



ΣΧΟΛΉ ΘΕΤΙΚΏΝ ΕΠΙΣΤΗΜΏΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ ΚΑΙ ΕΦΑΡΜΟΓΕΣ»

**«MASTER OF SCIENCE IN ARTIFICIAL INTELLIGENCE
AND APPLICATIONS»**

**«Εξηγήσιμη τεχνητή νοημοσύνη για
αποκατάσταση της προκατάληψης σε μοντέλα
εφαρμοσμένα στον τομέα της υγείας»**

Παναγιώτης Βλαχαΐτης Δημήτριος

Σεπτέμβριος 2026



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΉ ΘΕΤΙΚΏΝ ΕΠΙΣΤΗΜΏΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ ΚΑΙ ΕΦΑΡΜΟΓΕΣ»

«Master of Science in Artificial Intelligence and Applications»

**«Εξηγήσιμη τεχνητή νοημοσύνη για αποκατάσταση της
προκατάληψης σε μοντέλα εφαρμοσμένα στον τομέα της υγείας »**

Διπλωματική Εργασία που υποβλήθηκε στο Τμήμα Πληροφορικής και
Τηλεπικοινωνιών του Πανεπιστημίου Θεσσαλίας ως μέρος των απαιτήσεων
για την απόκτηση Διπλώματος Μεταπτυχιακών Σπουδών Τεχνητή
Νοημοσύνη και Εφαρμογές από τον

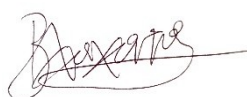
Παναγιώτη Βλαχαΐτη Δημήτρη

Δήλωση Αυθεντικότητας, ζητήματα Copyright

«Ο/Η μεταπτυχιακός/ή φοιτητής/τρια που εκπόνησε την παρούσα διπλωματική εργασία φέρει ολόκληρη την ευθύνη προσδιορισμού της δίκαιης χρήσης του υλικού, η οποία ορίζεται στη βάση των εξής παραγόντων: του σκοπού και χαρακτήρα της χρήσης (μη-εμπορικός, μη-κερδοσκοπικός, αλλά εκπαιδευτικός-ερευνητικός), της φύσης του υλικού που χρησιμοποιεί (τμήμα του κειμένου, πίνακες, σχήματα, εικόνες κ.λπ.), του ποσοστού και της σημαντικότητας του τμήματος που χρησιμοποιεί σε σχέση με το όλο κείμενο υπό copyright, και των πιθανών συνεπειών της χρήσης αυτής στην αγορά ή την γενικότερη αξία του υπό copyright κειμένου».

Ο Μεταπτυχιακός Φοιτητής

ΠΑΝΑΓΙΩΤΗΣ ΒΛΑΧΑΪΤΗΣ



Μήνας και Έτος Συγγραφής

Σεπτέμβριος 2026

Τριμελής Εξεταστική Επιτροπή

Η παρούσα Μ.Δ.Ε. εγκρίθηκε ομόφωνα από την τριμελή εξεταστική επιτροπή η οποία ορίστηκε από την Σ.Ε. / Συνέλευση του Τμήματος Πληροφορικής και Τηλεπικοινωνιών του Πανεπιστημίου Θεσσαλίας, σύμφωνα με το νόμο και τον εγκεκριμένο Οδηγό Σπουδών του Π.Μ.Σ. «Τεχνητή Νοημοσύνη και Εφαρμογές». Τα μέλη της Επιτροπής ήταν:

1 ^ο Μέλος (Επιβλέπων):	κ. Κωνσταντίνος Κολομβάτσος
2 ^ο Μέλος:	κ. Νικόλαος Τζιρίτας
3 ^ο Μέλος:	κ. Αντώνιος Δαδαλιάρης

Η έγκριση της Μ.Δ.Ε. από το Τμήμα Πληροφορικής και Τηλεπικοινωνιών του Πανεπιστημίου Θεσσαλίας, δεν υποδηλώνει αποδοχή των απόψεων του συγγραφέα.

Περίληψη

Τα μοντέλα βαθιάς μάθησης έχουν σημειώσει μεγάλη επιτυχία στη διάγνωση δεδομένων δερματικών παθών, με ορισμένα από αυτά να φτάνουν ακόμη και σε επίπεδο απόδοσης δερματολόγου. Ωστόσο, τα μοντέλα αυτά ενδέχεται να παρουσιάζουν μεροληψία έναντι υποεκπροσωπούμενων υποομάδων στα δεδομένα, συγκεκριμένα ατόμων με πιο σκούρο χρώμα δέρματος. Συνήθως, οι παραδοσιακές τεχνικές επαύξησης δεδομένων δεν καταφέρνουν να επιλύσουν αυτό το ζήτημα, καθώς συχνά μπορεί να είναι βαθιά ριζωμένο στο σύνολο των δεδομένων. Στην παρούσα διατριβή, προτείνουμε ένα εξηγήσιμο μοντέλο εξουδετέρωσης μεροληψίας με βάση την ανταγωνιστική μάθηση (LAD), το XLAD, το οποίο αξιοποιεί την αρχική αρχιτεκτονική LAD μαζί με ορισμένες κρίσιμες τροποποιήσεις, προκειμένου να επιτευχθεί πραγματική αμεροληψία στη λήψη αποφάσεων του μοντέλου, με κόστος την απόδοση της ακρίβειας. Στόχος είναι να επιτύχουμε υψηλές μετρικές δικαιοσύνης ενώ παράλληλα, να μην έχουμε μεγάλες μειώσεις στην απόδοση του μοντέλου. Ενσωματώνουμε επίσης την εξήγηση Grad-CAM στη ροή εργασίας μας, ώστε να προσθέσουμε ένα επίπεδο διαφάνειας στο μοντέλο μας, καθώς οι παραδοσιακές προσεγγίσεις βαθιάς μάθησης συχνά θεωρούνται αρχιτεκτονικές «μαύρου κουτιού» και οι κλινικοί γιατροί διστάζουν να τις χρησιμοποιήσουν. Για την αξιολόγηση των αποτελεσμάτων της μεθοδολογίας μας, καθιερώνουμε μια ισχυρή βάση αναφοράς που αποτελείται από τρεις βασικές αρχιτεκτονικές, το ResNet, το DenseNet και το EfficientNet, καθώς και συγκρίνουμε μεθόδους αιχμής στον τομέα της δικαιοσύνης στην τεχνητή νοημοσύνη, όπως το FairDisCo και το PatchAlign, στο σύνολο δεδομένων Fitzpatrick17k. Τα αποτελέσματα μετρούν τη δικαιοσύνη με τις μετρικές Equal Opportunity, της Disparate Impact και της

Accuracy Parity, όπου παρατηρείται σημαντική αύξηση, που σε ορισμένες περιπτώσεις φτάνει έως και το 10% στη μεθοδολογία του μοντέλου μας

Abstract

Deep learning models succeeded in diagnosing skin lesion data over the years, with some even reaching dermatologist level performance. However, these models may prove to be biased towards underrepresented subgroups in data, specifically individuals with darker skin tone. Typically, traditional data augmentation techniques fail to resolve this issue, as it can often be deeply rooted inside the dataset. In this thesis, we propose an explainable latent adversarial debiasing model XLAD that utilizes the original LAD architecture along with some crucial modification to achieve true fairness in model decision making at the cost of accuracy performance. We also incorporate Grad-CAM explaining in our pipeline to add a layer of transparency to our model, as traditional approaches are regarded as uninterpretable “black box” architectures, which in turn makes clinicians hesitant about their use. Results of our methodology are compared on a strong baseline of three core architectures, ResNet, DenseNet and EfficientNet in addition to State-of-the-Art methods in AI fairness such as FairDisCo and PatchAlign on the Fitzpatrick17k dataset. The results measure fairness in Equal Opportunity, Disparate Impact and Accuracy Parity, all of which see a significant increase, as high as 10% in some cases in our model methodology.

Table of Contents

1 Introduction.....	11
1.1 Background.....	11
1.1.1 AI in Healthcare.....	11
1.1.2 Skin Cancer Diagnosis.....	11
1.1.3 Role of Deep Learning.....	12
1.2 Problem Statement.....	13
1.2.1 Dataset Bias.....	13
1.2.2 Demographic Imbalance.....	13
1.2.3 Lack of Model Transparency.....	14
2 Related Work.....	15
2.1 Overview.....	15
2.2 Medical Imaging.....	17
2.2.1 Image Classification.....	18
2.2.2 Image Segmentation.....	19
2.2.3 Object Detection.....	20
2.2.4 Image Enhancement.....	20
2.3 Common Architectures for Medical Imaging.....	21
2.3.1 Convolutional Neural Networks.....	22
2.3.2 Auto Encoders & Variational Auto Encoders.....	24
2.3.3 Generative Adversarial Networks.....	25
2.3.4 Pre-trained Architectures.....	27
2.4 Fairness and Explainability Approaches.....	28
2.4.1 Fairness Approaches.....	29
2.4.2 Explainability Approaches.....	30
2.5 AI in Dermatology.....	32
2.5.1 Applications of AI in Dermatology.....	33
2.5.2 Fair AI in Dermatology.....	34
2.5.3 Explainable AI in Dermatology.....	35
3 Methodology.....	37
3.1 Explainable Latent Adversarial Debiasing XLAD.....	37
3.1.1 VQ-VAE.....	38
3.1.2 Adversarial Walk.....	39

3.1.3 GRAD-CAM Explaining.....	40
3.2 Datasets and Evaluation.....	41
3.2.1 Fitzpatrick17k.....	41
3.2.2 Evaluations Metrics.....	42
4 Experiments.....	44
4.1 Implementation Details.....	44
4.1.1 Environment Settings.....	44
4.1.2 Image Preprocessing and Augmentation.....	44
4.1.3 Model Settings.....	45
4.2 Results.....	47
4.2.1 Baseline.....	47
4.2.2 FairDisCo.....	48
4.2.3 LAD.....	49
4.2.4 XLAD.....	50
4.2.5 Model Comparisons and Visualizations.....	51
5 Conclusions & Future Work.....	59
5.1 Conclusion.....	59
5.2 Future Work.....	59
6 Acknowledgements.....	60
References.....	61

Table of Figures

Figure 1: ResNet18 final results. Artifacts and text in images disrupt model predictions.....	54
Figure 2: DenseNet121 final results.....	54
Figure 3: EfficientNetB0 final results. Disturbed predictions due to artifacts	55
Figure 4: XLAD ResNet18 architecture. The model does not rely on artifacts or text on the images.....	57
Figure 5: XLAD ResNet18 part 2.....	57
Figure 6: XLAD EfficientNetB0 architecture. The model does not rely on artifacts or text on the image.....	58
Figure 7: XLAD DenseNet121 architecture. The model does not rely on artifacts or text on the images.....	58

Index of Tables

Table 1: Comparison table of all tested architectures. PatchAlign was not manually tested, we instead trained our models with their settings and simply compared them.....	53
--	----

Table of Equations

Equation 1: Convolution Output Equation.....	24
Equation 2: Generative Adversarial Network Loss function.....	27
Equation 3: VQ-VAE Loss function and codebook contribution.....	38
Equation 4: Focal Loss Function.....	39
Equation 5: Adversarial Walk Equation.....	39
Equation 6: Perturbation strength calculation.....	40
Equation 7: Grad-CAM importance weight calculation.....	41
Equation 8: Final Grad-CAM heatmap generation.....	41
Equation 9: Accuracy calculation.....	42
Equation 10: F1 score calculation.....	42
Equation 11: Disparate Impact Calculation.....	43
Equation 12: Equal Opportunity Calculation.....	43
Equation 13: Accuracy Parity Calculation.....	43

1 Introduction

1.1 Background

1.1.1 AI in Healthcare

With the introduction of Machine Learning and Deep Learning, the field of Artificial Intelligence (AI) has advanced drastically. These methods provide exceptional performance in tasks such as image recognition, decision making, and natural language processing. Many researchers from different fields use these structures and algorithms to gain valuable insights from the data they are working with. With the recent archival digitization in healthcare, AI technologies are being adopted to support clinicians by providing decision support systems, improving diagnostic accuracy, assisting in treatment planning, and enhancing patient care. Oncology specifically, is the most active research fields in healthcare AI. Machine learning, and subsequently, Deep Learning methods are developed to assist diagnosis and treat cancer. World Health Organization statistics show that cancer is regarded as one of the leading causes of death worldwide, with nearly 10 million deaths in 2020 and approximately 145.4 deaths per 100,000 cancer cases according to recent statistics [1], [2].

