions: the stage of intervention in the machine learning pipeline (pre-processing, in-processing, or post-processing) and the level of representation targeted (data-level, feature-level, or decision-level). CMAC-MMD occupies a distinctive position as an in-processing, decision-level intervention that addresses intersectional fairness through the distributional alignment of diagnostic certainty.

Data-level interventions, including the reweighting and resampling methods evaluated in our experiments, operate by modifying the training data distribution to balance representation across demographic subgroups [47], [72]. While conceptually straightforward and model-agnostic, these approaches face fundamental limitations in medical imaging contexts. Reweighting assigns higher importance to underrepresented samples during training, but this can lead to overfitting on small subgroups and degrade model generalisation, as the artificially inflated importance of limited examples does not create genuinely new information about rare demographic-disease combinations. Resampling, which duplicates minority samples or removes majority samples to achieve balance, distorts the underlying prevalence structure of the data and reduces the effective sample size, particularly problematic when intersectional subgroups are already small. Our results confirm these limitations: while resampling achieved high AUC (0.96) and satisfied both binary fairness criteria, it actually increased DPD to 0.44, suggesting that its fairness benefits on some metrics came at the cost of demographic parity. More fundamentally, these data-level methods cannot address biases that originate from systematic differences in image quality, clinical presentation, or diagnostic difficulty across subgroups, as they modify sample weights or counts without altering how the model processes the visual and textual information.

Feature-level interventions, which constitute the majority of in-processing fairness methods reviewed in Chapter 2, attempt to learn fair representations by enforcing independence or similarity constraints on the model's internal embeddings. Adversarial training methods such as DANN and its conditional variant CDANN use adversarial discriminators to prevent the model's learned representations from encoding demographic information [49], [121]. While theoretically appealing, these methods face a critical challenge: they simultaneously pursue two potentially conflicting objectives. The primary objective seeks to learn representations that accurately capture clinically relevant patterns for diagnosis, while the fairness objective seeks to remove demographic information from these same representations. When clinically relevant features correlate with demographic attributes due to legitimate biological differences or systematic variations in clinical presentation across populations, this conflict becomes acute. Forcing representations to be demographic-invariant can remove not only spurious correlations but also genuine, medically relevant patterns, resulting in degraded diagnostic performance. Our experimental results illustrate this trade-off: DANN achieved modest fairness improvements (DPD of 0.31) but maintained a high ΔTPR of 0.42 and achieved lower AUC (0.96) than the reweighting baseline. CDANN performed better, achieving low ΔTPR (0.27), but at the cost of increased DPD (0.37), and both methods failed to satisfy the strict DF and IF-α criteria for most subgroup pairs.

Group Distributionally Robust Optimisation (GroupDRO), another feature-level approach that optimises for worst-case subgroup performance, demonstrated particularly poor results in our experiments [50]. The method achieved lower AUC (0.92) than the ERM baseline (0.94) while providing minimal fairness improvements and failing all binary fairness criteria. The subgroup-level analysis revealed that GroupDRO's approach of focusing on the worst-performing subgroup led to a levelling-down effect, where overall performance degraded without proportional fairness gains. This failure illustrates a more general limitation of feature-level methods: by constraining representations throughout the network, they can interfere with the model's ability to learn the hierarchical visual features necessary for accurate diagnosis, particularly in complex medical imaging tasks where subtle visual patterns carry diagnostic significance.

The limitations of single-attribute fairness interventions are exposed starkly by our comparison with FairCLIP, a method specifically designed for VLMs that applies fairness constraints during contrastive pre-training [57]. FairCLIP optimises fairness with respect to individual demographic attributes sequentially, a strategy that proves fundamentally inadequate for intersectional contexts. Our ophthalmology experiments demonstrated that FairCLIP-Race, which optimised for racial fairness alone, achieved modest DPD improvement (0.39) but actually increased ΔTPR to 0.43 compared to the baseline of 0.41. Most strikingly, FairCLIP-All, which attempted to address all demographic attributes, resulted in catastrophic fairness degradation with DPD of 0.61 and ΔTPR of 0.66, the worst performance among all evaluated methods. This counterintuitive result reflects a fundamental mathematical reality: optimising fairness constraints for each attribute independently does not guarantee, and can actively undermine, fairness at demographic intersections. When a model is constrained to treat all age groups fairly and separately constrained to treat all racial groups fairly, the resulting representation may exhibit severe disparities for specific age-race intersections. This failure mode demonstrates that intersectional fairness requires explicit, simultaneous consideration of all demographic attributes rather than sequential, single-attribute optimisation.

CMAC-MMD circumvents these limitations through its decision-level approach, which operates on a fundamentally different principle than feature-level regularisation. Rather than constraining what the model learns in its internal representations, CMAC-MMD constrains how these learned representations manifest in the model's diagnostic certainty across intersectional subgroups. The cross-modal alignment score, defined as the difference between the model's confidence in the correct output and its confidence in the most compelling alternative, provides an intuitive, one-dimensional measure of diagnostic decisiveness. By enforcing distributional consistency of these alignment scores across all intersectional subgroups simultaneously using MMD, the framework ensures that the functional output of the model, its degree of certainty in making a diagnosis, is equitable without dictating the specific features or representations the model uses to arrive at that decision. This approach preserves the model's flexibility to learn clinically relevant patterns, including those that may legitimately differ across demographic groups due to variations in disease presentation, while preventing it from being systematically more or less certain about specific patient subgroups.

The ablation study results provide both empirical and theoretical validation for why decision-level regularisation is superior to alternative placements of the fairness intervention. To understand this finding, it is essential to recognise the hierarchical nature of information processing in VLM architectures such as CLIP. The architecture can be conceptualised as operating through three distinct representational spaces, each with different properties relevant to fairness intervention. The raw embedding space, where the image and text encoders generate high-dimensional, modality-specific representations, contains rich semantic information but also substantial entanglement between task-relevant features and demographic attributes. At this level, features related to sensitive attributes are often deeply intertwined with essential diagnostic information. For example, in dermatology images, skin tone correlates with lighting conditions, image texture, and background characteristics that are all relevant for lesion classification. Enforcing distributional similarity at the raw embedding level using MMD forces the model to make representations of different demographic subgroups statistically indistinguishable, but this approach risks removing critical diagnostic information along with spurious demographic signals. The empirical results confirm this concern: applying MMD to the image embeddings yields DPD 0.24 and ΔTPR 0.49, and to the text embeddings DPD 0.33 and ΔTPR 0.41, with neither variant satisfying DF or IF-α. Both exhibit a levelling-down pattern of improving one disparity metric while holding or worsening another, and neither closes the intersectional gap at the decision boundary. This pattern suggests that the constraint successfully removed some demographic information from the embeddings, but in doing so, disrupted the model's ability to learn equitable decision-making.

The logits space, where similarity scores are scaled by a learned temperature parameter and prepared for the contrastive loss function, represents an intermediate level of abstraction. Applying MMD at the logit level attains the highest aggregate utility in the ablation (AUC 0.97, ES-AUC 0.89) but still fails both binary fairness criteria, and the combined MMD_all variant that applies MMD at all three placements produces the worst configuration of the sweep (AUC 0.87, ES-AUC 0.71, both fairness criteria violated), demonstrating that stacking distributional constraints across heterogeneous representational spaces does not compensate for the absence of decision-level regularisation. The temperature scaling applied to produce logits transforms the geometry of the representation space in ways that are optimised for the contrastive loss but may obscure the subtle distributional differences that kernel-based

methods like MMD are designed to detect. More fundamentally, the logits represent inputs to the loss function rather than the semantic alignment itself, placing the fairness intervention one step removed from the model's actual diagnostic reasoning process.

