Learning to See Lesions, Not Skin Tone: Counterfactual Multimodal Learning for Fair, Trustworthy, and Text-Free Dermatology AI

Shivam Jangid

Under the Supervision of

Prof. Swagatam Das

Electronics and Communication Sciences Unit

Indian Statistical Institute

A thesis submitted in partial fulfillment
of the requirements for the degree of

Masters of Technology
in Computer Science

June, 2026

To my mother and father,
For your endless love, support, and belief in me.
With gratitude and love.

Acknowledgements

Firstly, I would like to express my deepest gratitude to my supervisor, Prof. Swagatam Das, for his unwavering support and guidance throughout this journey. I am immensely grateful to him, not only for accepting me as his student but also for patiently teaching me the fundamental steps of independently conducting quality research. From identifying a meaningful problem to presenting a well crafted solution in a written article, he was always there to share his wealth of knowledge and experience. Without his invaluable insights and advice, this dissertation would never have become a reality.

I would also like to extend my heartfelt thanks to Dr. Faizanuddin Ansari for his constant and generous support throughout the course of this dissertation. His encouragement, thoughtful guidance, and timely counsel played an instrumental role in shaping the direction of my work. I feel truly fortunate to have had his presence and mentorship alongside me at every step of this academic journey.

I would like to take this opportunity to thank the rest of the faculty members of the Electronics and Communication Sciences Unit, Indian Statistical Institute, Kolkata, for offering their helpful suggestions. I also wish to thank my fellow colleagues, who over the course of two years have become my treasured friends. I would also like to acknowledge the Indian Statistical Institute and the Electronics and Communication Sciences Unit for providing the required infrastructure.

Last but not least, I would like to thank my parents and friends, the people who form my immediate and extended family. I am immensely grateful for their selfless love, constant support, and encouragement through every high and low of this journey.

Abstract

Recent advances in deep learning have significantly improved the performance of automated skin lesion classification systems, enabling accurate detection of various dermatological conditions from medical images. Despite these achievements, concerns regarding fairness and generalization remain a major challenge for the deployment of such systems in real-world clinical settings. A key factor contributing to these challenges is the presence of bias in training datasets, particularly with respect to skin-tone representation. Most publicly available skin lesion datasets contain a disproportionate number of samples from individuals with lighter skin tones. As a result, deep learning models trained on these datasets often learn representations that perform well for majority populations while exhibiting degraded performance on underrepresented groups. Such disparities can lead to unequal diagnostic outcomes and raise important concerns regarding the reliability and fairness of artificial intelligence systems in healthcare. The problem is especially relevant in the Indian context, where substantial diversity exists in skin tone, ethnicity, and geographical distribution. Variations in lesion appearance across different skin types, combined with the limited availability of representative datasets, make it difficult to develop models that generalize effectively to the broader population. Consequently, addressing dataset bias and improving model robustness across diverse demographic groups has become an important research objective in skin lesion analysis. Motivated by these challenges, this dissertation investigates fairness-aware approaches for skin lesion classification. The work focuses on the creation and analysis of diverse skin lesion datasets, the study of bias mitigation techniques, and the development of robust deep learning models capable of learning equitable representations. Furthermore, the dissertation explores invariant and equivariant learning paradigms as potential mechanisms for improving generalization across heterogeneous data distributions. To enhance model transparency and support trustworthy decision-making, explainability techniques are also examined to provide insights into the features utilized by the learned models during classification. Through these investigations, the dissertation aims to contribute toward the development of more reliable, interpretable, and fair deep learning systems for skin lesion analysis, with particular emphasis on improving performance across diverse skin-tone distributions.

Contents

List of Figures

List of Tables

List of Notations

$\mathbb{S}$	Dataset consisting of all samples.
s	A sample belonging to $\mathbb{S}$.
N	Total number of samples in a dataset.
d	Dimensionality of a sample.
c	Number of classes.
$\mathcal{C}$	Set of class labels, $\{1, 2, \ldots, c\}$.
P_i	Prior probability of the i-th class.
x	Input skin lesion image.
y	Ground-truth class label.
$\hat{y}$	Predicted class label.
g	Demographic (skin-tone) group.
$\mathbf{V}$	Sequence of Patch Embeddings.
$\mathbf{Q}$	Sequence of Token Embeddings.
$\mathcal{A}$	Cross-modal Attention map
$\mathcal{G}$	Set of demographic groups.
$\mathbf{T}$	Factual Note.
$\mathbf{T}_{cf}$	Counterfactual Note.
Q_{fact}	Factual clinical query embedding.
Q_{cf}	Counterfactual clinical query embedding.
q_{inf}	Global Learnable Inference Query (GLIQ).
Z_{patch}	Patch-level visual feature representation.
A_{fact}	Attention map generated from the factual query.
A_{cf}	Attention map generated from the counterfactual query.
A_{inf}	Attention map generated from the inference query.
τ	Temperature parameter used during text generation.
σ	Cosine similarity between factual and counterfactual descriptions.
λ_{fair}	Weight of the fairness regularization term.
λ_{inf}	Weight of the inference-query loss term.
α	Hyperparameter controlling auxiliary loss terms.
$\mathcal{L}_{\text{AC}-\text{QCQ}}$	Training objective for AC-CQC instance.
$\mathcal{L}_{\text{PL}-\text{CQC}}$	Training objective for PL-CQC instance.
$\mathcal{L}_{\text{total}}$	Total training objective.

List of Abbreviations and Acronyms

kNN	k-Nearest Neighbor.
SVM	Support Vector Machine.
MLP	Multi-Layer Perceptron.
CNN	Convolutional Neural Network.
ABCD	Asymmetry, Border irregularity, Colour variation, and Diameter
AI	Artificial Intelligence
SMOTE	Synthetic Minority Over-sampling Technique
ADASYN	Adaptive Synthetic Sampling
GAN	Generative Adversarial Network
TPR	True Positive Rate
FPR	False Positive Rate
FST	Fitzpatrick Skin Tone
DDI	Diverse Dermatology Images
ViT	Vision Transformer
VAE	Variational Auto Encoder
XAI	Explainability AI
LLM	Large Language Model
PQD	Predictive Quality Disparity.
DPM	Demographic Parity Measure.
EOM	Equal Opportunity Measure.

List of Publications

The following manuscripts were prepared in connection with this dissertation and are currently under review.

1. **Shivam Jangid**, [Co-authors anonymized], "[Title anonymized for double-blind review]," *Submitted to a top-tier international conference (anonymized for double-blind review). Under review.*

2. **Shivam Jangid**, Faizanuddin Ansari, Swagatam Das, "FitzDerm-CF: A Controlled Multimodal Dermatology Resource for Skin-Tone Counterfactual and Leakage Evaluation," *Submitted to the ACM International Conference on Information and Knowledge Management* **(CIKM 2026)**. *Under review.*

Research Artefacts. The FitzDerm-CF dataset and associated code are publicly available at:

- **Dataset:** https://doi.org/10.7910/DVN/OZKS1L
- **Research Code:** https://github.com/Shivamjan/FairVision-Research-Code
- **Dataset Evaluation Code:** https://github.com/Shivamjan/FitzDerm-CF-A-Controlled-Multimodal-Dermatology-Resource

Chapter 1

Introduction to Fairness and Class Imbalance in Medical Imaging

Summary

A classifier expects to be trained on an equal number of distinct labeled examples from all the classes. This expectation makes the learner to learn about all the classes equally. However, in many real-world scenario, we find it common for the events to not occur with equal probability which makes it difficult to gather enough examples for every events for the classifier to learn from. Thus, we observe an imbalance between the number of representatives for different classes while forming the training set. When the classifier learns from such a dataset, it tends to learn the pattern to better recognize the majority classes with larger number of labeled instances, leading to a deteriorating performance on the minority class. This issue has prompted both the classical machine learning and emerging deep learning communities to recognize class imbalance as a persistent challenge. Accordingly, this introductory chapter formally defines the problem, discusses its relevance in medical imaging, and surveys major research directions and developments. Furthermore, we outline the research problems that motivated the objectives of this thesis and provide a roadmap of the dissertation, highlighting the primary contributions of subsequent chapters.

1.1 The problem of class imbalance: An overview

Over the past decade, the integration of artificial intelligence and deep learning has revolutionized healthcare and medical image analysis. The convergence of large-scale medical imaging data and enhanced computational capabilities has catalyzed the development of automated diagnostic systems designed to augment clinician decision-making. Within this flourishing field, skin lesion detection has emerged as one of a critical research imperative, driven by the escalating global incidence of skin cancer and the clinical necessity for early, accurate diagnosis.

Deep learning architectures, especially Convolutional Neural Networks(CNNs) have shown remarkable effectiveness in the classification and segmentation of skin lesions. By analyzing dermoscopic and clinical imagery, these systems can help distinguish between malignant pathologies such as melanoma and benign conditions, thus optimizing diagnostic workflows and reducing the manual burden on dermatologists. Nonetheless, even though these models show great potential, their performance and reliability in real-world clinical environments is frequently undermined by issues pertaining to algorithmic fairness and bias.

A fundamental challenge within the field of medicine is the presence of dataset bias and class imbalance; skin lesion datasets, in particular, demonstrates significant skewedness towards specific skin tones, demographic cohorts, and lesion categories being underrepresented. As a result, the models frequently achieve superior performance on majority populations while demonstrating diminished accuracy for minority groups, leading to disparate outcomes. In a clinical context, such inequitable predictions are deeply concerning, as they risk delaying diagnosis or compromising the quality of patient care (de Vere Hunt et al., 2023).

Machine learning algorithms are inherently constrained by the quality and representativeness of their training data. When trained on imbalanced corpora, they are prone to learning biased feature representations that fail to generalize effectively across diverse skin types and pathological distributions. Addressing this lack of inclusivity and improving the robustness of skin lesion detection systems has, therefore, become a pivotal research challenge in the field of medical image analysis.

This dissertation centers on the mitigation of bias in deep learning based skin lesion classification; primarily on supervised deep learning models. In supervised learning, a model is trained using labeled examples to learn the mapping between medical images and their corresponding diagnostic categories. In the context of skin lesion analysis, the objective of a classifier is to correctly identify lesion categories such as melanoma, benign, etc.. This research investigates the implications of dataset imbalance and representation bias on model performance while exploring methodologies to enhance fairness and ensure robust performance across heterogeneous skin-tone distributions and diagnostic categories.

Formally, let $\mathbb{S}$ denote a dataset consisting of N number of d-dimensional data samples, where each sample represents a medical image that belongs to one of c predefined diagnostic classes. The classificaton task can therefore be described as learning a function H that maps the input data space $\mathbb{S}$ to a set of class labels C, where $C = \{1, 2, \cdots, c\}$. As discussed, a classifier is a data-driven modeling technique and the performance is largely dependent on the quality of the training set. Ideally, this training set should adhere to some regularity conditions based on the classifier's inherent assumptions one of which is equal number of training examples from each of the classes for fair learning. However, in real world scenarios, this may not always be the case and this leads to what we call the imbalanced dataset where some classes contain a larger number of labeled instances than the other ones. The classes

with abundance of labeled training instances are known as majority classes while those with a scarcity of labeled examples are known as minority classes. A classifier exposed to such class imbalance in the training set is likely to perform well on the majority classes while failing to achieve good accuracy on the minority classes (Kubat et al., 1997; Japkowicz, 2000). Let us assume that the i-th class contains N_i number of training samples in the training set X. Thus, the prior probability of the i-th class is denoted as $P_i = \frac{N_i}{N}$. We can now define class imbalance in a more formal manner as in Definition 1.1.

Definition 1.1. *A training set X is called class imbalanced if there exists a pair of distinct classes (i, j) such that $i, j \in C$, and $N_i \neq N_j$ or alternatively $P_i \neq P_j$.*

To observe how such a class imbalance in the training set can adversely affect the performance of a classifier, let us consider the following illustrative example.

It is observed from Example 1.1 that class imbalance may significantly deteriorate the accuracy of an MLP on the minority classes. However, minority classes usually correspond to rare and significant events whose accurate classification is often a crucial necessity. We can think of the example of a computer-aided diagnosis system which needs to accurately segregate the fatal malignant tumors from the harmless ones to avoid treatment delays if not unfortunate fatalities. Thus, the challenge appears in form of accurately classifying the minority class instances which are usually prone to misclassification due to their scarcity in the training set. Evidently, the problem of class imbalance gained sufficient attention from the machine learning community resulting in a plethora of approaches to efficiently combat its adverse effects concerning the performance on the minority class (He and Garcia, 2009; He and Ma, 2013; Branco et al., 2016; Krawczyk, 2016). The relevance of class imbalance only further increased with the advent of deep end-to-end learning systems (Johnson and Khoshgoftaar, 2019) directly handling complex data such as images, audios, videos, etc., and with the ever-growing importance of applications such as semantic segmentation, object detection, few-shot classification (Li et al., 2019), etc.

1.2 Class Imbalance in Medical Imaging

As discussed in the section1.1, class imbalance can substantially bias the learning process and degrade predictive performance on underrepresented classes. In medical imaging, whether across MRI, CT, ultrasound, or dermoscopy; one of the most persistent structural obstacles towards building a reliable diagnostic models is class imbalance: a statistically significant disparity in the number of samples available per diagnostic category (Cui et al., 2025). One would be wrong to assume this as a mere data collection failure that can simply be engineered away. It is an inherent reflection of clinical reality. Healthy tissue and common benign presentations vastly outnumber rare or malignant pathologies in any population, so the datasets

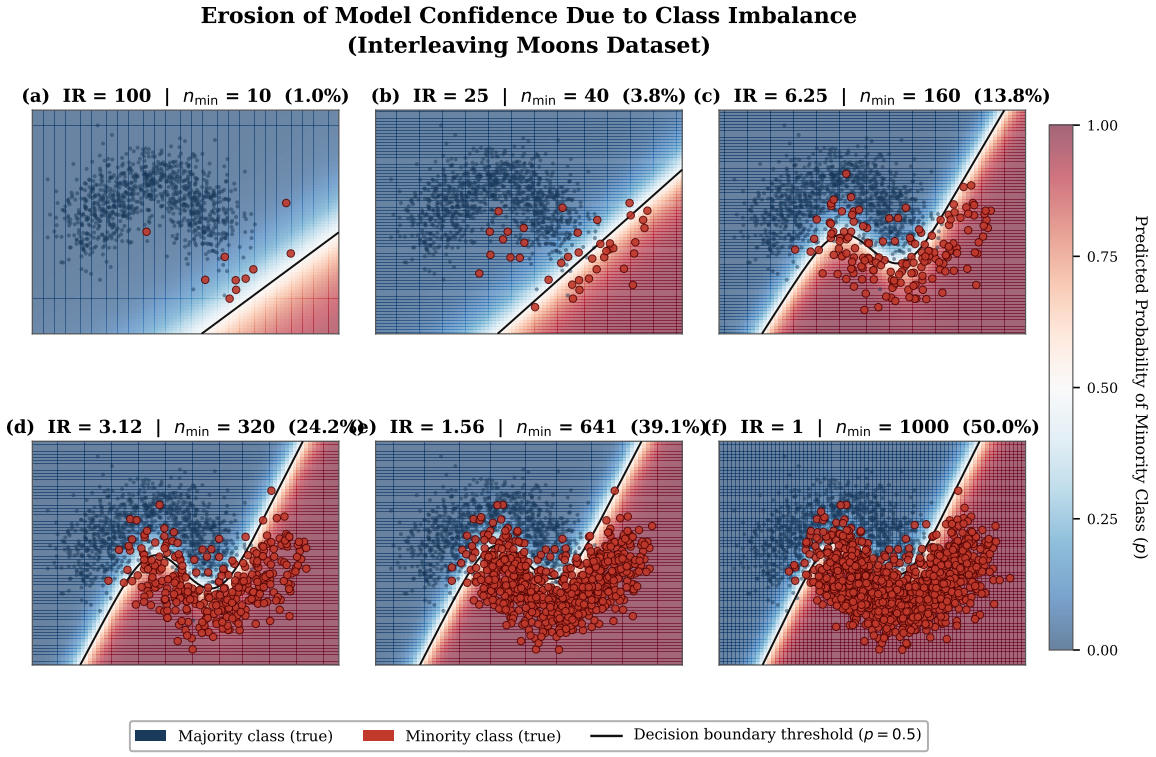


Figure 1.1: **Erosion of model confidence due to class imbalance.** The effect of class imbalance on the probability landscape and decision boundary learned by a Multi-Layer Perceptron (MLP) on the interleaving moons dataset. Background shading indicates the network's predicted probability for the minority class, ranging from 0.0 (deep blue, confident majority) to 1.0 (deep red, confident minority), with the solid black line representing the $p = 0.5$ decision threshold. **(a)** At an extreme imbalance ratio ($IR = 100$, 1.0% minority instances), the network loses almost all confidence in the minority class, resulting in a washed-out probability landscape and a severely collapsed decision boundary. **(b)–(e)** As the dataset becomes progressively more balanced ($IR = 25$ down to 1.56), the network's confidence gradually recovers. The decision boundary visibly shifts and bends to accommodate the complex non-linear geometry of the minority class. **(f)** When trained on a perfectly balanced dataset ($IR = 1$, 50.0% minority instances), the classifier displays high confidence across the feature space and smoothly separates the two interleaving clusters.

built from real patient encounters faithfully reproduce that asymmetry (Zhou et al., 2026).

The consequences for model training are well understood. A neural network minimising empirical risk over a highly skewed distribution can achieve a deceptively low loss by defaulting to the majority class for almost every prediction (Cui et al., 2024). The model's feature extractors become well-calibrated for the dominant class while the minority class feature space remains underdeveloped. In clinical terms, this translates directly into an elevated false-negative rate for the very conditions tumours, lesions, haemorrhages; that the system most urgently needs to detect. Standard accuracy is therefore a misleading evaluation

metric in imbalanced contexts; robust clinical assessment requires distribution-invariant measures such as the Dice Similarity Coefficient, Precision-Recall, and the Matthews Correlation Coefficient (Cui et al., 2025).

Researchers have developed two broad families of remedies. Data-centric approaches tends to modify the training distribution before the training loop. Naive random oversampling provides numerical balance but offers no new morphological information and frequently causes the network to memorise specific minority examples rather than learn generalizable features (Zhou et al., 2026). Other sophisticated methods such as SMOTE, ADASYN, and GAN-based synthesis generates plausible synthetic samples, though in domains characterized by high intra-class visual variability, such as dermoscopy; although affective sometimes, the synthesised instances can introduce boundary noise that blurs the very decision regions they were intended to clarify (Khandaker et al., 2025). Model-centric methods leave the dataset itself unchanged and instead modify the learning objective or architecture. Modified loss functions impose asymmetric penalties that steer the optimizer towards minority class sensitivity (Scholz et al., 2024).

In the specific context of skin lesion classification, these challenges are particularly acute. The HAM10000 benchmark (Tschandl et al., 2018), a widely used image corpus spanning seven diagnostic classes, allocates over 67% of its images to benign melanocytic nevi, while melanoma, arguably the most clinically dangerous category accounts for just 11%. A naive classifier trained on this distribution could achieve a baseline accuracy of roughly 67% by predicting "benign nevus" for every single input; a result while statistically impressive, is clinically catastrophic. The subsequent ISIC 2019 challenge expanded the corpus to over 25,000 images across eight classes, yet the fundamental dominance of benign pathology remained structurally unchanged (Kassem et al., 2020). Addressing this imbalance is a prerequisite for clinically deployable models, but it is, by itself, insufficient. As the following section argues, resolving the statistical distribution of target labels does not guarantee that the resulting model will serve all patients equitably.

