
Automatic Skin Disease Classification: A Review

Ilenia Carboni

Master's Degree in Artificial Intelligence
Alma Mater Studiorum - University of Bologna
ilenia.carboni3@studio.unibo.it

Abstract

Artificial Intelligence is rapidly transforming healthcare, with dermatology poised to benefit significantly from automated tools for skin disease classification. This study explores the current state of AI research in dermatology, focusing on the challenges and advancements in automatic skin disease detection. A key issue is the lack of diverse, representative datasets that can ensure fairness and accuracy across different demographics, such as age, gender and skin tones. Additionally, legal and ethical constraints on data sharing complicate the development of comprehensive datasets. This systematic literature review evaluates approximately 20 recent studies, addressing two critical questions: the availability of public datasets tailored for skin disease classification, and how current research accounts for diversity to ensure trust and fairness in AI models. The findings aim to contribute to the ongoing development of accessible, equitable and reliable AI-driven dermatological care.

1 Introduction

Artificial Intelligence (AI) has emerged as a transformative tool in various domains, with healthcare being one of the most promising areas of application. In particular, dermatology stands to benefit significantly from the integration of AI technologies, especially in the automated analysis of skin conditions. This is crucial given the increasing demand for dermatological care and the ongoing challenges in the field.

Globally, an estimated 3 billion people lack adequate access to medical care for skin diseases[7]. The shortage and unequal distribution of dermatologists exacerbate the issue, often leading to prolonged wait times for skin evaluation, causing delayed diagnoses and increased mortality rates. AI-based diagnostics tools hold the potential to address these disparities by supporting the triage decisions, assisting general practitioners in diagnosing skin diseases and enabling early identification of malignancies. These advancements have gained significant attention in recent years, sparking rapid development in the field of automated skin disease detection.

However, the adoption of AI in dermatology faces several challenges. One of the primary hurdles is the considerable variability in the presentation of skin diseases across different demographics, including age, gender and skin tones. This variability underscores the need for diverse, representative datasets to train AI models that can reliably detect and diagnose conditions across all patient groups. Furthermore, ensuring user trust and minimizing the risk of misdiagnosis are critical for the successful implementation of AI systems in clinical settings. A key aspect of this is developing strategies that enhance fairness, transparency and accountability in AI-based dermatological tools.

Another challenge is the complexity of building comprehensive, general-use datasets. The use of medical images is often subject to stringent regulations that govern data sharing, making it difficult to create large, diverse datasets for training AI models. Ethical and legal constraints further complicate the process, as patient privacy must be maintained and consent obtained for the use of medical data.

The objective of this study is to conduct a systematic literature review to assess the current state of research in automatic skin disease classification, focusing on the key developments in the field in recent years. Specifically, the study aims to answer two central questions:

- Are there publicly available datasets tailored specifically for skin disease classification?
- How does current research in Automatic Skin Disease detection accounts for the diversity present in dermatology, ensuring fairness and inclusivity in AI models?

In order to address these questions, this review covers approximately 20 papers from reputable journals and conferences in the fields of Biomedical Imaging and Signal Processing. By synthesizing the findings of these studies, the review will offer a comprehensive overview of the major advancements in automatic skin disease detection and provide insights into the strategies being developed to overcome existing challenges. This work aims to contribute to the ongoing effort to make AI-driven dermatological care more accessible, equitable and reliable for all patients.

2 Gather Studies, Selection and Refinement

This section outlines the process followed for gathering, selecting and refining the studies included in this review.

2.1 Study Collection

Relevant studies were retrieved from Scopus and IEEE Xplore, using predefined search queries. The tested queries that gave positive results were:

- (IEEEXplore): (Skin Disease Detection) $\wedge$ (Datasets) $\wedge$ (Fairness)
- (IEEEXplore): (Dermatology Imaging) $\wedge$ (Data Bias) $\wedge$ (Demographic Bias)
- (IEEEXplore): (Skin Disease Detection) $\wedge$ (Public Datasets)
- (Scopus): $ALL((\text{skin disease detection} \vee \text{dermatology imaging} \vee \text{skin lesion analysis}) \wedge (\text{datasets} \vee \text{data bias} \vee \text{dataset diversity} \vee \text{public datasets}) \wedge (\text{fairness} \vee \text{demographic bias} \vee \text{bias mitigation} \vee \text{ethical considerations})) \wedge \text{PUBYEAR} > 2017 \wedge \text{PUBYEAR} < 2024$

Only open-access papers or those accessible with a license from the University of Bologna were considered. Additionally, only studies published after 2014 were included, with a stricter constraint in Scopus to refine the research scope. Furthermore, all related works available in the ISIC Archive[14] that met the publication year and availability criteria were included.

To ensure relevance, an initial screening on the 103 resulting papers was conducted. The abstracts of the retrieved papers were reviewed to confirm they focused Skin Disease Classification in humans. Plus, duplicate studies were removed.

After this process, a total of 52 papers were shortlisted for further analysis.

2.2 Study Selection and Criteria

For final selection, the introduction of each shortlisted paper was reviewed to determine its main topics. Studies were included if they explicitly addressed one or more of the following:

- Skin Disease Recognition and Classification in Humans;
- Dataset construction;
- Fairness considerations in dataset curation or classification models.

Selected papers were cataloged in a structured table capturing metatata such as title, DOI, year and venue, facilitating further analysis.

Following this screening, a total of 19 papers were selected for inclusion in the review.

2.3 Additional Studies

To ensure comprehensive coverage of notable results in the field, the reference lists of the 19 selected papers were examined for significant citations. This additional check aimed to mitigate potential gaps in the initial query search.

As a result of this process, four more papers recognized as notable works, complying with both publication year and availability criteria were identified and included, bringing the total number of selected studies to 23.

An abridged version of the study selection table is available in the supplemental materials in the GitLab repository.