1.1.2 Skin Cancer Diagnosis

Skin cancer remains a critical global health challenge, driven largely by the severity of melanoma. Because this aggressive variant can quickly spread to other parts of the body if unaddressed, early detection and intervention are vital. When caught in these localized initial stages, the prognosis is

overwhelmingly positive, boasting a five-year survival rate greater than 99%. However, survival rates decrease significantly once the disease spreads to lymph nodes or distant organs, dropping to approximately 75% and 35%, respectively [3]. To support this, the availability of open-access dermoscopic image datasets has enabled the development of diagnostic tools that utilize computers and help dermatologists in evaluating skin lesions, ultimately streamlining workflows and accelerating the identification of malignant growths.

1.1.3 Role of Deep Learning

Skin lesion classification systems have seen significant advancements with recent deep learning approaches, enhancing diagnostic accuracy and automating tasks that traditionally required manual evaluation. Convolutional Neural Networks (CNNs) have dominated visual recognition tasks, excelling at extracting intricate patterns from medical imagery. Through this capacity for feature-extraction, deep learning architectures have achieved performance similar to experienced clinicians in various specialized tasks. The most used and well documented models are ResNet, DenseNet, and EfficientNet which have become essential benchmarks in analyzing medical imagery, owing to their capacity to learn discriminative features directly from raw input data. Despite these successes, DL methods are regarded as black-box systems, meaning that, it is difficult for humans to understand the reasoning behind their predictions. Furthermore, the performance of a model depends highly on the quality and representativeness of the training data, making these systems vulnerable to biases that already exist in the given dataset and may affect clinical reliability and fairness.

1.2 Problem Statement

1.2.1 Dataset Bias

Despite these notable achievements, artificial intelligence systems remain inherently vulnerable to biases embedded within their underlying training data. In medical imaging, dataset bias may arise from uneven sampling procedures, limited demographic representation, or inconsistencies in data collection protocols. As a result, models may learn spurious patterns that do not reflect clinically relevant characteristics. In response, recent research underscores the necessity of fairness-aware AI frameworks capable of maintaining consistent and dependable performance across diverse patient demographics. The extent of this limitation was highlighted in a 2024 survey of 54 open-access skin cancer repositories, where just three provided explicit skin tone annotations, obstructing systematic evaluation of demographic disparities [4].

1.2.2 Demographic Imbalance

Demographic imbalance poses a real challenge in skin lesion classification. Models trained on datasets dominated by specific demographic population often fail to achieve superior performance when applied to the underrepresented group. In dermatology, this issue is particularly concerning because skin lesions can present differently across skin tones. Unequal representation can lead to disparities in diagnostic accuracy and reinforce existing healthcare inequalities. Consequently, ensuring balanced performance across demographic groups has become an important

objective in AI systems that can be trusted by experts in their respected fields.

1.2.3 Lack of Model Transparency

In addition to fairness concerns, deep learning systems often suffer from a lack of transparency. Their decision-making processes are typically difficult to interpret, deeming it challenging for experts to comprehend why a specific decision was made. This limitation can reduce trust in AI-assisted diagnostic systems and hinder their adoption in clinical environments. Furthermore, hidden biases may remain undetected when model decisions cannot be adequately explained. In skin lesion classification, models may inadvertently rely on non-causal features such as skin tone or image acquisition characteristics instead of focusing on lesion-specific information. This study addresses these challenges by developing the Explainable Latent Adversarial Debiasing (XLAD) model based on LAD [5], a bias mitigation framework designed to separate causal signals from confounding factors while maintaining strong classification performance. This approach is evaluated using the Fitzpatrick17k dataset across multiple CNN architectures, including ResNet, DenseNet, and EfficientNet, as well as other bias mitigation strategies like FairDisCo with the goal of developing fairer and more reliable skin cancer classification systems.

2 Related Work

2.1 Overview

Artificial Intelligence (AI) has surfaced as a fundamental tool for modern healthcare, offering algorithms for analyzing complex medical data and supporting the decision-making of clinicians. Advances in ML and DL have facilitated the development of scalable, automated systems capable of identifying patterns in medical images, electronic health records, genomic data, and other medical information. These technologies demonstrate considerable success in tasks such as disease diagnostics, prognostic decisions, patient risk assessment, and planning treatments, often achieving performance comparable to that of medical experts [6], [7], [8], [9]. Within healthcare AI, medical imaging represents a core investigated domain. DL architectures like Convolutional Neural Networks (CNNs) have illustrated superior performance in tasks such as classification, detection, segmentation, and enhancement across diverse fields including radiology, pathology, ophthalmology and dermatology [10], [11], [12], [13], [14], [15]. More recently, advanced architectures such as autoencoders, generative models, and transfer learning frameworks have further augmented the capabilities of AI systems by enhancing feature extraction, representation learning, and image synthesis [16], [17], [18], [19]. Despite these advances, the integration of AI systems in clinical environments is still not adopted due to several important concerns. The most significant being the limited transparency of complex deep learning models. A primary limitation of modern deep networks is their “black box” nature. To elaborate, while these models are exceptionally accurate, they lack transparent mechanisms to

explain how specific decisions are derived. In the medicine field settings, where clinical decisions can directly affect patient diagnostics, the lack of interpretability may discourage clinicians and hinder the widespread adoption of AI-based decision-support systems [8], [20]. Recognizing this critical shortfall in transparency, Explainable Artificial Intelligence (XAI) has rapidly developed into an important subdiscipline of research. XAI's goal is to translate the uninterpretable nature of pure computations into human-understandable rationales, in order for the end user to interpret model inference. By revealing which features influence model decisions, explainability techniques enable clinicians and researchers to better evaluate model reliability, identify potential errors, and support transparent decision-making processes. As a result, explainability plays an important part in developing safe and trustworthy AI systems in healthcare [21], [22], [23], [24]. Alongside concerns regarding interpretability, researchers have increasingly recognized the presence of bias in medical AI systems. Bias can arise from imbalanced datasets, demographic underrepresentation, data collection practices, and confounding variables embedded within training data [21], [22], [25]. As a result, AI models may exhibit unequal performance across patient populations, potentially reinforcing existing healthcare disparities. These concerns are particularly evident in dermatology, where several studies have reported substantial imbalances in skin lesion datasets, with lighter skin tones being significantly overrepresented compared to darker skin types [4, p. 2024], [26], [27], [28], [29], [30], [31]. Recent research has therefore focused on using both explainability and fairness methods, investigating how interpretable machine learning techniques are used to understand model decisions in addition to identify and mitigate sources of bias. Explainability methods can reveal whether models rely on clinically meaningful features or on uncorrelated features that may contribute to

unfair predictions. By supporting bias detection and informing mitigation strategies, explainable AI will be able to provide robustness for both the transparency and fairness of medical AI systems [21], [22], [24]. We review relevant studies of explainable and fair medical AI in the rest of this chapter. Section 2.2 examines the DL architectures commonly employed in medical imaging, including convolutional neural networks (CNNs), autoencoders (AE), generative adversarial networks (GANs), and transfer learning approaches. Section 2.3 focuses on dermatology as a representative application domain, discussing advances in skin lesion classification, explainable AI techniques, and lastly Section 2.4 addresses fairness challenges and reviews existing bias mitigation strategies.

2.2 Medical Imaging

With the transformation of healthcare data into digital formats, deep learning has become the new standard for a plethora of task in medicine. These techniques are applied to critical tasks such as image segmentation [14], [26], object detection [32], [33] and disease classification [29], [34], [35], [36]. Furthermore, DL has critical role in the development of computer assisted diagnostic (CAD) systems by enabling more precise and consistent interpretations of medical data.

2.2.1 Image Classification

Image classification stands as an important and basic task within the computer vision field of study and is defined as the systematic process of assigning one or multiple meaningful labels to a given image. Traditional workflows relied on manual, low-level feature extraction techniques, and ultimately feeding these engineered features into conventional classifiers such as Support Vector Machines (SVMs) [37]. However, modern DL

techniques have consistently outperformed their traditional counterparts. By eliminating manual feature engineering in favor of automated, hierarchical representation learning, CNNs have set the gold standard in image classification, delivering superior empirical results [10], [38]. Building upon these advancements, various deep learning architectures have been widely adopted across diverse clinical domains. Specifically, they are important in classifying patterns of lung disease, utilizing high-resolution computed tomography (HRCT) [38], detecting focal liver lesions across multiphase computed tomography (CT) scans [10], evaluating and classifying breast cancer from medical imaging capturing methods like mammography and MRI [39], and automating the precise classification of histopathology slides [40], alongside a broad selection of other critical medical imaging applications.

2.2.2 Image Segmentation

Image segmentation constitutes the critical computational process of dividing an image into multiple distinct, information packed areas or structures, enabling precise pixel level analysis. Within the field of medical imaging, segmentation is a valuable tool for accurately identifying vital anatomical structures, pathological lesions, and various regions that need to be examined across a diverse range of diagnostic techniques, such as magnetic resonance imaging (MRI), computed tomography (CT), ultrasound and positron emission tomography (PET). Historically, classical segmentation frameworks relied on heuristic techniques such as region growing, edge detection, as well as the thresholding technique. However, these conventional approaches frequently encountered significant limitations when tested on actual medical data that display high variability

and noise characteristics [41]. To address these limitations, deep learning architectures, specifically Fully Connected Networks and U-Net frameworks, have become the benchmark for capturing complex hierarchical features across diverse imaging modalities. Practical clinical implementations of deep learning-based segmentation encompass a wide array of critical tasks, including tumor and lesion detection [42], precise organ and tissue delineation, radiotherapy and surgical treatment planning, longitudinal disease progression monitoring, and advanced cellular analysis within microscopy imaging [8], [9], [14], [41]. By facilitating highly accurate spatial localization and quantitative volumetric assessment, deep learning-driven segmentation serves a fundamental role in advancing computer-assisted diagnosis (CAD) platforms while accelerating the advancement of precision medicine.

2.2.3 Object Detection

Object detection attempts to recognize the presence of multiple items in images and localize them using bounding boxes and confidence scores. This task is a crucial in medical imaging, as abnormalities like tumors, lesions or nodules need precise localization in order to identify the disease and create a plan to treat it. Early object detection practices include hand-crafted features along with sliding-window techniques, that had proven to be computationally expensive and often inaccurate [33]. State of the art object detection frameworks include YOLO (You Only Look Once) [43], the R-CNN series [44], and Single Shot Multi box Detector (SSD) [45], have transitioned from general computer vision tasks to medical imaging. Their utility has been validated in screening X-ray abnormalities, isolating tissue anomalies,

and tracking lesions during endoscopic procedures [32], laying a solid groundwork for deep learning applications in healthcare.

2.2.4 Image Enhancement

The definition of image enhancement is the procedure of improving the overall fidelity of an image. This task often finds appliance in medical imaging challenges as clinical scans and other diagnostic tools often produce images that can be obscured with unwanted noise. The primary way of achieving this higher image clarity is to maximize image contrast, filter out said background noise and sharpen the critical anatomical areas in order to be more recognizable. In many advanced objectives, image enhancement plays a crucial part in the preprocessing phase. Early enhancement strategies involved spatial domain modifications, including techniques like histogram equalization, gamma correction, and adaptive contrast adjustment. These were often paired with several frequency domain transforms like the Fourier and wavelet transforms. However, with the rapid improvement of DL architectures, dramatic improvements have been noted across many fields including image enhancement. These algorithms utilize data-driven mathematical mappings to perform image denoising, achieve remarkable super-resolution, and execute detailed contrast optimizations [17], [20], [39], [46]. Within the diverse landscape of healthcare, these enhancement techniques have been seamlessly integrated across a wide variety of medical imaging modalities. For instance, they are actively utilized to clarify mammography screenings for the premature detection of breast cancer, to comprehensively refine cardiac MRIs and echocardiograms to ensure more definitive diagnoses of complex cardiovascular conditions, to amplify the clarity of neurological MRI scans to better expose hidden brain

tumors and subtle lesions, and to upgrade the visual quality of histopathology slides for meticulous, cellular-level evaluations [17]. Consequently, by producing sharper, clearer, and more informative image representations, these techniques assist both human medical experts during manual interpretations as well as artificial intelligence systems during automated processing, contributing to more accurate and reliable clinical decision-making [46].

2.3 Common Architectures for Medical Imaging

Deep learning models in medical imaging rely on specialized neural networks that are modeled to process complex spatial hierarchies, extract high-level feature representations, and learn domain-specific transformations. Rather than using generic machine learning pipelines, modern clinical AI tasks like the aforementioned classification, object detection, enhancement and segmentation, depend on specific model architectures that are optimized for medical data characteristics. The following subsections examine the foundational architectures that are essential in medicine data analysis: Autoencoders and Variational Autoencoders (Aes/VAEs), Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and Pre-trained architectures.