1.3 Algorithmic Fairness in Medical AI

A model may be statistically well-calibrated and yet be *clinically inequitable*. Resolving class imbalance within the dataset ensures that the algorithm has a better chance at detecting rare diseases; but it says nothing about whether that detection capability is distributed evenly across the diverse patients who present with those diseases. This distinction is the central concern of algorithmic fairness (Xu et al., 2022). While imbalanced classes are a characteristic of the distribution of the labels, fairness refers to the distribution of the performance over certain demographic sensitive attributes such as race, complexion, gender, age, and so forth. The AI system can be highly precise on average while being highly prejudiced against one

particular marginalized group within the population.

The significance of ensuring fairness in the application of medical AI has attracted growing attention among scientists and regulating agencies. As per the guidance from the FDA regarding the use of AI-enabled medical devices and the WHO's framework of trusted AI in healthcare, bias assessment is required before deployment. This is in acknowledgment of the fact that any disparity in performance may have implications for the quality of clinical decision-making (Boland et al., 2025).

Formalising Fairness

Translating the intuitive notion of equity across different classes of patients into quantifiable criteria is more complicated and involves mathematical formalism. Let us consider a classifier that transforms the input data X into an output prediction $\hat{Y}$, which is compared with the ground truth label Y, and the dataset records are divided according to some sensitive attribute $A \in \{0, 1\}$. In this context, the various statistical criteria that define fairness are usually divided into two main paradigms. *Individual fairness* stipulates that if two patients share a similar clinical presentation, they will be treated equally despite being from different demographics. Despite the appeal, its implementation in practice in medicine is difficult since it necessitates a measure of patient similarity that is non-trivial to determine and is likely to vary significantly from disease to disease and patient cohort to patient cohort. Therefore, individual fairness in medical applications of machine learning systems has yet to gain widespread traction (Xu et al., 2022). *Group fairness* is instead the standard employed in large-scale clinical auditing. There are three measures that are most often applied (Pessach and Shmueli, 2022):

Demographic Parity requires that the probability of a positive prediction be equal across groups, irrespective of the true label:

$$P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1). \tag{1.1}$$

In simpler terms, demographic parity entails requiring the algorithm to have equal positive prediction rates among all demographic groups. Although this definition is meant to mitigate any kind of discrimination that can happen with regards to model outputs, it may not apply to many cases within health care settings. There could be differences in disease occurrence rates among different populations depending on biology, environment, and social conditions. This, consequently, means that the use of such a criterion would prompt the model to ignore important medical information linked to both protected features and outcomes, potentially reducing predictive performance and leading to less accurate decisions across patient groups (Xu et al., 2022).

Equalized Odds, formalised by Hardt et al. (Hardt et al., 2016), conditions the fairness

requirement on the true outcome. A classifier satisfies equalized odds if both its TPR and FPR are equal across groups for every value of Y:

$$P(\hat{Y} = 1 \mid A = 0, Y = y) = P(\hat{Y} = 1 \mid A = 1, Y = y) \quad \forall\, y \in \{0, 1\}. \tag{1.2}$$

This criterion doesn't demand that disease be predicted at equal rates across demographic subgroups; it demands that the model be equally accurate for both the diseased patients and the healthy ones, regardless of their respective demographic groups, making it a far more clinically defensible criterion.

Equal Opportunity is a relaxation to Equalized Odds1.2 that focuses solely on the positive class:

$$P(\hat{Y} = 1 \mid A = 0, Y = 1) = P(\hat{Y} = 1 \mid A = 1, Y = 1). \tag{1.3}$$

In a cancer-screening context, this guarantees that every patient who genuinely has a malignancy has the same probability of being correctly flagged for intervention, regardless of their race or sex. It is the minimum fairness standard a triage system should satisfy (Hardt et al., 2016).

While these metrics provide a formal means of evaluating fairness, they do not explain where unfairness originates. In practice, disparities in model performance often arise from biases present in the underlying data itself. In dermatological AI, one of the most widely studied examples is skin-tone bias, where unequal representation of different skin tones during training can lead to systematic differences in diagnostic performance.

Skin Tone Bias: From Theory to Clinical Consequence

The disconnect between theoretical equity and the reality of clinical results may be most starkly evidenced by the demonstrated lack of generalization of dermatology AI for use across varying skin tones. The skin tones are categorized based on the Fitzpatrick Skin Type (FST) classification, which ranges from Types I through VI (Fitzpatrick, 1988a). Upon examination of the datasets utilized for training dermatology AI including ISIC (Kassem et al., 2020) and HAM10000 (Tschandl et al., 2018), the FSTs V and VI prove to be extremely scarce due to the nature of these particular clinical patient pools.

Indeed, the Fitzpatrick 17k dataset (Groh et al., 2021b) one of the most widely used benchmarks for studying fairness and skin-tone representation in dermatological AI, was assembled specifically to audit this gap. It contains approximately 7,755 images of FST I–II, 6,089 of FST III–IV, and only 2,168 of FST V–VI. Models trained on this distribution exhibit reliable sensitivity for light skin and systematic degradation for darker skin, directly violating equalized odds.

When the cutting-edge models trained on ISIC-like datasets were tested against the bal-

anced Diverse Dermatology Images(DDI) (Daneshjou et al., 2021) benchmark data set, which contained balanced proportions in all classes in the FST scale, their AUC-ROC scores dropped sharply from their test set baselines, especially for FST V-VI. Importantly, even human dermatologists failed to diagnose dark skin types correctly. Therefore, the bias problem *is intrinsic to the ground-truth annotations themselves* (Daneshjou et al., 2021).

Such disparities have significant clinical implications. In spite of the fact that melanoma is rare among patients with dark skin types, delays in diagnosis can be responsible for worse prognosis and higher mortality (Boland et al., 2025). Consequently, AI algorithms that have low accuracy in detecting melanoma among patients with dark skin will contribute to exacerbation of current medical health care disparities. Even though several approaches have been suggested to address the problem of biased results in recent literature, it remains difficult to guarantee equality of performance.

1.4 Motivation and objective

Despite deep learning seeing significant progress for classifying skin lesion, the clinical deployment of AI based dermatology applications is hindered due to the presence of biased data and which leads to unequal performance across different demographic groups and limit the reliability of such models. This happens because of the skewed nature of the current datasets towards individuals with lighter skin, leading the model to learn these spurious correlations between skin type and the diagnosis.

A more fundamental problem, however, exists at the core of such superficial inequalities. Conventional fairness measures may indicate *that* an algorithm performs more poorly on darker skin tones; however, they provide no information about *why*. Specifically, it is unknown whether the algorithm makes use of any skin tone characteristics in selecting appropriate visual features to make predictions, as well as whether such decision making would be different in the case of patients with different skin tones. Solving such a problem calls for two ingredients: firstly, carefully curated data sets allowing demographics to be varied independently; and secondly, training criteria that actively encourage the algorithm to base its decisions on the same visual features independently of the skin tone.

Motivated by these gaps, this dissertation pursues following interrelated objectives:

1. to construct a *controlled multimodal dataset*; pairing skin lesion images with factual and counterfactual clinical notes that differ only in Fitzpatrick related descriptors, enabling principled evaluation of demographic sensitivity;

2. to develop a fairness-aware framework for dermatological AI that enforces counterfactual consistency in lesion-focused visual evidence selection in excess of providing

improved predictive performance; wihle simultaneously supporting strictly text-free, privacy-preserving inference; and

Overall, we aim to establish a principled framework for studying, mitigating, and evaluating demographic bias in multimodal dermatology AI systems.

1.5 Contributions of the Dissertation

The principal contributions of this work are as follows.

1. **FitzDerm-CF: A counterfactual multimodal dermatology dataset.** We construct a dataset extending Fitzpatrick17k and DDI with paired factual and counterfactual clinical notes generated by BioMistral-7B. A four-stage validation pipeline enforces semantic similarity, causal invariance, and skin-tone lexical intervention. Leakage analyses confirm that disease and malignancy signals are preserved across generation temperatures while demographic recoverability decreases, yielding a favourable privacy–utility trade-off.

2. **CARE-Net: A Counterfactual Multimodal Training Objective for Fairness-Aware Dermatology Diagnosis.** We propose CARE-Net, a multimodal training framework that mitigates skin-tone bias by enforcing counterfactual consistency at the level of model's visual evidence selection process, rather than at the level of global embeddings or final predictions. Two regularisation approaches form the basis of the two instantiation of the CARE-Net approach. **AC-CQC**, an asymmetric stop-gradient loss-based regularisation approach that ensures consistency of the cross-attention map generated by factual and counterfactual text queries and penalises any use of demographic bias in the selection of the visual evidence of the model. **PL-CQC** is an approach that relaxes AC-CQC and focuses on consistency of patch logit.

3. **Global Learnable Inference Query (GLIQ).** A crucial component of CARE-Net that can be leveraged to address the modality gap between training and inference time. It is a learned, static query vector obtained by distillation of the fair attention mechanism via L_1-based teacher-student training. GLIQ allows us to perform counterfactually invariant inference without access to any textual data or sensitive attributes.

Organization of the Dissertation

The remainder of this dissertation is structured as follows. Chapter 2 reviews the relevant literature on skin lesion classification, demographic bias in dermatological AI, bias mitigation

strategies, multimodal and counterfactual learning, and explainability methods, and identifies the research gaps that motivate CARE-Net. Chapter 3 describes the construction and analysis of FitzDerm-CF, covering the generation pipeline, validation criteria, temperature sweep, and leakage experiments. Chapter 4 presents the CARE-Net framework, detailing its instances(AC-CQC and PL-CQC) and GLIQ which leads to text-free model deployment. And reports experimental results on Fitzpatrick17k and DDI, including ablation studies and explainability analyses. Chapter 5 presents the exploratory equivariant extension and discusses its preliminary findings and limitations. Chapter 6 concludes the dissertation, evaluates the contributions, and outlines open problems and future directions.

Chapter 2

Literature Review

Summary

Chapter 1 introduced the challenges of class imbalance and demographic bias in skin lesion classification, highlighting the need for diagnostic systems that are both accurate and equitable across diverse patient populations. Building upon this motivation, this chapter reviews the existing literature related to skin lesion analysis, fairness in medical AI, multimodal learning, counterfactual reasoning, and explainable AI.

This chapter begins with an overview of developments in automated skin lesion classification and subsequently examines the sources and consequences of dataset bias in dermatological AI. Existing bias mitigation approaches are then discussed, followed by a review of multimodal learning techniques, counterfactual reasoning frameworks, and explainability methods. Finally, the limitations of current approaches are discussed alongside the research gaps motivating the our proposed framework.

2.1 Skin Lesion Classification

Skin cancers form one of the common types of cancer in the world today, and early detection is key to ensuring successful treatment. Even though melanoma constitutes a small percentage of all skin cancers, it contributes to the largest number of fatalities from skin cancer; thus, underlining the importance of early detection (Siegel et al., 2024). In light of such real-world issues, various researchers have directed their efforts towards developing machine learning methods capable of classifying skin lesions. These models could be used by

healthcare practitioners in making early identification of suspicious skin lesions, leading to reduced time-to-diagnosis and increased access to dermatological screening in underserved areas. Moreover, they may be useful as decision support tools, helping physicians make correct diagnoses in case of any doubts. However, transferring promising research results to practical applications turns out to be more difficult than what is usually expected, and explaining the underlying reasons, calls for tracing the entire path of scientific progress in this area.

2.1.1 Deep Learning Based Classification

Before the rise of Deep learning, the field was characterized by handcrafted feature pipelines where lesion descriptors where manually designed by the domain expert and were then fed into classical classifiers such as SVMs or random forests (Barata et al., 2013). The performance for these pipelines were ceiling-limited by the expressiveness of the hand engineered features, and generalization across different conditions, and populations was consistently poor.

From a clinical perspective, dermatologists had always organized their visual examinations using well-known heuristics. With regard to the assessment of pigmented lesions, the ABCD method, comprising Asymmetry, Border irregularity, Colour variation, and Diameter, was used (Nachbar et al., 1994). Despite the fact that visual cues mentioned are good predictors of the presence of cancer, such examination is not always objective. Therefore, different conclusions might be made when looking at the same lesion, especially in difficult cases to assess. The results obtained would depend on the level of qualification and experience of the examining physician.

The adoption of CNNs marked a major advancement in automated skin lesion classification. Unlike traditional approaches, CNNs learn hierarchical representations directly from image data, enabling end-to-end lesion analysis (Kawahara et al., 2016). Their success was further supported by the increasing availability of large annotated dermatology datasets. A notable example is the work of (Esteva et al., 2017), who demonstrated that a CNN trained on over 129,000 clinical images could achieve performance comparable to that of board-certified dermatologists on selected classification tasks. These results stimulated further research in the field and coincided with the development of several publicly available datasets and benchmarking efforts, including HAM10000, ISIC, DDI, and Fitzpatrick17k (Tschandl et al., 2018; Kassem et al., 2020; Daneshjou et al., 2021; Groh et al., 2021b).

In spite of this, early unimodal architectures were characterised by significant drawbacks. CNNs, for all their superior capability in detecting local spatial patterns, works in an informational void when used as standalone dermoscopic classifiers, based purely on the very small dermoscopic patches and devoid of auxiliary contextual information which clinicians use in making diagnosis every day: patients' ages, anatomical locations of the lesions, histories of melanoma within patients' families, evolution of the lesions over time, or even accompanying clinical signs of bleeding or itching. While a lesion might trigger a biopsy request for a patient aged 70 years and having a melanoma in his family history, the same would probably require only follow-up observation for a young patient aged 25 years and having no risk factors associated.

Furthermore, CNNs may learn spurious correlations present in their training set. Instead of considering the actual attributes of lesions, CNNs may use artifacts in images, such as ruler lines, hair, skin marks, or other specific attributes related to a certain diagnosis (Winkler et al., 2019). Although such biases might not harm their accuracy on similarly distributed data, they do compromise their accuracy in clinical practice where the distributions differ from the training datasets.

To address these challenges, recent studies have explored a range of architectural improvements, including attention mechanisms, multi-task learning, and fairness-aware training strategies (Kawahara et al., 2016; Zhang et al., 2023). More recently, transformer-based architectures and foundation models have attracted significant interest due to their ability to learn richer and more transferable representations from large-scale datasets.

2.1.2 Foundation Models for Skin Lesion Analysis

Some of the recent studies focused on Vision Transformer architecture Dosovitskiy et al. (2020) and Foundation models in the task of skin lesion segmentation. In contrast to CNNs, ViTs employ self-attention to find dependencies between regions in images and shows promising results in numerous visual recognition tasks. Advances in self-supervised learning, such as Masked Autoencoders, have further improved their effectiveness in medical imaging by reducing reliance on large labelled datasets (He et al., 2022).

Foundation models have increased the capabilities of AI models in skin lesion detection even more. With their ability to pre-train on large datasets and then adapt themselves based

on specific tasks, these models are able to learn more universal representation compared to classical supervised learning models (Bommasani et al., 2021). Recently, research in multimodal foundation models was done to help systems leverage both visual and text data in performing various clinical tasks, not limited to lesion classification alone (Liu et al., 2023). In spite of promising results, issues such as interpretability and clinical robustness are still important limitations in applying these models to practical use. There were recent developments in creating domain-specific foundation models for analyzing skin lesions based on large skin lesion databases (Yan et al., 2024).

2.2 Bias in Dermatology AI Systems

The effectiveness of machine learning systems is strongly influenced by the quality and representativeness of training data. In dermatology, where visual characteristics of skin conditions can vary considerably across patient populations, this dependency has important clinical implications. When training corpora fail to reflect the full diversity of patients who will ultimately be examined by these systems, the resulting models do not simply underperform, they underperform *selectively*, and the populations most affected are the ones already facing structural disadvantages in healthcare access. Demographic and skin-tone imbalances have raised serious concerns regarding the equitable performance of AI-based diagnostic systems, and addressing these imbalances has become one of the central challenges in translating dermatological AI from research benchmarks to clinical practice (Adamson and Smith, 2018; Groh et al., 2021a).

2.2.1 Skin Tone and Demographic Bias

As discussed in Section 1.3, disparities in skin-tone representation remain the most widely documented source of bias in dermatological AI. Publicly available skin lesion datasets disproportionately represent lighter Fitzpatrick skin types, and models trained on these corpora have consistently shown reduced diagnostic performance on underrepresented populations (Adamson and Smith, 2018; Groh et al., 2021a; Daneshjou et al., 2021). Skin tone has attracted the most scrutiny, but analogous disparities have been observed across age, sex, and geographic groups; a pattern that points to demographic bias as a systemic problem rather

than a quirk of any single dataset.

These findings expose a meaningful limitation of conventional evaluation practice. A model can achieve strong aggregate performance while behaving very differently across demographic subgroups; headline accuracy figures simply do not reveal this. In response, recent work has placed growing emphasis on fairness-aware evaluation frameworks and training strategies designed to promote more consistent performance across diverse patient populations, an acknowledgement that optimizing for the average patient is not the same as building a system that works for all of them.

2.2.2 Bias through Differences in Dataset Acquisition

Bias might also emerge due to differences in data acquisition strategies, clinical environments, imaging devices, and labelling processes; factors that cause bias indirectly rather than through an explicit under representation of certain groups, but nonetheless have significant impacts. (Oakden-Rayner et al., 2020) refers to this as *hidden stratification*, a problem that occurs when models achieve excellent results on aggregate metrics while being biased against clinically relevant groups that are unlabelled and unstudied. Because of their unlabelled nature, these biases might not even be detected until the moment the model is applied in a real-life clinical context.

This is because of the way different acquisition techniques might generate different spurious statistical artifacts that models tend to focus on in the process of training (Oakden-Rayner et al., 2020). An algorithm trained using high-quality dermoscopy images collected at tertiary care hospitals might perform poorly at assessing the same skin lesion types acquired in primary care with much poorer quality images; likewise, the differences in labelling procedures between various institutions might lead to problems caused by label noise, further aggravating acquisition bias.

One aspect of the issue that is perhaps not well understood is the ability of these deep AI systems to identify characteristics such as race through images even when clinicians cannot. As such, the systems may contain some kind of proxy for these characteristics within their representation (Adamson and Smith, 2018). The implications of having a system that is able to determine the characteristic of patients from images is that there may be different decision rules that are used to classify the population.

Clinical Consequences of Bias

Inequitable distribution of the performance of models among patients' subgroups may be the cause of delayed diagnosis, as well as distrust towards the AI-infrastructure of healthcare. The concept of underdiagnosis proposed by (Seyyed-Kalantari et al., 2021) should capture the most obvious clinical impact of biases in algorithms, it is the misdiagnosis of patients who are sick. Deep learning-based classifiers have been shown to misdiagnose and underdiagnose those who belong to historically underrepresented populations: black population, Hispanics, women, poor socio-economically deprived patients. In addition to misdiagnosing patients, the algorithmic bias in clinical tools can exacerbate existing healthcare disparities at a population level. Hence, errors may become more systematic and thus harder to rectify than those of clinicians. Trust is an equally important dimension of this problem. Clinicians working in settings that serve diverse patient populations may reasonably hesitate to rely on diagnostic tools whose performance characteristics are poorly understood for their patient mix.

2.3 Bias Mitigation Techniques

2.3.1 Traditional Approaches to Handling Class Imbalance

Class imbalance has been extensively studied in traditional machine learning, and a broad range of mitigation strategies have been proposed as a result. These are generally organised into three categories: data-level, algorithm-level, and hybrid approaches, depending on whether they modify the training data, the learning algorithm, or both (Krawczyk, 2016; Das et al., 2018).

Data-Level Techniques

Data-level methods seek to redress class imbalance directly in the training set, either by removing majority-class samples, generating additional minority-class examples, or a combination of both.

- *Undersampling techniques:* It reduces the majority class count by selectively or randomly removing instances. Early methods such as Condensed Nearest Neighbour and Tomek Links (Hart, 1968; Tomek, 1976) remove redundant or boundary-ambiguous

samples, while Random Undersampling (Japkowicz, 2000) takes a simpler approach at the risk of discarding useful information. More informed strategies based on clustering, density, or neighbourhood criteria have since been proposed to better preserve informative samples.