3 Datasets

3.1 Introduction

Melanoma, a skin cancer, occurs on the skin surface, and can be detected through simple visual examination. To support diagnostic performance, dermoscopy has been introduced - an imaging technique that eliminates surface reflection of skin, allowing deeper layers to be visually enhanced. This improves recognition performance by approximately 50%[12]. As access to dermatological care remains inadequate, even in developed countries, there is a growing interest in Artificial Intelligence (AI) diagnostics and decision-support tools to aid in lesion triage and assist non-expert clinicians in identifying potential malignancies[7]. A common challenge in AI-driven dermatology is the reliance on siloed, private clinical image datasets. Building large medical datasets presents ethical, regulatory and legal complexities particularly with the increasing involvement of major technology companies in the healthcare domain[8]. Recent studies on predicting clinical outcomes in patients with COVID-19 used Federated Learning (FL)[8] to train models on disparate clinical sources, without the need for data to be transported or exposed outside their original location.

However, no studies in the search mention FL as a potential method for expanding training datasets for skin disease classification models. Instead, efforts have focused on establishing large-scale, centralized and institutionally coordinated public skin image datasets, particularly through the International Skin Imaging Collaboration (ISIC)[12]. ISIC also hosts an annual challenge to support the development of automated melanoma diagnosis algorithms using dermoscopic images.

3.2 The ISIC Challenges at ISBI

To advance research on automated skin disease diagnosis, particularly for melanoma and skin cancer, the International Skin Imaging Collaboration (ISIC) has hosted a challenge at the International Symposium on Biomedical Imaging (ISBI) since 2016. The challenge focuses on three key tasks in lesion analysis:

- Lesion Segmentation.
- Lesion Dermoscopic Feature Detection (identifying clinically defined patterns).
- Lesion Classification (distinguishing benign from malignant lesions).

Participants are provided with a database of dermoscopic images from the ISIC Data Archive[14], which includes number of images collected from leading clinical centers worldwide using various devices. These images undergo privacy screening and quality assurance, with clinical metadata vetted by recognized melanoma experts to ensure a representative and clinically relevant dataset.

Metadata include segmentation masks, expert pixel annotations for feature detection and malignancy labels. When available, approximate age and gender are also provided[5].

Despite its strengths, the organizers acknowledge dataset biases, such as limited representation of diseases, age groups, devices, or ethnicities, as well as incomplete dermoscopic feature annotations[5]. However, the challenge is continuously refined each year to enhance evaluation metrics and encourage broader participation and collaboration.

ISIC is widely regarded as the most comprehensive, technically advanced and accessible resource for digital dermatoscopy[26]. As a result, most publicly available datasets are provided using the guidelines from the archive and are also used as part of the challenge.

3.3 HAM10000

The "Human Against Machine with 10000 training images" (HAM10000) dataset comprises 10015 dermoscopic images collected over 20 years from two sources: the Department of Dermatology at the Medical University of Vienna, Austria, and Cliff Rosendahl's skin cancer practice in Queensland, Australia. Image storage and metadata formats vary between the sites, ranging from PowerPoint and Excel to diapositives.

The dataset was constructed through a combination of automated and manual extraction methods. A final manual quality review removed images that were close-up, out of focus, or contained potentially identifiable content (e.g. garments, jewelry, tattoos). Additionally, images affected by gel bubbles or those depicting non-pigmented, mucosal, ocular or subungual lesions were excluded. The dataset underwent technical validation by expert dermatopathologists.

HAM10000 includes images spanning seven common diagnostic categories:

- Actinic Keratoses and Intraepithelial Carcinoma,
- Basal Cell Carcinoma
- Benign Keratosis
- Dermatofibroma
- Melanocytic Nevi
- Melanoma
- Vascular Skin Lesions.

3.4 PAD-UFES-20

The PAD-UFES-20 dataset[18] consists of clinical images of skin lesions along with patient clinical data, collected using various smartphone devices. It includes six types of skin conditions - three skin cancers and three non-cancerous skin diseases.

Each image is accompanied by up to 21 patient clinical attributes, such as age, lesion location, Fitzpatrick Skin Type and lesion diameter, though not all attributes are available for every case.

The data collection workflow is illustrated in Figure 1.

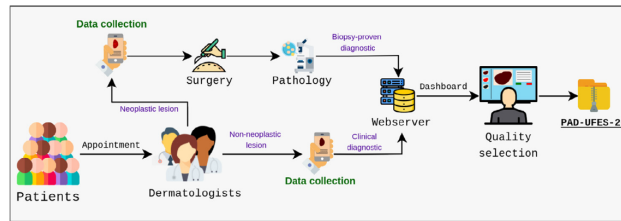


Figure 1: Data collection workflow for PAD-UFES-20 dataset, from [18]

3.5 Fitzpatrick 17k Dataset

The Fitzpatrick 17k Dataset[10] aims to answer the question "How does the accuracy of deep neural network models trained to classify clinical images of skin condition vary across skin color?". At the time of its release, only PAD-UFES-20 included any information on skin type, color or other race-related attributes. This dataset seeks to address bias in dermatology imaging, as skin conditions appear differently across skin types, potentially leading to delayed detection of malignant lesions and higher mortality rates in underrepresented categories.

The dataset contains 16577 clinical images labeled with both skin condition and Fitzpatrick skin type classification. The Fitzpatrick scale categorizes skin into six types (I to VI) based on sun reactivity. The images were sourced from two open-access dermatology atlases, DermaAmin and Atlas Dermatologico. Since neither source guaranteed diagnostic accuracy, 3% of the dataset was re-examined by board-certified dermatologists. The observed error rate was found to be consistent with those in widely used benchmark datasets across computer vision, natural language processing and audio processing.

The dataset includes 114 skin conditions, along with two additional aggregation levels to improved explainability. Fitzpatrick Skin Type (FST) labels were assigned by human annotators from Scale AI.

However, the dataset is imbalanced, both in terms of image distribution and skin condition representation. The majority of images depict lighter skin tones, and 25 of the 114 conditions have no representation for darker skin type.