2.3.1 Convolutional Neural Networks

Originally influenced by the biological architecture in human visual perception, Convolutional Neural Networks (CNNs) were first conceptualized in early neuro-computing research [47] and practically demonstrated for pattern recognition by LeCun and colleagues [48]. Since these foundational developments, CNNs have established absolute dominance across computer vision, pattern recognition, and medical image processing [49].

Rather than treating input scans as flat vector arrays, CNNs explicitly exploit the 2D and 3D spatial grid structures of images and clinical data, such as radiological projections, MRI slices, and histological stains, in order to learn localized, hierarchical feature representations. At a high level view, CNN architectures use three fundamental, sequential components, the convolutional layers, the pooling layers and the classification layer (head). The first layer, called the convolutional layer, acts as the primary feature-extraction engine, making it an important part of the network. Given an image input represented as a matrix of pixels, the layer applies a series of learnable filter kernels. These kernels slide across the boundaries of the input matrix, performing matrix multiplications and summations to generate specialized feature maps that isolate critical visual features like edges, boundaries, shapes, and other complex textures [49], [50]. The transformation within a convolutional layer depends on three hyperparameters [51]: Depth (D), is the total amount of distinct filters to be applied in a single layer, Stride (S) defines the step size in which the kernel moves through the input grid, and Padding (P) adds a zero-valued border around the perimeter. Mathematically, for an input with dimension W , kernel size K , padding P and stride S , the outputs can be computed as follows:

$$O = \frac{W - K + 2P}{S} + 1$$

Equation 1: Convolution Output Equation

The pooling layer executes localized down-sampling across the extracted feature maps, systematically reducing spatial dimensions, memory consumption, and the total number of downstream trainable parameters. Applied independently across every channel, pooling operations aggregate local neighborhood information. The most widely adopted configuration is max pooling [51], [52], which extracts the maximum feature intensity within

a sliding window (e.g., 2×2 with stride 2). In medical imaging, max pooling provides mild local translation invariance and isolates dominant signals, though at the cost of detail in spatial resolution. Finally, after the stacks of convolution, activation functions (most commonly ReLU) and pooling layers, the Fully Connected Linear layers or Classification Head receives a single flattened continuous vector that contains all the information and abstractions that were learned across the earlier representation to compute the final class probability distributions (usually via softmax) [50]. In modern clinical practice, CNNs have a variety of different use cases and tasks to complete. Guo et al. proposed a multimodal image segmentation model with feature fusion at the learning level, targeting tumor segmentation [41]. Esteva et al. proposed a CNN network that was extensively trained on large datasets and was tested to measure its performance to that of an expert dermatologist and concluded that the CNNs could in fact be almost as competent as a dermatologist [42]. Sun et al. introduced a hierarchical Conditional Random Field (CRF) methodology, designed for detecting cancerous regions within gastric histopathology imagery. To capture multi-scale spatial context, their method utilizes multiple CNN architectures to model pixel-level potentials alongside fine-tuned CNNs to generate patch-level potentials for precise segmentation [53]. Finally, addressing pulmonary pathology, Li et al. introduced a CNN with shallow convolutions to classify interstitial lung disease demonstrating high efficiency in this task as well as a general model for other medical tasks [54].

2.3.2 Auto Encoders & Variational Auto Encoders

Autoencoders (AE) represent a foundational architecture of unsupervised neural networks designed to perform non-linear reduction of dimensions by compressing high-dimensional input image to a lower dimension, with the help of a latent bottleneck, while also having the ability to reverse this

reduction and reconstruct the target signal from this latent representation [16]. Initially invented as a generalization of Principal Component Analysis (PCA)[55], [56], [57], an autoencoder relies on two distinct phases: an encoder module that condenses incoming information into a constraint latent vector, and a decoder tasked with rebuilding the original input from that exact compressed state. Model parameters are optimized by minimizing structural differences via task-specific objective functions, such as Mean Squared Error (MSE) for image reconstruction [58] or Binary Cross-Entropy for normalized mask creation [59]. To preserve localized spatial features, Convolutional Autoencoders substitute the dense vector layer with 2D or 3D convolutions and transposed deconvolutions making them suited for image restoration, artifact suppression, and super resolution [17], [60], [61]. Variational Autoencoders (VAE) extend the premise of the traditional autoencoder by remodeling the bottleneck as an approximate posterior distribution, parameterizing a mean and variance vector. Paired with the reparameterization trick and Kullback-Leibler divergence regularization in Evidence Lower Bound (ELBO) optimization, Variational Autoencoders construct a more well-structured latent space. Consequently, this smooth manifold serves as a foundation for unsupervised anomaly detection [62], synthetic data synthesis, and the characterization of complex clinical variability [63]. Finally, autoencoders in medicine have also seen some traction, where Miotto et al. created a network that utilized three autoencoders to diagnose patients using digital health records and Pundhir et al. utilized variational autoencoder to combat the skin tone bias in dermatology that stems from traditional CNNs [30], [64], [65].

2.3.3 Generative Adversarial Networks

The Generative Adversarial Network (GANs) is considered to be a well-known and well-used architecture in generative artificial intelligence, first proposed

by Goodfellow et al. [18]. Applications of GANs encompass a plethora of computer vision challenges, including medical imaging for synthetic data generation, image super-resolution, domain adaptation and attribute editing. Standard GANs operate as a zero-sum game between a Generator module (G) and a Discriminator (D). Without direct exposure to ground-truth data, the first module projects a random noise vector into a synthetic data space, attempting to replicate the underlying target distribution. At the same time, the discriminator receives both authentic training images and images generated from the generator network, and is then tasked to identify which image belongs to the generator's synthetic images and which belongs to the real dataset images. The discriminator's loss quantifies its accuracy in identifying the synthetic samples, whereas the generator uses this evaluation as a learning signal to refine its output and "fool" the discriminator [66]. Formally, this adversarial optimization scheme is governed by a minimax objective function, wherein the discriminator seeks to minimize its binary classification error, while the generator attempts to maximize the probability of the discriminator misidentifying synthetic data as real [66].

$$\min_G \max_D \text{Net}(D, G) = E_{x \sim P_{data}(x)} [\log D(x)] + E_{z \sim p_z(x)} [\log (1 - D(G(z)))]$$

Equation 2: Generative Adversarial Network Loss function

Additionally, GANs have been growing rapidly in bibliography with the introduction of Deep Convolution GANs (DCGANs), using convolutional and transposed convolution layers in place of fully connected networks, stabilizing adversarial learning and enabling the extraction of spatial features from imagery [67]. Another core GAN architecture is the conditional GAN (cGAN) where an auxiliary label like a label or segmentation maps are fed into both the generator and the discriminator, allowing targeted generation and domain-specific image synthesis, at the cost of making the

task supervised from the previously unsupervised methods [68]. Consequently, GANs and their architectural variants have been widely adopted within the medical field, effectively resolving critical challenges such as training data scarcity and suboptimal segmentation performance on tomography, MRI, and mammography, in addition to facilitating cross-modality image translation and enhancing the resolution of low-dose diagnostic scans [12], [14].

2.3.4 Pre-trained Architectures

Transfer learning is a powerful optimization method that utilizes the generalized knowledge acquired by robust models on massive source domains to address novel, target-domain challenges. Typically, these foundation architectures are pre-trained on large-scale visual corpora, in particular the ImageNet dataset, which is composed from over 15 million annotated common images [10]. Because these highly parameterized networks learn rich, hierarchical feature representations, ranging from basic edges to complex textures, they can be efficiently repurposed for specific tasks using only a fraction of the data and computational resources through a process known as fine-tuning [19]. Within computer vision, ResNet, DenseNet, and EfficientNet represent three of the most widely adopted pre-trained architectures. Introduced by He et al. [69], the Residual Network (ResNet) revolutionized deep learning by introducing residual learning. By utilizing identity shortcut connections to propagate inputs forward and bypass intermediate layers, ResNet effectively mitigates the vanishing gradient problem, enabling the successful optimization of exceptionally deep networks (exceeding 150 layers). The concept of skip connections has since become a foundational architectural mechanism, heavily influencing subsequent innovations such as U-Nets, DenseNets, and modern Transformer models. The architecture is categorized by its depth, with

standard variants including small sized architectures like ResNet-18, medium sized like ResNet-34 and ResNet-50 in addition to larger models like ResNet-101, and ResNet-152. Building upon the concept of residual pathways, the Densely Connected Convolutional Network (DenseNet) [70] maximizes information flow by establishing direct connections from each layer to all of its subsequent layers. This dense connectivity explicitly encourages structural feature reuse, substantially reduces the total parameter count, and improves overall optimization efficiency. Popular configurations of this architecture include the inherently large models like DenseNet-121, DenseNet-161, DenseNet-169, and the largest architecture of DenseNet-201. Finally, EfficientNet, proposed by Tan and Le [71], prioritizes computational efficiency through the use of mobile inverted bottleneck convolutions. Furthermore, it introduces a novel “compound scaling” factor methodology that systematically scales network depth, channel width, and input image resolution using a carefully balanced, fixed scaling coefficient. This controlled resource management yields a family of highly optimized models, scaling sequentially from the baseline EfficientNet-B0 up to EfficientNet-B7. While applying these models from natural scenes to clinical datasets introduces big difference in data distribution, this technique remains crucial in the field of medicine, as clinical datasets are known to be extremely small due to strict patient privacy rules and the cost of having an expert radiologist to annotate the dataset labels. In addition, pre-trained architectures have recently been expanded beyond traditional convolutional methods with the introduction of Vision Transformers (ViTs). By partitioning images into sequential patches and utilizing self-attention mechanisms, heavily pre-trained ViTs excel at capturing global anatomical context and long-range spatial dependencies that the localized receptive fields of standard CNNs often miss. Therefore, Transformer models and hybrid

architectures have emerged alongside CNNs as the new standard for robust, high-performance medical image classification and segmentation [36], [72].

2.4 Fairness and Explainability Approaches

Deep learning networks have proven to be very efficient on a plethora of problems encompassing many research fields, however, these approaches have proven to have two major flaws should be addressed. First being bias, an inclination of generate biased outputs due to misleading predictions [21]. Bias can be present inherently inside the data themselves, the algorithms applied or in the engineer that develops the algorithm and evaluates it. Either way, its a flaw that should be considered when developing models whose decisions may impact underrepresented groups, especially in the field of medicine. Furthermore, another critical flaw in DL networks is the lack of model interpretability. These architectures are by nature hard to interpret, since they are consisted of the outputs of element wise matrix multiplications, users as well as developers cannot get a clear explanation as to why the model reached to a decision, thus these models are often deemed as 'black boxes' [73], [74]. The rest of the section covers several generally applied methodologies proposed for fairness and explainability in AI models.

2.4.1 Fairness Approaches

Researchers are actively studying the nature of deep learning models and are trying to find ways to mitigate bias in model predictions. There three main phases in which mitigation techniques can be applied. Firstly, preprocessing techniques are methods applied as a preparation of the data before they are fed into the artificial model. A more straightforward solution to bias is to simply remove the sensitive information from the data like age,

sex or income, although potentially risking model performance [21]. When dealing with image data another particularly useful approach is data augmentation and synthetic data augmentation, as there is no sensitive attribute we can easily remove from the image, it is best that we try to regularize the parity between different groups. Calmon et al. proposed a probabilistic framework for convex optimization for learning a data transformation to control utility and discrimination [75]. Furthermore, in-processing methodologies are referred to as the operations that are taking place inside the training phase, often applied directly on the loss function or on the algorithm used. Zemel et al. proposed an algorithm that mapped each individual feature as a distribution with the goal to “unlearn” the protected data information [76]. Other common types of in-processing mitigation include implicit and explicit regularization techniques. Implicit regularization is applied on the loss function of a model and is used to disentangle association between model prediction and protected attributes. This type of regularization also needs additional labels that are tailored to the dataset, to be passed into the loss function. Conversely, explicit regularization modifies the loss function to penalize performance across distinct dataset subgroups [77]. Agarwal et al. proposed a constraint that included equal odds and accuracy parity in the optimization function [77]. In another study, Darlow et al. utilized adversarial debiasing to disentangle the confounding bias in images by learning a discrete latent space from a vq-vae encoder and then proceed to change the original image based on the outcome of a biased prediction [5]. Lastly, postprocessing techniques include after prediction calibration and the implementation of fairness metrics such as equal opportunity, disparate impact and accuracy parity [25]

2.4.2 Explainability Approaches

With the rapid advancement of “black box” DL architectures, a need for an explanation in model decisions emerged. As a result, a large portion of researchers are actively working towards a more explainable path for these architecture, with the first proposals dating back to 1977. [79]. Since then, models have been meticulously studied and more explainability techniques have been introduced. For instance, a widely used explanation technique is Local Interpretable Model-Agnostic Explanations (LIME), proposed in Ribeiro et al. [80]. Needless to say, LIME has been easily applied to many studies thanks to its model agnostic methodology, meaning that, it can operate on any architecture by solely processing its output. LIME’s objective is to create a smaller, more explainable, surrogate model on classification outputs of the complex DL network. This approach, while it has proven to be effective and computationally inexpensive, the fact that it relies on the fitting process of another model makes it unreliable at times. Furthermore, another well documented approach consists of Shapley Values (SHAP) and their variations. Originally proposed by Lundberg and Lee [81], they use equations from cooperative game theory to compute explanations for model outputs. Methods from this family are concerned with a fixed point in time where the model will assess the expected value for a conditional distribution in relation to a subset of variables present in the data. Subsequently, this approach calculates the importance of each variable in a subset by evaluating its expected value with respect to the contribution of each other variable that is already sampled from the subset [82], [83]. Lastly, in the area of CNNs, Zhou et al. proposed the Class Activation Mapping (CAM) methodology for identifying descriptive regions in images on pure convolutional neural nets (not applied on fully connected layers) [84]. This study found that the global average pooling layers used in CNNs enable the model to be highly

localizable, and expose the attention of a CNN in a given image. Furthermore, Selvaraju et al. [85] proposed Gradient-weighted CAM (Grad-CAM) further extending this methodology by integrating gradient flow of a given target on the CAM architecture. This enables the model to produce a localization map of the model's attention while still being applicable in a wide variety of model CNN architectures that may include fully connected layers [85]. In this study, we chose to incorporate Grad-CAM methods, as the method showcased a really low computational overhead while providing meaningful explanations in CNNs compared to other methods and as stated in Cheng et al. [86] Grad-CAM proved to be the fastest method.