- *Oversampling techniques:* Here we instead augment the minority class. The seminal contribution here is SMOTE (Chawla et al., 2002), which generates synthetic samples by interpolating between a minority instance and its nearest minority class neighbours, yielding richer class coverage without duplication. Despite the widespread adoption, SMOTE can produce samples that deviate from the true data distribution and may neglect difficult decision-boundary regions. These limitations have motivated numerous extensions, including deep generative approaches such as VAEs and GANs, though the effectiveness of any oversampling strategy ultimately depends on how faithfully the synthetic samples reflect the underlying distribution.

Algorithm-Level Methods

Algorithm-level approaches address class imbalance without altering the training data and instead modifies the learning process itself. The most widely used strategy is cost-sensitive learning, which assigns higher misclassification penalties to minority-class errors (Palacios et al., 2014), thereby directing the optimiser's attention toward underrepresented classes. These asymmetric costs can be incorporated through loss function modification, output adjustment, or adaptive weighting during training (Kukar et al., 1998), and may be fixed via parameter tuning, derived from class priors, or learned dynamically (Zhang et al., 2018).

Class Imbalance in Deep Learning

Deep neural networks are equally susceptible to class imbalance (Johnson and Khoshgoftaar, 2019; Oksuz et al., 2020). Since the loss is aggregated across all training samples, the majority class dominates the gradient signal, pushing the model toward majority class accuracy while minority class performance quietly degrades; often masked by a high overall accuracy score. The mitigation strategies outlined in the Section 2.3.1 carry over to deep learning, though not always cleanly. Undersampling is particularly ill-suited here, as the deep networks are

data hungry by nature. Oversampling and augmentation are therefore more common, and generative models such as GANs (Goodfellow et al., 2014) and VAEs in particular have extended this further by learning the minority class distribution outright. These approaches are powerful, but prone to mode collapse and distributional distortion when labelled minority data is scarce. On the algorithmic side, cost-sensitive losses transfer more naturally to deep learning, and modern frameworks allow these weights to be adapted dynamically rather than fixed in advance. A notable deep learning specific contribution is focal loss (Lin et al., 2017), which down-weights confidently classified samples so the network focuses on hard and rare examples. Related threads of research include long-tailed recognition, decision-boundary adjustment for minority classes, and ensemble methods designed to improve robustness under severe imbalance.

2.3.2 Bias Mitigation in Medical Imaging and Recent Fairness-Oriented Frameworks

The two main approaches of fairness in medical imaging involve:

- Disentangled representations where sensitive information is separated from the pathology-specific characteristics.
- Alignment-based solutions, where visual representations are grounded based on semantic supervision.

Two recent frameworks from dermatology are the current benchmarks of the two approaches mentioned above.

FairDisCo (Du et al., 2022) aims to tackle bias due to demographics using a contrastive learning technique that is based on disentanglement. This approach learns representations of the disease features independently of those of the sensitive variables (i.e., skin color). It minimizes the statistical relationship between the two representations via correlation regularization. FairDisCo achieves increased fairness by minimizing the demographic information from disease representations, without relying on any sensitive attributes while making predictions.

PatchAlign Aayushman et al. (2024) on the other hand adopts a multimodal skin lesion classification framework where the semantics of the disease labels are matched with the

local image patches through Graph Optimal Transport Chen et al. (2020). Additionally, the introduction of the Masked GOT helps highlight regions that are useful for diagnosis. While it performs well in dermatology benchmarks, it focuses on improving cross-modal alignment and accuracy rather than demographic fairness. In addition, it lacks constraints for the change in visual evidence selection across demographics.

2.4 Multimodal Learning for Skin Lesion Analysis

In real world, dermatologists rarely make diagnostic decision based on lesion morphology alone. In routine clinical practice, multiple factors such as patient age, sex, lesion location, medical history, etc. are interpreted alongside visual assessment which leads to the final conclusion. Models operating only on the dermoscopic images often lack access to these contextual information. Multimodal learning addresses this limitation by integrating heterogeneous data sources, including images, clinical text, and structured metadata, into a unified representation that more closely reflect real world clinical decision making (Yan et al., 2025).

(Pham et al., 2025) suggests that metadata variables such as age, sex, and lesion location provide clinically relevant information that image features alone cannot fully capture. Modern multimodal architectures encodes metadata alongside visual features, enabling interactions between clinical context and image representations through mechanisms such as cross attention (Yan et al., 2024). However, because demographic and clinical variables often correlate with sensitive attributes and healthcare access, multimodal systems must also be evaluated for potential bias amplification and fairness concerns (Adamson and Smith, 2018).

Advancements in vision-language learning have further spurred the development of multimodal medical AI. The effectiveness of joint pre-training of multimodal representations in tasks like classification can be illustrated by examples like CLIP (Radford et al., 2021) and BiomedCLIP (Zhang et al., 2024). Pre-training specific to a certain domain may lead not only to improvement in diagnosis but also to the elimination of proxy demographic biases and thereby increase fairness in medical applications (Seyyed-Kalantari et al., 2021).

Models specific to the field of dermatology such as PanDerm (Yan et al., 2024) show impressive results across several diagnostic tasks thanks to the inclusion of multiple modalities and clinical data. However, multimodal approaches also bring about challenges associated

with increased complexity, decreased interpretability, and reasoning failure between modalities. Addressing these limitations remains an important research direction for the safe and equitable deployment of multimodal dermatological AI.

2.5 Counterfactual Learning and Fairness

Traditional fairness evaluation in machine learning is typically based on aggregate performance metrics measured across different demographic groups. While such metrics can reveal disparities in accuracy, sensitivity, or FPRs, they provide limited insight into whether a model's predictions are actually influenced by sensitive attributes. In medical AI, this distinction is particularly important, as a model may appear fair at the population level while still relying on demographic characteristics when making predictions for individual patients.

Counterfactual reasoning offers a causal framework for investigating this problem. Rather than asking whether performance differs across groups, it asks whether a model's prediction would change if a sensitive attribute were different while all other relevant factors remained unchanged (Pearl, 2009). For example, suppose a skin lesion classifier produces different predictions for the same lesion when only the patient's skin tone is altered; the model here responds to demographic information rather than pathological evidence. By examining such hypothetical scenarios, counterfactual methods move fairness evaluation beyond statistical description and towards causal understanding of model behaviour (Mothilal et al., 2020).

Counterfactual Fairness

Grounded in Pearl's Structural Causal Models (SCMs) and do-calculus (Pearl, 2009), the concept of counterfactual fairness was formally introduced by (Kusner et al., 2017), who defined it as follows:

Definition 2.1. *A predictor $\hat{Y}$ is counterfactually fair with respect to a sensitive attribute A if, for any individual with observed covariates $X = x$ and $A = a$:*

$$P(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a) \tag{2.1}$$

where U represents exogenous background variables and a' denotes a counterfactual value

of the sensitive attribute A.

Intuitively, the prediction should remain invariant when the sensitive attribute is changed, provided all other relevant characteristics of the individual remain unchanged. Which is a stricter condition than statistical fairness metrics 1.3.

More importantly, a model can satisfy demographic parity while still remaining causally biased at the individual level, since over-prediction in one subgroup can be offset by under-prediction in another. In such cases, group-level statistics may appear fair even though individual predictions remain influenced by demographic characteristics. Counterfactual fairness addresses this limitation by anchoring the fairness criterion to causal structure rather than population-level statistics (Kusner et al., 2017).

A practical limitation is that it requires specification of a causal graph encoding assumptions about which pathways from sensitive attributes to predictions are clinically permissible. In dermatology, skin tone has real biological correlations to the concentration of melanin and some presentation of lesions. So not all paths from A to Y are spurious. An essential, non-delegable aspect of model design is domain expertise. This is because clinical judgement, which cannot be provided by the statistical framework, is needed to decide which to block.

Counterfactual Data Augmentation

Much recent studies tends towards counterfactual augmentation of training samples where new training instances are generated with varying sensitive attributes while pathological content is held fixed. This prevents the model from using demographic signal as a predictive shortcut (Kaushik et al., 2019). In medical imaging, particularly in the context of skin tone bias in skin lesion classification, generating such counterfactual examples is far from trivial. Altering only skin tone while preserving the underlying lesion morphology requires high-fidelity generative modeling, a task for which diffusion models have emerged as the most effective approach. Ktena et al. (2024) demonstrated across dermatology, chest radiography, and histopathology that training on such counterfactually balanced datasets reduced demographic sensitivity without sacrificing diagnostic accuracy. The approach is more principled than resampling or reweighting, which adjust sample importance but leave demographic signal embedded in learned representations intact. Counterfactual augmentation breaks this

entanglement by construction. Its limitation is generative fidelity: artifacts or unintended pathological alterations introduced during image synthesis corrupts the training signal, and fairness gains that stem from imperfect generation may not reflect genuine causal decoupling.

Counterfactual Auditing in Dermatological AI

Beyond training, counterfactual methods provide a rigorous mechanism for post-hoc model auditing. By comparing predictions across counterfactually generated variants of the same lesion — identical pathology, systematically varied skin tone — researchers can determine whether diagnostic decisions are driven by disease features or demographic proxies (Pfohl et al., 2019). If a model predicts malignancy for a lesion presented on lighter skin but reverses to benign when the same lesion is projected onto darker skin via a generative model, the causal bias is immediately exposed and individually attributable (Mothilal et al., 2020). This individual-level auditing is qualitatively different from group-level evaluation: it identifies not merely that disparities exist, but which model components and training signals produce them. The combination of SCM-based training constraints and diffusion-based counterfactual auditing represents one of the most technically rigorous approaches currently available for building and certifying fair dermatological classifiers (Pfohl et al., 2019; Ktena et al., 2024). Its demands are correspondingly high: careful causal graph specification, generative models of sufficient quality, and empirical validation that synthesised images faithfully preserve pathological content. Where these conditions are met, counterfactual learning offers a principled path from identifying demographic bias to causally eliminating it.

2.6 Explainable Artificial Intelligence in Medical Imaging

For medical practitioners, a diagnostic recommendation without a supporting rationale is of limited value; it is neither a tool the they can interrogate nor a finding they can take responsibility for. In regulated medical environments, leads to direct consequences for patient safety, medico-legal liability, and regulatory approval (Samek et al., 2017). Explainable AI(XAI) has emerged in response, developing methods that attempt to render complex model decisions legible to practitioners, though the gap between producing a visual explanation and a clinically meaningful one has proven stubbornly persistent (Guidotti et al., 2018).

Explainability Techniques

Some of the most widely adopted post-hoc explanation methods in medical imaging are gradient based. Grad-CAM (Selvaraju et al., 2017) produces a heatmap of influential image regions by weighting convolutional feature maps against the class score gradient, while LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) attribute importance through input perturbation and game theory based marginal contributions respectively. Despite their prevalence, all three explains *where* a model looked rather than *why* looking there produced a particular prediction something which is more critical in clinical decision making.

2.6.1 Explainability in Dermatology

Indeed, dermatology presents a prime example for the application of XAI since diagnosis relies on an auditable language of visual cues such as asymmetry, irregular borders, colour variation, and dermoscopic structures, against which explanations can then be directly evaluated. As was shown by (Tschandl et al., 2020), the specific way in which AI explains itself, beyond predictive accuracy, greatly influences the value that clinicians gain from their support, depending on their level of experience. In addition to techniques based on saliency mapping, prototype analysis techniques have the additional advantage of indicating not just what part of the image the AI model focused on when making its diagnosis, but which previous images were the basis for the decision (Chen et al., 2019). It should also be noted, however, that the trade-off between performance and interpretability continues to present a core challenge in AI dermatology.

2.7 Research Gaps and Motivation for the CARE-Net Framework

The limitations discussed in the previous sections reveal real structural weaknesses in the conceptualization of fairness, multimodal reasoning and explainability in dermatological AI. A synthesis of the literature identifies four inter-connected gaps that motivate the design of CARE-Net.

1. **Absence of counterfactual fairness in multimodal dermatology.** Current state-

of-the-art multimodal dermatology systems attain high-level benchmark results by associatively learning the image and text representations alignment, but lack any way of separating the causes of skin diseases from confounds related to patient demographics. Given that the patient's skin color is associated with the same disease in the training set, the algorithm has no way of telling if that relationship is biologically true or simply a consequence of the way the data was collected. The issue becomes compounded when considering intersectional bias, as removing each type of demographic bias separately still results in systemic undiagnosing of patients that fall into multiple minority groups (Seyyed-Kalantari et al., 2021). Moreover, no existing dermatology framework integrates multimodal processing with counterfactual reasoning mechanisms that explicitly separate disease-relevant morphology from demographic confounding across both image and text streams.

2. **Clinically misaligned multimodal attention.** The integration of clinical text and dermoscopic image modalities leads to an interesting yet dangerous mode of failure, i.e., cross-modal attention patterns can become misaligned with diagnostically relevant visual evidence, allowing models to rely on textual context or spurious image cues despite producing correct predictions (Ben Melech Stan et al., 2024). A model that arrives at the right answer for the wrong reasons is inherently fragile; its reliability collapses precisely in the cases where the demographic shortcut is absent or misleading.

3. **Insufficiency of controlled counterfactual evaluation resources.** Current dermatological datasets and literature on fairness tend to focus mainly on overall differences in the performance of the system, not on any controlled counterfactual resources which allow for the study of the isolated influence of skin color without changing the lesion's morphology. The absence of such counterfactual evaluation materials makes it harder to estimate how sensitive a system is to the skin color confounding factor on an individual-prediction basis.

4. **Lack of clinical significance in explanations.** Post hoc techniques like Grad-CAM and LIME show what parts the model attended to without any assurance that those parts were causal in determining the outcome (Guidotti et al., 2018). Even more importantly, current explainability approaches used in dermatology fail to analyze if

the areas contributing to an outcome would still be relevant if different demographic attributes were considered as counterfactuals. There is little research on explaining models which simultaneously account for fairness and interpretability.

However, all of these represent different aspects of the same underlying challenge - dermoscopic AI models that are trained through correlation, react to medically insignificant cues, cannot be evaluated using reliable resources, and provided post-hoc explanations cannot be trusted from a medical perspective regardless of their overall benchmarked performance. Our CARE-Net model incorporates each of these aspects as part of the same framework that addresses counterfactual attention regularization, multivariate fairness, counterfactual robustness analysis, and clinically meaningful visualization.

Though explainability plays an important role in ensuring the clinical viability of a model, fairness and robustness under demographic distribution change remain the main focus of this work, and not the introduction of new explainability techniques. In that sense, a full causal and clinically relevant explanation approach would represent an important avenue of future work.

Chapter 3

Development of the FitzDerm-CF Dataset

3.1 Introduction

Multimodal dermatology models pair lesion images with clinical text for retrieval, triage, and risk stratification, yet such text can leak diagnoses, amplify malignancy cues, or encode shortcuts linking skin tone to disease prevalence—a critical risk given persistent skin-tone disparities in AI skin-lesion diagnosis, with lower accuracy on darker skin types (Tjiu and Lu, 2025). Fitzpatrick17k annotates 16,577 clinical images with Fitzpatrick types but is skewed toward lighter skin (Groh et al., 2021b), while DDI is a curated, biopsy-supported diverse-skin-tone dataset with documented degradation on darker tones and uncommon diseases (Daneshjou et al., 2021). Yet both resources primarily pair images with structured metadata, not clinical notes, leaving a modality gap for auditing demographic bias and diagnostic leakage in dermatology vision–language models; Table **??** summarizes the motivating metadata and skin-tone imbalance.

Dermatology vision–language models remain vulnerable to skin-tone bias and reduced reliability on underrepresented skin types (Nijjer et al., 2025). Auditing these failures requires paired vision–language data with controlled text informativeness, label disclosure, and protected-attribute cues; we address this by using existing metadata to generate controlled factual and counterfactual dermatology notes.

In this chapter we present a curated multimodal dermatology dataset that extends the

Fitzpatrick17k (Groh et al., 2021b) and DDI (Daneshjou et al., 2021) image databases with clinically-relevant synthetic texts generated using BioMistral-7B (Labrak et al., 2024). Conditioned synthesis takes into account the Fitzpatrick skin type, disease type, and malignancy state simultaneously, with temperature-controlling decoding used to yield descriptions that are true to clinically-meaningful morphology and explicitly avoid disease names, identifying patient information, and diagnostic certainty language.

The creation of the dataset was motivated by the need to analyse the effect of demographic bias and counterfactual robustness in multimodal dermatology models. In particular, this involved the generation of factual clinical notes based on structured metadata for every image, and creating a separate counterfactual dataset where all appearance descriptors depended solely on the FST while retaining all disease related information. This allows for analysis of whether the models rely on the lesion appearance or on any demographic hints present in the clinical notes. Under this design, the shift in model predictions between factual and counterfactual conditions serves as a direct measure of sensitivity to skin-tone language rather than to lesion morphology. The final resource has been used throughout the thesis as the main multimodal dataset for evaluating the CARE-Net framework explained in further Section 4.

3.2 Sources Dermatology Datasets

This work evaluates on two publicly available dermatology benchmarks that together provide complementary coverage of skin-tone diversity, label granularity, and diagnostic noise characteristics.

Fitzpatrick17k Fitzpatrick17k (Groh et al., 2021b) is a large-scale clinical dermatology dataset comprising 16,577 images sourced from two open-access dermatology atlases and annotated with Fitzpatrick skin types (FST I–VI) (Fitzpatrick, 1988a). Following prior work (Aayushman et al., 2024; Du et al., 2022), we adopt a three-class taxonomy — *Benign*, *Malignant*, and *Non-Neoplastic* — and retain the 16,012 usable records after filtering incomplete entries.

As shown in Table 3.2, the dataset exhibits two pronounced structural imbalances that are directly relevant to fairness evaluation. First, the class distribution is heavily skewed:

Table 3.1: Representative factual and counterfactual note pairs from the Fitzpatrick17k-derived portion of FitzDerm-CF. The image and non-skin-tone morphology remain fixed, while Fitzpatrick/skin-tone-dependent descriptors are perturbed.

Image	FST	Label group	Factual note excerpt	Counterfactual note excerpt	Sim.
	I → VI	Non-neoplastic	Inflammatory lesions typically present as **red or brown** patches, plaques, papules, or nodules. They may be scaly, crusty, or ulcerated. The Fitzpatrick skin type **1** may result in a lighter color presentation.	Inflammatory lesions typically present as **dark brown, black, or deeply pigmented** patches, plaques, papules, or nodules. They may be scaly, crusty, or ulcerated. The Fitzpatrick skin type **VI** may result in a darker color presentation.	0.911
	II → VI	Benign dermal	This patient has a benign dermal lesion. Lesions of this type are typically **skin-colored to brown** and have a smooth, well-defined border. They may be flat or slightly elevated.	This patient has a benign dermal lesion. Lesions of this type are typically **dark brown to black** and have a smooth, well-defined border. They may be flat or slightly elevated.	0.922
	VI → I	Genodermatosis	Fitzpatrick skin type **6** indicates a darker skin color. Lesions of this type typically present as flat or slightly elevated, well-defined, and **brown or black** macules or papules.	Fitzpatrick skin type **1** indicates a pale white, fair, ivory skin color. Lesions of this type typically present as flat or slightly elevated, well-defined, and **bright red or pink** macules or papules.	0.908
	VI → IV	Malignant melanoma	This lesion is a malignant melanoma characterized by a rapidly growing, **dark-colored brown-to-black** discoloration. The Fitzpatrick skin type **6** is associated with a darker skin tone.	This lesion is a malignant melanoma characterized by a rapidly growing, **olive-brown to dark-brown** discoloration. The Fitzpatrick skin type **IV** is associated with a darker skin tone.	0.925
	III → IV	Malignant epidermal	This patient has a malignant epidermal lesion, typically presenting as a rapidly growing, **red or brown**, scaly or crusty plaque or tumor with a raised, irregular border.	This patient has a malignant epidermal lesion, typically presenting as a rapidly growing, **dark or olive-colored**, scaly or crusty plaque or tumor with a raised, irregular border.	0.906

Non-Neoplastic conditions dominate with 11,692 samples (≈73% of the total), while Benign and Malignant conditions each contribute only 2,160 samples (≈13.5% each). Within the Malignant class, sample counts decline monotonically from FST I–II (453 and 742, respectively) through to FST V–VI (147 and 61), meaning the rarest and most clinically critical skin-tone groups are simultaneously the most underrepresented in the disease-positive subset. Second, there is a substantial skin-tone imbalance at the dataset level: FST II is the largest group with 4,808 images, whereas FST VI contains only 635, a ratio exceeding 7.5:1. The three lighter skin-tone groups (T1–T3) together account for 11,063 of 16,012 samples (≈69%), while the three darker groups (T4–T6) contribute only 4,949 (≈31%). This compounding of class imbalance and demographic imbalance means a model optimising aggregate accuracy

Table 3.2: Distribution of skin conditions across Fitzpatrick skin types in Fitzpatrick17k.