3.6 Diverse Dermatology Images

The Diverse Dermatology Images (DDI) Dataset[7] shares the same objective as the Fitzpatrick 17k Dataset - addressing bias in dermatology imaging. Additionally, DDI tackles a key limitation of online datasets: the lack of histopathological confirmation for malignancies.

This dataset was retrospectively curated from histopatologically confirmed lesions diagnosed in Stanford Clinics from 2010 to 2020. All cases were reviewed by two board-certified dermatologists with 1 and 5 years of experience. Like Fitzpatrick 17k, DDI uses the FST scale to classify skin tones and assess disparities in treatments and bias in common Skin Disease Classification models.

DDI contains 656 images - not enough to train a model from scratch, as noted by the authors. However, their experiments demonstrate that fine-tuning existing models on DDI helps close the performance gap between lighter (FST I-II) and darker skin tones (FST V-VI).

Dataset	License	Total Images	Categories	Protected Attributes
ISIC 2017 Challenge Dataset	CC-0	2000	Benign/Malignant; Diagnosis	Sex, Approx. Age
HAM10000 (in ISIC Archive)	CC BY-NC 4.0	11,720	Benign/Malignant; Diagnosis	Sex, Approx. Age
ISIC Archive (at 2/02/2025)	Multiple	401059	Benign/Malignant; Diagnosis	Sex, Approx. Age
PAD-UFES-20	-	2298	Diagnosis (Six Categories)	Sex, Age, Fitzpatrick Skin Type (FST)
Fitzpatrick 17K Dataset	CC BY-NC 3.0	16557	114 Skin Diseases (with different levels of aggregations)	FST
Diverse Dermatology Images (DDI)	-	656	Benign/Malignant; Histopathological Diagnosis	FST, Sex, Age

Table 1: Summary of Datasets presented in this section. Where the information is unknown, it is reported as (-). Unless specified, the Diagnosis is referred either as the seven categories that are reported in the HAM10000 paragraph, or as a general annotation with no additional indication.

3.7 Other notable works

While this research primarily focuses on dataset construction, there are also notable studies analyzing the influence of clinical context on melanoma recognition. One such study [23] explores the "ugly duckling" concept - a diagnostic approach used by dermatologists to evaluate one skin lesion by comparing with other lesions from the same patient. This study applies Machine Learning (ML) to build a dataset of comparative lesions from the same individual. This dataset contains 33126 dermoscopy images from 2056 patients, collected between 1998 to 2020 across multiple international centers. To maintain dataset integrity, lesions previously included in ISIC challenges were excluded. The dataset is available in two formats, the Digital Imaging and Communication in Medicine (DICOM), a standard for dermatology imaging, and JPEG images with an associated CSV file for metadata. However, the authors acknowledge the underrepresentation of darker skin types in the dataset.

Another notable work is [21], which presents the first public database of dermoscopy and clinical images created in a hospital in Argentina. This dataset includes 1616 images representing 1246 unique lesions from 623 patients, making it the first of its kind in Hispanic America. The dataset aims to address geographic disparities in medical imaging data which contribute to health data poverty - a phenomenon where certain populations are disadvantaged in benefiting from medical innovations and research due to the scarcity of representative data. Data collection follow HIPAA guidelines for de-identification and involved review by three expert dermatologists. The dataset is publicly available through the ISIC Archive with a CC-BY license.

Given that larger datasets improve classification performance, researchers have explored data augmentation as a solution to the challenges of building large medical datasets. These challenges include the rarity of diseases, privacy concerns, high costs and the need for medical experts to annotate data. One such study [4] introduces a novel image augmentation algorithm based on a pulse-coupled neural network fusion strategy in a nonsubsampling shearlet transform (NSST) domain. The method aims to blend the structure of one dermoscopic image with the color characteristics of another, generating augmented images that incorporate a diverse range of skin tones.

4 Fairness Analysis

This section explores the second key question of this literature review: How does research in automatic skin disease detection accounts for the disparate biases present in Dermatology? This section is structured in two main parts:

- Demographic Bias and Algorithmic Fairness
- Skin Type annotations

The datasets used in these studies are shown in Figure 2. Only datasets mentioned in previous sections are cited.

Due to limited computational resources, which prevent the re-execution of the studies on uniform datasets, it was not possible to make quantitative comparisons between different methods.

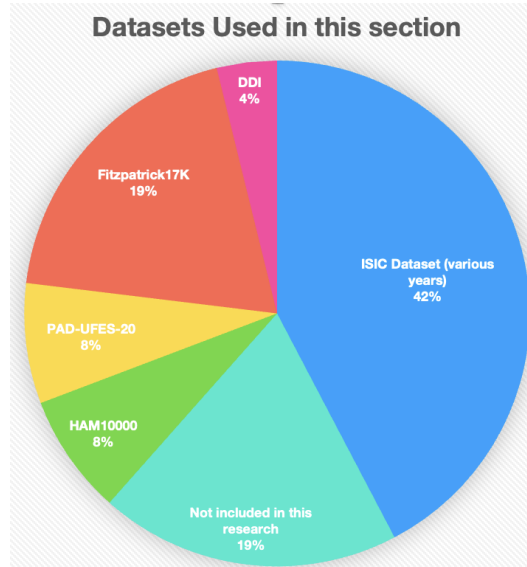


Figure 2: Datasets used in the analyzed studies

This analysis begins with an overview of algorithmic fairness, examining how studies address the high variability of dermatology datasets in terms of protected attributes. It then focuses on skin tone annotations for datasets, highlighting different methods. When possible, the different contributions are addressed in chronological order.

4.1 Demographic Bias and Algorithmic Fairness

As Machine Learning (ML) algorithms surpass expert performance in medical image analysis, the need for reliable generalization becomes crucial. While large, representative datasets could help address this challenge, such datasets are often lacking in the medical domain.[1]. Instead, dataset bias is a common issue, with certain patient groups being overrepresented. This necessitates strategies to mitigate potential biases.