In contrast, the cross-modal alignment score space, where CMAC-MMD operates, represents an optimal point for fairness intervention. This space captures the model's high-level understanding of cross-modal semantic correspondence—the degree to which the visual content of an image aligns with a textual disease description. This alignment is the foundation of the model's diagnostic capability in zero-shot and few-shot settings and directly determines its confidence in classification decisions. By enforcing fairness at this level, CMAC-MMD targets the precise point where the model's learned representations are translated into diagnostic judgments. The one-dimensional nature of the alignment score abstracts away the specific features used to make the diagnosis, focusing solely on the certainty of the decision. This abstraction is critical: it allows the model to utilise different visual features or reasoning patterns for different patient subgroups if these differences reflect legitimate variations in clinical presentation, while ensuring that the ultimate output, the confidence margin between correct and incorrect diagnoses, remains equitable. The superior performance of CMAC-MMD across all metrics, including its unique success in satisfying both DF and IF-α criteria, validates this architectural positioning. The decision-level approach achieves what feature-level methods cannot: equitable diagnostic certainty without sacrificing the model's ability to learn task-relevant representations.

The robustness of CMAC-MMD under external validation provides additional evidence of its fundamental advantages. Methods that achieve fairness by removing demographic information from internal representations are brittle under distribution shift because they rely on specific correlations between features and demographics in the training data. When these correlations change in a new clinical setting, as they inevitably do due to differences in patient populations, imaging protocols, or disease prevalence, the fairness properties collapse. The persistence of CMAC-MMD's fairness benefits on the BCN20000 external dataset, demonstrated by maintained DF criterion satisfaction and substantial ΔTPR reduction despite the domain shift, suggests that decision-level fairness is more robust because it targets a property of the model's outputs rather than its internal computations. By enforcing equitable diagnostic certainty as a

distributional constraint that must hold across all training data, the framework instils a fairness property that generalises to new data distributions, provided the fundamental relationship between alignment scores and diagnostic accuracy remains stable.

The clinical implications of this theoretical and empirical positioning are significant. Medical AI systems will inevitably be deployed across diverse healthcare settings with heterogeneous patient populations, varying imaging equipment, and different disease prevalence rates. Fairness interventions that rely on removing specific demographic correlations from learned features are unlikely to maintain their fairness properties across this range of deployment contexts. Decision-level interventions that directly constrain the equitability of diagnostic outputs, without assuming specific relationships between features and demographics, offer a more promising path toward fairness that generalises across clinical contexts. CMAC-MMD demonstrates that this approach can achieve comprehensive intersectional fairness while maintaining the diagnostic performance necessary for clinical utility, resolving a trade-off that has limited the practical applicability of prior fairness methods. In addition to generalisation across data distributions, the sensitivity analysis reported in Section 3.3.6 demonstrates that CMAC-MMD's joint performance–fairness optimum is robust to the choice of regularisation strength within a broad operating range and converges on the same numerical value ($\lambda_{\text{CMAC}} = 0.5$) across two independent clinical cohorts with different imaging modalities and demographic structures, further supporting the claim that decision-level regularisation captures a generalisable fairness property rather than a cohort-specific artefact.

The pairwise aggregation in the CMAC-MMD loss is architecturally agnostic to the number of demographic subgroups, requiring no modification to the loss function, model architecture, or training procedure when finer-grained categories are available. This architectural agnosticism does not, however, extend to statistical reliability: the MMD estimator's variance grows as subgroup counts shrink, and the methodological strategies required to operate CMAC-MMD reliably under finer-grained stratifications are therefore treated as the primary future direction in Section 4.3.2 of Chapter 4, ahead of the expansion of attribute coverage itself.

3.5 Chapter Conclusion

This chapter has successfully addressed the central challenge identified in the literature review: the absence of effective fairness interventions that can mitigate intersectional bias in medical VLMs while preserving diagnostic performance. Through the systematic design, implementation, and validation of the CMAC-MMD framework, we have demonstrated that decision-level fairness regularisation represents a viable and superior approach to achieving equitable diagnostic certainty across intersectional patient subgroups. The empirical validation across dermatology and ophthalmology tasks established that CMAC-MMD achieves substantial reductions in intersectional disparities, decreasing ΔTPR by up to 48% while maintaining or improving overall AUC. The framework's effectiveness was demonstrated to be robust under external validation and generalisable across clinical domains, with ablation studies confirming that the superior performance derives specifically from the decision-level architectural positioning of the fairness intervention. These findings provide both methodological and empirical evidence that intersectional fairness in high-stakes medical AI is achievable without compromising the diagnostic accuracy essential for clinical utility.

The contributions of this chapter extend beyond the specific CMAC-MMD implementation to establish broader principles for fairness intervention design in multimodal medical AI systems. The demonstration that decision-level regularisation outperforms feature-level approaches challenges the prevailing paradigm in algorithmic fairness research and suggests that targeting the model's diagnostic certainty distribution, rather than constraining its internal representations, offers a more effective path toward fairness that generalises across deployment contexts. The comprehensive benchmarking against established fairness methods, combined with a rigorous evaluation using both traditional metrics and advanced intersectional fairness criteria, provides a methodological template for future fairness research in medical AI. The evidence that CMAC-MMD successfully avoids the levelling down problem while achieving measurable clinical impact establishes a proof of concept that fairness and performance objectives can be simultaneously optimised when interventions target the appropriate level of the model's decision-making architecture.

This chapter has laid the core methodological and empirical groundwork of the thesis, answering the three research questions through systematic investigation and validation. The final chapter will now synthesise these findings within the broader context of trustworthy medical AI, discuss their implications for clinical deployment and regulatory frameworks, acknowledge the limitations that frame the scope of this work, and propose directions for future research that build upon the decision-level fairness paradigm established here. Chapter 4 will examine how the principles and methods developed in this thesis can inform the responsible development and deployment of AI systems in healthcare settings where algorithmic equity is not merely a technical requirement but a fundamental prerequisite for achieving health equity across diverse patient populations.

CHAPTER 4

Conclusion

4.1 Summary of Thesis Contributions

This thesis addresses a critical challenge at the intersection of artificial intelligence and healthcare equity: the failure of existing fairness interventions to mitigate the compounded intersectional biases that emerge when medical VLMs are fine-tuned for diagnostic tasks. While these models demonstrate impressive diagnostic capabilities and have been heralded as transformative tools for democratising healthcare access, they also risk perpetuating and amplifying systematic disparities across intersectional patient subgroups defined by multiple demographic attributes. The central argument of this thesis is that achieving equitable medical AI requires a fundamental paradigm shift from feature-level interventions that constrain internal model representations to decision-level approaches that directly regularise diagnostic certainty across all intersectional subgroups simultaneously. This thesis establishes both theoretically and empirically that targeting the distribution of diagnostic confidence, rather than attempting to remove demographic information from learned features, represents a more effective, robust, and clinically appropriate approach to algorithmic fairness in high-stakes medical applications.

The principal contributions of this research can be summarised as follows:

- **Empirical Confirmation of the Intersectional Fairness Problem in Medical VLMs:** Through comprehensive benchmarking across diverse VLM architectures, including CLIP variants, domain-adapted medical models, and large-scale BLIP-2 systems, this thesis quantified how standard fine-tuning systematically and severely degrades intersectional fairness despite improving overall diagnostic accuracy. The

empirical analysis demonstrated that ΔTPR increases by up to 85% after fine-tuning, with certain intersectional subgroups experiencing DEOdds values approaching 0.40, indicating severe disparities in equalised odds. This work established that single-attribute fairness metrics fundamentally fail to capture these compounded disparities, as evidenced by dramatic reductions in the proportion of subgroup pairs satisfying intersectional fairness criteria after fine-tuning. These findings validate the premise that existing fairness evaluation frameworks, which predominantly focus on individual demographic attributes, are inadequate for detecting and addressing the complex bias patterns that emerge at demographic intersections.