Condition	T1	T2	T3	T4	T5	T6	Total
Benign	444	671	475	367	159	44	2160
Malignant	453	742	456	301	147	61	2160
Non-neoplastic	2050	3395	2377	2113	1227	530	11692
Total	2947	4808	3308	2781	1533	635	16012

is strongly incentivised to rely on skin-tone as a spurious correlate of disease, rather than attending to true lesion morphology — precisely the failure mode that fairness-aware training seeks to correct.

Diverse Dermatology Images (DDI) The Diverse Dermatology Images (DDI) dataset (Daneshjou et al., 2021) is a smaller, curated benchmark of 656 biopsy-proven clinical images specifically assembled to evaluate model performance across underrepresented skin tones. Unlike Fitzpatrick17k, all labels are histopathologically confirmed, substantially reducing annotation noise and making DDI a higher-fidelity ground truth for binary *Malignant* versus *Non-Malignant* classification.

As reported in Table 3.3, DDI presents a markedly different distributional profile from Fitzpatrick17k. Across the three skin-tone strata — FST I–II (T12), FST III–IV (T34), and FST IV–VI (T45) — group sizes are 208, 241, and 207 respectively, yielding a near-uniform ratio of approximately 1:1.16:1. This relative balance across skin-tone groups makes DDI a considerably more demanding testbed for cross-group generalisation than Fitzpatrick17k, where darker-tone groups are severely underrepresented. At the class level, however, a label imbalance persists: Non-Malignant cases account for 485 of 656 samples (≈74%), with Malignant cases comprising only 171 (≈26%). Notably, the Malignant subset is itself unevenly distributed across strata, with T34 containing 74 malignant cases compared to only 48 and 49 in T45 and T12 respectively, a disparity that can exacerbate per-group recall differences. DDI has been specifically highlighted in prior literature as exposing performance degradation on darker skin tones and uncommon conditions (Daneshjou et al., 2021); its controlled skin-tone balance and biopsy-verified labels therefore make it an essential out-of-distribution complement to Fitzpatrick17k for stress-testing fairness under realistic clinical conditions.

Table 3.3: Distribution of skin conditions across skin-tone groups in DDI.

Condition	T12	T34	T45	Total
Malignant	49	74	48	171
Non-malignant	159	167	159	485
Total	208	241	207	656

3.3 Design Objectives

The dataset is constructed under five active constraints that shape every stage of the generation and validation pipeline.

O1 — Semantic capacity control. The amount of diagnostic information encoded in a note must be controllable via a single generation parameter. A note at low temperature should be semantically rich and morphologically specific; a note at high temperature should be more stochastic and less diagnostically informative. This spectrum enables experiments that isolate the effect of textual semantic content without altering the visual data.

O2 — Disease-agnostic supervision. No generated note may name, imply, or allow a trained linear probe to confidently recover the fine-grained disease category. Disease names, near-synonyms, and hallmark diagnostic phrases are excluded at the prompt level and verified post-hoc via probing experiments.

O3 — Malignancy preservation. While disease identity is excluded, coarse malignancy status is preserved. The generated text must retain a clinically meaningful signal at the level of risk assessment—benign versus malignant morphological patterns—without enabling label-specific leakage.

O4 — Counterfactual consistency. For every factual note, a counterfactual variant must be producible in which only Fitzpatrick-related appearance descriptors are altered. The counterfactual must not introduce or remove statements about disease risk, susceptibility, or prevalence; only visual descriptors that legitimately vary with skin tone may change.

O5 — Leakage minimization. Beyond disease identity, spurious correlations between skin tone and disease prevalence must not be reinforced by the text. Fitzpatrick-aware morphological language is permitted (e.g., visibility of erythema on darker skin tones); causal or epidemiological statements linking skin type to disease likelihood are not.

3.4 Dataset Construction Pipeline

3.4.1 Image Backbone and Metadata

We provide controlled text companions for images in Fitzpatrick17k (Groh et al., 2021b) and DDI (Daneshjou et al., 2021). Each image is associated with three structured categorical metadata fields: ***(i)*** **Fitzpatrick skin type (I–VI)**: the six-point phototype scale encoding appearance (e.g., ivory, olive, deeply pigmented) and sun-reactivity (e.g., always burns, never burns) (Fitzpatrick, 1988a). ***(ii)*** **Disease label**: the fine-grained diagnostic category from the source dataset. ***(iii)*** **Malignancy status**: binary (benign/malignant) or three-partition label. These structured records are passed through the generation pipeline to produce controlled synthetic notes. The images themselves are not modified; We add a controlled text layer on top of existing visual data.

3.4.2 Factual Note Generation

Clinical notes are generated using **BioMistral-7B** (Labrak et al., 2024), a 7B-parameter biomedical language model, loaded in 4-bit NF4 quantization with `bfloat16` compute precision to enable generation on a single consumer GPU. Text is extracted from after the `[/INST]` prompt delimiter to remove prompt reflection from the output.

The prompt conditions the model on the structured metadata with four explicit exclusion constraints: no disease names, no patient-specific attributes (age, sex, lesion size, location, duration), no management recommendations, and no diagnostic certainty language. The disease label is provided solely as a latent conditioning signal that shapes general morphological description; it must not appear in, nor be inferable from, the output.

To construct a spectrum of semantic capacity, each record is processed at five generation temperatures $\tau \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, with up to `MAX_RETRIES`$= 5$ attempts per temperature until a valid note is produced. This yields five text variants per image, stored as `Description_1`–`Description_5` in the output JSON.

A validity check rejects any note that terminates with a truncated section header (e.g., "`clinical impression:`" without following content), which indicates token budget exhaustion before completion.

The complete prompt template used for factual note generation is shown in Listing 3.4.2.

The prompt is designed to elicit purely morphological descriptions while suppressing diagnostic leakage and patient-specific information. By conditioning on structured metadata and explicitly prohibiting disease names, demographic attributes, lesion location, management recommendations, and certainty statements, the generated notes remain clinically descriptive without revealing the underlying class label. Varying the generation temperature further encourages lexical and semantic diversity while preserving factual consistency.

Prompt Template

Factual Note Generation Prompt

```
[INST]
    You are a highly experienced dermatologist specializing in clinical
    documentation.

    Generate a clinical description of a skin lesion based ONLY on the
    diagnostic and phenotypical information provided below.
    Do NOT include any specific details that are not derivable from these
    categories, such as:
    - Patient age, gender, or demographics
    - Exact lesion size, dimensions, or measurements
    - Specific anatomical location
    - Duration or history of the lesion
    - Symptoms (pain, itching, bleeding, etc.)

    Input Categories:
    - Fitzpatrick Scale (Sun-reactive skin type): {fitzpatrick}
    - Nine-Partition Label (Detailed diagnosis category): {nine_partition_label}
    - Three-Partition Label (High-level malignancy status):
    {three_partition_label}

    Required Output:
    1. Clinical Impression: Based on the diagnosis category, describe the
    typical morphological features, color characteristics, and border patterns
```

```
associated with this type of lesion. Reference how the Fitzpatrick skin
type may affect the lesion's appearance.

2. Malignancy Assessment: State the three-partition classification and what
this implies clinically.

3. Differential Considerations: List 2-3 alternative diagnoses that share
similar features with the nine-partition category provided.

Important: Frame your description using general medical terminology (e.g.,
"lesions of this type typically present with..." rather than "this 2.5cm
lesion on the left arm..."). Only describe what can be reasonably inferred
from the categorical labels provided.
[/INST]
```

3.4.3 Counterfactual Note Generation

For an accepted factual note, a counterfactual note is generated by perturbing the Fitzpatrick attribute while holding all other metadata fixed. Formally, let s denote the original structured record and s^{cf} its counterfactual such that: $s^{\text{cf}} = (s \setminus \{a_p\}) \cup \{a'_p\}$, where a_p is the Fitzpatrick skin type and a'_p is its counterfactual value. All other metadata remain unchanged, and the associated image is never modified.

Target skin type selection. A mixed sampling strategy provides gradient diversity across the counterfactual collection: 40% of targets use the diametric opposite ($1 \leftrightarrow 6$, $2 \leftrightarrow 5$, $3 \leftrightarrow 4$), 30% use the nearest neighbor (± 1 on the Fitzpatrick scale), and 30% use a uniformly random alternative. This design ensures that the benchmark contains pairs with a range of skin-tone contrast levels.

Generation prompt. The counterfactual prompt instructs BioMistral-7B(Labrak et al., 2024) to rewrite the original note so that skin-tone-dependent appearance descriptors match a specified target Fitzpatrick profile, while preserving all other clinical content exactly. Each of the six target profiles specifies appearance adjectives and sun-reactivity descriptions (e.g., for Type VI: *dark brown, black, deeply pigmented; never burns, deeply pigmented, least sun-*

sensitive). The model is explicitly prohibited from altering disease risk statements, susceptibility language, or causal attributions.

The counterfactual rewriting prompt is shown in Listing 3.4.3. Unlike factual note generation, the objective is not to create new clinical content but to minimally edit an existing note so that appearance descriptors align with a target Fitzpatrick profile. The prompt explicitly constrains the model to preserve all non-skin-tone-related information, thereby isolating skin-tone appearance as the sole intervention variable. This controlled rewriting process produces counterfactual text pairs suitable for fairness evaluation and causal analysis.

Prompt Template

Factual Note Generation Prompt

```
[INST] You are an expert dermatologist rewriting patient notes for a medical
    dataset augmentation task.

    Goal: Rewrite the Original Text to change the patient's skin type to
    Fitzpatrick Type {target['id']}.

    Target Profile (Type {target['id']}):
    - Appearance: {target['adjectives']}
    - Sun Reaction: {target['reaction']}

    Instructions:
    1. Identify sentences describing skin tone, color, or sun sensitivity.
    2. Rewrite those sentences to match the Target Profile.
       - You MUST change descriptive adjectives (e.g., "fair" -> "dark") and
    visual symptoms (e.g., "bright redness" -> "subtle darkness") to be
    medically consistent with the new skin tone.
    3. If NO skin tone is mentioned: You must INSERT a brief description of the
    patient's skin (e.g., "Patient has {target['adjectives']} skin") at the
    beginning of the note.
    4. Preserve all other medical facts and morphology (Lesion size, shape,
    diagnosis, location) exactly as they are.
    5. Output the full, revised medical note. Do not include introductory text.
```

```
IMPORTANT CONSTRAINTS (MUST FOLLOW):
- Do NOT change disease risk, susceptibility, prevalence, or causal
explanations.
- Do NOT change statements about whether a patient is more or less prone to
a condition.
- Skin tone changes must affect ONLY visual appearance, not underlying
biology or risk.

Original Text:
"{ORIGINAL_NOTE}"

Rewritten Counterfactual:
[/INST]
```

3.4.4 Four-Stage Validation Pipeline

Each candidate counterfactual is accepted only if it passes all four validation stages sequentially. Up to `MAX_RETRIES`= 5 attempts are made before a record is marked as failed. **Stage 1 — No-op rejection.** Candidates identical to the original note are immediately rejected. This filters cases where the model ignores the perturbation instruction entirely. **Stage 2 — Semantic similarity bounds.** Cosine similarity σ between the original and counterfactual note is computed using `sentence-transformers/all-MiniLM-L6-v2` (Wang et al., 2020). Candidates are accepted only if $\sigma \in (0.85, 0.93)$. The lower bound ensures the counterfactual is meaningfully different; the upper bound ensures that global clinical content is not substantially rewritten. **Stage 3 — Causal invariance check.** A keyword list covering causal and epidemiological phrases—*more prone*, *higher risk*, *susceptibility*, *predisposed*, *protective*, *risk of*, and four related terms—is checked for asymmetric presence between original and counterfactual. Any asymmetry constitutes a causal invariance violation and triggers rejection. **Stage 4 — Skin-tone intervention check.** The counterfactual must contain at least one lexical change involving a skin-tone indicator term (e.g., *fair*, *dark*, *erythema*, *olive*, *pigmented*), verified for asymmetric presence. A candidate that passes all prior stages

but makes no skin-tone-related lexical change is rejected as a failed intervention. Records for which no valid counterfactual is found within the retry budget are stored with `cf_status: "failed"` and are excluded from the counterfactual benchmark split. Accepted records carry `cf_status: "valid"` and a recorded similarity score.

3.4.5 Data and Code Availability

FitzDerm-CF is publicly archived on Harvard Dataverse and includes generated factual notes, valid counterfactual JSON records, and metadata. Data loaders, generation code, and benchmark evaluation scripts are available at GitHub.

3.5 Dataset Analysis

3.5.1 Temperature and Semantic Capacity Analysis

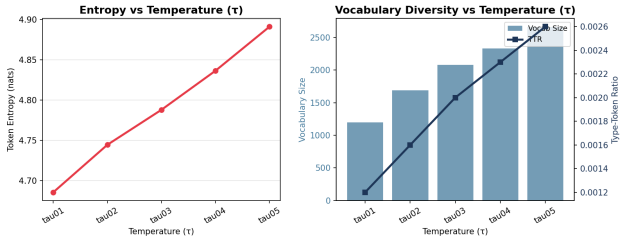


Table 3.4: Dataset statistics across temperature settings.

(τ)	Entropy	Vocab Size	TTR
0.1	4.6858	1198	0.0012
0.2	4.7447	1692	0.0016
0.3	4.7881	2077	0.0020
0.4	4.8366	2329	0.0023
0.5	4.8913	2658	0.0026

Table 3.5: Effect of temperature on lexical diversity. Increasing temperature results in higher token entropy, larger vocabulary size, and increased type-token ratio, indicating richer and more diverse generated dermatology descriptions.

To characterize the effect of temperature on the generated dermatology descriptions, we measured token entropy and vocabulary size, and type-token ratio (TTR) across all temperature settings. As shown in Table 3.4 and Figure 3.5, increasing the temperature from $\tau = 0.1$ to $\tau = 0.5$ resulted in a consistent increase in token entropy (4.69 to 4.89), vocabulary size (1,198 to 2,658 unique tokens), and TTR (0.0012 to 0.0026). These trends indicate that higher temperatures produce more lexically diverse and less repetitive descriptions. Notably,

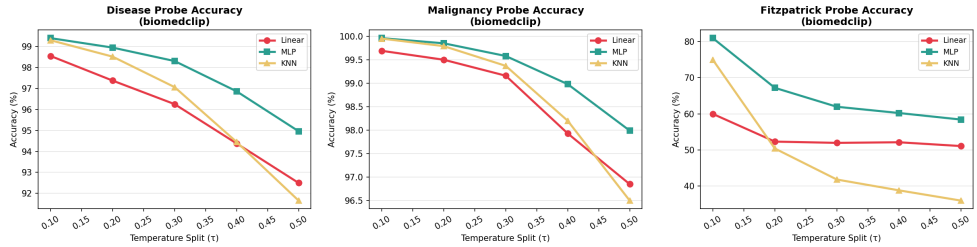


Figure 3.1: Probe accuracy across temperature settings for BioMedCLIP embeddings. Disease and malignancy prediction remain highly recoverable across temperatures, while Fitzpatrick skin-type prediction decreases substantially as temperature increases. Results are shown for Linear, MLP, and KNN probes.

Table 3.6: MLP probe accuracy (%) across temperature settings. Disease and malignancy remain highly recoverable, while Fitzpatrick prediction declines with increasing temperature.

	BioMedCLIP			MiniLM		
τ	Disease	Malig.	Fitz.	Disease	Malig.	Fitz.
0.1	99.3	99.9	80.9	99.3	99.9	84.7
0.2	98.9	99.8	67.0	99.1	99.9	77.8
0.3	98.3	99.6	61.8	98.8	99.8	77.3
0.4	96.9	99.0	60.2	97.5	99.3	79.0
0.5	94.9	98.0	58.4	96.4	98.7	80.7

the average description length remained largely unchanged across temperatures, suggesting that the increase in diversity arises from richer word choice rather than longer outputs. This analysis confirms that the temperature sweep successfully modulates the semantic capacity of the generated text while preserving comparable description lengths.

3.5.2 Leakage Analysis

3.5.2.1 Experiment 1: Probing Semantic Leakage in Generated Clinical Text

We quantify diagnostic and demographic leakage by probing frozen text embeddings, treating probe performance as attribute recoverability under different generation temperatures. For each temperature, we extract MiniLM (Wang et al., 2020) and BioMedCLIP (Zhang et al., 2024) embeddings and train three frozen-representation probes with 5-fold stratified cross-validation: logistic regression for linear separability, MLP for non-linear boundaries, and KNN for local neighborhood structure. We evaluate disease classification, malignancy prediction, and Fitzpatrick skin-type classification, where disease/malignancy are clinically

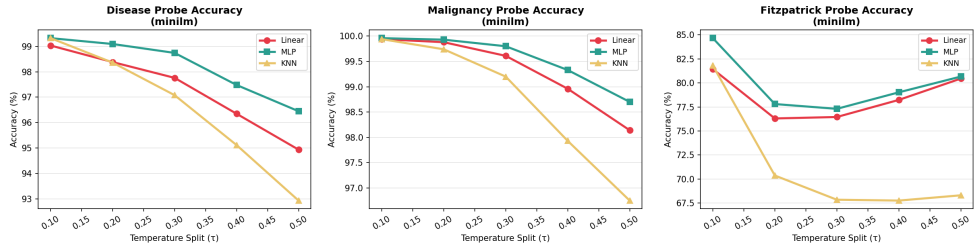


Figure 3.2: Retrieval leakage(Purity@10) across temperature settings. Disease and malignancy remain highly clustered, while Fitzpatrick clustering decreases at higher temperatures.

relevant attributes and Fitzpatrick type is a demographic attribute. We report accuracy for all tasks, Macro-F_1 for imbalanced disease and Fitzpatrick classification (Sokolova and Lapalme, 2009), and ROC-AUC for threshold-independent malignancy prediction.

Leakage is assessed by how recoverability changes with temperature: high probe performance indicates encoded attribute information, while a favorable privacy–utility trade-off preserves clinical recoverability and reduces demographic recoverability. As Table 3.6 shows, disease and malignancy remain highly recoverable across temperatures for both encoders: BioMedCLIP disease accuracy decreases modestly from 99.3% to 94.9%, with malignancy above 98%, and MiniLM keeps disease and malignancy accuracies above 96% and 98%, respectively. In contrast, BioMedCLIP Fitzpatrick prediction drops from 80.9% to 58.4% as τ increases from 0.1 to 0.5, indicating that higher-temperature generation reduces demographic recoverability while preserving clinically relevant information.

3.5.2.2 Experiment 2: Retrieval-Based Analysis of Semantic Leakage

Supervised probes test whether attributes are decodable, whereas retrieval analysis tests whether generated text embeddings naturally cluster by clinical or demographic attributes as temperature increases. For each temperature, we L2-normalize embeddings, compute the exhaustive pairwise cosine-similarity matrix, and measure embedding anisotropy via average off-diagonal cosine similarity to rule out representation collapse. After excluding self-similarity, each sample retrieves nearest neighbors from the remaining dataset, and retrieval is evaluated for disease, malignancy, and Fitzpatrick labels over $k \in \{1, 5, 10\}$.