2.5 AI in Dermatology

Artificial Intelligence models have been adopted, among other fields, in dermatology and dermoscopic image analysis. As mentioned in the previous chapters, deep learning networks have excelled in image recognition tasks over the last years. Additionally, digitization of documents has been a major trend in the healthcare field, enabling access to intelligent algorithms to aid in the better performance and care of patients [87], [88]. Despite these notable achievements, the effectiveness of AI models across different skin tones still remains a challenge for researchers. Since early disease detection is crucial to saving a patient's life, these models must attempt to be as effective and accurate as possible. With the absence of at least a skin tone label in the skin cancer focused dataset publicly available, only a small amount of datasets provide such information. Consequently, models tend to cause disparities and discriminate against the minority population [89]. Furthermore, while AI systems have achieved substantial performance and aid in dermatological diagnosis, they lack transparency which makes doctors reluctant to using them. As the responsibility of the decision lies with the

clinician, not understanding the reason of the output of a given model discourages them from using these models [90].

2.5.1 Applications of AI in Dermatology

Recent developments have positioned artificial intelligence as a primary catalyst for advancing analytical frameworks in dermatology, demonstrating a plethora of clinical appliances across a spectrum of neoplastic, inflammatory, and infectious skin diseases. In cutaneous oncology, AI systems have been extensively deployed to address the rising incidence of skin cancers, most notably melanoma and non-melanoma skin cancer (NMSC)[91], [92]. Algorithms trained on massive dermatoscopic datasets, such as the ISIC archive [93], [94], have achieved diagnostic and classification accuracies that match or surpass those of experienced dermatologists when differentiating between benign melanocytic lesions and malignant tumors [42]. Furthermore, AI facilitates advanced clinical management through whole-body 3D mapping for longitudinal lesion tracking, and utilizes Artificial Neural Networks (ANNs) for predicting NMSC risks based on patient health records [95], [96]. Beyond oncology, AI is revolutionizing the evaluation of inflammatory and autoimmune dermatoses, including psoriasis, eczema, atopic dermatitis, acne, and vitiligo [88]. For these conditions, AI models provide objective, standardized disease severity assessments (such as automated Body Surface Area and Psoriasis Area and Severity Index scoring), which effectively overcome the inherent subjectivity of traditional human evaluation. Advanced machine learning architectures, including Support Vector Machines (SVMs) and Generative Adversarial Networks (GANs), are also being leveraged for precise lesion segmentation, differentiating complex morphological manifestations of eczema, and synthesizing high-resolution diagnostic images [97], [98], [99], [100]. Additionally, deep learning frameworks have been successfully integrated

into broader diagnostic workflows, from segmenting MRI volumes to differentiate benign from malignant neurofibromas [101], to automating microscopic fungal detection with near-perfect sensitivity [102]. These diverse applications underscore AI's profound potential to reduce clinician workload, standardize subjective diagnostic criteria, and facilitate the early, accurate detection of critical dermatological conditions.

2.5.2 Fair AI in Dermatology

Recent research in skin cancer detection using AI methods has shown that prediction fairness is a critical issue [103]. The root cause of this bias predominantly stems from severe disproportionate representation of dark skin phototypes (Fitzpatrick types IV–VI) in foundational training corpora, leading deep learning models to extract feature embeddings optimized heavily for low-melanin backgrounds. Consequently, ensuring algorithmic fairness has become a mandatory factor in skin cancer diagnosis, driving researchers to develop robust mitigation strategies against prejudiced predictions. To this end, Du et al. proposed the FairDisCo framework which actively removes skin type bias from the model representations by disentangling diagnostic features from demographic attributes [104]. FairDisCo utilized adversarial techniques as well as custom loss functions like confusion and contrastive loss to promote diversity and intra-class cohesion. Similarly, Aayushman et al. also proposed PatchAlign, a framework that improves skin condition diagnostic accuracy by incorporating Graph Optimal Transport (GOT) loss as a regularizing term. Their method achieves cross-domain alignment by framing the correspondence between image patches and clinical conditions as a graph-matching problem[34]. In addition to these representation learning methods, adversarial training techniques are also widely used to counter bias. For instance, Li et al. proposed an adversarial framework incorporating a discrimination module against bias

and a critic module to predict unfairness on top of the main classification framework of a GAN [105]. Ultimately, evaluating the efficacy of these diverse technical interventions requires strict adherence to mathematical fairness constraints such as Equal Opportunity and Equalized Odds to guarantee that diagnostic sensitivity for malignant lesions does not degrade across minority demographic groups. Furthermore, our model builds upon Latent Adversarial Debiasing framework [106] which utilizes an external, low capacity, biased critic, that learns to predict the latent representations of a Vector Quantizing Variational Autoencoder (VQ-VAE), in order to produce images that calibrate model's focus on the more essential parts of each image [106].

2.5.3 Explainable AI in Dermatology

Explainability and lack of model transparency are crucial weaknesses of artificial neural network models. The fact that the output of most architectures are by nature uninterpretable, makes clinicians hesitant about utilizing these techniques in practice, as they are unable to reason and support their decisions with methods that humans deem acceptable. In dermatology, many approaches have been applied successfully. In this regard, Nigar et al. [74] proposed training pipeline utilizing the ResNet-18 pretrained architecture paired with local interpretable model-agnostic explanations (LIME) and proved to be effective on the ISIC 2019 skin lesion dataset. Abbas et al. proposed an explanation pipeline that utilizes the layer-wise relevance propagation (LRP) method to determine the relevance scores of each input [107]. Munjan et al. introduced SkinSage XAI, a framework that utilizes both gradient based Class Activation Mappings [85] (Grad-CAM) and LIME methods paired with a variety of pretrained models like InceptionV3, ResNet-50, Dense-Net-201 and concluded that InceptionV3 adapted better to their framework [108]. Anand et al. proposed a Combine Hybrid

Architecture for Scalable High-performance in Neural Iterations (CHASHNIt) that elevated the effectiveness of GANs to create synthetic images and combined SHAP [109] and LIME for model transparency [110]. A different approach that uses machine learning techniques was introduced in [111] where they attempted to isolate features from preprocessed images on the PH2 dataset. Specifically, they utilized the feature extraction capabilities the CNN offers, to carefully map important features and then train a machine learning algorithm like XGBoost, KNN, decision trees and random forests. Finally, Naz et al. [112] proposed a federated learning based methodology, respecting patient privacy and data protection regulations. They employed a VGG19 model and Grad-CAM explanations to interpret their model's decision-making process.

3 Methodology

We propose a novel framework to actively mitigate bias from skin cancer datasets and add an explainability layer to make model predictions more interpretable. Ultimately, producing an efficient workflow that could function in situations where time constraints are strict and the resources are not abundant. The remaining section describes each component of the framework.

3.1 Explainable Latent Adversarial Debiasing XLAD

To simultaneously tackle demographic disparity and the uninterpretable nature of deep neural networks in medical imaging, we present Explainable Latent Adversarial Debiasing. (XLAD). This framework extends the standard Latent Adversarial Debiasing (LAD) approach by explicitly integrating Explainable AI (XAI) verification and dynamically addressing dataset class imbalance. Medical datasets, particularly in dermatology, are notoriously skewed, with benign lesions vastly outnumbering rare malignancies. Furthermore, the joint probability of a darker skin tone which is already rare in dermatology datasets, combined with the scarcity of malignancies, accounts for a fraction of data points. To prevent the network from latching onto highly irrelevant confounding artifacts (e.g., clinical rulers, surgical markings, or specific skin phototypes), XLAD employs a focal loss objective during classification. Furthermore, by integrating XAI techniques, XLAD provides visual, interpretative proof that the causal debiasing process successfully redirects the model's attention away from spurious background confounders and toward true pathological tissue. The XLAD framework is constructed upon three core modules: a Vector-Quantized Variational

Autoencoder (VQ-VAE), an adversarial walk mechanism operating in the latent space, and Gradient-weighted Class Activation Mapping (Grad-CAM) for spatial interpretability.

3.1.1 VQ-VAE

The VQ-VAE is the first core component of XLAD, which operates on a foundational autoencoder architecture distinguished by its utilization of a quantized, discrete latent space rather than a continuous one. To achieve this discretization, the VQ-VAE employs a fixed-size, trainable codebook containing a static number of embedding vectors. During the forward pass, the encoder maps inputs to a continuous representation, which is subsequently mapped to the nearest codebook vector using an L_2 norm nearest-neighbor lookup. Because this argmin quantization operation is non-differentiable, backpropagation is achieved via a straight-through estimator, which directly copies the gradients from the decoder back to the encoder [63]. Post-quantization, the decoder reconstructs the input image using these discrete codebook representations. Over successive training iterations, the dictionary vectors are continuously adjusted toward representations produced by the encoder, thereby optimizing codebook usage, reducing reconstruction error, and capturing the target data distribution. Consequently, the overall VQ-VAE loss function is expressed as:

$$L_{VQ} = \log p(x|z_q(x)) + \|sg[z_e(x)] - e_q\|_2^2 + \beta \|z_e(x) - sg[e_q]\|_2^2$$

Equation 3: VQ-VAE Loss function and codebook contribution

Where $sg[\cdot]$ represents the stop-gradient operator (acting as a non-updatable constant during backpropagation), x denotes raw input, $z_e(x)$ is the continuous output of the encoder, z_q represents the closest selected codebook embedding, e_q stands for the codebook embedding, β is a

hyperparameter dictating the magnitude of the commitment loss and $\|\cdot\|_2^2$ is the operator denoting the square Euclidean distance.

3.1.2 Adversarial Walk

An auxiliary classifier is attached directly to the encoder's latent space. Because this space initially encapsulates both causal physiological signals and confounding artifacts, the classifier naturally gravitates toward "easy-to-learn" spurious features, according to Darlow et al. [5]. To counteract the heavy class imbalance of dermatological datasets during this phase, XLAD optimizes this classifier using focal loss rather than standard Cross-Entropy. Focal loss incorporates a modulating factor $(1 - p_t)^\gamma$ that reduces the relative loss for easily classified examples (e.g., prominent artifacts), thereby prioritizing model training on hard, complex causal features:

$$L_{focal} = -a_t (1 - p_t)^\gamma \log(p_t)$$

Equation 4: Focal Loss Function

With p_t denoting the model's estimated probability for the ground truth class, a_t is a weighting factor that balances positive and negative class frequencies, and $\gamma \geq 0$ represents the focusing parameter. Once the classifier is established, XLAD executes an adversarial walk entirely within the latent space. Its explicit objective is to maximize the predictive entropy H of the auxiliary classifier, forcing it into a state of maximum uncertainty regarding the confounding features it previously relied upon:

$$H = \max H(\text{Classifier}(\text{quantize}(z_e(x))))$$

Equation 5: Adversarial Walk Equation

The iterative adversarial perturbation is governed by normalizing entropy's gradient:

$$\delta = a \cdot \frac{\frac{\partial H}{\partial h} - \text{mean}(\frac{\partial H}{\partial h})}{\text{std}(\frac{\partial H}{\partial h})}$$

Equation 6: Perturbation strength calculation

Where a denotes a hyperparameter tuning the perturbation strength, H denotes the classifier's prediction entropy, h is encoder's latent representation and δ is the overall perturbation of the image. The VQ-VAE decoder then reconstructs this perturbed latent vector into a synthesized image. The resulting image is effectively scrubbed of its confounding artifacts, serving as an intentionally debiased augmentation for training the final diagnostic model.