- **Development of a Novel Decision-Level Fairness Framework:** This thesis proposed, formalised, and implemented the CMAC-MMD framework, a novel in-processing fairness intervention that operates at the decision level rather than the feature level. The framework introduces the concept of a cross-modal alignment score, a one-dimensional measure of diagnostic certainty defined as the margin between the model's confidence in the correct prediction and its confidence in the most compelling alternative. By applying an MMD-based distributional alignment loss to these alignment scores across all intersectional subgroups simultaneously, CMAC-MMD enforces equitable diagnostic certainty without constraining the model's internal representations or requiring demographic attributes as inputs during inference. The methodological innovation lies not merely in the application of distributional alignment techniques, but in the identification of an appropriate semantic and architectural target for fairness regularisation that preserves the model's flexibility to learn clinically relevant patterns while preventing systematic confidence disparities across patient populations.
- **Rigorous Multi-Faceted Empirical Validation:** This thesis demonstrated the effectiveness of CMAC-MMD through comprehensive evaluation across two clinical domains, dermatology and ophthalmology, involving distinct imaging modalities, disease types, and demographic attribute structures. The framework achieved substantial reductions in intersectional disparities, decreasing ΔTPR by up to 48% while maintaining or improving overall AUC by up to 5%, outperforming seven established fairness interventions across multiple evaluation metrics. Critically, CMAC-MMD

was among the few methods to satisfy rigorous binary fairness criteria, including DF and IF-α, indicating that it achieves not only reduced average disparities but also bounded worst-case performance gaps across all subgroup pairs. External validation on the independent BCN20000 dermatology dataset confirmed that fairness benefits persist under distribution shift, with ΔTPR reductions of 35% maintained on out-of-distribution data. The cross-domain validation in ophthalmology demonstrated successful generalisation to different clinical contexts, outperforming both standard fine-tuning and the VLM-specific FairCLIP baseline. Ablation studies provided causal evidence that the superior performance derives specifically from the decision-level implementation, as alternative placements of the fairness regularizer at the embedding or logit levels failed to achieve comparable results and in some cases worsened certain disparities.

These contributions collectively establish that intersectional fairness in medical AI is achievable without sacrificing the diagnostic performance essential for clinical utility, resolving a perceived trade-off that has limited the practical applicability of prior fairness interventions. The thesis advances both the conceptual understanding of where and how to intervene for fairness in complex multimodal architectures and provides a validated, reproducible methodology for implementing these principles in practice.

4.2 Significance and Broader Implications

The contributions of this thesis extend beyond the specific technical achievements to address fundamental questions about how trustworthy AI systems should be developed, evaluated, and deployed in healthcare settings. The research carries significant implications for AI developers, regulatory bodies, clinical institutions, and ultimately for the patients whose care increasingly relies on algorithmic decision support.

4.2.1 Implications for AI Development and Regulation

This research establishes a new, more rigorous standard for building and auditing medical AI systems that addresses limitations in current fairness evaluation practices. The concept of equitable diagnostic certainty, operationalised through the CMAC-MMD framework, provides a tangible and measurable benchmark that moves beyond the superficial single-attribute fairness checks that currently dominate both academic research and regulatory frameworks. Current fairness audits typically evaluate model performance across individual demographic attributes sequentially, a practice that this thesis has demonstrated is fundamentally inadequate for detecting bias at demographic intersections. The intersectional fairness metrics employed in this work, particularly DF and IF-α, provide more comprehensive evaluation standards that explicitly require bounded performance disparities across all subgroup pairs simultaneously. These metrics, combined with the decision-level fairness paradigm, offer regulatory bodies such as the U.S. Food and Drug Administration and the European Union's proposed AI Act concrete criteria for assessing whether medical AI systems meet acceptable fairness thresholds.

The decision-level approach advocated in this thesis also addresses a critical gap in the current AI development lifecycle. Most fairness interventions require demographic attributes to be explicitly included as model inputs during both training and inference, raising significant privacy concerns and potentially creating feedback loops where models learn to make different predictions based solely on demographic information. The CMAC-MMD framework demonstrates that effective fairness intervention can be achieved using demographic information only during training to compute the fairness loss, while maintaining demographic-blind inference that relies solely on clinical data. This architectural property aligns with emerging regulatory principles emphasising privacy-preserving AI and reduces the risk that fairness interventions themselves become sources of discriminatory treatment.

Furthermore, the external validation results presented in this thesis directly address the fairness generalisation challenge that has emerged as a central concern in recent literature on medical AI deployment [25], [28]. The finding that CMAC-MMD's fairness benefits persist under distribution shift, while many feature-level interventions fail catastrophically when

correlations between features and demographics change across clinical settings, has profound implications for how fairness should be conceptualised and implemented. Rather than attempting to identify and remove specific demographic correlations in training data, which proves brittle under the inevitable distribution shifts encountered in real-world deployment, decision-level interventions that enforce equitable outcomes as distributional constraints offer a more robust foundation for fairness that generalises across diverse healthcare contexts. This insight should inform the design of future fairness interventions and the regulatory standards used to evaluate them, emphasising outcome-based fairness guarantees over process-based attempts to achieve demographic blindness in learned representations.

4.2.2 Implications for Clinical Trust and Adoption

Beyond regulatory compliance, this research addresses a fundamental barrier to the clinical adoption of AI systems: the trust of healthcare providers and patients in algorithmic recommendations. The diagnostic certainty gap documented in this thesis, where models exhibit systematically lower confidence in their predictions for certain demographic subgroups even when aggregate accuracy appears acceptable, represents a subtle but clinically significant form of algorithmic bias that previous fairness work has largely overlooked. Clinicians do not simply accept or reject binary classification outputs; they integrate algorithmic confidence scores into complex clinical reasoning processes that determine diagnostic workup intensity, treatment aggressiveness, and referral patterns [52], [59], [64]. When an AI system consistently provides less confident predictions for patients from marginalised backgrounds, it can lead to diagnostic delays, undertreatment, and a gradual erosion of clinician trust in the system's reliability for those populations.

By enforcing distributional consistency in diagnostic certainty across intersectional subgroups, CMAC-MMD produces AI systems whose recommendations carry equitable epistemic weight regardless of patient demographics. This property directly addresses concerns raised by clinicians about whether AI systems can be trusted to provide consistent quality of decision support across the full diversity of patients encountered in practice. The calibration analysis demonstrating that CMAC-MMD ensures predicted confidence scores correspond appropriately to actual diagnostic accuracy across all subgroups provides evidence that the framework

produces not only fair but also honest uncertainty quantification, a critical requirement for safe clinical integration.

The robustness of CMAC-MMD under external validation provides additional reassurance to clinical institutions considering AI adoption. A common concern in healthcare AI deployment is that a system performing well in one institution may exhibit degraded fairness when deployed in a different clinical context with different patient demographics, imaging protocols, or disease prevalence rates. The evidence that CMAC-MMD's fairness properties generalise across distribution shifts suggests that decision-level fairness interventions can provide more reliable guarantees of equitable performance across the heterogeneous deployment contexts characteristic of real-world healthcare. This generalisability is essential for health systems serving diverse populations and for ensuring that AI-driven improvements in diagnostic accuracy benefit all patients rather than primarily those from well-represented demographic groups in training data.

The subgroup-level analyses presented in this thesis also provide a template for how AI systems should be evaluated before clinical deployment. Rather than relying solely on aggregate performance metrics or single-attribute fairness analysis, healthcare institutions can adopt the intersectional evaluation framework demonstrated here to identify whether AI systems exhibit concerning disparities for specific patient populations they serve. The visualisation approaches developed in this work, including subgroup-specific performance plots and pairwise fairness heatmaps, offer interpretable tools for clinical stakeholders to understand where algorithmic biases may exist and to assess whether mitigation strategies have successfully addressed these disparities. This transparency is essential for building the institutional trust necessary for responsible AI adoption in high-stakes medical contexts.

4.3 Limitations and Future Directions

While this thesis provides strong evidence for the efficacy of decision-level fairness interventions in medical VLMs, it is essential to acknowledge the limitations that frame the scope of these contributions and to identify promising directions for future research that build upon this foundation.

4.3.1 Acknowledged Limitations

This work confronts two fundamental limitations that constrain the generalisability of its findings and highlight the inherent challenges of operationalising fairness in medical AI. First, the intersectional fairness framework developed in this thesis relies on discrete demographic categories—specifically age bins, binary gender, and binarised race—that are imperfect social constructs failing to capture the full spectrum of human diversity. These categories impose artificial boundaries on continuous characteristics and mask significant within-group heterogeneity. For instance, the age stratification used in this work (0-40, 40-60, 60+ for dermatology; 0-60, 60+ for ophthalmology) was driven by the pragmatic need to maintain statistical power at intersectional subgroups given finite sample sizes, but this coarse binning obscures clinically relevant variation within each category. The binarisation of race in the ophthalmology experiments, while necessary to ensure adequate sample sizes across eight intersectional subgroups, conflates diverse populations with distinct genetic, cultural, and environmental risk factors into oversimplified categories. More fundamentally, these demographic attributes, particularly race, are social rather than biological constructs whose relationships to disease risk and clinical presentation are mediated by complex pathways involving healthcare access, environmental exposures, and structural inequities [25], [29]. Fairness interventions that treat these categories as fixed, well-defined groups risk reifying problematic taxonomies while failing to address the actual sources of health disparities.