We report Purity@k, Excess Purity@k, Hit@k, and NDCG@k, using Purity@10 as the primary retrieval-leakage measure because it captures attribute clustering in moderately sized

Table 3.7: Multimodal malignancy classification performance (%) across temperature settings.

τ	Balanced Acc	F_1
0.1	100.0	100.0
0.2	99.9	99.9
0.3	99.5	99.5
0.4	98.0	97.7
0.5	96.3	95.3

neighborhoods(Palacio-Niño and Berzal, 2019). Figure 3.2 shows that disease and malignancy remain strongly clustered as temperature increases, with Purity@10 decreasing only from 98.8% to 87.8% and from 99.9% to 94.5%, respectively. Fitzpatrick purity drops more sharply from 67.8% to 29.1%, indicating that temperature-induced diversity weakens demographic organization more than clinical organization; however, Fitzpatrick purity remains above random baseline, showing skin-type information is reduced but not eliminated. These findings align with the probe results and indicate that higher-temperature generation preserves clinically meaningful structure while attenuating demographic clustering.

Benchmark Task: Multimodal Classification

In addition to the leakage analyses, a simple multimodal classification experiment confirmed that the generated descriptions retained substantial diagnostic utility across all temperature settings. Classification performance remained above 96% Balanced Accuracy even at the highest generation temperatures, as shown in Table 3.7, indicating that reduced demographic recoverability did not substantially compromise clinically relevant information.

3.6 Limitations

Although FitzDerm-CF constitutes a valuable effort in developing a framework for controlled multimodal assessment of skin tone fairness in dermatological AI, there are various shortcomings to note that limit the generalizability of findings based on the present dataset and point to future avenues of research.

Clinical Validity of Generated Texts All clinical notes in FitzDerm-CF are synthesized using the BioMistral-7B (Zhang et al., 2024) model and have not been reviewed by any experts. Although the validation process ensures the semantic plausibility, causal invariance, and lexical intervention, these are algorithmic criteria that cannot be taken for clinical reality. Although qualitative examination indicated that the notes appeared to be logical and medically reasonable, no dermatologist conducted systematic review was performed. Therefore, the medical credibility of all descriptions produced is not completely assured.

Coverage of Protected Attributes The construction of counterfactual examples was restricted to Fitzpatrick Skin Type (FST), the primary demographic attribute considered in this study. However, in actual clinical deployments, demographic fairness involves a much wider range of protected attributes like patient age cohort, biological sex, and patient ethnicity/origin, each one of which has the potential to be a source of spurious correlation between vision-language pairs. This limitation means that the present benchmark only partially addresses the issue of demographic confounding that impacts dermatology AI models. An extension of the counterfactual approach to include multiple protected attributes in one test while maintaining clinical relevance and imposing similar causal invariance constraints would expand the applicability and ecological validity of this benchmark significantly.

Dependency on One Generation Backbone The FitzDerm-CF benchmark is entirely dependent on BioMistral-7B as the text generation model. In other words, all the leakage features, lexical characteristics, and morphological description patterns found in the benchmark data are a product of the particular model's inductive bias, data used for training, and generative process. Notes generated with any other biomedical language model, or any instruction-finetuned version, might have a fundamentally different leakage pattern and linguistic characteristics that might influence the probe performance. Comparing leakage features of several biomedical language models in their prompts would improve the generalizability of the benchmark findings.

Chapter 4

CARE-Net: Counterfactual Attention Regularized Fair and Text-Free Dermatology Diagnosis

Summary

Deep learning models for dermatology diagnosis exhibit significant performance disparities across skin tones. Existing fairness approaches typically operate on global representations or final predictions, leaving the model's internal visual reasoning unconstrained. In this work, we introduce ***CARE-Net****, a multimodal framework that mitigates demographic bias by encouraging* counterfactual invariance at the attention level. *CARE-Net introduces* Counterfactual Attention Consistency *(AC-CQC), utilizing an asymmetric stop-gradient loss to regularize cross-attention maps between factual and LLM-generated counterfactual clinical queries. This ensures that lesion-focused attention remains invariant to demographic perturbations. To enable privacy-preserving, text-free deployment, we propose a* Global Learnable Inference Query *(GLIQ). Distilled from the fairness-aware multimodal teacher, GLIQ allows the model to retain counterfactually invariant attention during inference without requiring clinical text or sensitive attributes. Additionally, we present PL-CQC, a relaxed variant enforcing consistency at the patch-logit level via confidence-aware gating. Experiments on Fitzpatrick17k and biopsy-proven DDI demonstrate CARE-Net maintains or improves accuracy while substantially advancing demographic parity and equality of opportunity.*

4.1 Introduction

While deep learning models rival dermatologists in diagnosing lethal skin cancers (Gessert et al., 2019; El-Khatib et al., 2020), their clinical deployment is stalled by profound demographic performance disparities (Groh et al., 2021a; Li et al., 2021; Daneshjou et al., 2021). Trained on structurally imbalanced datasets predominantly featuring light skin (Paxton et al., 2025), these models exhibit severe skin-tone bias, systematically underperforming on darker skin types. To optimize aggregate accuracy, they learn demographic shortcuts leveraging skin color as a spurious correlate rather than focusing on true lesion morphology (Yang et al., 2024). Because algorithms can infer protected attributes like the Fitzpatrick skin type from images even when imperceptible to clinicians (Gichoya et al., 2022), there is an urgent need for models that achieve demographic fairness by strictly prioritizing disease-specific visual evidence across all skin tones. Prior efforts to mitigate skin lesion classification biases include data-centric worst-group risk optimization (Group-DRO (Sagawa et al., 2019)) and representation-level contrastive learning (FairDisCo (Du et al., 2022)). However, because methods operating on global embeddings lack explicit constraints on internal spatial reasoning, diagnostic attention consistency across varying skin tones remains uncertain. Concurrently, Vision-Language Models (VLMs) restore semantic grounding via clinical reports (GLoRIA (Huang et al., 2021), MedCLIP (Wang et al., 2022)) or local patch alignment (PatchAlign (Aayushman et al., 2024)). Yet, these fail to enforce counterfactual invariance regarding the selected visual evidence and typically require unavailable test-time textual inputs, limiting clinical applicability. Although counterfactual frameworks assess demographic perturbations (Vlontzos et al., 2022) via post-hoc explanations (ACAT (Fontanella et al., 2023)), and recent architectures explore learned queries for text-free inference (Xu et al., 2024; Ye et al., 2025), uniting attention-level counterfactual invariance with text-free deployment remains a critical open challenge. To address this gap, we propose CARE-Net, a multimodal framework mitigating skin-tone disparities by regularizing internal feature selection. Because standard benchmarks (Fitzpatrick17k, DDI) lack text, we render categorical metadata into clinical notes to generate precise factual and counterfactual pairs differing solely by the protected Fitzpatrick skin-type. Hypothesizing that fair models must attend to identical morphological evidence regardless of skin tone, we explicitly enforce visual reason-

ing invariance under these demographic perturbations. ***The main contributions of this chapter are:***

- ***(i)*** **CARE-Net**, a fairness framework utilizing Counterfactual Attention Consistency (AC-CQC) and a relaxed patch-logit variant (PL-CQC) to enforce demographic invariance across multiple internal reasoning levels.
- ***(ii)*** **GLIQ** (Global Learnable Inference Query), an architectural primitive distilling this fairness-aligned multimodal attention into a static query for strictly privacy-preserving, text-free deployment.
- ***(iii)*** Demonstration that CARE-Net achieves competitive diagnostic accuracy while substantially improving fairness under in-domain and out-of-domain skin-tone shifts on Fitzpatrick17k and DDI.

4.2 Methodology

4.2.1 Problem Formulation and Notation

Problem Formulation : Let $x \in \mathcal{X} \subset \mathbb{R}^{H\times W\times C}$ denote an input dermatology image, $y \in \mathcal{Y} = \{1, \ldots, D\}$ the disease label, and $s \in \mathcal{S}$ the skin tone group (e.g., Fitzpatrick scale (Fitzpatrick, 1988a)).

Our primary goal is to learn a model that maximizes the predictive likelihood $P(y|x)$. To ensure that the diagnosis is robust and not biased by the skin tone attribute, we impose a constraint that the prediction should be conditionally independent of s. Formally, we aim to maximize $P(y|x)$ subject to the condition:

$$P(y|x, s) = P(y|x). \tag{4.1}$$

This implies that explicitly knowing the skin tone s provides no additional information for predicting y once x is known, thereby mitigating reliance on spurious correlations associated with s.

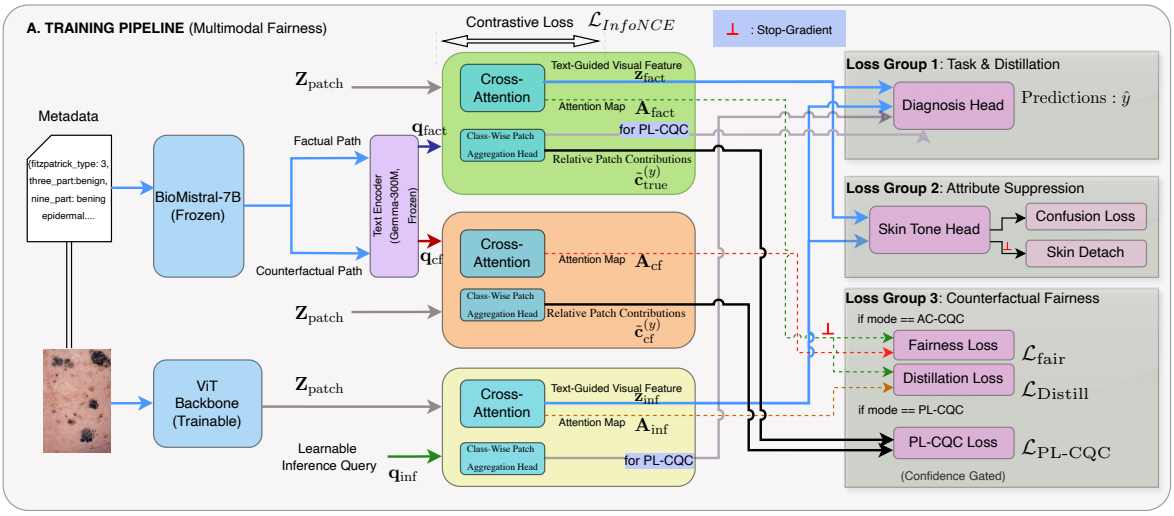


Figure 4.1: CARE-Net framework (see Sec. 4.3 for pipeline breakdown).

Inference Constraints. It is important to note that while auxiliary information such as ground-truth skin tone annotations or LLM-generated textual descriptions may be leveraged during training, neither the sensitive attribute s nor any textual inputs are available at inference. Consequently, the deployed model must rely exclusively on the image x to predict the disease label y.

4.2.2 Patch-Level Attention Backbone

We adopt a dual-encoder vision-language architecture, conceptually extending the PatchAlign (Aayushman et al., 2024) to dense semantic alignment. The visual stream employs a Vision Transformer(ViT) backbone, E_v, which processes the input image $x \in \mathbb{R}^{H\times W\times C}$ into a sequence of patch embeddings $\mathbf{V} = \{v_1, v_2, ..., n_N\} \in \mathbb{R}^{N\times D}$, where N represents the number of patches and D is the feature dimension. Unlike standard classifiers that pool these patches into a single token, we retain the spatial patch sequence to preserve localized morphological features. The textual stream utilizes a pre-trained large language model (*Google-300m*), E_t, to encode the clinical description $\mathbf{T}$ into a sequence of token embeddings $\mathbf{Q} = \{q_1, q_2, ..., q_M\} \in \mathbb{R}^{M\times D}$. The core interaction point is a scaled dot-product cross-attention mechanism, where the text embeddings act as queries ($\mathbf{Q}$ to attend to the visual patch keys ($\mathbf{K} = W_k\mathbf{V}$ and values ($\mathbf{V} = W_v\mathbf{V}$. This mechanism generates a cross-modal attention map $\mathcal{A} \in \mathbb{R}^{H_{heads}\times M\times N}$, representing the alignment probability between each text token and image patch, which is

then subsequently aggregated fot disease classification.

4.2.3 AC-CQC: Counterfactual Attention Consistency

To ensure that the model's diagnostic reasoning is invariant to demographic attributes, we introduce the Counterfactual Attention Consistency mechanism. Existing methods like ACAT (Fontanella et al., 2023) rely on generative adversarial perturbations to create visual counterfactuals, a process prone to hallucinating artifacts in medical imaging. Conversely, Rao et.al. (Nan et al., 2021) utilize causal intervention to maximize attention sensitivity for fine-grained classification. We invert this paradigm: instead of maximizing sensitivity, we enforce *causal invariance* against sensitive attributes by leveraging the semantic disentanglement capabilities of Large Language Models. We posit that a fair model must effectively attend to the same morphological evidence regardless of the patient's skin tone.

The counterfactual supervision approach is instantiated via the factual and counterfactual clinical notes described in Section 3.4. Specifically, for every skin lesion image in our dataset, there is a factual note T_{fact} and a counterfactual note T_{cf}, which has been created through a perturbation of only Fitzpatrick skin type while keeping all clinical details relevant for disease intact as discussed in Section 3.3. Since both notes have been produced through a consistent controlled generation pipeline, they maintain similarity on linguistic level as well as clinical context, differing only in terms of the protected demographic feature. The constrained nature of their generation process makes sure that any observed discrepancies in behavior can be attributed to skin-tone manipulation alone.

For a pair (T_{fact}, T_{cf}) representing one image, the two descriptions are encoded to form alternative query representations over visual features. Hence, a different attention mechanism is triggered by every query, which allows a measure of visual reasoning stability to be captured. For each training image, these queries induce two attention distributions over the same visual features:

$$\mathbf{A}_{\text{fact}} = \text{Softmax}\left(\frac{\hat{Q}_{\text{fact}}\hat{K}^{\top}}{\sqrt{d}\,\tau}\right), \quad \mathbf{A}_{\text{cf}} = \text{Softmax}\left(\frac{\hat{Q}_{\text{cf}}\hat{K}^{\top}}{\sqrt{d}\,\tau}\right), \tag{4.2}$$

where $Q_{\text{fact}} = E_t(T_{\text{fact}})$, $Q_{\text{cf}} = E_t(T_{\text{cf}})$, $\hat{Q}$ and$\hat{K}$ denote ℓ_2-normalized queries and visual patch embeddings respectively, d is the embedding dimension, and τ is a learnable temperature parameter.

We treat the factual attention $\mathcal{A}_{true}$ as the internal control signal, representing the ground truth visual search anchored in reality. To enforce process-level invariance using an asymmetric L_1 consistency loss with a stop-gradient (sg) operator, inspired by SimSiam (Chen and He, 2021):

$$\mathcal{L}_{fairness} = \mathbb{E}_{(\mathbf{I},T_{fact},T_{cf})\sim\mathcal{D}} \left[\|\text{sg}(\mathcal{A}_{true}) - \mathcal{A}_{cf}\|_1 \right]. \tag{4.3}$$

The stop-gradient operator prevents degenerate mode collapse by fixing the factual attention as the target, forcing the counterfactual branch to align its reasoning process. We deliberately select the L_1 norm over the Frobenius norm to induce sparsity: this penalizes diffuse attention *smearing*, forcing the model to maintain sharp focus on lesion morphology even when primed with contradictory demographic information.

Additionally, we employ an InfoNCE style contrastive loss $\mathcal{L}_{InfoNCE}$ to stabilize cross-modal grounding between retrieved visual features and textual queries. This loss serves as a regularizer and does not contribute directly to fairness enforcement. To further suppress sensitive-attribute information at the representation level, we adopt a confusion-based disentanglement objective without adversarial gradient reversal. Following prior work (Du et al., 2022), we apply a confusion loss $\mathcal{L}_{conf}$ on a skin-type prediction head that maximizes the entropy of sensitive-attribute predictions, encouraging uniform outputs and discouraging the encoding of skin-type information in the learned representation. To avoid degenerate solutions where the classifier collapses without removing sensitive information, we additionally employ a standard cross-entropy loss $\mathcal{L}_s$ on a detached skin-type classifier, which optimizes only the classifier parameters and serves as a diagnostic measure of residual demographic leakage.

4.2.4 Global Learnable Inference Query for Fairness-Aware Inference

A critical limitation of existing Multi-Modal Models in dermatology is the inference modality gap. While textual clinical reports provide essential guidance for fair feature extraction during training, access to these texts in usually restricted or not readily available during deployment. Current methods either hallucinate text(cite- SGSeg (Ye et al., 2024), LERG), which is expensive and risky, or use constrastive alignment(MedCLIP (Wang et al., 2022)), which often fails to capture fine-grained counterfactual nuances because it operates on global

embeddings. Drawing inspiration from the object query mechanics, in the semantic slots in Slot-VLM (Xu et al., 2024), introduce the GLIQ, a lightweight architectural intervention designed to enable text-free, fairness-aligned inference. We define the inference query as a static, learnable parameter $\mathbf{q}_{inf} \in \mathbb{R}^D$, initialized as a random normal vector. Unlike dynamic text queries that vary per sample, $\mathbf{q}_{inf}$ acts as a universal diagnostic probe.

During the multimodal traninig phase, we employ a *Teacher-Student Attention Distillation* strategy. The Teacher is the text-guided attention map $\mathbf{A}_{text}$, generated by querying the visual features with the true clinical report embeddings. The Student is the attention map $\mathbf{A}_{inf}$, generated by the static inference query:

$$A_{\text{inf}} = \text{Softmax}\left(\frac{\hat{q}_{\text{inf}}\hat{K}^\top}{\sqrt{d}\,\tau}\right) \tag{4.4}$$

where $\hat{\mathbf{q}}_{\text{inf}}$ and $\hat{\mathbf{K}}$ denote ℓ_2-normalized inference queries and visual patch embeddings respectively, d is the embedding dimension, and τ is a learnable temperature parameter that regulates the entropy of the attention distribution, preventing the static query from collapsing into a uniform average.

Standard attention uses a fixed scaling factor. However, when distilling a dynamic query into a static query, there is a risk of Entropy Mismatch. By making τ a learnable parameter constrined to a low range(0.03-0.2), the model is forced to make sharp, low-entropy selections over visual patches (Oorloff et al., 2025). The optimization objective is an L_1 distillation loss that forces the inference query to mimic the text-guided attention distribution:

$$\mathcal{L}_{distill} = \|sg(\mathbf{A}_{text}) - \mathbf{A}_{inf}\|_1 \tag{4.5}$$

Crucially, we detach the visual features $\mathbf{V}$ when computing $\mathbf{A}_{inf}$. This ensures that gradients from $\mathcal{L}_{distill}$ update only the query $\mathbf{q}_{inf}$, forcing it to adapt to the fixed visual backbone rather than distorting the backbone to suit a weak query.

Crucially, causal fairness is enforced during multimodal training via AC-CQC, which constrains the text-guided attention to be invariant under counterfactual perturbations of sensitive attributes. GLIQ does't introduce new fairness constraints, instead, it distills this counterfactually invariant attention policy into a static inference query, enabling fairness-

preserving inference without access to text or sensitive attribute at test time. By enforcing, $\mathbf{A}_{inf} \approx \mathbf{A}_{text}$ via distillation, we transitively enforce $\mathbf{A}_{inf} \approx \mathbf{A}_{cf}$. Consequently, the learned query $\mathbf{q}_{inf}$ effectively *caches* the fairness-aware attention patterns of the teacher. Importantly, the query encodes a task-level searcg strategy *what visual cues to prioritize*, rather than sample-specific spatial priors, allowing the resulting attention to remain image-adaptive despite the query being static. This allows the model to deploy an un-biased attention mechanism that selectively highlights disease morphology while ignoring skin-tone artifacts, achieving the fairness utility of multimodal guidance with the efficiency of unimodal inference.