3.1.3 GRAD-CAM Explaining

To ensure clinical transparency and validate the efficiency of the latent adversarial debiasing, XLAD incorporates Gradient-weighted Class Activation Mapping (Grad-CAM) as its primary Explainable AI (XAI) mechanism. While the adversarial walk removes confounders mathematically in the latent space, Grad-CAM provides a highly interpretable visual verification in the pixel space. To identify the image regions driving a network's decision, Grad-CAM leverages the spatial representations retained in the final convolutional layer. Specifically, the gradients of the target class y^c are calculated with respect to the activation map of the target layer A^k , subsequently applying global averages pooling to these gradients to determine the importance weights a_k^c for each feature channel:

$$a_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

Equation 7: Grad-CAM importance weight calculation

These activation maps are weighted and subsequently processed through a ReLU activation. This step isolates features that positively contribute to the classification class score. As a result, the localized class-activation heatmap is produced:

$$L_{Grad-CAM}^c = ReLU\left(\sum_k a_k^c A^k\right)$$

Equation 8: Final Grad-CAM heatmap generation

Within the XLAD framework, Grad-CAM serves as a critical evaluative tool. Prior to debiasing, Grad-CAM heatmaps typically reveal that the biased model anchors its predictions on unrelated artifacts, such as dark skin pigmentation, surgical ink, or dark corners of the dermatoscope lens.

3.2 Datasets and Evaluation

For our train and test dataset, despite the scarcity of skin tone labeled datasets, we concluding upon using the Fitzpatrick17k dataset [113] for both the train and test splits. Additionally, we establish three core metrics that point to model fairness according to Canton and Haas, Dhabliya et al. , [25], [114] equal opportunity, disparate impact and overall accuracy parity.

3.2.1 Fitzpatrick17k

Dermatology datasets often lack dark skin tone representation, in addition to the absence of skin tone labeling. This limitation sparked the creation of the Fitzpatrick17k dataset [113]. This dataset is consisted of 16,577 images and its images are obtained from two online repositories, DermaAmin and Atlas Dermatologico. The dataset also contains labels for 22 categories of

the most common dermatology conditions, labels for a nine label partition and a three label partition exist, each being increasingly more generalized to include a variety of the conditions. Furthermore, the dataset also provides skin type labels that uses six distinct points as label, initially developed to classify sun reactivity of skin and treat patients accordingly. We chose to take on the three label classification problem of the dataset. Lastly, due to dataset imbalance, we resulted in using methods to balance the dataset like data augmentation and weighted samplers.

3.2.2 Evaluations Metrics

For our evaluation metrics we chose to include F1 score and Accuracy. These widely used metrics are utilized to measure purely model performance, complementary to the rest of the fairness metrics. These scores are calculated as shown:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 9: Accuracy Calculation

$$F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

Equation 10: F1 Score Calculation

Where TP and TN denote accurately predicted positive and negative instances, respectively; FP represents negative samples erroneously labeled as positive; and FN signifies positive samples mistakenly categorized as negative. Furthermore, we incorporate three distinct evaluation metrics for the fairness validation of the models, that measure the discrimination between demographic groups in the data. We consider two groups to be applied in these metrics: the protected group (g_{prot}) being the minority class of data points with darker skin tone, and the non-protected group (g_{non-p}) being the majority of the dataset [25]. Firstly, to analyze the model's

predictive capabilities, we utilize the “disparate impact” metric (DI) that showcases positive classification ratio, compared to the other two groups.

$$DI = \frac{Pr(\hat{y}=1|g_{prot})}{Pr(\hat{y}=1|g_{non-prot})}$$

Equation 11: Disparate Impact Calculation

Where these probabilities measures the positive classification given the image to classify belongs to the protected or the non-protected group. A $DI \approx 1$ denotes that both groups are being equivalently classified as positive. Continuing, to quantify classification performance and ensure that there is no performance decrease on minority groups, the equal opportunity metric (EO) is used which compares the probabilities of the True Positive Rates (TPR) for both demographics.

$$EO = \frac{Pr(\hat{y}=1|y=1, g_{prot})}{Pr(\hat{y}=1|y=1, g_{non-p})}$$

Equation 12: Equal Opportunity Calculation

An $EO \approx 1$ denotes that our model is capable of correctly identifying malignancies regardless of subgroup membership.

Lastly, the overall accuracy parity (AP) simply measures the accuracy of the classifications between the two groups using the aforementioned accuracy formula.

$$AP = \frac{Accuracy_{g_{prot}}}{Accuracy_{g_{non-prot}}}$$

Equation 13: Accuracy Parity Calculation

An $AP \approx 1$ indicates that, for both subgroups, the model’s overall accuracy is actually the same.

4 Experiments

The performed experiments are tested on the aforementioned Fitzpatrick17k dataset where we split the demographic groups by fitzpatrick

skin tone. Specifically, the protected group includes data points where $Fitzpatrick < 4$ and the majority group contains data where $Fitzpatrick \geq 4$. In the effort to prove that our method fairly predicts all groups we train the following models: Baseline architectures without focal loss, Baseline architectures with focal loss, XLAD without focal loss, XLAD with focal loss and FairDisCo for comparison. The implementation details of our experiments are stated in the rest of this chapter along with hardware, hyperparameters and approaches and then present this study's outcomes.

4.1 Implementation Details

4.1.1 Environment Settings

Architectures were modeled utilizing the PyTorch framework, were all pretrained model weights, the backend implementations of core components used like convolutional layers, data augmentations and optimizers are included. The machine that trained the models is hosted in a Linux native OS along with Ubuntu 24.04 distribution, with RTX 4090 24GB VRAM GPU, 128GB of RAM, NVMe M.2 storage and an Intel Xeon W-2235 CPU at 3.80GHz. Additionally, the proposed MLOps framework includes model saving and loading functions, gradient accumulation for loss computation to account for model size resulting in fewer images per batch and visualizations.

4.1.2 Image Preprocessing and Augmentation

Image preprocessing and augmentations are crucial to the model's classification performance and can decide the outcome between a well-trained model and an overfitted model. Firstly, the dataset is split with a ratio of 80:20, train and validation split, and is then processed through the Pillow and pandas libraries, complementary to PyTorch, to enable fast

computations thanks to their reliance on numpy's C pre-compiled array code for the image and dataframe metadata. In the beginning, images are loaded based on their name in the dataset folders, and proceed through a series of transformations. The first being image resize, forces all the images to be consistent in size and align with model fixed input sizes. Additionally, we converged on resizes of 224 as they are a good middle-ground between image fidelity and resource consumption. Furthermore, we include a series of random crops and horizontal/vertical flips to add more diversity to the images, as some images are rare and are bound to be redundant. To continue, we also implement Gaussian blur and color jitter to further enhance the robustness of our model. Lastly, the PyTorch dataloader that inputs the data into the model uses 8 workers for speed and pinned GPU memory.

4.1.3 Model Settings

In the present subchapter we will define some of the hyperparameters for the models and why we chose them. For the VQ-VAE model, we utilized the following implementation [115] with the following differences from Darlow et al.'s implementation: embedding dimension = 128, number of embeddings 2048, commitment cost = 0.15, 6 residual layers inside the encoder/decoder with 512 neurons in hidden layers and mean squared error loss. Taking into consideration that our images are large (224x224) and possess complex features, a need for an expanded codebook arises. To combat that, we use a higher amount of dimensions together with a higher number of embeddings. Training pipeline specification utilize, the AdamW optimizer with its regularized computations, 64 images per batch and a 0.001 for learning rate, reduced to approximately 0.0 by a Linear learning

rate scheduler. Model stabilization was achieved by epoch ~200 and achieved high perplexity (codebook usage 70%) and 0.0001 reconstruction loss. After tuning the VQ-VAE, we proceed with the simple classifier used in the adversarial walk to disentangle the confounding bias. We follow Darlow et al.'s proposed simple classifier, which is a plain two layer model of 100 hidden dimension of linear units and relu activation functions. Furthermore, for the final downstream implementation, the three strong classifiers ResNet, DenseNet and EfficientNet, utilized 32 images per batch, neuron dropout with probability of 0.5, a 20 epoch warm-up CosineAnnealing learning rate scheduler, in order to tune the new classifier head and subsequently, fine-tune the model's feature extraction layers. Furthermore, for the focal loss, a brief hyperparameter search resulted in $\alpha=1.5$ and $\gamma=2$ for most models. Additionally, for the adversarial walk parameters, we tried to find the best alpha parameter for the perturbation strength that increases fairness while simultaneously keeping accuracy relatively high. The tested alpha spectrum includes values between 0.01 and 0.07 and walk steps were set to 4 for all models, since a larger for loop would significantly hinder overall performance. Finally, in order to perform model selection, we set F1 score as model criterion as it is a more meaningful score than plain accuracy. This means that, the best saved model with early stopping is derived from F1 score, not accuracy.

4.2 Results

This chapter showcases the experimental results and evaluates their outcomes. We focus on four main testings. The first test evaluation involves the baseline models which include the standard procedure of training pre-trained models like ResNet18, DenseNet121 and EfficientNetB0. Models training includes standard cross entropy loss and a scheduler for reducing learning rate, in addition to utilizing AdamW optimizer. Furthermore, we also evaluate the performance of FairDisCo architecture on the validation data, using the model's custom loss function along with the AdamW optimizer and a Linear learning rate scheduler. Furthermore, we include the basic LAD network, the Linear scheduler and CE as our loss function. Finally, we test our proposed architecture with focal loss and warmup scheduler tuned for 20 epochs of linear scheduler and then a cosine annealing scheduler with $T_{\max} = 50$ and minimum learning rate of $1e-6$. After the tests, we visualize all the grad-cam explanations of the models and compare them, along with their fairness and accuracy metrics of the whole epoch were the model performed the best to reach our conclusions.

4.2.1 Baseline

The standard training procedure across the three pre-trained backbones serves as the baseline for our evaluation. Among these, the ResNet18 architecture achieved high overall efficiency, with accuracy and F1 score reaching 79.00% and 77.00% respectively, after a 400 epoch convergence. While the model exhibited high parity in Equal Opportunity (91.16%) and Accuracy Parity (92.60%), its Disparate Impact (DI) was notably low at 61.82%. Because this DI value falls well below the standard 80% fairness threshold, it demonstrates that despite solid predictive capabilities, the baseline model

incorporates significant confounding bias into its decision-making process. The DenseNet121 architecture, despite being a larger network, with dense skip connections, surpassing resnet in other tasks, its performance on the test data falls short of ResNet's. The model scored a high accuracy of 87.00%, but at the cost of F1 score being at 67.57%, possibly denoting that the model can only classify part of the dataset because of skin labels. For the fairness metrics, DenseNet underperformed on equal opportunity (87.78%) and disparate impact (60.11%) while achieved a slightly higher accuracy parity (93.2%), with high overall accuracy accounting for high AP ratio. Similar to ResNet, due to the low fairness metrics, this baseline model may prove to also be biased in certain predictions, failing to distinguish between classes and disentangle the confounder from the images. Lastly, EfficientNet managed to overtake the other two baseline architectures on both performance and fairness metrics, scoring 87.21% accuracy and 77.25% f1 score, while also dominates among the fairness of the baseline model with 93.11% EO, 61.58% DI and an AP score of 92.98% which is only 0.22% lower than DenseNet. While EfficientNetB0 metrics show increased metrics compared to the rest of the baseline models, the low DI score indicates that this model too may suffer from biased decision-making.

4.2.2 FairDisCo

The second evaluation test assesses the FairDisCo framework, which utilizes adversarial learning to disentangle the skin type along with confusion loss and contrastive loss for feature extraction enhancement. We decided to use the whole FairDisCo network and not the variation without the contrastive loss as it showcased better overall performance in the original paper [104]. This architecture achieved similar classification results as the baseline models with an accuracy of 80.5% and an F1 score of 76.86%, an increase of 1.5% in accuracy and a slight decrease of -0.14% compared to the baseline

ResNet model. Fairness metrics equal opportunity and accuracy parity are over the 90% threshold with 91.35% and 92.18% respectively, while disparate impact score falls short of the even the baseline model with a lower score at 60.29%, a slight but important decrease of -1.53%. Consequently, while FairDisCo achieved a slight gain in overall classification accuracy, it fails to disentangle confounding biases from the data and instead it subtly reinforces the current bias and existing decision disparities as shown in the fairness results.