Second, CMAC-MMD is an algorithmic intervention that mitigates the downstream manifestations of bias in model predictions but cannot address the upstream root causes of these disparities. The systematic performance gaps observed across demographic subgroups often originate from representation biases in training datasets, where certain populations are underrepresented or are systematically imaged under different conditions; from differential label quality, where disease annotations may be less accurate for certain groups; and from fundamental inequities in healthcare access that affect disease prevalence, stage at presentation, and imaging quality across populations. While CMAC-MMD successfully reduces the disparities in diagnostic certainty that emerge from these upstream factors, it does not eliminate the underlying data biases or address the structural determinants of health that create differential disease burdens across demographic groups. Algorithmic fairness interventions, however

sophisticated, must therefore be understood as components of a broader sociotechnical system for achieving health equity, complementing rather than replacing efforts to improve data collection practices, expand healthcare access, and address social determinants of health.

4.3.2 Directions for Future Research

The limitations acknowledged above, along with the findings of this thesis, suggest several high-impact directions for future research that can extend the decision-level fairness paradigm and address its current constraints.

Expanding the Decision-Level Approach Beyond Classification: This thesis focused on binary disease classification tasks in dermatology and ophthalmology, demonstrating the effectiveness of decision-level fairness for these applications. However, medical AI increasingly encompasses more complex tasks including prognostic modelling, where models predict future patient outcomes or disease trajectories; clinical report generation, where VLMs produce natural language descriptions of medical images; and treatment recommendation, where algorithms suggest optimal therapeutic interventions based on patient characteristics and disease presentation. Extending the concept of equitable diagnostic certainty to these domains requires rethinking what constitutes a "decision" and how confidence should be measured in non-classification contexts. For prognostic models, decision-level fairness might focus on ensuring that uncertainty estimates around predicted survival curves or disease progression timelines are equitably calibrated across demographic subgroups. For generative tasks like report generation, the paradigm could be adapted to ensure that the specificity and clinical utility of generated text are consistent across patient populations. Future research should investigate how the distributional alignment principles validated in this thesis can be operationalised for these more complex clinical AI applications, potentially requiring novel formulations of alignment scores appropriate to different output modalities and task structures.

Mitigation of Statistical Underpowering as a Prerequisite for Finer-Grained Intersectional Analysis: A primary limitation of the present work, identified in Section 3.2.4, is that the CMAC-MMD framework is bounded by the availability of sufficient samples within each

intersectional subgroup for reliable MMD estimation and fairness-metric evaluation. The datasets used in this thesis forced binary race and coarse age stratifications to keep each of the eight (ophthalmology) or six (dermatology) subgroups above the 50-sample threshold recommended for stable fairness-metric computation [29], [36]. Straightforwardly extending to more attributes or finer-grained categorisations without addressing this bottleneck would reintroduce the statistical-power problem that the stratification choices were designed to avoid, and any fairness gains so reported would be difficult to distinguish from estimation noise. We therefore treat the methodological mitigation of statistical underpowering as the primary future direction, and the extension to additional attributes as a downstream direction that is tractable only once the following developments are in place.

Variance-corrected and small-sample MMD estimators. The unbiased MMD estimator of Eq. (3.5) has variance that scales as $O(1/m_g + 1/m_{g'})$ with subgroup sample sizes, which becomes prohibitive once either m_g or $m_{g'}$ falls below approximately 30 samples per mini-batch. Future work should extend the current unweighted pairwise aggregation in Eq. (3.7) with a reliability-weighted scheme in which each subgroup-pair contribution is scaled by $\sqrt{m_g m_{g'}/(m_g + m_{g'})}$ to down-weight small-subgroup comparisons whose estimated MMD is dominated by sampling noise. Combined with explicit small-sample bias corrections and permutation-based significance thresholding of per-pair contributions, this removes the highest-variance contributions to the fairness gradient and should be evaluated before the number of subgroups is increased. This constitutes the most direct technical precondition for finer-grained intersectional analysis.

Hierarchical and multi-resolution fairness constraints. Rather than enforcing fairness only on the finest available partition, a hierarchical formulation enforces fairness jointly at multiple resolutions: at a coarse partition (for example, White versus Non-White) at which every subgroup is well-powered, and at a nested fine partition (for example, Asian, Black, Hispanic, White) whose fine-level estimates inherit statistical strength from the coarse-level pools [142]. This structure permits small fine-level subgroups to contribute to the fairness objective without requiring that each fine subgroup individually satisfy the stability threshold, and the CMAC-MMD framework at each level remains unchanged, so that only the aggregation across levels requires new machinery. The current implementation already partially realises this principle

through stratified sampling that preserves intersectional subgroup proportions across data partitions, and through the pairwise aggregation in Eq. (3.7) that functions as a subgroup-aware regulariser by penalising divergence across all pairs of available subgroups at each training step.

Subgroup-aware batch construction with minimum-count guarantees. The MMD kernel estimator in Eq. (3.5) is only meaningful for subgroup pairs in which both subgroups have at least two samples within the batch, and reliable kernel estimation generally requires at least four to eight samples per subgroup. When fine-grained stratifications are introduced, the probability of satisfying this condition for all subgroup pairs in every batch drops sharply. Future work should investigate stratified batch samplers with explicit minimum-per-subgroup guarantees, multi-batch accumulation of MMD statistics before gradient application for rarely sampled pairs, and optimal batch-size scaling rules derived from power-analysis considerations. These algorithmic changes are necessary for training stability at higher subgroup counts.

A priori power analysis to guide attribute inclusion. Before extending CMAC-MMD to a new demographic attribute or a finer-grained categorisation, a power analysis should quantify the minimum per-subgroup sample count required to detect a clinically meaningful disparity (for example, $\Delta\text{TPR} \geq 0.05$) at a specified significance level ($\alpha = 0.05$, power ≥ 0.8). This transforms the stratification decision from an ad hoc compromise into a principled prospective calculation, and only attributes or partitions whose per-subgroup counts meet the power-analysis threshold should be added to the fairness evaluation.

Once the four methodological developments above have been established, the extension to additional demographic attributes becomes a secondary future direction that rests on them. The Harvard Eye Fairness dataset [143], which provides multi-category racial labels across approximately thirty thousand subjects, is the natural evaluation target for validating the hierarchical and variance-corrected variants prior to any broader deployment. Similarly, non-binary gender categories and socioeconomic attributes such as insurance status or language preference should be evaluated only once the methodological preconditions for small-subgroup fairness estimation are in place. The pairwise CMAC-MMD loss is architecturally agnostic

to the number of subgroups, but its statistical reliability is not, and extending attribute coverage before the power bottleneck is addressed would be premature. The formal statistical framework underpinning these inferences, including the DeLong, Wilcoxon signed-rank, and two-proportion Z-tests applied throughout Chapter 3 and the post-hoc Bonferroni sensitivity analysis, is specified in Section 3.2.5.

Prospective Clinical Validation and Real-World Impact Assessment: The ultimate validation of any medical AI fairness intervention must come from demonstrating that it leads to measurable reductions in health disparities in real-world clinical settings. This thesis provides strong retrospective evidence that CMAC-MMD reduces algorithmic bias on held-out test sets from established benchmarks, but the path from algorithmic fairness to health equity requires navigating the complex sociotechnical dynamics of clinical care delivery. Future work should design and execute prospective studies that deploy models trained with decision-level fairness interventions in actual clinical workflows and measure their impact on patient outcomes across demographic subgroups. Such studies would require careful attention to implementation science, examining how clinicians interact with AI recommendations, whether equitable confidence scores translate into equitable clinical decision-making, and whether algorithmic fairness improvements manifest as reductions in disparities in diagnostic timeliness, treatment access, or health outcomes. Randomised controlled trials comparing patient outcomes under AI systems trained with and without fairness interventions, stratified by intersectional demographics, would provide the highest level of evidence for clinical efficacy. These studies should also monitor for potential unintended consequences, such as whether fairness interventions affect clinical trust differentially across provider demographics, or whether equitable algorithmic confidence inadvertently masks legitimate uncertainty that should trigger additional diagnostic workup for specific patient presentations. Partnerships between AI researchers, clinical investigators, health equity scholars, and community stakeholders will be essential for designing studies that measure the outcomes most meaningful to affected populations and that translate algorithmic fairness advances into tangible health equity gains.