Final training objective, where classification, fairness enforcement, and inference-query distillation are optimized jointly(we keep $L_{distill} = 0.1$. L_{fair} is tuned linearly. alpha=1.0 for InfoNCE(goes in Technical Specification)).:

$$\mathcal{L}_{total} = \mathcal{L}_{cls}^{text} + \lambda_{inf}\mathcal{L}_{cls}^{inf} + \alpha\mathcal{L}_{conf} + \mathcal{L}_{skin}^{detach} + \mathcal{L}_{AC\text{-}CQC} \tag{4.6}$$

where,

$$\mathcal{L}_{AC\text{-}CQC} = \mathcal{L}_{InfoNCE} + \lambda_{fair}\mathcal{L}_{fair} + \lambda_{distill}\mathcal{L}_{distill} \tag{4.7}$$

4.2.5 Patch-Logit Counterfactual Consistency (PL-CQC)

While the Counterfactual Attention Consistency defined in Section4.2.3 enforces invariance in the model's visual focus, it relies on the assumption that attention equates to reasoning. A model can have a correct attention map, but still classify based on *spurious* features encoded in the latent vector that are spatially coincident with the lesion. To bridge this gap, we introduce **Logit-Grounded CQC**, a risk-aware objective that enforces consistency directly on the decision forming evidence. Drawing inspiration from interpretable-by-design architectures like BagNet (Brendel and Bethge, 2019) and MedicalPatchNet (Wienholt et al., 2025), PL-CQC decomposes the final disease-specific classification logit into an additive sum of patch-level contributions. Unlike AC-CQC, PL-CQC does not enforce gradient detachment between inference queries and visual representations. The design choice of joint optimization is necessary for aligning patch-level evidence with query-conditioned importance weights, enabling effective logit-level decomposition. Let $\mathbf{v}_p \in \mathbb{R}^D$ be the visual embedding for patch

p. We define the disease specific evidence $\ell_p^{(d)}$ and the query-conditioned importance $\alpha_p^{(d)}$ as:

$$\ell_p^{(d)} = f_{\text{patch}}(\mathbf{v}_p)^{(d)}, \quad \alpha_p^{(d)} = \text{softmax}_p\Big(f_{\text{imp}}([\mathbf{v}_p; \mathbf{q}])^{(d)}\Big) \tag{4.8}$$

where f_{cls} is the shared classifier head and $\mathbf{q}$ is the text query (factual or counterfactual). The final logit for disease d is the weighted aggregation of local evidence: $z_d = \sum_{p=1}^{P} \alpha_p^{(d)} \cdot \ell_p^{(d)}$. This formulation explicitly attributes the diagnosis to specific visual regions. To enforce fairness, we encourage consistency of the attributed evidence for the ground-truth disease y under demographic perturbation. We define the normalized contribution map $\tilde{c}_p^{(y)} = (\alpha_p^{(y)} \ell_p^{(y)})/Z$, where $Z = \sum_{p'=1}^{P} |\alpha_{p'}^{(y)} \ell_{p'}^{(y)}| + \epsilon$ is a normalization constant. We compute the divergence between the contributions derived from the factual query (T_{fact}) and the counterfactual query (T_{cf}):

$$\mathcal{D}_{attr} = \sum_{p=1}^{P} \left\| \tilde{c}_p^{(y)}(T_{fact}) - \tilde{c}_p^{(y)}(T_{cf}) \right\|_1 \tag{4.9}$$

Crucially, enforcing consistency on ambiguous or low-confidence samples can lead to optimization instability. We therefore introduce a risk-aware gating mechanism. The consistency penalty is modulated by the model's confidence in the factual prediction, ensuring fairness is emphasized primarily when a decisive diagnostic opinion is formed. Concretely, we define a soft confidence gate:

$$g_i = \max\left(0, \ \frac{\sigma(z_{y,i}) - \tau}{1 - \tau}\right), \tag{4.10}$$

which smoothly increases the strength of the regularization as the model's confidence exceeds the threshold τ. The logit grounded loss is then computed as a gated expectation over the batch:

$$\mathcal{L}_{\text{PL-CQC}} = \frac{\sum_{i=1}^{B} g_i \, \mathcal{D}_{attr,i}}{\sum_{i=1}^{B} g_i + \epsilon}. \tag{4.11}$$

where $\sigma(\cdot)$ denotes the sigmoid function and τ is a confidence threshold. This formulation intentionally relaxes the strict inference-time invariance constraint of AC-CQC by allowing joint optimization at the logit level, while employing confidence-gated regularization to emphasize fairness on high-confidence predictions and suppress noisy gradients from uncertain

cases.

The overall training objective combines supervised diagnosis, skin-tone prediction, and the proposed PL-CQC regularization. Specifically, the model is trained using a weighted sum of classification losses and the gated fairness objective, ensuring predictive accuracy while promoting decision-level consistency under counterfactual perturbations.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}}^{\text{text}} + \lambda_{\text{inf}}\mathcal{L}_{\text{cls}}^{\text{inf}} + \alpha\mathcal{L}_{\text{skin}} + \mathcal{L}_{\text{skin}}^{\text{detach}} + \lambda_{\text{fair}}\mathcal{L}_{\text{PL-CQC}}, \tag{4.12}$$

4.3 Putting It All Together: The CARE-Net Pipeline

To summarize, the CARE-Net framework (Fig. 4.1) unifies multiscale fairness enforcement. During **Multimodal Fairness Training**, frozen BioMistral-7B (Labrak et al., 2024) and Gemma-300M (Schechter Vera et al., 2025) models transform structured metadata into semantic embeddings representing factual ($\mathbf{q}_{\text{fact}}$) and counterfactual ($\mathbf{q}_{\text{cf}}$) clinical notes. Concurrently, a trainable ViT extracts spatial image patches ($\mathbf{Z}_{patch}$), feeding three parallel cross-attention streams: factual, counterfactual, and GLIQ distillation ($\mathbf{q}_{\text{inf}}$). Optimization relies on three loss groups: **(1) Task & Distillation:** Supervise classification ($\mathcal{L}_{\text{cls}}^{\text{text}}, \mathcal{L}_{\text{cls}}^{\text{inf}}$) and distill fairness-aligned factual attention ($\mathbf{A}_{\text{fact}}$) into the static inference query ($\mathbf{A}_{\text{inf}}$) via $\mathcal{L}_{\text{distill}}$. **(2) Attribute Suppression:** Erase visual skin-tone leakage via confusion and detached cross-entropy losses ($\mathcal{L}_{\text{conf}}, \mathcal{L}_{\text{skin}}^{\text{detach}}$). **(3) Counterfactual Fairness:** Enforce demographic invariance via **AC-CQC** (penalizing attention divergence between $\mathbf{A}_{\text{fact}}$ and $\mathbf{A}_{\text{cf}}$) or **PL-CQC** (penalizing gated patch-logit divergence between $\mathbf{c}_{\text{true}}^{(y)}$ and $\mathbf{c}_{\text{cf}}^{(y)}$). To stabilize training, we apply stop-gradient to anchor the factual pathway against mode collapse, while detaching intermediate features prevents gradients from propagating into the visual backbone. For **Text-Free Deployment**, text pathways are entirely discarded. The efficient, privacy-preserving unimodal classifier predicts diagnosis y directly from the image utilizing the fairness-distilled static query ($\mathbf{q}_{\text{inf}}$).

4.4 Experimental Results

4.4.1 Experimental Setup

4.4.1.1 Datasets & Evaluation Protocol:

Datasets & Evaluation Protocol: We evaluate on Fitzpatrick17k Groh et al. (2021a) (16,577 images, 6 skin-types Fitzpatrick (1988b), 3-class taxonomy: Benign, Malignant, Non-Neoplastic) and Diverse Dermatology Images (DDI) Daneshjou et al. (2021) (656 biopsy-proven images mitigating Fitzpatrick17k noise, grouped into skin-types I-II, III-IV, V-VI for binary Malignant vs. Non-Malignant classification), both extended with the FitzDerm-CF clinical notes described in Chapter 3. Following Du et al. (2022); Aayushman et al. (2024), both datasets undergo in-domain evaluation via 80:20 stratified splits. Fitzpatrick17k additionally includes an out-domain setting to assess cross-group generalization (training on specific skin-tone subsets, testing on unseen groups). Results report mean ± standard deviation across 5 random seeds.

4.4.1.2 Training Details

CARE-Net variants employ an ImageNet-pretrained ViT backbone. We compare against FairDisCo Du et al. (2022), PatchAlign Aayushman et al. (2024), and classical baselines (BASE, REWT, RESM) reproduced exactly from Du et al. (2022). Models are trained for 20 (Fitzpatrick17k) or 15 (DDI) epochs using Adam ($lr = 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$), StepLR (step size 2, $\gamma = 0.9$), batch size 32, 10 workers, gradient clipping 1.0, inverse-class frequency sampling, and skin-tone confusion $\alpha = 1.0$. The fairness weight λ_{fair} warms up from 0 (epochs 1-2) to 0.2 by epoch 8, remaining fixed thereafter. For text-free validation, both variants utilize a learnable inference query guided by a precomputed text-embedding bank. In AC-CQC, inference-query attention is computed on detached visual patches to isolate backbone gradients, adding an alignment loss ($\lambda_{inf} = 0.3$). Conversely, PL-CQC excludes this loss and detachment, permitting visual backbone gradient flow.

4.4.1.3 Metrics Reported

We report overall accuracy and three fairness metrics adapted from Du et al. (2020). Buildin on the definitions of fairness discussed in Section1.3 we report these fairness metric: Let $\mathcal{G}$ be the set of skin-tone groups, $\mathcal{Y}$ the set of disease classes ($|\mathcal{Y}| = M$), and $g \in \mathcal{G}$. Higher values (closer to 1) indicate reduced disparity.

- Predictive Quality Disparity (PQD) measures prediction quality differences between each skin-type group, calculated as ratio between the lowest accuracy to the highest accuracy across different such groups:

$$\text{PQD} = \frac{\min_{g \in \mathcal{G}} \text{Acc}_g}{\max_{g \in \mathcal{G}} \text{Acc}_g}$$

- Demographic Parity Metric (DPM) assesses class prediction bias, i.e. diversity of positive outcomes for each skin-tone group:

$$\text{DPM} = \frac{1}{M} \sum_{i=1}^{M} \frac{\min_{g \in \mathcal{G}} P(\hat{y} = i \mid g)}{\max_{g \in \mathcal{G}} P(\hat{y} = i \mid g)}$$

- Equality of Opportunity (EOM) evaluates recall disparity, i.e., different groups should have similar TPR:

$$\text{EOM} = \frac{1}{M} \sum_{i=1}^{M} \frac{\min_{g \in \mathcal{G}} P(\hat{y} = i \mid y = i, g)}{\max_{g \in \mathcal{G}} P(\hat{y} = i \mid y = i, g)}$$

4.4.2 Results and Discussion

4.4.2.1 Fitzpatrick17k Evaluation

In-Domain Table 4.1 reports in-domain performance on Fitzpatrick17k. AC-CQC achieves the highest average accuracy, improving over Patch -Align by +0.85%, with consistent gains across most skin-tone groups. Fairness metrics also improve, with DPM increasing by +2.40% (95% Confidence Interval (CI) [-1.97, 6.77]) and EOM by +1.25% (95% CI [-2.45, 4.95]) across five seeds. Worst-group accuracy rises from 84.9% to 85.9%, indicating improved demographic stability. AC-CQC(no $\lambda_{distill}$) attains comparable accuracy with smaller fairness gains, while PL-CQC yields competitive performance but higher variance across seeds.

Table 4.1: In-Domain classification on Fitzpatrick17k Dataset

Model	Accuracy(↑)							PQD(↑)	DPM(↑)	EOM(↑)
	Avg	T1	T2	T3	T4	T5	T6			
BASE	85.14 (±0.21)	81.93 (±0.80)	82.62 (±0.50)	85.64 (±0.84)	89.04 (±0.34)	90.68 (±1.36)	84.74 (±2.05)	90.17 (±1.32)	48.58 (±3.78)	64.82 (±3.34)
RESM	84.82 (±0.36)	81.28 (±1.61)	81.96 (±0.27)	85.53 (±0.59)	88.97 (±1.30)	89.94 (±1.56)	88.39 (±2.91)	89.19 (±1.85)	55.51 (±6.24)	72.54 (±7.74)
REWT	85.88 (±0.46)	82.29 (±1.99)	83.36 (±0.57)	87.03 (±0.35)	89.50 (±0.35)	89.93 (±2.13)	89.42 (±2.52)	89.97 (±2.55)	55.38 (±5.29)	77.95 (±4.61)
FairDisCO	84.93 (±0.26)	82.12 (±1.45)	82.42 (±0.81)	85.71 (±0.44)	88.41 (±0.82)	89.10 (±1.40)	87.49 (±1.70)	90.98 (±2.54)	58.28 (±4.24)	77.19 (±7.08)
PatchAlign	87.61 (±0.20)	84.86 (±1.26)	84.94 (±0.77)	88.49 (±0.85)	90.92 (±1.59)	91.51 (±1.88)	91.17 (±2.11)	90.63 (±2.37)	60.60 (±5.27)	79.42 (±3.56)
AC-CQC(ours) (No $\lambda_{distill}$)	88.44 (±0.41)	85.65 (±1.11)	85.87 (±0.75)	89.42 (±0.26)	91.53 (±1.31)	**92.81** (±1.26)	91.16 (±1.79)	**91.61** (±1.15)	60.32 (±5.47)	80.59 (±4.89)
AC-CQC(ours)	**88.46** (±0.13)	**85.86** (±2.39)	**86.08** (±0.74)	89.23 (±0.83)	**91.86** (±1.18)	91.53 (±1.70)	**91.58** (±2.73)	90.46 (±1.87)	**63.01** (±7.14)	**80.68** (±2.22)
PL-CQC(ours)	88.25 (±0.42)	85.43 (±1.57)	85.17 (±0.77)	**89.95** (±0.51)	91.81 (±0.91)	91.67 (±1.39)	91.28 (±2.13)	90.80 (±1.75)	59.16 (±5.02)	78.95 (±3.86)

Out-of-domain evaluation. Fig 4.2 reports cross-demographic generalization under three shifts: $\mathcal{A}$ (train T1–T2, test T3–T6), $\mathcal{B}$ (train T3–T4, test others), and $\mathcal{C}$ (train T5–T6, test T1–T4). Across settings, AC-CQC(no $\lambda_{distill}$ and AC-CQC maintain competitive or highest average accuracy under restricted-tone training. Fairness varies by shift: in $\mathcal{A}$, PL-CQC attains stronger fairness metrics, whereas in $\mathcal{C}$, performance differences narrow across methods. We evaluate AC-CQC(mul_query) as an ablation with multiple inference queries; it shows comparable accuracy and fairness across shifts, with minor stability gains in Group C but no consistent improvement over the base variant. Overall, attention-level alignment favors predictive stability, while decision-level consistency can yield stronger fairness under certain shifts at the cost of increased variance.

4.4.2.2 DDI Evaluation

Table 4.2 presents in-domain results on DDI. While PatchAlign achieves the highest average accuracy and AUC, AC-CQC remains competitive and improves fairness metrics, including stronger performance on the most challenging subgroup (T56) and higher DPM and EOM. Variance across runs is reduced relative to prior methods, indicating improved fairness stability. These results reflect a modest accuracy trade-off for improved subgroup balance on this curated dataset.

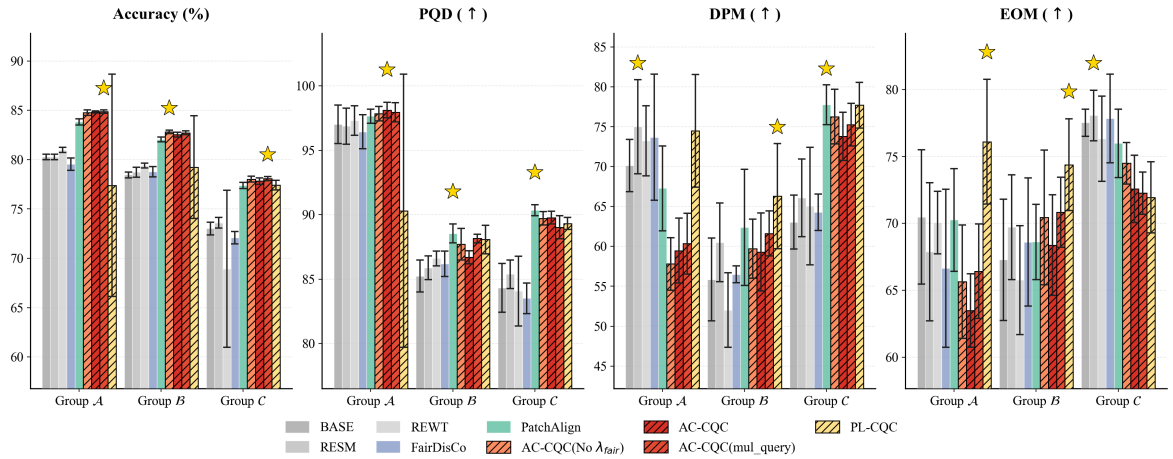


Figure 4.2: **Out-domain Accuracy, PQD, DPM, and EOM on Fitzpatrick17k.** Bars display mean ± std over 5 runs. Hatched patterns distinguish our proposed methods (*AC-CQC w/o* λ_{fair}, *AC-CQC*, *AC-CQC(*`mul_query`*)* and *PL-CQC*). Gold stars denote state-of-the-art performance.