The following sections will explore various techniques for addressing these biases.

4.1.1 Dataset Shift Detection

Dermatology AI systems are data-driven models that learn from examples, making them highly dependent on the quality and diversity of their training data. However, achieving comprehensive data coverage is particularly challenging in the medical domain, where unpredictability on unseen data is critical[20]. To ensure trustworthy AI, research has focused on out-of-distribution (OOD) detection methods, which contribute to making models causal, explainable, fair, robust and privacy-preserving.

The goal of Dataset Shift Detection is to leverage the output of the models to identify OOD inputs, which differ from the known in-distribution training data[20]. Literature distinguish between:

- Pre-training detection, where prior knowledge of the OOD samples is incorporated during training.
- Post-training detection, which requires no prior knowledge of OOD samples[15].

Additionally, OOD methods can be categorized into ensemble-based approaches and deep learning-based approaches[2].

Park et al.[20] focus on OOD detection methods that work on post-training detection, using pre-trained discriminative classifiers without requiring retraining or prior dataset knowledge. They evaluate the performance of state-of-the-art techniques, including:

- Mahalanobis Distance,
- Gram Matrices, and
- Energy-based scoring functions, less susceptible to over-confidence.

These methods are tested on a DenseNet-121-based skin lesion classifier, demonstrating the feasibility of OOD detectors in healthcare AI applications.

Kim et al.[15] propose a novel post-training OOD detector that can be applied to any pre-trained network. Their method combines:

- Subset scanning, which identifies the most anomalous subset of each layer using scoring functions; in particular, to avoid an exhaustive search and at the same time impose minimal assumptions on the underlying distribution, they use Non-Parametric Scan Statistics (NPSS).
- ODIN_{low} , a modification of ODIN, input perturbation strategy that consists on preprocessing the input by applying a small gradient-based perturbation, followed by temperature scaling on the Softmax layer, tuned to obtain the lowest softmax score performance.

Their results, tested on the ISIC-2019 dataset[14] as in-distribution dataset, evaluate AUROC (Area Under Receiver Operating Characteristics Curve) and F1-Score to measure performance on samples of unknown collection protocols and unknown diseases. Their findings indicate that while the method successfully detects unknown collection samples, ODIN_{low} struggles with unknown diseases, likely due to the blurring of unique features caused by the input noise.

Kim et al.[15] also provide an initial analysis of OOD fairness across skin tone, reporting an unstable performance for darker skinned samples, which they attribute to either a lack of training representation or differences in disease manifestation.

A more comprehensive fairness evaluation is conducted by Benmalek et al.[2], who developed a framework for assessing OOD detection performance in dermatology. Their key contributions include:

- Manual annotation of the ISIC-2019 dataset for skin tone, incorporating domain-expert supervision. They highlight that only 0.02% of dermoscopic images represent darker skin tones, as opposed to datasets like the Fitzpatrick 17K[10], which have 13.5%.
- Performance evaluation using F1-score, AUROC and Representation Gap (RG) - the difference in OOD detector performance across different skin types.

Their findings show that traditional ensemble methods and autoencoders outperform deep learning-based networks in mitigating fairness disparities. RG measurements reveal significant performance differences, emphasizing the need for improved fairness-aware OOD detection methods in dermatology AI.

4.1.2 Domain Generalization

The second family of methods, exemplified by [3], draws from Computer Vision (CV) research on Domain Generalization (DG) to improve the identification of benign vs. malignant skin lesions. DG studies how models become vulnerable to spurious correlations, a known issue in medical image analysis. However, its adoption in dermatology AI has been limited due to the common lack of labeled environments, which are essential for many DG techniques.

The approach in [3] has three key steps:

1. Trap Set Creation - Training and test sets are crafted to obtain amplified correlations between artifacts (non-disease-related visual cues, e.g. markers, patches...) and class labels, with the aim of simulating a high-bias scenario. Since seven artifact types can co-occur in lesions, this results in $2^7 \cdot 2$ potential environments, though only non-empty ones are used.
2. Group Distributional Robust Optimization (GroupDRO) - Learning strategy that minimizes the maximum risk across groups, preventing small, underrepresented groups from being overshadowed by dominant ones; chosen because of the setting, which includes many few-samples environments.
3. Feature Selection for Test Images (NoiseCrop), via Foreground Segmentation - To mitigate spurious correlations, the model removes background skin areas from test images using segmentation masks (either ground truth or deep-learning inferred). The lesion foreground is then rescaled to maximize its area, while the remaining pixels are set to random noise, ensuring that only disease-relevant features contribute to classification.

The method is evaluated across multiple datasets, maintaining binary classification of melanoma vs. benign lesions. Results indicate that this pipeline outperforms existing DG methods, with ablation studies confirming that the model learns task-relevant features rather than spurious artifacts.

Despite its success on the task, the review finds that no other dermatology AI studies employing Domain Generalization to address algorithmic fairness surfaced during the literature search.

4.1.3 Contrastive Learning

A third family of methods follows the *Fairness through blindness* approach to mitigate unfairness in Deep Neural Networks (DNNs). This method enforces blindness to sensitive attributes is achieved by introducing an additional branch alongside the primary DNN task. The auxiliary branch is specifically designed to classify the sensitive attribute, while the main model learns representations agnostic to it. This forces the framework to capture semantic information necessary for the target task while minimizing the accuracy of the sensitive attribute classifier, effectively reducing bias. However, a key limitation of this approach is that it may negatively impact representation learning for the primary task. By forcing the model to ignore sensitive attributes, it might also discard essential feature combinations that serve as proxies for those attributes. For instance, as noted in [9], both lesion color and visual features, play a crucial role in skin lesion classification, yet they are also indicative of the individual's skin type. Ignoring these attributes could hinder performance on the primary classification task.