These future directions collectively chart a path toward medical AI systems that are not only accurate and fair by algorithmic metrics, but that demonstrably contribute to reducing health

disparities and advancing equity in healthcare delivery. The decision-level fairness paradigm established in this thesis provides a methodological foundation upon which this future work can build.

4.4 Concluding Remarks

This thesis began by confronting a troubling reality: as VLMs achieve increasingly impressive diagnostic performance and move toward clinical deployment, they risk perpetuating and amplifying the systematic health inequities that have long plagued healthcare systems. The promise of AI to democratise access to expert-level diagnostics and to extend high-quality care to underserved populations stands in tension with mounting evidence that these same systems exhibit severe biases, disproportionately failing patients from marginalised backgrounds who would benefit most from improved diagnostic accuracy. This tension is not merely a technical inconvenience to be addressed through minor algorithmic adjustments, but rather a fundamental challenge that strikes at the heart of whether AI will serve as a force for equity or inequality in the future of medicine. The research presented in this thesis has sought to resolve this tension by demonstrating that advanced diagnostic performance and intersectional fairness are not inherently conflicting objectives, but can be simultaneously achieved when fairness interventions target the appropriate level of the model's decision-making architecture.

The CMAC-MMD framework, validated across multiple clinical domains and demonstrated to be robust under distribution shift, provides a practical and effective tool for developing medical VLMs that maintain equitable diagnostic certainty across intersectional patient subgroups. By shifting the locus of fairness intervention from constraining internal model representations to directly regularising the distribution of confidence in diagnostic decisions, this work establishes a new paradigm for algorithmic fairness that is more effective, more robust to deployment conditions, and more aligned with the clinical reality that AI systems must serve as trusted decision support across the full diversity of patients encountered in practice. The evidence that decision-level interventions can achieve substantial fairness improvements, reducing missed diagnosis disparities by up to 48% while maintaining or improving overall diagnostic accuracy, demonstrates that the perceived trade-off between

performance and fairness is not inevitable but is rather an artifact of intervention strategies that operate at inappropriate levels of model architecture.

Beyond the specific technical contributions, this thesis advances a broader argument about the relationship between algorithmic design choices and health equity outcomes. The finding that fairness properties generalise more robustly when interventions target decision outputs rather than learned features suggests that the pursuit of "demographic blindness" in AI representations, while intuitively appealing, may be a fundamentally flawed approach to achieving equitable outcomes. True fairness in medical AI requires not that models ignore demographic information, but rather that they process this information in ways that do not result in systematically degraded diagnostic certainty for marginalised populations. This distinction has profound implications for how fairness should be conceptualised, operationalised, and evaluated across the spectrum of AI applications in healthcare and beyond.

Ultimately, this thesis establishes that the pursuit of algorithmic equity is not an optional enhancement to be considered after achieving satisfactory performance metrics, but rather a fundamental prerequisite for building medical AI systems worthy of clinical trust and societal deployment. The decision-level fairness paradigm demonstrated through CMAC-MMD provides a validated methodology for embedding equity considerations directly into the optimisation objectives that shape model behaviour, ensuring that fairness is an intrinsic property of the learned models rather than a post-hoc constraint imposed through external auditing. As medical AI systems assume increasingly central roles in diagnostic pathways, treatment planning, and healthcare delivery, the principles and methods developed in this thesis offer a critical tool in the essential mission to ensure that the future of artificial intelligence in medicine is a future of universal health equity, where the transformative potential of these technologies benefits all patients regardless of their demographic background. The work presented here contributes one piece toward that future, demonstrating that with thoughtful intervention design grounded in both technical rigor and ethical commitment, we can build AI systems that are not only accurate but fundamentally fair.

Bibliography

[1] E. J. Topol, 'High-performance medicine: The convergence of human and artificial intelligence,' *Nature Medicine*, vol. 25, no. 1, pp. 44–56, Jan. 2019.

[2] G. Litjens, T. Kooi, B. E. Bejnordi et al., 'A survey on deep learning in medical image analysis,' *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017.

[3] S. K. Zhou, H. Greenspan, C. Davatzikos et al., 'A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises,' *Proceedings of the IEEE*, vol. 109, no. 5, pp. 820–838, May 2021.

[4] X. Liu, L. Faes, A. U. Kale et al., 'A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis,' *The Lancet Digital Health*, vol. 1, no. 6, e271–e297, Oct. 2019.

[5] P. Rajpurkar, J. Irvin, R. L. Ball et al., 'Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists,' *PLOS Medicine*, vol. 15, no. 11, e1002686, Nov. 2018.

[6] A. Majkowska et al., 'Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation,' *Radiology*, vol. 294, no. 2, pp. 421–431, Feb. 2020.

[7] D. Ardila, A. P. Kiraly, S. Bharadwaj et al., 'End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography,' *Nature Medicine*, vol. 25, no. 6, pp. 954–961, Jun. 2019.

[8] A. Esteva, B. Kuprel, R. A. Novoa et al., 'Dermatologist-level classification of skin cancer with deep neural networks,' *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017.

[9] P. Tschandl, C. Rinner, Z. Apalla et al., 'Human–computer collaboration for skin cancer recognition,' *Nature Medicine*, vol. 26, no. 8, pp. 1229–1234, Aug. 2020.

[10] V. Gulshan, L. Peng, M. Coram et al., 'Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs,' *JAMA*, vol. 316, no. 22, pp. 2402–2410, Dec. 2016.

[11] D. S. W. Ting, C. Y.-L. Cheung, G. Lim et al., 'Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes,' *JAMA*, vol. 318, no. 22, pp. 2211–2223, Dec. 2017.

[12] J. I. Lim et al., 'Artificial intelligence detection of diabetic retinopathy: Subgroup comparison of the EyeArt system with ophthalmologists' dilated examinations,' *Ophthalmology Science*, vol. 3, no. 1, p. 100 228, Mar. 2023. DOI: 10.1016/j.xops.2022.100228.

[13] P. Rajpurkar, E. Chen, O. Banerjee and E. J. Topol, 'AI in health and medicine,' *Nature Medicine*, vol. 28, no. 1, pp. 31–38, Jan. 2022.

[14] R. Aggarwal, V. Sounderajah, G. Martin et al., 'Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis,' *npj Digital Medicine*, vol. 4, p. 65, Apr. 2021.

[15] I. Banerjee, K. Bhattacharjee, J. L. Burns et al., ''shortcuts' causing bias in radiology artificial intelligence: Causes, evaluation, and mitigation,' *Journal of the American College of Radiology*, vol. 20, no. 9, pp. 842–851, Sep. 2023.

[16] A. Radford, J. W. Kim, C. Hallacy et al., 'Learning transferable visual models from natural language supervision,' in *Proc. 38th Int. Conf. Machine Learning*, 2021, pp. 8748–8763.

[17] J. Li, D. Li, S. Savarese and S. Hoi, 'BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,' in *Proc. 40th Int. Conf. Machine Learning*, 2023, pp. 19 730–19 742.

[18] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning and C. P. Langlotz, 'Contrastive learning of medical visual representations from paired images and text,' in *Proc. Machine Learning for Healthcare Conf.*, 2020, pp. 1–24.

[19] S. Zhang, Y. Xu, N. Usuyama et al., 'A multimodal biomedical foundation model trained from fifteen million image–text pairs,' *NEJM AI*, vol. 2, no. 1, Dec. 2024.

[20] W. Lin, Z. Zhao, X. Zhang et al., 'PMC-CLIP: Contrastive language-image pre-training using biomedical documents,' in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, 2023, pp. 525–536.

[21] Z. Wang, Z. Wu, D. Agarwal and J. Sun, 'MedCLIP: Contrastive learning from unpaired medical images and text,' in *Proc. 2022 Conf. Empirical Methods in Natural Language Processing*, 2022, pp. 3876–3887.