Table 4.2: In-Domain classification on DDI Dataset

Model	Accuracy(↑)				PQD(↑)	DPM(↑)	EOM(↑)	AUC(↑)
	Avg	T12	T34	T56				
BASE	81.67 (± 2.55)	83.03 (± 6.34)	77.03 (± 5.28)	86.07 (± 3.27)	87.16 (± 6.97)	80.44 (± 16.45)	73.61 (± 10.13)	74.53 (± 4.45)
REWT	79.55 (± 2.83)	81.52 (± 5.07)	73.15 (± 5.63)	85.51 (± 2.90)	84.48 (± 7.18)	70.01 (± 15.12)	68.08 (± 15.59)	68.50 (± 12.41)
RESM	81.82 (± 2.49)	82.54 (± 6.18)	77.95 (± 3.33)	86.02 (± 2.58)	88.67 (± 4.54)	64.59 (± 11.62)	69.77 (± 7.27)	74.49 (± 6.26)
FairDisCO	81.97 (± 1.75)	83.82 (± 4.48)	78.42 (± 3.52)	84.37 (± 4.03)	89.55 (± 5.07)	76.45 (± 15.41)	75.77 (± 8.88)	78.23 (± 2.81)
PatchAlign	**83.78** (± 1.41)	82.53 (± 3.91)	**81.28** (± 2.11)	88.74 (± 2.57)	**90.53** (± 5.65)	78.72 (± 10.65)	77.43 (± 8.62)	78.98 (± 3.94)
AC-CQC (λ_{fair} =0.20)	82.43 (± 1.88)	80.50 (± 3.02)	78.94 (± 2.33)	**89.63** (± 3.10)	87.58 (± 2.23)	**85.93** (± 7.54)	77.74 (± 8.41)	78.40 (± 4.47)
AC-CQC (λ_{fair} =0.15)	83.64 (± 1.40)	**83.75** (± 2.11)	79.64 (± 2.17)	88.85 (± 3.30)	89.71 (± 4.66)	77.96 (± 6.42)	**79.52** (± 4.95)	77.39 (± 4.30)
PL-CQC	83.18 (± 2.31)	81.95 (± 3.39)	81.17 (± 4.66)	87.27 (± 3.51)	90.18 (± 4.84)	72.69 (± 7.59)	77.37 (± 8.91)	**79.05** (± 3.80)

4.4.2.3 Ablation Studies

Fairness Regularization and Distillation Analysis. We ablate the fairness weight $\lambda_{\text{fair}} \in 0, 0.1, 0.5, 1.0$ to assess sensitivity to regularization strength. As shown in Fig. 4.3, accuracy remains stable across settings, while EOM improves for moderate values (0.1–0.5) without

Table 4.3: **Effect of attention distillation loss L_{distill} for Inference Query.**

Setting	Acc. (%)	DPM	EOM
w/o L_{distill}	88.41	60.78	77.93
Random Teacher	**88.52**	56.37	78.33
Full model	88.46	**63.01**	**80.68**

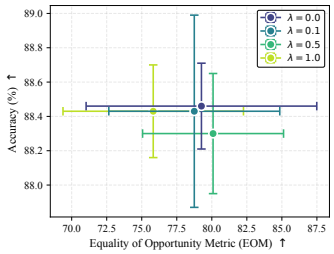


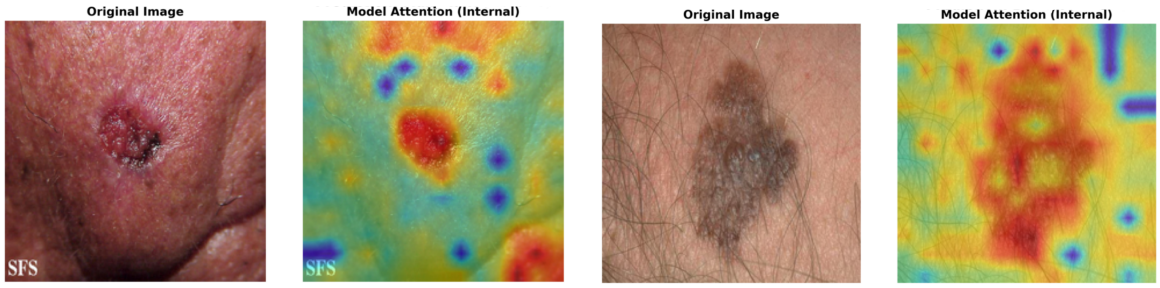
Figure 4.3: **Accuracy–fairness trade-off for λ_{fair}.** Points show mean accuracy and EOM (5 seeds); error bars denote ±1 std.

degrading performance. Larger values yield no consistent additional gains. Based on this stability, we adopt a scheduled λ_{fair} during training, which improves overall results (Table 4.1). We further compare no distillation, random-teacher, and text-guided distillation. While accuracy is similar across variants (Table 4.3), fairness improves only under counterfactually invariant text-guided supervision, indicating that gains stem from semantic attention transfer.

4.4.2.4 Attention Map Visualisation

CARE-Net 4 naturally produces spatial attention maps through the GLIQ, allowing visual inspection of the image regions attended to during prediction. Since AC-CQC explicitly regularises attention consistency under counterfactual demographic perturbations during training, these maps tend to provide a useful qualitative tool for inspecting individual model predictions. Cases where attention spreads broadly beyond the lesion boundary may warrant closer examination, as this could suggest reliance on contextual rather than lesion-specific cues.

Figure 4.4 presents two representative cases of malignant lesions from different skin-tone groups. In the darker-skin case (left), the attention map tracks the lesion structure along the nose area with little weights to the surrounding anolamies, suggesting that the model attends to local morphological variation rather than broad skin-tone information. In the relatively lighter-skin (right), attention concentrates tightly on the lesion core, with substantially lower weight assigned to the surrounding healthy tissue. In both cases, high-attention regions correspond to the diagnostically relevant parts of the image, and surrounding skin receives comparatively low weight regardless of skin tone, consistent with the AC-CQC objective of



(a) FST V–VI (b) FST III–IV

Figure 4.4: Representative GLIQ attention maps from CARE-Net. Each panel shows the original image (left) and attention map (right), with warmer colors indicating higher attention weight. Across both skin-tone groups, the model attends to morphologically relevant regions.

attending to the lesion morphology and not the skin tone of a patient.

4.5 Contribution of the Chapter 4

In this chapter we introduced CARE-Net as a multimodal fairness algorithm for dermatological diagnosis that deals with skin color bias through enforcing counterfactual invariance directly in the attention stage. Unlike previous work that regularized the global representations and predictions of a model, CARE-Net encourages the model to maintain consistent visual attention patterns under demographic shifts, ensuring that the regions it prioritises for diagnosis remain stable across skin tones.

We present three key technical contributions. The first is AC-CQC: a novel asymmetric consistency loss designed to enforce consistency between cross-attention maps derived from factual and counterfactual clinical queries, which explicitly imposes the constraint that in order for a model to be fair, it needs to attend to the exact same morphological features independent of skin color. The second contribution is GLIQ: a small yet statically learned query derived from the fairness aware multimodal teacher, allowing for strict textless, privacy preserving inference without compromising the learned counterfactually invariant attention policy. The third contribution is PL-CQC, which relaxes CQC by enforcing consistency at patch-level decision-making.

CARE-Net exhibits comparable diagnostic performance, while significantly enhancing demographic fairness and equal opportunity among skin-tone groups, as evidenced by ex-

periments performed on the Fitzpatrick17k dataset and biopsy-confirmed DDI. Experiments showed that fairness was due to counterfactual invariant attention transfer and not merely architectural considerations.

Chapter 5

Beyond Invariance: A Look into Equivariance

5.1 Towards Equivariant Counterfactual Fairness: An Exploratory Extension

The work discussed in the previous chapters has been concerned with the basic architecture of CARE-Net and its associated training objective. This chapter presents a brief exploratory extension developed alongside the primary work, motivated by a question raised by the fairness and causal-representation literature: *can the strong invariance objective used by CARE-Net be complemented by a controlled form of equivariance?* Rather than treating equivariance as a weaker replacement for invariance, we view it as a different structural constraint. Invariance asks the model to preserve lesion-focused reasoning under skin-tone interventions, whereas equivariance asks whether some skin-tone-conditioned changes in the visual or query representation can be transformed in a predictable and controlled way.

5.1.1 Motivation

CARE-Net encourages its attention distributions to remain consistent under skin-tone interventions, reflecting the intuition that a fair model should attend to the same diagnostic regions regardless of whatever demographic background the patient belongs to. This is a reasonable and practically motivated objective. However, it rests on an implicit assumption that skin tone is entirely orthogonal to the visual signal relevant for classification, such that

it can be removed from the representation without affecting diagnostic capacity.

In dermatology imaging, skin pigmentation can affect the way a lesion is visually expressed, including lesion–background contrast, apparent colour, and the visibility of some morphological attributes. CARE-Net addresses this issue by encouraging the model to focus on lesion-specific evidence that remains stable under skin-tone interventions, thereby reducing reliance on demographic shortcuts. Nevertheless, it remains an open question whether every skin-tone-associated change should be treated only as a nuisance factor to be suppressed. An alternative view is that some changes may be better understood as structured transformations of the visual signal. This possibility motivates the exploratory investigation in this chapter.

Given this, our study aims at exploring whether there is potential in augmenting CARE-Net's invariance property with equivariance as well. The objective is not to undermine the initial design principle but rather to examine whether a slight module would learn to map from factual queries to counterfactuals given certain skin tone variations. In this regard, the invariant path retains the basic idea that diagnostic information should continue to be lesion-centered, while the other side of the coin attempts to capture the remaining variation.

Moreover, the exploratory experiments provided some indication that combining invariance and equivariance objectives may influence the fairness–accuracy trade-off. However, the observed improvements were modest and were obtained from a limited set of experiments, making it difficult to determine whether they reflect a genuine advantage of the proposed formulation or normal experimental variation. For this reason, the results are best interpreted as preliminary evidence motivating further investigation rather than as a conclusive demonstration of superiority over the original CARE-Net objective.

5.1.2 A Formal Equivariance View

Let Q_s denote a query representation conditioned on skin-tone group s, and let $T_{s \to s_{cf}}$ denote a learned transformation that maps representations from a factual skin-tone condition s to a counterfactual condition s_{cf}. A minimal equivariant formulation should satisfy

$$T_{s \to s_{cf}}(Q_s) \approx Q_{s_{cf}}. \tag{5.1}$$

This makes explicit what is meant by equivariance in the present setting: the representation is not forced to be identical after the intervention, but its change should be predictable from the source and target skin-tone attributes.

For a stronger and more principled formulation, the transformation should also satisfy three consistency properties:

$$T_{s\to s}(Q_s) \approx Q_s, \tag{5.2}$$

$$T_{s_{cf}\to s}\big(T_{s\to s_{cf}}(Q_s)\big) \approx Q_s, \tag{5.3}$$

$$T_{s_2\to s_3}(T_{s_1\to s_2}(Q_{s_1})) \approx T_{s_1\to s_3}(Q_{s_1}). \tag{5.4}$$

These correspond to identity, inverse consistency, and composition consistency. The exploratory implementation below only partially enforces these properties through an identity regulariser; therefore, the current method should be understood as a first step toward controlled equivariance rather than a complete geometric equivariance framework.

5.1.3 Query Transformation Module

To investigate the feasibility of equivariance to the query, a transformation model T_θ was added as a simple post-processing layer alongside CARE-Net. This model maps the factual query Q_{fact} and the factual and counterfactual skin-tone labels s and s_{cf} into a counterfactual query prediction:

$$Q_{cf}^{\text{pred}} = T_\theta(Q_{\text{fact}},\, s,\, s_{cf}). \tag{5.5}$$

This counterfactual query then passes through the attention computation stage, resulting in a counterfactual attention distribution prediction A_{cf}^{pred}. This distribution is used as another consistency loss in addition to the original attention-invariance objective.

Equivalently, the module can be written as $T_{s\to s_{cf}}$, making the source and target skin-tone conditions explicit:

$$Q_{cf}^{\text{pred}} = T_{s\to s_{cf}}(Q_{\text{fact}}). \tag{5.6}$$

In the present implementation, T_θ acts only on the query representation. It does not directly transform the visual keys, values, or image patches. This choice keeps the extension lightweight and compatible with CARE-Net, but it also limits the extent to which the method

can model skin-tone-dependent changes in lesion visibility or image contrast. A more complete extension would decompose the visual representation into invariant lesion structure and skin-conditioned interaction components.

T_θ was trained as a small residual MLP based on the source and target skin tone attributes. The residual was added so that the module starts out as an approximate identity mapping function only learning deviations as soon as the loss has enough gradient information. This was also done in order to reduce the chance of the module getting in the way of the already learned attention path by CARE-Net. Since this extension was used for an exploratory purpose, the model was made intentionally small so that no additional parameters would be added unnecessarily.

At inference time, T_θ is not required. The extension is used only as a training-time regulariser, and the deployed CARE-Net model retains the original text-free and sensitive-attribute-free inference pathway through the Global Learnable Inference Query (GLIQ). Thus, the equivariant module is intended to shape the fairness-aware attention learned during training rather than to introduce a new test-time dependency on skin-tone labels.

5.1.4 Training Objective

The overall loss function used in the exploratory experiments extended the standard CARE-Net objective with two additional terms. The first was an equivariance consistency loss:

$$\mathcal{L}_{\text{eqv}} = \left\| \text{sg}(A_{cf}) - A_{cf}^{\text{pred}} \right\|_1, \tag{5.7}$$

where $\text{sg}(\cdot)$ represents the stop-gradient operation. In the current exploratory implementation employed by this dissertation, the attention prediction path will be detached while calculating $\mathcal{L}_{\text{eqv}}$. As a result, $\mathcal{L}_{\text{eqv}}$ serves more as an additional attention alignment criterion and indicator than the optimisation target of T_θ. This is because the residual design of the transformation module and identity constraint serve to stabilise it. Future implementations may consider using a stop-gradient for A_{cf} such that $\mathcal{L}_{\text{eqv}}$ trains T_θ.

The original CARE-Net attention-invariance objective is retained separately:

$$\mathcal{L}_{\text{inv}} = \left\| \text{sg}(A_{\text{fact}}) - A_{cf} \right\|_1. \tag{5.8}$$

This separation clarifies the role of each branch: $\mathcal{L}_{\text{inv}}$ preserves lesion-focused invariance under demographic perturbation, while $\mathcal{L}_{\text{eqv}}$ trains the auxiliary module to model the structured query-to-attention transformation.

Because CARE-Net explicitly generates both factual and counterfactual text queries during training, a direct query-prediction regulariser can also be included(which was excluded for all the results for experiments conducted in this chapter):

$$\mathcal{L}_Q = \left\| T_\theta(\text{sg}(Q_{\text{fact}}),\, s,\, s_{cf}) - \text{sg}(Q_{cf}) \right\|_1. \tag{5.9}$$

This term gives T_θ a more explicit target in query space. If this term is not used in a particular experimental run, its coefficient can be set to zero.

The second additional term was an identity regulariser:

$$\mathcal{L}_{\text{id}} = \left\| T_\theta(\text{sg}(Q_{\text{fact}}),\, s,\, s) - \text{sg}(Q_{\text{fact}}) \right\|_1. \tag{5.10}$$

This penalizes the transformation network because it makes a nontrivial transformation while the factual and counterfactual skin-tone classes are equal. It means that T_θ is expected to be an identity mapping when there are no demographic interventions. Thus, we can achieve stability of the transformation model, which relates to the identity property from Eq. 5.2.

The complete objective keeps the original CARE-Net components, including classification, inference-query classification, contrastive alignment, GLIQ distillation, and sensitive-attribute suppression:

$$\begin{aligned}\mathcal{L} = {} & \mathcal{L}_{\text{cls}}^{\text{text}} + \lambda_{\text{inf}}\mathcal{L}_{\text{cls}}^{\text{inf}} + \alpha\mathcal{L}_{\text{conf}} + \mathcal{L}_{\text{skin}}^{\text{detach}} \\ & + \lambda_{\text{fair}}(\lambda_{\text{inv}}\,\mathcal{L}_{\text{inv}} + \lambda_{\text{eqv}}\,\mathcal{L}_{\text{eqv}}) + \lambda_{\text{id}}\,\mathcal{L}_{\text{id}} + \lambda_Q\,\mathcal{L}_Q + \mathcal{L}_{\text{contrastive}} + \lambda_{\text{distill}}\,\mathcal{L}_{\text{distill}}.\end{aligned} \tag{5.11}$$

For the experiments in this chapter, λ_Q. The attribute confusion and detachment of skin prediction terminology is used to denote that the equivariant extension is not meant to substitute the attribute suppression module in CARE-Net. In order to maintain the core of fairness in CARE-Net, the invariance path has been kept intact and the equivariance path is an additional training signal.

Table 5.1: In-Domain classification on Fitzpatrick17k Dataset for equivariant mode

Model	Accuracy(↑)							PQD(↑)	DPM(↑)	EOM(↑)
	Avg	T1	T2	T3	T4	T5	T6			
AC-CQC	**88.46** (±0.13)	**85.86** (±2.39)	**86.08** (±0.74)	89.23 (±0.83)	91.86 (±1.18)	91.53 (±1.70)	**91.58** (±2.73)	90.46 (±1.87)	63.01 (±7.14)	**80.68** (±2.22)
PL-CQC	88.25 (±0.42)	85.43 (±1.57)	85.17 (±0.77)	**89.95** (±0.51)	91.81 (±0.91)	91.67 (±1.39)	91.28 (±2.13)	90.80 (±1.75)	59.16 (±5.02)	78.95 (±3.86)
AC-CQC-EAM	88.23 (±0.23)	85.56 (±1.57)	85.70 (±0.60)	88.91 (±0.89)	**91.86** (±1.16)	**92.01** (±0.69)	90.90 (±1.76)	**91.69** (±1.45)	**63.97** (±3.87)	79.59 (±3.89)

5.2 Results and Preliminary Observations

A limited set of exploratory experiments was conducted on the Fitzpatrick17k and DDI datasets to assess whether the equivariant extension could be incorporated into the CARE-Net framework without destabilising training or substantially degrading predictive performance. The results are summarised in Tables 5.1 and 5.2.

Fitzpatrick17k. On Fitzpatrick17k, AC-CQC-EAM performed comparably to the standard AC-CQC baseline, but it did not clearly outperform it in overall accuracy. AC-CQC-EAM achieved an average accuracy of 88.23%, compared with 88.46% for AC-CQC. This difference is small and falls within the reported standard deviations, suggesting that the additional equivariance-inspired losses did not materially disrupt the classification objective.

The fairness metrics show a mixed but potentially interesting pattern. AC-CQC-EAM obtained the highest PQD score (91.69) and the highest DPM score (63.97) among the variants in Table 5.1, while its EOM score (79.59) was slightly lower than AC-CQC (80.68). These results suggest that the equivariance signal may influence the fairness–accuracy trade-off, but the margins are modest and not sufficient to claim superiority over the original CARE-Net objective.

DDI. On the DDI dataset, the table includes AC-CQC-EAM along with the AC-CQC and PL-CQC variants. AC-CQC-EAM achieved the highest average accuracy (85.15%) and the highest AUC in this DDI comparison (79.23). However, the fairness results are mixed. AC-CQC-EAM improved over PL-CQC in DPM and EOM, but it did not outperform all AC-CQC settings: AC-CQC with $\lambda_{\text{fair}} = 0.20$ achieved the highest DPM (85.93), and AC-CQC with $\lambda_{\text{fair}} = 0.15$ achieved the highest EOM (79.52). This suggests that the equivariant extension

Table 5.2: In-Domain classification on DDI Dataset for equivariant model

Model	Accuracy(↑)				PQD(↑)	DPM(↑)	EOM(↑)	AUC(↑)
	Avg	T12	T34	T56				
AC-CQC (λ_{fair} =0.20)	82.43 (± 1.88)	80.50 (± 3.02)	78.94 (± 2.33)	**89.63** (± 3.10)	87.58 (± 2.23)	**85.93** (± 7.54)	77.74 (± 8.41)	78.40 (± 4.47)
AC-CQC (λ_{fair} =0.15)	83.64 (± 1.40)	**83.75** (± 2.11)	79.64 (± 2.17)	88.85 (± 3.30)	89.71 (± 4.66)	77.96 (± 6.42)	**79.52** (± 4.95)	77.39 (± 4.30)
PL-CQC	83.18 (± 2.31)	81.95 (± 3.39)	**81.17** (± 4.66)	87.27 (± 3.51)	**90.18** (± 4.84)	72.69 (± 7.59)	77.37 (± 8.91)	79.05 (± 3.80)
AC-CQC-EAM	**85.15** (± 2.31)	**87.56** (± 5.39)	80.67 (± 4.46)	88.25 (± 3.31)	88.18 (± 5.84)	81.18 (± 7.92)	78.19 (± 4.76)	**79.23** (± 4.17)

may improve predictive performance on DDI, but it does not yet provide a uniformly better fairness profile.

Across DDI variants, the standard deviations for DPM and EOM are large, frequently exceeding five percentage points. Therefore, the DDI results should be interpreted cautiously. Additional seeds, significance testing, and controlled hyperparameter sweeps are required before drawing strong conclusions about the effect of EAM on this smaller biopsy-proven dataset.

Overall. Taken together, the results suggest that the equivariance-inspired objective can be incorporated into CARE-Net without obvious training instability. On Fitzpatrick17k, it maintains comparable accuracy and slightly improves some fairness metrics, but the gains are small and not consistent across all metrics. On DDI, it improves average accuracy and AUC in the reported table, but the fairness improvements are not uniform. These observations support the use of EAM as a preliminary feasibility direction rather than as conclusive evidence that equivariance is superior to CARE-Net's original attention-invariance objective.

A stronger empirical claim would require a more systematic evaluation: larger numbers of runs, matched random seeds, confidence intervals or statistical tests, sweeps over λ_{inv}, λ_{eqv}, λ_{id}, and λ_Q, as well as out-of-domain Fitzpatrick17k evaluation under the same protocol used in the CARE-Net.