4.2.3 LAD

The third evaluation examines the Latent Adversarial Debiasing (LAD) designed to mitigate biases present in images by disentangling the latent space of a variational autoencoder using adversarial debiasing methods. We test the proposed model as is without any modifications other than the VQ-VAE encoder in which we used for our reconstructions. The ResNet architecture demonstrates a slight improvement of 3% (82.41% increase) accuracy compared to its baseline counterpart, which degraded the F1 score by almost 7% (yielded 70.14%). From a fairness perspective, model Equal Opportunity increased to 99.65%, a near perfect score that proves that true positives of each group are balanced, while Disparate Impact and Accuracy Parity see a slight decrease from their baseline counterpart with 61.19% and 92.03% respectively. The application of the LAD approach on the ResNet architecture clearly demonstrates an improvement in model fairness as well as a minor increase in overall accuracy, but fails to significantly decrease the level of bias in the data, as the DI metric still underperforms. To continue, the application of LAD on a deeper network like DenseNet121 enhanced fairness metrics, especially EO while slightly decreasing accuracy. The model scored 83.46% in accuracy, an approximate 4% decrease while it significantly increased EO score to 99.73%, a slight increase in DI to 60.54%, and also a

slight decrease in AP to 92.70%. These results denote that while the LAD framework contributes to fairer decisions, it is still not enough to balance the classes. Lastly, the EfficientNet architecture increased its DI by ~8% at 69.20% while all the other metrics suffered decreases, especially EO which had the biggest decrease at 88.60% as well as AP with 91.80%. Among the other baseline model paired with LAD, EfficientNet-B0 is the only architecture heavily increased its DI metric and managed to compete and surpass rivaling architectures in terms of F1 score and accuracy.

4.2.4 XLAD

The fourth and final evaluation inspects the performance of our proposed explainable Latent Adversarial Debiasing model (XLAD). Our model is build on top of the standard LAD architecture, intergrating focal loss among other balancing techniques to dynamically mitigate bias while also minimizing performance losses due to fairness constraints. We examine the three core pretrained architectures, ResNet, DenseNet and EfficientNet. For the ResNet architecture, we observe a noticeable improvement in fairness, at the expense of a trade-off in performance. Specifically, F1-Score decreased by approximately 4% (yielding 72.38%), while overall accuracy reached 81.50%. Despite this minor drop in F1-Score, the model achieved exceptional fairness metrics, recording 96.61% Equal Opportunity (EO), 70.23% Disparate Impact (DI), and 93.80% Accuracy Parity (AP). This performance surpasses both its specific baseline counterpart in addition to the baseline models regarding equitable prediction. Conversely, XLAD with DenseNet backbone demonstrated simultaneous gains in targeted performance and fairness. The proposed methodology achieved improved fairness metrics of 85.22% EO, 73.16% DI, and 91.22% AP, alongside an approximate 4% increase in the F1-Score. This robust performance gain was accompanied by a minor trade-off in overall accuracy, which settled at 85.93%. Ultimately, highlighting the

effectiveness of pairing Focal Loss with LAD in a deeper network architecture such as DenseNet-121. Finally, the EfficientNet architecture showcased a more complex trade-off dynamic between fairness constraints and performance. The model scored an 87.50% in overall accuracy, easily surpassing the baseline, but this was accompanied by a 7% reduction in the F1-Score (resulting in 70.01%). Regarding fairness, the model achieved 78.91% in EO, 81.73% in DI, and 89.67% in AP. While EO and AP experienced a decline of roughly 10% compared to the baseline, the DI metric saw a substantial 20% increase. This significant improvement in demographic parity (DI) underscores the model's enhanced ability to reduce systemic algorithmic bias and improve overall fairness, even when dealing with strict architectural constraints.

4.2.5 Model Comparisons and Visualizations

In Table 1, we summarize the performance and fairness metrics across all evaluated architectures and methodologies. To establish a benchmark, we included the baseline models along with some established State-of-the-art architectures: FairDisCo [104] and PatchAlign [34], the latter of which was evaluated using the authors' configurations and datasets on our tested models. After analyzing these results we uncover some key insights regarding the current landscape of bias in medical imaging. The observed trend across the baseline models, FairDisCo and PatchAlign is their inability to achieve fair demographic parity results. Regardless of the underlying architecture, these models consistently plateau at a DI score of 60.11% to 62.70%. While FairDisCo and PatchAlign achieve competitive accuracy results, they fail to disentangle the biases embedded in the training data. PatchAlign, in particular, suffers from a lower EO score of 72.60%, indicating

a severe imbalance in how it correctly diagnoses true positives across demographic groups. The application of the LAD methodology on the Baseline architectures indicates a partial solution to the problem. LAD successfully optimized EO for ResNet and EfficientNet but failed to impact the DI metric in a meaningful way. This suggests that while EO ensures malignancies are diagnosed equally when present, it still over-predicts or under-predicts the positive class at unequal rates overall across all protected groups. Additionally, with the application of our proposed XLAD framework we significantly reduce the amount of biases decisions made by the models compared to all other methods. By integrating focal loss to handle class imbalances paired with the LAD method, the model is forced to abandon confounding signals. This is proven by the significant increase in DI metrics of all compared architectures where we observe a positive change of roughly 10% in all models. More importantly, the EfficientNetB0 XLAD method, achieves a DI of 81.73%, making it the most bias resilient model among the tested architectures. While models in this category suffer from a moderate decrease in F1 score, the overall accuracy remains highly competitive with most models even surpassing SOTA methodologies. This trade-off is helpful in clinical deployment cases as it sacrifices a minor degree of raw predictive precision in exchange for a more equitable decision making model that does not disadvantage minority subgroups.

	Model	Accuracy (%)	F1 score	EO	DI	AP
Baseline	ResNet18	79.06	77.05	91.16	61.82	92.60
	DenseNet121	87.00	67.57	87.78	60.11	93.20
	EfficientNetb0	87.21	77.25	93.11	61.58	92.98
FairDisCo[104]	VGG-16	80.50	76.86	91.35	60.29	92.18

PatchAlign[34]	Transformer	<u>87.40</u>	-	72.60	62.70	92.70
	ResNet18	82.41	70.14	<u>99.65</u>	61.19	92.03
LAD	DenseNet121	83.46	70.38	<u>99.73</u>	60.54	92.45
	EfficientNetb0	<u>92.18</u>	74.70	88.60	69.20	91.80
XLAD (proposed)	ResNet18	81.50	72.38	<u>96.61</u>	<u>70.23</u>	<u>93.80</u>
	DenseNet121	85.93	71.41	85.22	<u>73.16</u>	91.22
	EfficientNetb0	<u>87.50</u>	70.01	78.91	<u>81.73</u>	89.67

Table 1: Comparison table of all tested architectures. PatchAlign was not manually tested, we instead trained our models with their settings and simply compared them

Beyond the metrics present in Table 1, the efficacy of the model is displayed through visual inspection utilizing Grad-CAM's heatmaps. After extracting spatial information from the model's last convolutional layer, heatmaps are generated (where red is no attention and deep blue is maximum attention). In the baseline models, the Grad-CAM heatmaps reveal that these methods are prone to collider bias. The models often provide their attention to spurious background artifacts that are unrelated to the actual lesions themselves. Dark background, image annotations and artifacts, all impact the models ability to focus on the lesion and instead direct its attention to more unimportant features inside the image, as demonstrated in Figures: Figure 1 Figure 2 Figure 3. This misdirected attention visually explains the low Disparate Impact scores observed in Table 1, as the models learn to associate artifacts with the positive class.



Figure 1: ResNet18 final results. Artifacts and text in images disrupt model predictions



Figure 2: DenseNet121 final results.



Figure 3: EfficientNetB0 final results. Disturbed predictions due to artifacts

Conversely, the Grad-CAM explanations for the XLAD architectures demonstrate a clear spatial shift in diagnostic attention. Following the latent adversarial walk and optimization via Focal Loss, the resulting heatmaps show that XLAD successfully ignores background confounders. The high-activation regions are tightly localized and uniformly concentrated on the

diagnostic characteristics of skin lesions (Figure 5 Figure 4 Figure 7 Figure 6). This visual comparison further enhances our claim that the XLAD framework can effectively reduce bias in unfair settings. It validates that the significant increases in Disparate Impact (DI) and Equal Opportunity (EO) are not statistical anomalies achieved through the loss function. Instead, the model has genuinely learned to disentangle and discard biased artifacts, grounding its classifications in true clinical pathology. This alignment between high quantitative fairness and interpretable, medically sound visual reasoning represents a substantial step toward trustworthy AI in dermatology.



Figure 4: XLAD ResNet18 architecture. The model does not rely on artifacts or text on the images.



Figure 5: XLAD ResNet18 part 2.



Figure 6: XLAD EfficientNetB0 architecture. The model does not rely on artifacts or text on the image.

Figure 7: XLAD DenseNet121 architecture. The model does not rely on artifacts or text on the images.

5 Conclusions & Future Work

5.1 Conclusion

In conclusion, we investigated fair and explainable approaches to mitigate skin tone bias from the skin lesion dataset Fitzpatrick17k. We established the foundations of medical imaging and AI in dermatology and surveyed applied methodologies applied on the subject both modern and traditional. Furthermore, we implored the Latent Adversarial Debiasing framework with some modifications along with Focal Loss and Grad-CAM explaining to propose a framework that provides model explanations based on the attention of the model and identify if it focuses on unrelated correlations like skin tone or actual lesion features. Secondly, we also examined and improved model fairness by incorporating fairness metrics in our evaluations like Equal Opportunity, Disparate Impact and Accuracy Parity, and also by using LAD's debiasing logic to create a robust framework for fair lesion classification tasks.

5.2 Future Work

Building upon these findings, several promising directions exist for future research. First, while this study focused on the Fitzpatrick17k dataset, expanding and evaluating the proposed framework across diverse, multi-center repositories will be essential to verify its cross-demographic generalization ability. Second, future work could advance beyond visual attribution maps like Grad-CAM by integrating quantitative explainability metrics and counterfactual explanations, paired with human-in-the-loop reader studies involving clinical dermatologists to assess diagnostic trust.

Moreover, future work could also extend the current methodology by implementing a more tailored loss for the current problem, possibly by adding fairness constraints directly in the model's loss calculations. Finally, while this work is still in early stages, it provides strong computational efficiency due to the usage of the light weight Grad-CAM model explanations. Clinicians could implement this framework and inference the model on local machines in hospitals to help doctors diagnose patients more accurately and efficiently without the for expensive servers and hardware.

6 Acknowledgements

To end this master's thesis, I would like to dedicate this chapter to all the people that made this work possible. First, I would like to thank my parents for supporting me throughout all my years as a student and helped me reach my goals. I would also like to thank my partner which has been there for me since the beginning, being at my side and supporting me always. I would like to thank all my friends, old and new, that have helped me each in their own way to continue and stay determined. I would like to thank my professors, especially Dr. Konstantinos Kolomvatsos for being a great mentor and helping me understand everything related to our field. If it were not for his enthusiasm and excellent way of teaching all his subjects, I would not be writing this thesis today. Words cannot express the appreciation I feel for everyone that has been with me all these years, but I can surely say, I could not have done it without them.