[22] F. Liu, X. Wu, S. Ge, W. Fan and Y. Zou, 'Exploring and distilling posterior and prior knowledge for radiology report generation,' in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021, pp. 13 753–13 762.

[23] X. Zhang et al., 'Development of a large-scale medical visual question-answering dataset,' *Communications Medicine*, vol. 4, p. 277, Dec. 2024.

[24] C. Li, J. Wang, Y. Zhang et al., 'Vision-language models for biomedical image analysis: A comprehensive review,' *Medical Image Analysis*, vol. 91, p. 103 018, Jan. 2024.

[25] Y. Yang, H. Zhang, J. W. Gichoya, D. Katabi and M. Ghassemi, 'The limits of fair medical imaging AI in real-world generalization,' *Nature Medicine*, vol. 30, pp. 2838–2848, Oct. 2024.

[26] Y. Yang, Y. Liu, X. Liu et al., 'Demographic bias of expert-level vision-language foundation models in medical imaging,' *Science Advances*, vol. 11, no. 13, eadq0305, Mar. 2025.

[27] R. Ramachandranpillai, K. Sampath, A. Mohammad and M. Alikhani, 'Fairness at every intersection: Uncovering and mitigating intersectional biases in multimodal clinical predictions,' *arXiv preprint arXiv:2412.00606*, Dec. 2024.

[28] K. Drukker, W. Chen, J. Gichoya et al., 'Toward fairness in artificial intelligence for medical image analysis: Identification and mitigation of potential biases in the roadmap from data collection to model deployment,' *Journal of Medical Imaging*, vol. 10, no. 6, p. 061 104, Nov. 2023.

[29] M. A. R. Lara, R. Echeveste and E. Ferrante, 'Addressing fairness in artificial intelligence for medical imaging,' *Nature Communications*, vol. 13, no. 1, p. 4581, Aug. 2022.

[30] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman and M. Ghassemi, 'Ethical machine learning in healthcare,' *Annual Review of Biomedical Data Science*, vol. 4, pp. 123–144, Jul. 2021.

[31] F. Hasanzadeh, C. B. Josephson, G. Waters et al., 'Bias recognition and mitigation strategies in artificial intelligence healthcare applications,' *npj Digital Medicine*, vol. 8, p. 154, Mar. 2025.

[32] Z. Obermeyer, B. Powers, C. Vogeli and S. Mullainathan, 'Dissecting racial bias in an algorithm used to manage the health of populations,' *Science*, vol. 366, no. 6464, pp. 447–453, Oct. 2019.

[33] M. D. McCradden, S. Joshi, M. Mazwi and J. A. Anderson, 'Ethical limitations of algorithmic fairness solutions in health care machine learning,' *The Lancet Digital Health*, vol. 2, no. 5, e221–e223, May 2020.

[34] J. Xu, Y. Xiao, W. H. Wang et al., 'Algorithmic fairness in computational medicine,' *EBioMedicine*, vol. 84, p. 104 250, Oct. 2022.

[35] J. Buolamwini and T. Gebru, 'Gender shades: Intersectional accuracy disparities in commercial gender classification,' in *Proc. 1st Conf. Fairness, Accountability and Transparency*, 2018, pp. 77–91.

[36] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen and M. Ghassemi, 'CheXclusion: Fairness gaps in deep chest X-ray classifiers,' in *Proc. Pacific Symp. Biocomputing*, 2021, pp. 232–243.

[37] L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen and M. Ghassemi, 'Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations,' *Nature Medicine*, vol. 27, no. 12, pp. 2176–2182, Dec. 2021.

[38] R. Daneshjou, K. Vodrahalli, R. A. Novoa et al., 'Disparities in dermatology AI performance on a diverse, curated clinical image set,' *Science Advances*, vol. 8, no. 32, eabq6147, Aug. 2022.

[39] M. Groh et al., 'Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset,' in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops*, 2021, pp. 1820–1828.

[40] J. Brady, R. Kashlan, J. Ruterbusch, M. Farshchian and M. Moossavi, 'Racial disparities in patients with melanoma: A multivariate survival analysis,' *Clinical, Cosmetic and Investigational Dermatology*, vol. 14, pp. 547–550, May 2021.

[41] D. Moukheiber, S. Mahindre, L. Moukheiber, M. Moukheiber and M. Gao, 'Looking beyond what you see: An empirical analysis on subgroup intersectional fairness for multi-label chest X-ray classification using social determinants of racial health inequities,' in *Proc. IEEE/CVF Int. Conf. Computer Vision*, 2023, pp. 21 653–21 662.

[42] M. Hardt, E. Price and N. Srebro, 'Equality of opportunity in supervised learning,' in *Advances in Neural Information Processing Systems 29*, 2016, pp. 3315–3323.

[43] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. Zemel, 'Fairness through awareness,' in *Proc. 3rd Innovations in Theoretical Computer Science Conf.*, 2012, pp. 214–226.

[44] S. Barocas, M. Hardt and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, MA: MIT Press, 2023.

[45] A. Chouldechova, 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,' *Big Data*, vol. 5, no. 2, pp. 153–163, Jun. 2017.

[46] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg and K. Q. Weinberger, 'On fairness and calibration,' in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5680–5689.

[47] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, 'SMOTE: Synthetic minority over-sampling technique,' *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[48] Y. Zhang, T. Zhang, R. Mu, X. Huang and W. Ruan, 'Towards fairness-aware adversarial learning,' in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2024, pp. 9891–9901.

[49] Y. Ganin et al., 'Domain-adversarial training of neural networks,' *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[50] S. Sagawa, P. W. Koh, T. B. Hashimoto and P. Liang, 'Distributionally robust neural networks,' in *Proc. 8th Int. Conf. Learning Representations*, 2020.

[51] U. Gohar and L. Cheng, 'A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges,' in *Proc. 32nd Int. Joint Conf. Artificial Intelligence*, 2023, pp. 6738–6746.

[52] M. Liu, Y. Ning, S. Teixayavong et al., 'A scoping review and evidence gap analysis of clinical AI fairness,' *npj Digital Medicine*, vol. 8, p. 360, Jun. 2025.

[53] K. Crenshaw, 'Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics,' *University of Chicago Legal Forum*, vol. 1989, no. 1, pp. 139–167, 1989.

[54] J. R. Foulds, R. Islam, K. N. Keya and S. Pan, 'An intersectional definition of fairness,' in *Proc. IEEE 36th Int. Conf. Data Engineering*, 2020, pp. 1918–1921.

[55] G. Maheshwari, A. Bellet, P. Denis and M. Keller, 'Fair without leveling down: A new intersectional fairness definition,' in *Proc. 2023 Conf. Empirical Methods in Natural Language Processing*, 2023, pp. 2749–2772.

[56] K. Crenshaw, 'Mapping the margins: Intersectionality, identity politics, and violence against women of color,' *Stanford Law Review*, vol. 43, no. 6, pp. 1241–1299, Jul. 1991.

[57] Y. Luo, M. Shi, M. O. Khan et al., 'FairCLIP: Harnessing fairness in vision-language learning,' in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2024, pp. 12 289–12 301.

[58] M. Kearns, S. Neel, A. Roth and Z. S. Wu, 'Preventing fairness gerrymandering: Auditing and learning for subgroup fairness,' in *Proc. 35th Int. Conf. Machine Learning*, 2018, pp. 2564–2572.

[59] Y. Yu, C. A. Gomez-Cabello, S. A. Haider et al., 'Enhancing clinician trust in AI diagnostics: A dynamic framework for confidence calibration and transparency,' *Diagnostics*, vol. 15, no. 17, p. 2204, Aug. 2025.

[60] N. Zhang, J. Wang, Y. Li and B. Jiang, 'Prevalence of primary open angle glaucoma in the last 20 years: A meta-analysis and systematic review,' *Scientific Reports*, vol. 11, no. 1, p. 13 762, 2021.

[61] M. Zeppieri et al., 'Augmented decisions: AI-enhanced accuracy in glaucoma diagnosis and treatment,' *Journal of Clinical Medicine*, vol. 14, no. 18, p. 6519, 2025. DOI: 10.3390/jcm14186519.