Du et al. propose "Fairer AI in Dermatology via Disentanglement Contrastive Learning" (FairDisCo)[9], a model designed to mitigate skin-type bias in skin lesion classification from RGB images. The sensitive attribute in this study is skin type. FairDisCo incorporates three key branches after the feature extractor:

- Target branch - Predicts skin conditions and serves as the primary classifier, compared against baseline models.
- Sensitive Attribute branch - Another classifier; uses two different, opposing losses:
 - A confusion loss to remove skin-type information from feature representations.
 - A skin-type predictive cross-entropy loss.

The second loss counters the possibility of the confusion loss pushing for a trivial solution like setting all weights to zero.

- Contrastive branch - Addresses the limitations of Fairness through Blindness, introducing a contrastive loss. This loss preserves essential features, promoting intra-class cohesion and inter-class diversity in a low-dimensional latent space.

The framework is evaluated on Fitzpatrick17k[10] and Diverse Dermatology Images (DDI)[7], grouping skin types into three broader categories. Fairness is measured using:

- Predictive Quality Disparity (PQD): ratio between the lowest accuracy and the highest accuracy across different groups.

$$PQD = \frac{\min(acc_j, j \in S)}{\max(acc_j, j \in S)} \quad (1)$$

- Demographic Disparity (DP): percentage diversities of positive outcomes for each sensitive group.

$$DP = \frac{1}{M} \sum_{i=1}^M \frac{\min(p(\hat{y} = i | s = j), j \in S)}{\max(p(\hat{y} = i | s = j), j \in S)} \quad (2)$$

- Equality of opportunity (EO): different sensitive groups should have similar true positive rates.

$$EO = \frac{1}{M} \sum_{i=1}^M \frac{\min(p(\hat{y} = i | y = i, s = j), j \in S)}{\max(p(\hat{y} = i | y = i, s = j), j \in S)} \quad (3)$$

A higher value in these metrics indicates a fairer model.

FairDisCo outperforms its baseline fairness algorithms, demonstrating superior fairness across skin types while maintaining predictive accuracy. The results reinforce previous findings on skin-type bias in dermatology AI and highlight the necessity of reducing disparities in medical AI models.

4.1.4 Calibration Bias Mitigation

The second-to-last group of studies focuses on mitigating Calibration Bias in dermatology AI. A model is considered well-calibrated if its predicted probabilities align with the actual probability of making a correct prediction[25]. While existing methods improve calibration on the overall population, they often fail to do so for sub-populations, leading to overconfidence and potentially incorrect diagnosis, thereby compromising fairness and equity for a subset of the population.

However, as Ricci Lara et al[22] highlight, improper use of evaluation metrics for calibration can lead to misleading results in healthcare AI. This requires attention in the selection of calibration metrics. Ricci Lara et al[22] cite the Expected Calibration Error (ECE), a widely used metric that is highly sensitive and heavily biased with respect to sample size, often worsening as the sample size decreases. Their study on light-skinned vs. dark-skinned individuals (Fitzpatrick Skin Type) using the PAD-UFES-20[18], suggests that previously observed calibration disparities might actually be spurious, merely reflecting unequal subgroup sample sizes rather than true bias, as also noted by [16].

To address calibration bias, Shui et al.[25] propose Cluster-Focal, a two-stage approach:

1. Clustering Stage: A model f_{id} is trained using Empirical Risk Minimization (ERM) without any fairness considerations. The gap between confidence score and correctness is computed, with larger gaps indicating poor calibration. K-means clustering is then applied to group instances based on the gap.
2. Focal Loss Stage: A new prediction model f_{pred} is trained using a group-wise focal loss, penalizing overconfident predictions, thereby reducing the chances of poor calibration.

Shui et al evaluate their approach on gender fairness in the HAM10000[26], using Q-ECE[17] as the metric. Their method effectively reduces calibration error in the worst-performing subgroup, demonstrating its potential for improving fairness.

4.1.5 Reinforcement Learning-based Models

While most studies address one attribute at a time, unfairness in dermatology can occur across multiple factors, including skin tone, age, gender and disease site. Moreover, a correlation exists between these attributes - improving fairness for one may inadvertently reduce fairness for another.

The "Multi-dimension AI Fairness Framework by Uniting Off-the-Shelf Models" (Muffin)[24] seeks to address this issue using Reinforcement Learning (RL). The framework selects and fuses off-the-shelf models, which are then trained on biased datasets to simultaneously maximize fairness across multiple attributes. Figure 3 provides an overview of this framework.

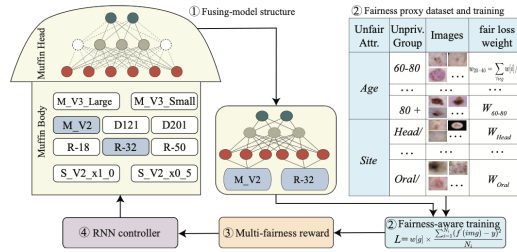


Figure 3: Overview of the Muffin framework, as shown in [24]

This work highlights the success of Reinforcement Learning using an unfairness score-based reward and a Monte Carlo policy gradient algorithm to obtain a fused-model for multi-dimension AI Fairness. Muffin-Nets significantly improve fairness across multiple attributes while maintaining accuracy. As shown in Figure 4, which presents the Pareto frontiers, Muffin-Nets outperform the other tested models in terms of fairness for both age and disease site, while retaining an accuracy over 82% on the ISIC 2019 dataset[14]. A similar trend can be observed when training with Fitzpatrick 17k dataset[10], where overall accuracy remains stable, while fairness for skin tone and type improves significantly, highlighting that the method can be applied to different datasets successfully.

4.2 Skin Type Annotation

Among the various protected attributes, skin tone has the most significant impact on fairness in dermatology AI. Although people of color are roughly 20 to 30 times less likely to develop melanoma than lighter-skinned individuals, their diagnosis tends to occur at more advanced stages, leading to lower survival rates and poorer outcomes. [16].

A key factor contributing to this disparity is dataset bias. Most publicly available dermatology datasets and atlases, including those discussed in the previous section, predominantly feature lighter-skinned individuals. This imbalance arises due to difference in disease prevalence and connected underrepresentation in medical imaging, which may result in worse model performance and potential outcome disparities across ethnic groups when using AI-driven diagnosis systems.