References

- [1] “Cancer.” [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/cancer>
- [2] “Cancer Statistics - NCI.” [Online]. Available: <https://www.cancer.gov/about-cancer/understanding/statistics>
- [3] G. P. Guy, S. R. Machlin, D. U. Ekwueme, and K. R. Yabroff, “Prevalence and costs of skin cancer treatment in the U.S., 2002-2006 and 2007-2011,” *Am. J. Prev. Med.*, vol. 48, no. 2, pp. 183–187, Feb. 2015, doi: 10.1016/J.AMEPRE.2014.08.036.
- [4] N. Alipour, T. Burke, and J. Courtney, “Skin Type Diversity in Skin Lesion Datasets: A Review,” *Curr. Dermatol. Rep.*, vol. 13, no. 3, p. 198, Sep. 2024, doi: 10.1007/S13671-024-00440-0.
- [5] L. Darlow, S. Jastrzębski, and A. Storkey, “Latent Adversarial Debiasing: Mitigating Collider Bias in Deep Neural Networks”.
- [6] H. Sheikh, C. Prins, and E. Schrijvers, “Artificial Intelligence: Definition and Background,” pp. 15–41, 2023, doi: 10.1007/978-3-031-21448-6_2.
- [7] K. Shailaja, B. Seetharamulu, and M. A. Jabbar, “Machine Learning in Healthcare: A Review,” *Proc. 2nd Int. Conf. Electron. Commun. Aerosp. Technol. ICECA 2018*, pp. 910–914, Sep. 2018, doi: 10.1109/ICECA.2018.8474918.
- [8] A. Rajkomar, J. Dean, and I. Kohane, “Machine Learning in Medicine,” *N. Engl. J. Med.*, vol. 380, no. 14, pp. 1347–1358, Apr. 2019, doi: 10.1056/NEJMRA1814259/SUPPL_FILE/NEJMRA1814259_DISCLOSURE_S.PDF.
- [9] R. C. Deo, “Machine learning in medicine,” *Circulation*, vol. 132, no. 20, pp. 1920–1930, Nov. 2015, doi: 10.1161/CIRCULATIONAHA.115.001593/ASSET/F82A0780-3EFF-4CE8-839D-5FF553B00121/ASSETS/GRAPHIC/1920FIG04.JPEG.
- [10] W. Wang *et al.*, “Medical Image Classification Using Deep Learning,” 2020, doi: 10.1007/978-3-030-32606-7_3.
- [11] H. P. Chan, R. K. Samala, L. M. Hadjiiski, and C. Zhou, “Deep Learning in Medical Image Analysis,” *Adv. Exp. Med. Biol.*, vol. 1213, p. 3, 2020, doi: 10.1007/978-3-030-33128-3_1.
- [12] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, “A review of deep learning based methods for medical image multi-organ

- segmentation,” *Phys. Med.*, vol. 85, pp. 107–122, May 2021, doi: 10.1016/J.EJMP.2021.05.003.
- [13] K. Suzuki, “Overview of deep learning in medical imaging,” *Radiol. Phys. Technol.*, vol. 10, no. 3, pp. 257–273, Sep. 2017, doi: 10.1007/S12194-017-0406-5/FIGURES/5.
- [14] X. Liu, L. Song, S. Liu, and Y. Zhang, “A Review of Deep-Learning-Based Medical Image Segmentation Methods,” *Sustain. 2021 Vol 13 Page 1224*, vol. 13, no. 3, p. 1224, Jan. 2021, doi: 10.3390/SU13031224.
- [15] R. Radhika and R. Mahajan, “Medical Image Enhancement: A Review,” *Lect. Notes Netw. Syst.*, vol. 288, pp. 105–118, 2022, doi: 10.1007/978-981-16-5120-5_9/TABLES/3.
- [16] D. Bank, N. Koenigstein, and R. Giryes, “Autoencoders,” *Mach. Learn. Data Sci. Handb. Data Min. Knowl. Discov. Handb. Third Ed.*, pp. 353–374, Jan. 2023, doi: 10.1007/978-3-031-24628-9_16/FIGURES/8.
- [17] Y.-W. Chen and L. C. Jain, “Intelligent Systems Reference Library 171 Deep Learning in Healthcare Paradigms and Applications”, [Online]. Available: <http://www.springer.com/series/8578>
- [18] I. J. Goodfellow *et al.*, “Generative Adversarial Nets,” *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, [Online]. Available: <http://www.github.com/goodfeli/adversarial>
- [19] A. Hosna, E. Merry, J. Gyalmo, Z. Alom, Z. Aung, and M. A. Azim, “Transfer learning: a friendly introduction,” *J. Big Data*, vol. 9, no. 1, pp. 1–19, Dec. 2022, doi: 10.1186/S40537-022-00652-W/FIGURES/6.
- [20] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, “Deep learning for healthcare: review, opportunities and challenges,” *Brief. Bioinform.*, vol. 19, no. 6, p. 1236, May 2017, doi: 10.1093/BIB/BBX044.
- [21] K. Mavrogiorgos, A. Kiourtis, A. Mavrogiorgou, A. Menychtas, and D. Kyriazis, “Bias in Machine Learning: A Literature Review,” *Appl. Sci. 2024 Vol 14 Page 8860*, vol. 14, no. 19, p. 8860, Oct. 2024, doi: 10.3390/APP14198860.
- [22] L. H. Nazer *et al.*, “Bias in artificial intelligence algorithms and recommendations for mitigation,” *PLOS Digit. Health*, vol. 2, no. 6, p. e0000278, Jun. 2023, doi: 10.1371/JOURNAL.PDIG.0000278.
- [23] D. Roselli, J. Matthews, and N. Talagala, “Managing bias in AI,” *Web Conf. 2019 - Companion World Wide Web Conf. WWW 2019*, pp. 539–544, May 2019, doi: 10.1145/3308560.3317590;PAGEGROUP:STRING:PUBLICATION.
- [24] M. A. Sufian, L. Alsadder, W. Hamzi, S. Zaman, A. S. M. S. Sagar, and B. Hamzi, “Mitigating Algorithmic Bias in AI-Driven Cardiovascular Imaging for Fairer Diagnostics,” *Diagnostics*, vol. 14, no. 23, p. 2675, Dec. 2024, doi: 10.3390/DIAGNOSTICS14232675.

- [25] D. Dhabliya, S. S. Dari, A. Dhablia, N. Akhila, V. Khetani, and A. Professor, “Addressing Bias in Machine Learning Algorithms: Promoting Fairness and Ethical Design”, doi: 10.1051/e3sconf/202449102040.
- [26] M. Benčević, M. Habijan, I. Galić, D. Babin, and A. Pižurica, “Understanding skin color bias in deep learning-based skin lesion segmentation,” *Comput. Methods Programs Biomed.*, vol. 245, p. 108044, Mar. 2024, doi: 10.1016/J.CMPB.2024.108044.
- [27] T. Kalb, K. Kushibar, C. Cintas, K. Lekadir, O. Diaz, and R. Osuala, “Revisiting Skin Tone Fairness in Dermatological Lesion Classification,” *Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.*, vol. 14242 LNCS, pp. 246–255, 2023, doi: 10.1007/978-3-031-45249-9_24.
- [28] P. J. Bevan and A. Atapour-Abarghouei, “Detecting Melanoma Fairly: Skin Tone Detection and Debiasing for Skin Lesion Classification,” *Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.*, vol. 13542 LNCS, pp. 1–11, 2022, doi: 10.1007/978-3-031-16852-9_1/TABLES/3.
- [29] A. A. Abdulredah, M. A. Fadhel, L. Alzubaidi, Y. Duan, M. Kherallah, and F. Charfi, “Towards unbiased skin cancer classification using deep feature fusion,” *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, pp. 1–22, Dec. 2025, doi: 10.1186/S12911-025-02889-W/FIGURES/16.
- [30] A. Pundhir, S. Verma, and B. Raman, “Towards Ethical Dermatology: Mitigating Bias in Skin Condition Classification,” *Proc. Int. Jt. Conf. Neural Netw.*, 2024, doi: 10.1109/IJCNN60899.2024.10650487.
- [31] A. Bissoto, M. Fornaciali, E. Valle, and S. Avila, “(De)Constructing Bias on Skin Lesion Datasets,” *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshop*, vol. 2019-June, pp. 2766–2774, Apr. 2019, doi: 10.1109/CVPRW.2019.00335.
- [32] C. Albuquerque, R. Henriques, and M. Castelli, “Deep learning-based object detection algorithms in medical imaging: Systematic review,” *Heliyon*, vol. 11, no. 1, p. e41137, Jan. 2025, doi: 10.1016/J.HELİYON.2024.E41137.
- [33] B. R. Solunke and S. R. Gengaje, “A Review on Traditional and Deep Learning based Object Detection Methods,” *2023 Int. Conf. Emerg. Smart Comput. Inform. ESCI 2023*, 2023, doi: 10.1109/ESCI56872.2023.10099639.
- [34] Aayushman, H. Gaddey, V. Mittal, M. Chawla, and G. R. Gupta, “Fair and Accurate Skin Disease Image Classification by Alignment with Clinical Labels,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, and J. A. Schnabel, Eds.,

- Cham: Springer Nature Switzerland, 2024, pp. 394–404. doi: 10.1007/978-3-031-72378-0_37.
- [35] R. Fakoor, F. Ladhak, the ... A. N.-P., and undefined 2013, “Using deep learning to enhance cancer diagnosis and classification,” *AdmistongjieducnR Fakoor F Ladhak Nazi M HuberProceedings Int. Conf. Mach. Learn. 2013•admistongjieducn*, 2013, [Online]. Available: https://admis.tongji.edu.cn/_upload/article/files/a1/28/1a8d7ac94b05860e2d1fe6c3bf2a/dffeeaae-5d26-4e84-9b60-42bb4b47ee41.pdf
 - [36] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, “Transfer learning for medical image classification: a literature review,” *BMC Med. Imaging* 2022 221, vol. 22, no. 1, pp. 1–13, Apr. 2022, doi: 10.1186/S12880-022-00793-7.
 - [37] N. Nayak, H. Chang, A. Borowsky, P. Spellman, and B. Parvin, “Classification of tumor histopathology via sparse feature learning,” *Proc. - Int. Symp. Biomed. Imaging*, pp. 410–413, 2013, doi: 10.1109/ISBI.2013.6556499.
 - [38] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, “Medical image classification with convolutional neural network,” *2014 13th Int. Conf. Control Autom. Robot. Vis. ICARCV 2014*, pp. 844–848, 2014, doi: 10.1109/ICARCV.2014.7064414.
 - [39] B. Shi *et al.*, “Prediction of Occult Invasive Disease in Ductal Carcinoma in Situ Using Deep Learning Features,” *J. Am. Coll. Radiol.*, vol. 15, no. 3, pp. 527–534, Mar. 2018, doi: 10.1016/J.JACR.2017.11.036.
 - [40] M. Talo, “Automated classification of histopathology images using transfer learning,” *Artif. Intell. Med.*, vol. 101, p. 101743, Nov. 2019, doi: 10.1016/J.ARTMED.2019.101743.
 - [41] Z. Guo, X. Li, H. Huang, N. Guo, and Q. Li, “Deep learning-based image segmentation on multimodal medical imaging,” *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 3, no. 2, pp. 162–169, Mar. 2019, doi: 10.1109/TRPMS.2018.2890359.
 - [42] A. Esteva *et al.*, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nat.* 2017 5427639, vol. 542, no. 7639, pp. 115–118, Jan. 2017, doi: 10.1038/nature21056.
 - [43] J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognit. CVPR 2017*, vol. 2017-January, pp. 6517–6525, Dec. 2016, doi: 10.1109/CVPR.2017.690.
 - [44] R. Girshick, “Fast R-CNN.” pp. 1440–1448, 2015. [Online]. Available: <https://github.com/rbgirshick/>
 - [45] W. Liu *et al.*, “SSD: Single shot multibox detector,” *Lect. Notes Comput. Sci. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.*, vol. 9905 LNCS, pp. 21–37, 2016, doi: 10.1007/978-3-319-46448-0_2/FIGURES/5.

- [46] R. Radhika and R. Mahajan, "Medical Image Enhancement: A Review", doi: 10.1007/978-981-16-5120-5_9.
- [47] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, Apr. 1980, doi: 10.1007/BF00344251/METRICS.
- [48] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998, doi: 10.1109/5.726791.
- [49] K. O'shea and R. Nash, "An Introduction to Convolutional Neural Networks".
- [50] D. Dai, "An Introduction of CNN: Models and Training on Neural Network Models," *Proc. - 2021 Int. Conf. Big Data Artif. Intell. Risk Manag. ICBAR 2021*, pp. 135–138, 2021, doi: 10.1109/ICBAR55169.2021.00037.
- [51] J. Wu, "Introduction to Convolutional Neural Networks," 2017.
- [52] "Deep Learning: An Overview of Convolutional Neural Network(CNN)," 2020.
- [53] C. Sun *et al.*, "Gastric histopathology image segmentation using a hierarchical conditional random field," *Biocybern. Biomed. Eng.*, vol. 40, no. 4, pp. 1535–1555, Oct. 2020, doi: 10.1016/J.BBE.2020.09.008.
- [54] Q. Li, W. Cai, X. Wang, Y. Zhou, D. D. Feng, and M. Chen, "Medical Image Classification with Convolutional Neural Network".
- [55] E. Oja, "Simplified neuron model as a principal component analyzer," *J. Math. Biol.*, vol. 15, no. 3, pp. 267–273, 1982, doi: 10.1007/BF00275687/METRICS.
- [56] K. P. F.R.S, "LIII. On lines and planes of closest fit to systems of points in space," *Lond. Edinb. Dublin Philos. Mag. J. Sci.*, vol. 2, no. 11, pp. 559–572, Nov. 1901, doi: 10.1080/14786440109462720.
- [57] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Internal Representations by Error Propagation," Sep. 1985, doi: 10.21236/ADA164453.
- [58] U. Michelucci, "An Introduction to Autoencoders," Jan. 2022, [Online]. Available: <https://arxiv.org/pdf/2201.03898>
- [59] A. Creswell, K. Arulkumaran, and A. A. Bharath, "On denoising autoencoders trained to minimise binary cross-entropy," Aug. 2017, [Online]. Available: <https://arxiv.org/pdf/1708.08487>
- [60] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy Image Compression with Compressive Autoencoders," *5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc.*, Mar. 2017, [Online]. Available: <https://arxiv.org/pdf/1703.00395>