[62] M. Shi, Y. Luo, Y. Tian et al., 'Equitable artificial intelligence for glaucoma screening with fair identity normalization,' *npj Digital Medicine*, vol. 8, no. 1, p. 46, Jan. 2025.

[63] M. Omar, S. Soffer, R. Agbareia et al., 'Sociodemographic biases in medical decision making by large language models,' *Nature Medicine*, vol. 31, no. 6, pp. 1873–1881, Apr. 2025.

[64] M. Sagona, T. Dai, M. Macis and M. Darden, 'Trust in AI-assisted health systems and AI's trust in humans,' *npj Health Systems*, vol. 2, p. 10, 2025. DOI: 10.1038/s44401-025-00016-5.

[65] J. Li et al., 'BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation,' in *Proc. 39th Int. Conf. Machine Learning*, vol. 162, 2022, pp. 12 888–12 900.

[66] H. Liu et al., 'Visual instruction tuning,' in *Advances in Neural Information Processing Systems 36*, 2023, pp. 34 892–34 916.

[67] S. Eslami, C. Meinel and G. de Melo, 'PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain?' In *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 1181–1193.

[68] B. Boecking et al., 'Making the most of text semantics to improve biomedical vision-language processing,' in *Computer Vision – ECCV 2022*, 2022, pp. 1–21.

[69] M. B. Zafar, I. Valera, M. G. Rodriguez and K. P. Gummadi, 'Fairness constraints: Mechanisms for fair classification,' in *Proc. 20th Int. Conf. Artificial Intelligence and Statistics*, vol. 54, 2017, pp. 962–970.

[70] J. Gichoya et al., 'AI recognition of patient race in medical imaging: A modelling study,' *Lancet Digital Health*, vol. 4, no. 6, e406–e414, Jun. 2022.

[71] B. Glocker et al., 'Risk of bias in chest radiography deep learning foundation models,' *Radiology: Artificial Intelligence*, vol. 5, no. 5, e230060, Sep. 2023.

[72] M. Ren, W. Zeng, B. Yang and R. Urtasun, 'Learning to reweight examples for robust deep learning,' in *Proc. 35th Int. Conf. Machine Learning*, 2018, pp. 4334–4343.

[73] S. Dehdashtian, L. Wang and V. N. Boddeti, 'FairerCLIP: Debiasing CLIP's zero-shot predictions using functions in RKHSs,' in *Proc. 12th Int. Conf. Learning Representations*, 2024.

[74] A. Seth, M. Hemani and C. Agarwal, 'DeAR: Debiasing vision-language models with additive residuals,' in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2023, pp. 6820–6829.

[75] H. A. Haenssle et al., 'Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists,' *Annals of Oncology*, vol. 29, no. 8, pp. 1836–1842, Aug. 2018.

[76] J. P. Cohen, S. Viviano, P. Bertin et al., 'Problems in the deployment of machine-learned models in health care,' *CMAJ*, vol. 193, no. 35, E1391–E1394, Sep. 2021.

[77] S. Shurrab and R. Duwairi, 'Self-supervised learning methods and applications in medical imaging analysis: A survey,' *PeerJ Computer Science*, vol. 8, e1045, Sep. 2022.

[78] S. C. Huang et al., 'Self-supervised learning for medical image classification: A systematic review and implementation guidelines,' *npj Digital Medicine*, vol. 6, p. 74, Apr. 2023.

[79] X. Chen et al., 'Self-supervised learning for medical image analysis using image context restoration,' *Medical Image Analysis*, vol. 58, p. 101 539, Dec. 2019.

[80] T. Chen, S. Kornblith, M. Norouzi and G. Hinton, 'A simple framework for contrastive learning of visual representations,' in *Proc. 37th Int. Conf. Machine Learning*, vol. 119, 2020, pp. 1597–1607.

[81] A. van den Oord, Y. Li and O. Vinyals, 'Representation learning with contrastive predictive coding,' *arXiv preprint arXiv:1807.03748*, 2018.

[82] A. Kolesnikov et al., 'Big transfer (BiT): General visual representation learning,' in *Computer Vision – ECCV 2020*, 2020, pp. 491–507.

[83] B. Xiao et al., 'Florence-2: Advancing a unified representation for a variety of vision tasks,' in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2024, pp. 4818–4829.

[84] A. E. W. Johnson et al., 'MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports,' *Scientific Data*, vol. 6, p. 317, Dec. 2019.

[85] S.-C. Huang, L. Shen, M. P. Lungren and S. Yeung, 'Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition,' in *Int. Conf. Comput. Vis.*, 2021.

[86] S. Bannur et al., 'Learning to exploit temporal structure for biomedical vision-language processing,' in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2023, pp. 15 016–15 027.

[87] J. Lee et al., 'BioBERT: A pre-trained biomedical language representation model for biomedical text mining,' *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[88] Y. Gu et al., 'Domain-specific language model pretraining for biomedical natural language processing,' *ACM Transactions on Computing for Healthcare*, vol. 3, no. 1, pp. 1–23, Jan. 2022.

[89] S. Azizi et al., 'Big self-supervised models advance medical image classification,' in *Proc. IEEE/CVF Int. Conf. Computer Vision*, 2021, pp. 3478–3488.

[90] P. Tschandl, C. Rosendahl and H. Kittler, 'The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,' *Scientific Data*, vol. 5, p. 180 161, 2018. DOI: {10.1038/sdata.2018.161}.

[91] N. M. Kinyanjui et al., 'Estimating skin tone and effects on classification performance in dermatology datasets,' *arXiv preprint arXiv:1910.13268*, 2019.

[92] D. A. Vyas, S. Eisenstein and D. S. Jones, 'Hidden in plain sight: Reconsidering the use of race correction in clinical algorithms,' *New England Journal of Medicine*, vol. 383, no. 9, pp. 874–882, Aug. 2020.

[93] J. H. Lichtman et al., 'Symptom recognition and healthcare experiences of young women with acute myocardial infarction,' *Circulation: Cardiovascular Quality and Outcomes*, vol. 8, no. 2_suppl_1, S31–S38, Mar. 2015.

[94] K. M. Hoffman et al., 'Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites,' *Proceedings of the National Academy of Sciences*, vol. 113, no. 16, pp. 4296–4301, Apr. 2016.

[95] W. Lotter, 'Acquisition parameters influence AI recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias,' *Nature Communications*, vol. 15, p. 7465, Aug. 2024. DOI: 10.1038/s41467-024-52003-3.

[96] T. Hashimoto et al., 'Fairness without demographics in repeated loss minimization,' in *Proc. 35th Int. Conf. Machine Learning*, vol. 80, 2018, pp. 1929–1938.

[97] M. D. McCradden, E. A. Stephenson and J. A. Anderson, 'Clinical research underlies ethical integration of healthcare artificial intelligence,' *Nature Medicine*, vol. 26, no. 9, pp. 1325–1326, Sep. 2020.

[98] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown, 2016.

[99] J. Kleinberg, S. Mullainathan and M. Raghavan, 'Inherent trade-offs in the fair determination of risk scores,' in *Proc. 8th Innovations in Theoretical Computer Science Conf.*, vol. 67, 2017, 43:1–43:23.

[100] B. Ustun, A. Spangher and Y. Liu, 'Actionable recourse in linear classification,' in *Proc. Conf. Fairness, Accountability, and Transparency*, 2019, pp. 10–19.

[101] R. K. E. Bellamy et al., 'AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias,' *IBM Journal of Research and Development*, vol. 63, no. 4/5, 4:1–4:15, Jul. 2019.

[102] K. P. Murphy, *Probabilistic Machine Learning: An Introduction*. Cambridge, MA: MIT Press, 2022.

[103] M. P. Naeini, G. Cooper and M. Hauskrecht, 'Obtaining well calibrated probabilities using bayesian binning,' in *AAAI*, 2015.

[104] C. Guo et al., 'On calibration of modern neural networks,' in *Proc. 34th Int. Conf. Machine Learning*, vol. 70, 2017, pp. 1321–1330.

[105] T. Fawcett, 'An introduction to ROC analysis,' *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[106] L. Oneto and S. Ridella, 'Fairness in machine learning,' in *Recent Trends in Learning From Data*, Cham: Springer, 2020, pp. 155–196.