Limitations and Future Directions

This exploratory extension has several important limitations. First, T_θ is a purely data-driven residual MLP and is not yet constrained by a formal geometric, causal, or group-theoretic structure. Although Eqs. 5.2–5.4 describe desirable equivariance properties, the current implementation explicitly enforces only identity behaviour. Therefore, there is no guarantee that the learned transformations are consistent across all skin-tone pairs or clinically interpretable.

Second, the transformation is applied only in query space. This is computationally convenient, but it does not fully address the visual mechanism described in the motivation, where skin pigmentation may alter lesion contrast, colour appearance, and background interaction. A more principled extension could decompose representations into an invariant lesion component and a skin-conditioned equivariant component:

$$Q = Q_{\text{inv}} + Q_{\text{eqv}}, \qquad V = V^{\text{inv}}_{\text{lesion}} + V^{\text{eqv}}_{\text{skin}}. \tag{5.12}$$

The invariant component would remain stable under demographic intervention, while the equivariant component would be transformed in a controlled way.

Third, the empirical evaluation is still limited. The current results should be treated as a feasibility check rather than a benchmark-level validation. Future work should include systematic hyperparameter tuning, additional random seeds, out-of-domain skin-tone shift evaluation, lesion-localisation analysis where annotations are available, and stronger baselines designed specifically for equivariant representation learning.

Despite these limitations, the results indicate that a query-level transformation module can be integrated into CARE-Net without clearly harming the main classification objective. This suggests a promising direction for future work: controlled counterfactual equivariance, where lesion-relevant reasoning remains invariant but skin-conditioned interaction components are allowed to transform in a structured and auditable manner.

Chapter 6

Conclusion

Bias against underrepresented skin tones remains a persistent problem in real world dermatological AI systems. This issue is particularly important in demographically diverse countries such as India, where substantial variation in skin tones can exacerbate the effects of dataset imbalance and limit the equitable performance of diagnostic models across these groups. Existing fairness oriented approaches operates either on global representations or final predictions, leaving the system's internal visual reasoning unconstrained. Motivated by this limitation, this dissertation explored whether fairness can be enforced by supervising *where* a model looks, rather than only *what* it predicts.

To this end, two primary contributions were developed. **FitzDerm-CF** introduced in Chapter 3 extends the Fitzpatrick17k and DDI datasets with paired factual and counterfactual clinical notes, generated by BioMistral-7B and validated through a four-stage pipeline. Generation of counterfactuals in a controlled manner affects only the Fitzpatrick appearance terms whereas all disease-relevant text remains intact. Leakage analysis has shown that while diagnostic leakage remains constant for all generation temperatures, demographic leakage tends to be reduced, giving rise to a favorable privacy–utility tradeoff.

CARE-Net leverages FitzDerm-CF to enforce counterfactual consistency at the attention level. AC-CQC uses an asymmetric stop-gradient loss to align cross-attention maps induced by factual and counterfactual queries, penalising demographic shortcuts within the model's spatial reasoning. The fairness-aligned attention was distilled into a static learnable query via GLIQ, which further enables text-free and privacy preserving inference at deployment. PL-CQC provides a relaxed alternative that enforces consistency on patch-logit

contributions via confidence gated regularization. Experiments on Fitzpatrick17k and DDI show that CARE-Net maintains diagnostic accuracy while improving demographic parity and equality of opportunity over prior work.

We additionally explored in Chapter 5 whether the demographic variation can be better modelled via equivariance rather than strict invariance. Although our findings are preliminary, they seem to indicate that considering skin-tone perturbation as a structural transformation, instead of simply a nuisance to eliminate, offers a richer view on representation fairness. This leads us to identify an interesting path of investigation for further research into the confluence of fairness, counterfactuals, and equivariance.

Collectively, the work supports the claim that *fairness improvements can be obtained by constraining where a model attends, not only what representation it learns.*

6.1 Limitations, Open Problems, and Future Directions

Some limitations of the work needs to be pointed out. FitzDerm-CF depends on synthetic note creation with the help of an LLM which is not validated by any professional dermatologist, and only plausibility can be ensured. Also, counterfactual interventions are limited to the Fitzpatrick skin type only. Other protected attributes like age, gender, and ethnicity were not given any consideration and remain outside the scope of this dissertation.

With CARE-Net, the enforcement of fairness would rely on the quality of the counterfactual pairs we generate. Moreover, attention consistency might fail to capture any bias present in the channels which do not manifest themselves in the cross-attention plot. Another point that must be made is that attention maps are not to be taken as definitive explanations for how models work. There are only two benchmarks used for evaluation. Hence broader prospective validation is needed before clinical deployment.

These limitations suggest three key areas where future research could head towards. First, the FitzDerm-CF framework could be further augmented with dermatologist-labeled data and synthesized counterfactual images and multiple attribute interventions. Second, the equivariance extension suggested in Chapter 5 needs further exploration, specifically with the incorporation of learned operators for query transformations and group-equivariant attention mechanisms that enable learning from variable skin tones without trying to mitigate it. Third,

future efforts will need to validate the efficacy of these gains through clinical evaluation and uncertainty quantification would help determine whether the observed benchmark gains translate to real-world diagnostic equity.

Overall, this dissertation shows that counterfactual multimodal supervision can guide a model to focus on lesion morphology rather than demographic appearance cues. Through FitzDerm-CF and CARE-Net, we demonstrate how controlled text-based interventions can be used both to audit and to improve fairness in dermatology AI. While much work remains, the results suggest that fairness-aware attention modelling is a promising step toward systems that learn to see lesions, not skin tone.

References

Aayushman, H. Gaddey, V. Mittal, M. Chawla, and G. R. Gupta. Patchalign:fair and accurate skin disease image classification by alignment with clinical labels, 2024. URL https://arxiv.org/abs/2409.04975. 18, 27, 42, 44, 51

A. S. Adamson and A. Smith. Machine learning and health care disparities in dermatology. *JAMA dermatology*, 154(11), 2018. 14, 15, 19

C. Barata, M. Ruela, T. Mendonça, and J. S. Marques. A bag-of-features approach for the classification of melanomas in dermoscopy images: The role of color and texture descriptors. In *Computer vision techniques for the diagnosis of skin cancer*, pages 49–69. Springer, 2013. 12

G. Ben Melech Stan, E. Aflalo, R. Y. Rohekar, A. Bhiwandiwalla, S.-Y. Tseng, M. L. Olson, Y. Gurwicz, C. Wu, N. Duan, and V. Lal. Lvlm-intrepret: An interpretability tool for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8182–8187, 2024. 24

C. Boland, S. Tsaftaris, and S. Dahdouh. Preventing shortcut learning in medical image analysis through intermediate layer knowledge distillation from specialist teachers. *arXiv preprint arXiv:2511.17421*, 2025. 6, 8

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 14

P. Branco, L. Torgo, and R. P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2):31, 2016. 3

W. Brendel and M. Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet, 2019. URL https://arxiv.org/abs/1904.00760. 48

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 17

C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019. 23

L. Chen, Z. Gan, Y. Cheng, L. Li, L. Carin, and J. Liu. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pages 1542–1553. PMLR, 2020. 19

X. Chen and K. He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 46

L. Cui, D. Li, X. Yang, C. Liu, and X. Yan. A unified approach addressing class imbalance in magnetic resonance image for deep learning models. *IEEE Access*, 12:27368–27384, 2024. 4

L. Cui, M. Xu, C. Liu, T. Liu, X. Yan, Y. Zhang, and X. Yang. Towards reliable healthcare imaging: A multifaceted approach in class imbalance handling for medical image segmentation. *Interdisciplinary Sciences: Computational Life Sciences*, 17(3):614–633, 2025. 3, 5

R. Daneshjou, K. Vodrahalli, W. Liang, R. A. Novoa, M. Jenkins, V. Rotemberg, J. Ko, S. M. Swetter, E. E. Bailey, O. Gevaert, et al. Disparities in dermatology ai: Assessments using diverse clinical images. *arXiv preprint arXiv:2111.08006*, 2021. 8, 12, 14, 26, 27, 29, 31, 42, 51

S. Das, S. Datta, and B. B. Chaudhuri. Handling data irregularities in classification: Foundations, trends, and future challenges. *Pattern Recognition*, 81:674–693, 2018. 16

I. de Vere Hunt, S. Owen, A. Amuzie, V. Nava, A. Tomz, L. Barnes, J. K. Robinson, J. Lester, S. Swetter, and E. Linos. Qualitative exploration of melanoma awareness in black people in the usa. *BMJ open*, 13(1):e066967, 2023. 2

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 13

M. Du, F. Yang, N. Zou, and X. Hu. Fairness in deep learning: A computational perspective, 2020. URL https://arxiv.org/abs/1908.08843. 52

S. Du, B. Hers, N. Bayasi, G. Hamarneh, and R. Garbi. Fairdisco: Fairer ai in dermatology via disentanglement contrastive learning, 2022. URL https://arxiv.org/abs/2208.10013. 18, 27, 42, 46, 51

H. El-Khatib, D. Popescu, and L. Ichim. Deep learning–based methods for automatic diagnosis of skin lesions. *Sensors*, 20(6):1753, 2020. 42

A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542 (7639):115–118, 2017. 12

T. B. Fitzpatrick. The validity and practicality of sun-reactive skin types I through VI. *Archives of Dermatology*, 124(6):869–871, 1988a. doi: 10.1001/archderm.1988.01670060015008. 7, 27, 31, 43

T. B. Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124 6:869–71, 1988b. URL https://api.semanticscholar.org/CorpusID:29991932. 51

A. Fontanella, A. Antoniou, W. Li, J. Wardlaw, G. Mair, E. Trucco, and A. Storkey. Acat: Adversarial counterfactual attention for classification and detection in medical imaging, 2023. URL https://arxiv.org/abs/2303.15421. 42, 45

N. Gessert, T. Sentker, F. Madesta, R. Schmitz, H. Kniep, I. Baltruschat, R. Werner, and A. Schlaefer. Skin lesion classification using cnns with patch-based attention and diagnosis-guided loss weighting. *IEEE Transactions on Biomedical Engineering*, 67(2):495–503, 2019. 42

J. W. Gichoya, I. Banerjee, A. R. Bhimireddy, J. L. Burns, L. A. Celi, L.-C. Chen, R. Correa, N. Dullerud, M. Ghassemi, S.-C. Huang, et al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, 2022. 42

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 18

M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, and O. Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1820–1828, 2021a. 14, 42, 51

M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, and O. Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1820–1828, 2021b. 7, 12, 26, 27, 31

R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018. 22, 24

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016. 6, 7

P. Hart. The condensed nearest neighbor rule (corresp.). *IEEE transactions on information theory*, 14(3):515–516, 1968. 16

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. 3

H. He and Y. Ma. *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley-IEEE Press, 1st edition, 2013. 3

K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 13

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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3942–3951, 2021. 42

N. Japkowicz. The class imbalance problem: Significance and strategies. In *Proceedings of the 2000 ICAI*, pages 111–117, 2000. 3, 17

J. M. Johnson and T. M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27, 2019. 3, 17

M. A. Kassem, K. M. Hosny, and M. M. Fouad. Skin lesions classification into eight classes for isic 2019 using deep convolutional neural network and transfer learning. *IEEE access*, 8:114822–114832, 2020. 5, 7, 12

D. Kaushik, E. Hovy, and Z. C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*, 2019. 21

J. Kawahara, A. BenTaieb, and G. Hamarneh. Deep features to classify skin lesions. In *2016 IEEE 13th international symposium on biomedical imaging (ISBI)*, pages 1397–1400. IEEE, 2016. 12, 13

A. I. Khandaker, A. Al Shafi, and M. Ahmad. Handling class imbalance problem in skin lesion classification: finding strengths and weaknesses of various balancing techniques. In *2025*

International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN), pages 1–6. IEEE, 2025. 5

B. Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016. 3, 16

I. Ktena, O. Wiles, I. Albuquerque, S.-A. Rebuffi, R. Tanno, A. G. Roy, S. Azizi, D. Belgrave, P. Kohli, T. Cemgil, et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine*, 30(4):1166–1173, 2024. 21, 22

M. Kubat, S. Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *ICML*, volume 97, pages 179–186, 1997. 3

M. Kukar, I. Kononenko, et al. Cost-sensitive learning with neural networks. In *ECAI*, volume 98, pages 445–449, 1998. 17

M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017. 20, 21

Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the association for computational linguistics: acl 2024*, pages 5848–5864, 2024. 27, 31, 33, 50

A. Li, T. Luo, T. Xiang, W. Huang, and L. Wang. Few-shot learning with global class representations. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

X. Li, Z. Cui, Y. Wu, L. Gu, and T. Harada. Estimating and improving fairness with adversarial learning. *arXiv preprint arXiv:2103.04243*, 2021. 42

T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE ICCV*, pages 2980–2988, 2017. 18

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 14

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 23

R. K. Mothilal, A. Sharma, and C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020. 20, 22

F. Nachbar, W. Stolz, T. Merkle, A. B. Cognetta, T. Vogt, M. Landthaler, P. Bilek, O. Braun-Falco, and G. Plewig. The abcd rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, 30(4):551–559, 1994. 12

G. Nan, R. Qiao, Y. Xiao, J. Liu, S. Leng, H. Zhang, and W. Lu. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2765–2775, 2021. 45

K. Nijjer, R. Bui, D. Jiu, A. Ahmed, P. Wang, K. Zhu, and L. Zhu. Adapting large language models to mitigate skin tone biases in clinical dermatology tasks: A mixed-methods study. *arXiv preprint arXiv:2510.00055*, 2025. 26

L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pages 151–159, 2020. 15

K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas. Imbalance problems in object detection: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. doi: 10.1109/TPAMI.2020.2981890. 17

T. Oorloff, Y. Yacoob, and A. Shrivastava. Mitigating hallucinations in diffusion models through adaptive attention modulation, 2025. URL https://arxiv.org/abs/2502.16872. 47

J.-O. Palacio-Niño and F. Berzal. Evaluation metrics for unsupervised learning algorithms. *arXiv preprint arXiv:1905.05667*, 2019. 39

A. Palacios, K. Trawiński, O. Cordón, and L. Sánchez. Cost-sensitive learning of fuzzy rules for imbalanced classification problems using furia. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 22(05):643–675, 2014. 17

K. Paxton, K. Aslansefat, D. Thakker, and Y. Papadopoulos. Evaluating fairness and mitigating bias in machine learning: A novel technique using tensor data and bayesian regression. *arXiv preprint arXiv:2506.11627*, 2025. 42

J. Pearl. *Causality*. Cambridge university press, 2009. 20

D. Pessach and E. Shmueli. A review on fairness in machine learning. *ACM computing surveys (CSUR)*, 55(3):1–44, 2022. 6

S. R. Pfohl, T. Duan, D. Y. Ding, and N. H. Shah. Counterfactual reasoning for fair clinical risk prediction. In *Machine Learning for Healthcare Conference*, pages 325–358. PMLR, 2019. 22

N. L. T. Pham, D. D. Pham, T. D. Le, and K. T. Huynh. A multimodal deep ensemble framework for skin lesion classification. In *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making*, pages 100–111. Springer, 2025. 19

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 19

M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016. 23

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 42

W. Samek, T. Wiegand, and K.-R. Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017. 22

H. Schechter Vera, S. Dua, B. Zhang, D. Salz, and e. a. Mullins. Embeddinggemma: Powerful and lightweight text representations. 2025. URL https://arxiv.org/abs/2509.20354. 50

D. Scholz, A. C. Erdur, J. A. Buchner, J. C. Peeken, D. Rueckert, and B. Wiestler. Imbalance-aware loss functions improve medical image classification. In *Medical imaging with deep learning*, 2024. 5

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 23

L. Seyyed-Kalantari, H. Zhang, M. B. McDermott, I. Y. Chen, and M. Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021. 16, 19, 24

R. L. Siegel, A. N. Giaquinto, and A. Jemal. Cancer statistics, 2024. *CA: a cancer journal for clinicians*, 74(1):12–49, 2024. 11

M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009. 38

J.-W. Tjiu and C.-F. Lu. Equity and generalizability of artificial intelligence for skin-lesion diagnosis using clinical, dermoscopic, and smartphone images: A systematic review and meta-analysis. *Medicina*, 61(12):2186, 2025. 26

I. Tomek. Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(11):769–772, 1976. 16

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*, 5 (1):180161, 2018. 5, 7, 12

P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature medicine*, 26(8):1229–1234, 2020. 23

A. Vlontzos, D. Rueckert, and B. Kainz. A review of causality for learning algorithms in medical image analysis. *arXiv preprint arXiv:2206.05498*, 2022. 42

W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788, 2020. URL https://arxiv.org/abs/2002.10957. 35, 37

Z. Wang, Z. Wu, D. Agarwal, and J. Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876, 2022. 42, 46

P. Wienholt, C. Kuhl, J. N. Kather, S. Nebelung, and D. Truhn. Medicalpatchnet: A patch-based self-explainable ai architecture for chest x-ray classification, 2025. URL https://arxiv.org/abs/2509.07477. 48

J. K. Winkler, C. Fink, F. Toberer, A. Enk, T. Deinlein, R. Hofmann-Wellenhof, L. Thomas, A. Lallas, A. Blum, W. Stolz, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 155(10):1135–1141, 2019. 13

J. Xu, Y. Xiao, W. H. Wang, Y. Ning, E. A. Shenkman, J. Bian, and F. Wang. Algorithmic fairness in computational medicine. *EBioMedicine*, 84, 2022. 5, 6

J. Xu, C. Lan, W. Xie, X. Chen, and Y. Lu. Slot-vlm: Slowfast slots for video-language modeling, 2024. URL https://arxiv.org/abs/2402.13088. 42, 47

S. Yan, Z. Yu, C. Primiero, C. Vico-Alonso, Z. Wang, L. Yang, P. Tschandl, M. Hu, G. Tan, V. Tang, et al. A general-purpose multimodal foundation model for dermatology. *arXiv preprint arXiv:2410.15038*, 2(6), 2024. 14, 19

S. Yan, Z. Yu, C. Primiero, C. Vico-Alonso, Z. Wang, L. Yang, P. Tschandl, M. Hu, L. Ju, G. Tan, et al. A multimodal vision foundation model for clinical dermatology. *Nature Medicine*, 31(8):2691–2702, 2025. 19

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*, 30(10):2838–2848, 2024. 42

S. Ye, M. Meng, M. Li, D. Feng, and J. Kim. *Enabling Text-Free Inference in Language-Guided Segmentation of Chest X-Rays via Self-guidance*, pages 242–252. Springer Nature Switzerland, 2024. ISBN 9783031721113. URL http://dx.doi.org/10.1007/978-3-031-72111-3_23. 46

S. Ye, U. Naseem, M. Meng, and J. Kim. Alleviating textual reliance in medical language-guided segmentation via prototype-driven semantic approximation, 2025. URL https://arxiv.org/abs/2507.11055. 42

C. Zhang, K. C. Tan, H. Li, and G. S. Hong. A cost-sensitive deep belief network for imbalanced classification. *IEEE transactions on neural networks and learning systems*, 30 (1):109–122, 2018. 17

L. Zhang, Z. Wang, X. Dong, Y. Feng, X. Pang, Z. Zhang, and K. Ren. Towards fairness-aware adversarial network pruning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5168–5177, 2023. 13

S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, A. Crabtree, B. Piening, C. Bifulco, M. P. Lungren, T. Naumann, S. Wang, and H. Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1), 2024. doi: 10.1056/AIoa2400640. URL https://ai.nejm.org/doi/full/10.1056/AIoa2400640. 19, 37, 40

D. Zhou, S. Gao, and X. Huang. Strategies for class-imbalanced learning in multi-sensor medical imaging. *Sensors*, 26(6):1998, 2026. 4, 5