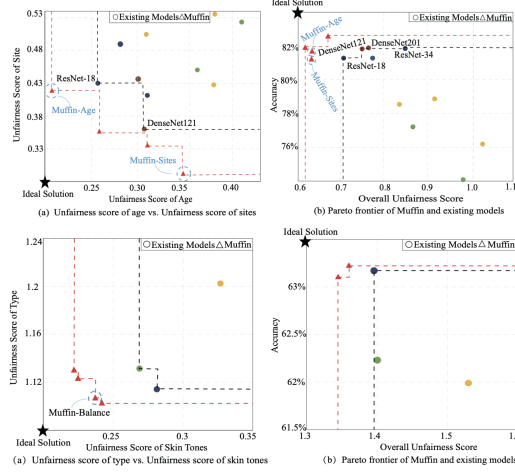


Figure 4: Pareto Frontiers for Muffin-Nets and comparison with existing neural networks, from [24]

ITA Range	Skin Tone Category	Abbreviation
$ITA > 55^\circ$	Very Light	very_lt
$48^\circ < ITA \leq 55^\circ$	Light 2	lt2
$41^\circ < ITA \leq 48^\circ$	Light 1	lt1
$34.5^\circ < ITA \leq 41^\circ$	Intermediate 2	int2
$28^\circ < ITA \leq 34.5^\circ$	Intermediate 1	int1
$19^\circ < ITA \leq 28^\circ$	Tanned 2	tan2
$10^\circ < ITA \leq 19^\circ$	Tanned 1	tan1
$ITA \leq 10^\circ$	Dark	dark

Figure 5: Skin Tone Categorization scheme in [16] based on ITA

4.2.1 Individual Typology Angle

Kinyanjui et al[16] propose an automated method for annotating skin tone in medical using Individual Typology Angle (ITA) when explicit labels are unavailable. ITA is calculated as:

$$ITA = \arctan\left(\frac{L - 50}{b}\right) \times \frac{180}{\pi} \quad (4)$$

where L represents luminance and b indicates the amount of yellow chromatic component. The ITA value is then mapped to the Fitzpatrick Skin Type (FST) scale, leveraging the correlation between the FST and the melanin index and the inverse correlation between the melanin index and the ITA.

To extract non-diseased skin regions for ITA calculation, the authors employ Mask R-CNN[13], pretrained on the COCO dataset and fine-tuned using lesion boundary segmentation data from the ISIC 2018 Challenge Task 1 (see 3.2). Contour extraction and thresholding further refine the segmentation.

Notably, this ITA-based annotation strategy provides a finer-grained representation of skin tone compared to the original FST scale (Figure 5).

Corbin et al[6] propose ITA-based skin type estimation method that eliminates the need for segmentation masks by leveraging patch-based sampling. Instead of relying on precisely segmented skin regions, their approach computes ITA as the median value across $N \times N$ pixel patches extracted from the image. They explored three patching strategies:

- Center Cropped: ignores the center portion of the image, assuming that the skin lesion would exist in the center.
- Structured Patches: select patches around the image border to avoid lesion interference.
- Random Patches across the image.

The patch size was hand-selected based on empirical performance.

Experiments on various ISIC datasets indicate that the Center Cropped approach yields reliable ITA estimations, provided that the lesion occupies less than 80% of the image. However, the authors acknowledge certain limitations, including:

- Exclusively testing on dermoscopic images without pre-processing and without FST labels
- Manual patch size selection, which may not generalize optimally across datasets

4.2.2 Fitzpatrick Skin Type

The Fitzpatrick Skin Type (FST) scale is a six-point classification system originally developed for assessing sun reactivity of skin and guiding clinical treatments according to skin phenotype. In computer vision, it serves as the default standard for evaluating algorithmic fairness and model accuracy across different skin tones [10].

Despite its widespread adoption, only three datasets at the Seventh ISIC Skin Image Analysis Workshop at ECCV 2022[19] contained FST labels:

- Fitzpatrick17[10], also the only dataset covering all six skin types.
- Diverse Dermatology Images (DDI)[7]
- PAD-UFES-20[18]

However Fitzpatrick17K has notable limitations[19]:

- Images are unverified by dermatologists
- Skin types are annotated by non-experts
- Varying lighting conditions and perspectives (as images are taken in real-world clinical settings and not through dermoscopy)
- Presence of erroneous images.

Pakzad et al. utilized Fitzpatrick17K[10] to develop CIRCLe[19](Color Invariant Representation learning for unbiased Classification of skin Lesions), a Generative Adversarial Network-based model. CIRCLe learns feature representations invariant to skin color by performing image-to-image transformations across different skin type domains.

4.2.3 Experts vs. Crowds vs. Algorithms

AI-powered image-based diagnosis holds great promise in dermatology, potentially enhancing clinical decision-making. However, incorrect predictions could mislead both specialists and generalists, sometimes causing greater than human error. To address the accuracy of FST annotations, Groh et al.[11] investigate the question: "Who is qualified to annotate images with estimated FSTs?".

Their study evaluates inter-rater reliability among

- Three medical experts.
- A computer vision algorithm (ITA-FST)
- Two crowdsourcing approaches

Results show that no statistical significant difference was found between medical expert and crowdsourced annotations. However, the ITA-FST algorithm showed statistically significant deviations from experts. While crowdsourcing performed similarly to experts, non experts made errors that clinicians would not typically make. Moreover, image-based estimated FST annotations is limited by factors like lighting, image quality and pose variability, which are not an issue for in-person assessments. The experts also emphasize that clinical FST classification involves factors beyond skin color (like for example eye color), making image-based FST estimation inherently limited.

Groh et al. also introduce a dynamic consensus protocol to improve crowdsourced annotations. This protocol aggregates crowdsourced annotations dynamically, flags low-agreement or high-difficulty images for expert review and enables expert adjudication for images with error report.