- [61] Y. Zhang, “A Better Autoencoder for Image: Convolutional Autoencoder”.
- [62] M. A. Kramer, “Autoassociative neural networks,” *Comput. Chem. Eng.* vol. 16, no. 4, pp. 313–328, Apr. 1992, doi: 10.1016/0098-1354(92)80051-A.
- [63] A. van den O. DeepMind, O. V. DeepMind, and K. K. DeepMind, “Neural Discrete Representation Learning,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
- [64] R. Miotto, L. Li, B. A. Kidd, J. D.-S. reports, and undefined 2016, “Deep patient: an unsupervised representation to predict the future of patients from the electronic health records,” *NaturecomR Miotto Li BA Kidd JT DudleyScientific Rep. 2016•naturecom*, vol. 6, May 2016, doi: 10.1038/srep26094.
- [65] R. Miotto, L. Li, and J. T. Dudley, “Deep learning to predict patient future diseases from the electronic health records,” *SpringerR Miotto Li JT DudleyEuropean Conf. Inf. Retr. 2016•Springer*, vol. 9626, pp. 768–774, 2016, doi: 10.1007/978-3-319-30671-1_66.
- [66] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, “Generative Adversarial Networks: An Overview,” *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018, doi: 10.1109/MSP.2017.2765202.
- [67] A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” *4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc.*, Nov. 2015, [Online]. Available: <https://arxiv.org/pdf/1511.06434>
- [68] M. Mirza and S. Osindero, “Conditional Generative Adversarial Nets,” Nov. 2014, [Online]. Available: <https://arxiv.org/pdf/1411.1784>
- [69] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-December, pp. 770–778, Dec. 2015, doi: 10.1109/CVPR.2016.90.
- [70] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely Connected Convolutional Networks.” pp. 4700–4708, 2017. [Online]. Available: <https://github.com/liuzhuang13/DenseNet>.
- [71] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.” PMLR, pp. 6105–6114, May 2019. [Online]. Available: <https://proceedings.mlr.press/v97/tan19a.html>
- [72] S. Takahashi *et al.*, “Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review,” *J. Med. Syst.*, vol. 48, no. 1, p. 84, Sep. 2024, doi: 10.1007/s10916-024-02105-8.

- [73] F. Doshi-Velez and B. Kim, “Towards A Rigorous Science of Interpretable Machine Learning,” Mar. 02, 2017, *arXiv*: arXiv:1702.08608. doi: 10.48550/arXiv.1702.08608.
- [74] N. Nigar, M. Umar, M. K. Shahzad, S. Islam, and D. Abalo, “A Deep Learning Approach Based on Explainable Artificial Intelligence for Skin Lesion Classification,” *IEEE Access*, vol. 10, pp. 113715–113725, 2022, doi: 10.1109/ACCESS.2022.3217217.
- [75] F. P. Calmon, D. Wei, K. N. Ramamurthy, and K. R. Varshney, “Optimized Data Pre-Processing for Discrimination Prevention”.
- [76] R. Zemel, Y. (Ledell,) Wu, K. Swersky, T. Pitassi, and C. Dwork, “Learning Fair Representations,” 2013.
- [77] M. Du, F. Yang, N. Zou, and X. Hu, “Fairness in Deep Learning: A Computational Perspective,” *IEEE Intell. Syst.*, vol. 36, no. 4, pp. 25–34, Jul. 2021, doi: 10.1109/MIS.2020.3000681.
- [78] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, “A Reductions Approach to Fair Classification,” Jul. 16, 2018, *arXiv*: arXiv:1803.02453. doi: 10.48550/arXiv.1803.02453.
- [79] A. C. Scott, W. J. Clancey, R. Davis, and E. H. Shortliffe, “Explanation Capabilities of Production-Based Consultation Systems,” *Am. J. Comput. Linguist.*, pp. 1–50, Feb. 1977.
- [80] 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*, in KDD ’16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.
- [81] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Aug. 29, 2026. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
- [82] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek, “Explainable AI Methods - A Brief Overview,” in *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, Eds., Cham: Springer International Publishing, 2022, pp. 13–38. doi: 10.1007/978-3-031-04083-2_2.
- [83] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, “Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges,” in *Natural Language Processing and Chinese Computing*,

- J. Tang, M.-Y. Kan, D. Zhao, S. Li, and H. Zan, Eds., Cham: Springer International Publishing, 2019, pp. 563–574. doi: 10.1007/978-3-030-32236-6_51.
- [84] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning Deep Features for Discriminative Localization,” presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929. Accessed: Aug. 29, 2026. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Zhou_Learning_Deep_Features_CVPR_2016_paper.html
- [85] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, “Grad-CAM: Why did you say that?,” Jan. 25, 2017, *arXiv*: arXiv:1611.07450. doi: 10.48550/arXiv.1611.07450.
- [86] Z. Cheng, Y. Wu, Y. Li, L. Cai, and B. Ihnaini, “A Comprehensive Review of Explainable Artificial Intelligence (XAI) in Computer Vision,” *Sensors*, vol. 25, no. 13, p. 4166, Jan. 2025, doi: 10.3390/s25134166.
- [87] D. Ho *et al.*, “Enabling Technologies for Personalized and Precision Medicine,” *Trends Biotechnol.*, vol. 38, no. 5, pp. 497–518, May 2020, doi: 10.1016/j.tibtech.2019.12.021.
- [88] Z. Li, K. C. Koban, T. L. Schenck, R. E. Giunta, Q. Li, and Y. Sun, “Artificial Intelligence in Dermatology Image Analysis: Current Developments and Future Trends,” *J. Clin. Med.*, vol. 11, no. 22, p. 6826, Jan. 2022, doi: 10.3390/jcm11226826.
- [89] L. N. Montoya, J. S. Roberts, and B. S. Hidalgo, “Towards Fairness in AI for Melanoma Detection: Systemic Review and Recommendations,” in *Advances in Information and Communication*, K. Arai, Ed., Cham: Springer Nature Switzerland, 2025, pp. 320–341. doi: 10.1007/978-3-031-84460-7_21.
- [90] T. Chanda *et al.*, “Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma,” *Nat. Commun.*, vol. 15, no. 1, p. 524, Jan. 2024, doi: 10.1038/s41467-023-43095-4.
- [91] K. Bera, K. A. Schalper, D. L. Rimm, V. Velcheti, and A. Madabhushi, “Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology,” *Nat. Rev. Clin. Oncol.*, vol. 16, no. 11, pp. 703–715, Nov. 2019, doi: 10.1038/s41571-019-0252-y.
- [92] H. A. Haenssle *et al.*, “Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists,” *Ann. Oncol.*, vol. 29, no. 8, pp. 1836–1842, Aug. 2018, doi: 10.1093/ANNONC/MDY166.

- [93] “ISIC Challenge.” [Online]. Available: <https://challenge.isic-archive.com/>
- [94] “About | ISIC.” [Online]. Available: <https://www.isic-archive.com/mission>
- [95] M. Rokuss *et al.*, “LesionLocator: Zero-Shot Universal Tumor Segmentation and Tracking in 3D Whole-Body Imaging,” in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2025, pp. 30872–30885. doi: 10.1109/CVPR52734.2025.02875.
- [96] D. Ahmedt-Aristizabal *et al.*, “Monitoring of Pigmented Skin Lesions Using 3D Whole Body Imaging,” *Comput. Methods Programs Biomed.*, vol. 232, p. 107451, Apr. 2023, doi: 10.1016/j.cmpb.2023.107451.
- [97] Z. Mirikharaji, K. Abhishek, S. Izadi, and G. Hamarneh, “D-LEMA: Deep Learning Ensembles from Multiple Annotations - Application to Skin Lesion Segmentation,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2021, pp. 1837–1846. doi: 10.1109/CVPRW53098.2021.00203.
- [98] T. L. Diepgen *et al.*, “Guidelines for diagnosis, prevention and treatment of hand eczema”, doi: 10.1111/ddg.12510_1.
- [99] A. Mahbod, G. Schaefer, C. Wang, G. Dorffner, R. Ecker, and I. Ellinger, “Transfer learning using a multi-scale and multi-network ensemble for skin lesion classification,” *Comput. Methods Programs Biomed.*, vol. 193, p. 105475, Sep. 2020, doi: 10.1016/j.cmpb.2020.105475.
- [100] H. Rashid, M. A. Tanveer, and H. Aqeel Khan, “Skin Lesion Classification Using GAN based Data Augmentation,” in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2019, pp. 916–919. doi: 10.1109/EMBC.2019.8857905.
- [101] C.-J. Wei *et al.*, “A multicenter study of neurofibromatosis type 1 utilizing deep learning for whole body tumor identification,” *Npj Digit. Med.*, vol. 8, no. 1, p. 56, Jan. 2025, doi: 10.1038/s41746-025-01454-z.
- [102] R. Erukulla *et al.*, “Deep learning-based classification of fungal and Acanthamoeba keratitis using confocal microscopy,” *Ocul. Surf.*, vol. 38, pp. 203–208, Oct. 2025, doi: 10.1016/j.jtos.2025.07.012.
- [103] T. Kalb, K. Kushibar, C. Cintas, K. Lekadir, O. Diaz, and R. Osuala, “Revisiting Skin Tone Fairness in Dermatological Lesion Classification,” in *Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging*, S. Wesarg, E. Puyol Antón, J. S. H. Baxter, M. Erdt, K. Drechsler, C. Oyarzun Laura, M. Freiman, Y. Chen, I. Rekik, R. Eagleson, A. Feragen, A. P. King, V. Cheplygina, M. Ganz-Benaminsen,

- E. Ferrante, B. Glocker, D. Moyer, and E. Petersen, Eds., Cham: Springer Nature Switzerland, 2023, pp. 246–255. doi: 10.1007/978-3-031-45249-9_24.
- [104] S. Du, B. Hers, N. Bayasi, G. Hamarneh, and R. Garbi, “FairDisCo: Fairer AI in Dermatology via Disentanglement Contrastive Learning,” Aug. 22, 2022, *arXiv*: arXiv:2208.10013. doi: 10.48550/arXiv.2208.10013.
- [105] X. Li, Z. Cui, Y. Wu, L. Gu, and T. Harada, “Estimating and Improving Fairness with Adversarial Learning,” May 11, 2021, *arXiv*: arXiv:2103.04243. doi: 10.48550/arXiv.2103.04243.
- [106] L. Darlow, S. Jastrzębski, and A. Storkey, “Latent Adversarial Debiasing: Mitigating Collider Bias in Deep Neural Networks”.
- [107] S. Abbas, F. Ahmed, W. A. Khan, M. Ahmad, M. A. Khan, and T. M. Ghazal, “Intelligent skin disease prediction system using transfer learning and explainable artificial intelligence,” *Sci. Rep.*, vol. 15, no. 1, p. 1746, Jan. 2025, doi: 10.1038/s41598-024-83966-4.
- [108] G. Munjal, P. Bhardwaj, V. Bhargava, S. Singh, and N. Nagpal, “SkinSage XAI: An explainable deep learning solution for skin lesion diagnosis,” *Health Care Sci.*, vol. 3, no. 6, pp. 438–455, 2024, doi: 10.1002/hcs2.121.
- [109] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Aug. 27, 2026. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
- [110] S. Anand *et al.*, “CHASHNIt for enhancing skin disease classification using GAN augmented hybrid model with LIME and SHAP based XAI heatmaps,” *Sci. Rep.*, vol. 15, no. 1, p. 31138, Aug. 2025, doi: 10.1038/s41598-025-13647-3.
- [111] T. Khater, S. Ansari, S. Mahmoud, A. Hussain, and H. Tawfik, “Skin cancer classification using explainable artificial intelligence on pre-extracted image features,” *Intell. Syst. Appl.*, vol. 20, p. 200275, Nov. 2023, doi: 10.1016/j.iswa.2023.200275.
- [112] N. S. Naz *et al.*, “Privacy preserving skin cancer diagnosis through federated deep learning and explainable AI,” *Sci. Rep.*, vol. 15, no. 1, p. 36094, Oct. 2025, doi: 10.1038/s41598-025-19905-8.
- [113] M. Groh *et al.*, “Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2021, pp. 1820–1828. doi: 10.1109/CVPRW53098.2021.00201.

- [114] S. Caton and C. Haas, “Fairness in Machine Learning: A Survey,” *ACM Comput Surv*, vol. 56, no. 7, p. 166:1-166:38, Apr. 2024, doi: 10.1145/3616865.
- [115] “vq-vae.ipynb - Colab.” [Online]. Available: <https://colab.research.google.com/github/zalandoresearch/pytorch-vq-vae/blob/master/vq-vae.ipynb#scrollTo=AV0X-P5m6HBM>