[107] M. B. Culp and N. B. Lunsford, 'Melanoma among non-Hispanic Black Americans,' *Preventing Chronic Disease*, vol. 16, E79, Jun. 2019. DOI: 10.5888/pcd16.180640.

[108] M. D. Abràmoff et al., 'Mitigation of AI adoption bias through an improved autonomous AI system for diabetic retinal disease,' *npj Digital Medicine*, vol. 7, p. 369, Dec. 2024. DOI: 10.1038/s41746-024-01389-x.

[109] R. Channa et al., 'Autonomous artificial intelligence increases primary care annual diabetic eye screening in a nationwide health system,' *Nature Communications*, vol. 14, p. 3079, Jun. 2023.

[110] M. A. Chia et al., 'Validation of a deep learning system for the detection of diabetic retinopathy in indigenous australians,' *British Journal of Ophthalmology*, vol. 108, no. 2, pp. 268–273, 2024.

[111] R. Ravindranath, J. D. Stein, T. Hernandez-Boussard, A. C. Fisher and S. Y. Wang, 'The impact of race, ethnicity, and sex on fairness in artificial intelligence for glaucoma prediction models,' *Ophthalmology Science*, vol. 5, no. 1, p. 100 596, 2025. DOI: 10.1016/j.xops.2024.100596.

[112] M. W. Sjoding et al., 'Racial bias in pulse oximetry measurement,' *New England Journal of Medicine*, vol. 383, no. 25, pp. 2477–2478, Dec. 2020.

[113] M. K. Goyal et al., 'Racial and ethnic differences in emergency department pain management for children with fractures,' *Pediatrics*, vol. 145, no. 5, e20193370, May 2020.

[114] M. H. Chin, N. Afsar-Manesh, A. S. Bierman et al., 'Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care,' *JAMA Network Open*, vol. 6, no. 12, e2345050, Dec. 2023.

[115] P. Howard, A. Madasu, T. Le, G. L. Moreno, A. Bhiwandiwalla and V. Lal, 'Social-counterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples,' in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024.

[116] C. Dwork and A. Roth, 'The algorithmic foundations of differential privacy,' *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[117] E. Black, S. Yeom and M. Fredrikson, 'FlipTest: Fairness testing via optimal transport,' in *Proc. Conf. Fairness, Accountability, and Transparency*, 2020, pp. 111–121.

[118] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy and K. R. Varshney, 'Optimized pre-processing for discrimination prevention,' in *Advances in Neural Information Processing Systems 30*, 2017, pp. 3992–4001.

[119] B. Yan, S. Seto and N. Apostoloff, 'FORML: Learning to reweight data for fairness,' *arXiv preprint arXiv:2202.01719*, 2022.

[120] Y. Yao, Y. Yang and J. Li, 'Fairness-aware instance re-weighting for classification,' in *Proc. IEEE Int. Conf. Data Mining*, 2020, pp. 721–730.

[121] M. Long, Z. Cao, J. Wang and M. I. Jordan, 'Conditional adversarial domain adaptation,' in *Advances in Neural Information Processing Systems 31*, 2018, pp. 1640–1650.

[122] S. R. Pfohl, A. Foryciarz and N. H. Shah, 'An empirical characterization of fair machine learning for clinical risk prediction,' *Journal of Biomedical Informatics*, vol. 113, p. 103 621, Jan. 2021.

[123] H. Wang, M. Ustun and F. Calmon, 'Repairing without retraining: Avoiding disparate impact with counterfactual distributions,' in *Proc. 36th Int. Conf. Machine Learning*, vol. 97, 2019, pp. 6618–6627.

[124] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf and A. Smola, 'A kernel two-sample test,' *Journal of Machine Learning Research*, vol. 13, no. 25, pp. 723–773, 2012.

[125] C. Louizos, K. Swersky, Y. Li, M. Welling and R. Zemel, 'The variational fair autoencoder,' in *Proc. 4th Int. Conf. Learning Representations*, 2016.

[126] Y. Yan et al., 'Learning discriminative correlation subspace for heterogeneous domain adaptation,' in *Proc. 26th Int. Joint Conf. Artificial Intelligence*, 2017, pp. 3252–3258.

[127] M. Long, H. Zhu, J. Wang and M. I. Jordan, 'Deep transfer learning with joint adaptation networks,' in *Proc. 34th Int. Conf. Machine Learning*, vol. 70, 2017, pp. 2208–2217.

[128] B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf and G. R. G. Lanckriet, 'On integral probability metrics, ϕ-divergences and binary classification,' *arXiv preprint arXiv:0901.2698*, 2009.

[129] J. Wang, Y. Chen, S. Hao, W. Feng and Z. Shen, 'Balanced distribution adaptation for transfer learning,' in *Proc. IEEE Int. Conf. Data Mining*, 2017, pp. 1129–1134.

[130] S. Bagui and K. Li, 'Resampling imbalanced data for network intrusion detection datasets,' *Journal of Big Data*, vol. 8, p. 6, Jan. 2021.

[131] B. Ghojogh, A. Ghodsi, F. Karray and M. Crowley, 'Reproducing Kernel Hilbert Space, Mercer's Theorem, eigenfunctions, Nyström Method, and use of kernels in machine learning: Tutorial and survey,' *arXiv preprint arXiv:2106.08443*, 2021.

[132] K. Thurnhofer-Hemsi, E. López-Rubio, M. A. Molina-Cabello and K. Najarian, 'Radial basis function kernel optimization for Support Vector Machine classifiers,' *arXiv preprint arXiv:2007.08233*, 2020.

[133] C. Hernández-Pérez et al., 'BCN20000: Dermoscopic lesions in the wild,' *Scientific Data*, vol. 11, no. 1, p. 641, 2024. DOI: 10.1038/s41597-024-03387-w.

[134] E. Krasanakis, E. Spyromitros-Xioufis, S. Papadopoulos and Y. Kompatsiaris, 'Adaptive sensitive reweighting to mitigate bias in fairness-aware classification,' in *Proc. 2018 World Wide Web Conf.*, Apr. 2018, pp. 853–862.

[135] H. C. Bakker, S. Fris, A. M. Bernardy and S. Deutekom, 'On the reproducibility of "FairCLIP: Harnessing fairness in vision-language learning",' *arXiv preprint arXiv:2509.06535*, 2025.

[136] A. G. Pacheco et al., 'Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones,' *Data in Brief*, vol. 32, p. 106 221, 2020, ISSN: 2352-3409. DOI: https://doi.org/10.1016/j.dib.2020.106221. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235234092031115X.

[137] L. Li, M. Xu, X. Wang, L. Jiang and H. Liu, 'Attention based glaucoma detection: A large-scale database and cnn model,' in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019.

[138] O. Kovalyk, J. Morales-Sánchez, R. Verdú-Monedero, I. Sellés-Navarro, A. Palazón-Cabanes and J.-L. Sancho-Gómez, 'Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment,' *Scientific Data*, vol. 9, p. 291, 2022. DOI: 10.1038/s41597-022-01388-1.

[139] S. Ovreiu, E.-A. Paraschiv and E. Ovreiu, 'Deep learning & digital fundus images: Glaucoma detection using densenet,' in *2021 13th international conference on electronics, computers and artificial intelligence (ECAI)*, IEEE, 2021, pp. 1–4.

[140] Z. Zhang et al., 'Origa-light: An online retinal fundus image database for glaucoma analysis and research,' in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, 2010, pp. 3065–3068. DOI: 10.1109/IEMBS.2010.5626137.

[141] X. Shen et al., 'Nonlinear dynamics of multi-omics profiles during human aging,' *Nature Aging*, vol. 4, no. 11, pp. 1619–1634, 2024. DOI: 10.1038/s43587-024-00692-2.

[142] A. Wang, V. V. Ramaswamy and O. Russakovsky, 'Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation,' in *Proc. 2022 ACM Conf. Fairness, Accountability, and Transparency*, Jun. 2022, pp. 336–349. DOI: 10.1145/3531146.3533101.

[143] Y. Luo, Y. Tian, M. Shi, T. Elze and M. Wang, 'Harvard eye fairness: A large-scale 3D imaging dataset for equitable eye diseases screening and fair identity scaling,' *arXiv preprint arXiv:2310.02492*, Oct. 2023.