Paper Title	DOI (either link or String)	Authors	Year	Venue	Method	Datasets
<i>Risk of Training Diagnostic Algorithms on Data with Demographic Bias</i>	https://doi.org/10.1007/978-3-030-61166-8_20	Abbasi-Sureshjani et al.	2020	IMIMIC 2020 MIL-3D 2020 LABELS 2020	Survey of practices, Bias Mitigation	ISIC 2017
<i>Fairness of Classifiers Across Skin Tones in Dermatology</i>	https://doi.org/10.1007/978-3-030-59725-2_31	Kinyanjui et al.	2020	MICCAI 2020	Estimate skin tone using ITA	ISIC 2018 SD-198
<i>Out-of-distribution detection in Dermatology using Input Perturbation and Subset Scanning</i>	10.1109/ISBIS2829.2022.9761412	Kim et al.	2021	2022 IEEE 19th ISBI	Novel Out-of-distribution detection strategy	ISIC 2019 SD-198
<i>Reliable and Trustworthy Machine Learning for Health using Dataset Shift Detection</i>	https://doi.org/10.48550/arXiv.2110.14019	Park et al.	2021	NeurIPS 2021	Out-of-distribution detection	HAM10000 ISIC 2017
<i>Artifact-Based Domain Generalization of Skin Lesion Models</i>	https://doi.org/10.1007/978-3-031-25069-9_10	Bissoto et al.	2022	ECCV 2022 Workshops	Domain Generalization	ISIC 2019 PH2 Derm7pt PAD-UFES-20
<i>FairDisCo: Fairer AI in Dermatology via Disentanglement Contrastive Learning</i>	https://doi.org/10.1007/978-3-031-25069-9_13	Du et al.	2022	ECCV 2022 Workshops	Contrastive Learning	Fitzpatrick17K DDI
<i>CIRCLE: Color Invariant Representation Learning for Unbiased Classification of Skin Lesions</i>	https://doi.org/10.1007/978-3-031-25069-9_14	Pakzad et al.	2022	ECCV 2022 Workshops	Generative Adversarial Networks for Regularization	Fitzpatrick 17K
<i>Exploring strategies to generate Fitzpatrick Skin Type Metadata for dermoscopic images using Individual Typology Angle Techniques</i>	https://doi.org/10.1007/s11042-022-14211-1	Corbin et al.	2022	-	ITA without segmentation masks	ISIC 2016 ISIC 2017 ISIC 2018
<i>Towards Transparency in Dermatology Image Datasets with Skin Tone Annotation by Experts, Crowds and an Algorithm</i>	https://doi.org/10.48550/arXiv.2207.02942	Groh et al.	2022	-	Inter-rater reliability for FST estimation	Fitzpatrick 17K Online Dermatology Atlases
<i>Towards Unraveling Calibration Biases in Medical Image Analysis</i>	https://doi.org/10.1007/978-3-031-45249-9_13	Ricci Lara et al.	2023	CLIP 2023 EPMI 2023 FAIMI 2023	Evaluation of Calibration Bias	PAD-UFES-20
<i>Mitigating Calibration Bias without Fixed Attribute Grouping for Improved Fairness in Medical Imaging Analysis</i>	https://doi.org/10.1007/978-3-031-43898-1_19	Shui et al.	2023	MICCAI 2023	Calibration Bias through Clustering and Focal Loss	HAM10000
<i>Mufin: A Framework Toward Multi-Dimension AI Fairness by Uniting Off-the-Shelf Models</i>	10.1109/DAC56929.2023.10247765	Yi Sheng et al.	2023	DAC 2023	Reinforcement Learning based Framework	ISIC 2019 Fitzpatrick17K
<i>Evaluating the Impact of Skin Tone Representation on Out-of-Distribution Detection Performance in Dermatology</i>	10.1109/ISBIS6570.2024.10635847	Benmalek et al.	2024	ISBI 2024	Evaluation framework for OOD detectors	ISIC 2019 (manually annotated) Fitzpatrick17K

Table 2: Summary of proposed methods

The study suggests that crowdsourcing, when properly managed, can match expert-level FST annotations. However, expert oversight remains crucial, and image-based estimation has inherent limitations compared to in-person assessments. For this reason, the authors invite to use expert-based and crowdsourced annotations to build future datasets.

5 Limitations of this Review

This literature review examined a diverse range of studies and methodologies, encompassing various datasets, methods, and evaluation metrics. Due to this high variability, a quantitative comparison of the presented algorithms was not feasible, as re-training and evaluating different deep neural networks would require substantial computational resources.

Additionally, the reviewed papers represent only a subset of the extensive medical AI literature. While most studies were accessible through university licenses, some were unavailable due to access restrictions and were therefore excluded from the analysis.

6 Conclusions

This literature review explored the advancements in automatic skin disease detection, with a particular focus on the availability of tailored datasets and the measures taken to ensure fairness and inclusivity in AI models.

To investigate this, over 100 studies were retrieved from two major scientific archives, IEEE Xplore and Scopus, as well as the ISIC Archive. After careful selection, approximately 20 papers were shortlisted to exemplify the state of the art over the past decade.

Research in this field has focused both on building new datasets and improving on current ones, with ISIC playing a pivotal role in providing the most comprehensive collection of dermatological images to date[14], and promoting challenges to further advancements in the field of lesion analysis. ISIC’s position as the principal database for Lesion Analysis research has pushed researchers to either innovate on it, e.g. by including missing annotations - such as Benmalek et al.[2] - or proposing alternative collections, focused on addressing fairness concerns, such as Daneshjou et al. [7]’s Diverse Dermatology Images (DDI), or the PAD-UFES-20[18] dataset by Pacheco et al, which incorporate skin type labels. The publication of datasets within the ISIC archive has also driven efforts toward standardization, fostering more rigorous method development and enabling meaningful comparisons across studies.

At the same time, research on fairness and demographic bias in dermatology AI has gained significant attention, with studies highlighting inherent biases in widely used datasets that reflect real-world healthcare disparities[16]. This review examines multiple strategies for mitigating these biases,

including Out-of-Distribution detection[2][15][20], Domain Generalization[3], Contrastive Learning[9], Calibration Bias Mitigation[25] and Reinforcement Learning frameworks like Muffin[24].

Despite significant advancements, fairness in automatic skin lesion diagnosis remains an open challenge in literature. The absence of standardized skin tone annotations further complicates fairness assessments related to skin color, emphasizing the need for robust and scalable annotation methods[16][10][6] and stringent quality control across dermoscopic and clinical datasets in literature[19][11]. Ensuring generalizability and fairness in dermatology AI requires continuous efforts in dataset curation and model evaluation.

The rapid emergence of fairness-focused methodologies within a short publication span underscores the growing recognition of bias-related challenges and the research community’s commitment to addressing them. As the push for more equitable healthcare continues, this review provides a foundation for assessing progress in AI-driven clinical decision support and expanding access to dermatological care globally.

References

- [1] Samaneh Abbasi-Sureshjani et al. “Risk of Training Diagnostic Algorithms on Data with Demographic Bias”. In: *Interpretable and Annotation-Efficient Learning for Medical Image Computing*. Ed. by Jaime Cardoso et al. Cham: Springer International Publishing, 2020, pp. 183–192. ISBN: 978-3-030-61166-8.
- [2] Assala Benmalek et al. “Evaluating the Impact of Skin Tone Representation on Out-of-Distribution Detection Performance in Dermatology”. In: *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*. 2024, pp. 1–5. DOI: 10.1109/ISBI56570.2024.10635847.
- [3] Alceu Bissoto et al. “Artifact-Based Domain Generalization of Skin Lesion Models”. In: *Computer Vision – ECCV 2022 Workshops*. Ed. by Leonid Karlinsky, Tomer Michaeli, and Ko Nishino. Cham: Springer Nature Switzerland, 2023, pp. 133–149.
- [4] Esther Chabi Adjobo et al. “Towards Accurate Skin Lesion Classification across All Skin Categories Using a PCNN Fusion-Based Data Augmentation Approach”. In: *Computers* 11.3 (2022).
- [5] Noel C. F. Codella et al. *Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC)*. 2018. arXiv: 1710.05006 [cs.CV]. URL: <https://arxiv.org/abs/1710.05006>.
- [6] Adam Corbin and Oge Marques. “Exploring strategies to generate Fitzpatrick skin type metadata for dermoscopic images using individual typology angle techniques”. In: *Multimedia Tools and Applications* 82.15 (2023), pp. 23771–23795.
- [7] Roxana Daneshjou et al. “Disparities in dermatology AI performance on a diverse, curated clinical image set”. In: *Science Advances* 8.32 (Aug. 2022). ISSN: 2375-2548. DOI: 10.1126/sciadv.abq6147. URL: <http://dx.doi.org/10.1126/sciadv.abq6147>.
- [8] Ittai Dayan et al. “Federated learning for predicting clinical outcomes in patients with COVID-19”. In: *Nature Medicine* 27.10 (2021), pp. 1735–1743.
- [9] Siyi Du et al. “FairDisCo: Fairer AI in Dermatology via Disentanglement Contrastive Learning”. In: *Computer Vision – ECCV 2022 Workshops*. Ed. by Leonid Karlinsky, Tomer Michaeli, and Ko Nishino. Cham: Springer Nature Switzerland, 2023, pp. 185–202.
- [10] Matthew 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)*. 2021, pp. 1820–1828. DOI: 10.1109/CVPRW53098.2021.00201.
- [11] Matthew Groh et al. *Towards Transparency in Dermatology Image Datasets with Skin Tone Annotations by Experts, Crowds, and an Algorithm*. 2022. arXiv: 2207.02942 [cs.CV]. URL: <https://arxiv.org/abs/2207.02942>.
- [12] David Gutman et al. *Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC)*. 2016. arXiv: 1605.01397 [cs.CV]. URL: <https://arxiv.org/abs/1605.01397>.

- [13] Kaiming He et al. *Mask R-CNN*. 2018. arXiv: 1703.06870 [cs.CV]. URL: <https://arxiv.org/abs/1703.06870>.
- [14] Visited: 2/02/2025. URL: <https://www.isic-archive.com/major-publications>.
- [15] Hannah Kim et al. “Out-of-Distribution Detection in Dermatology using Input Perturbation and Subset Scanning”. In: *CoRR* abs/2105.11160 (2021). arXiv: 2105.11160. URL: <https://arxiv.org/abs/2105.11160>.
- [16] Newton M. Kinyanjui et al. “Fairness of Classifiers Across Skin Tones in Dermatology”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Ed. by Anne L. Martel et al. Cham: Springer International Publishing, 2020, pp. 320–329.
- [17] Jishnu Mukhoti et al. “Calibrating Deep Neural Networks using Focal Loss”. In: *CoRR* abs/2002.09437 (2020). arXiv: 2002.09437. URL: <https://arxiv.org/abs/2002.09437>.
- [18] Andre G.C. Pacheco et al. “PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones”. In: *Data in Brief* 32 (2020), p. 106221.
- [19] Arezou Pakzad, Kumar Abhishek, and Ghassan Hamarneh. “CIRCLE: Color Invariant Representation Learning for Unbiased Classification of Skin Lesions”. In: *Computer Vision – ECCV 2022 Workshops*. Ed. by Leonid Karlinsky, Tomer Michaeli, and Ko Nishino. Cham: Springer Nature Switzerland, 2023, pp. 203–219.
- [20] Chunjong Park et al. “Reliable and trustworthy machine learning for health using dataset shift detection”. In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. NIPS ’21. Red Hook, NY, USA: Curran Associates Inc., 2021. ISBN: 9781713845393.
- [21] María Agustina Ricci Lara et al. “A dataset of skin lesion images collected in Argentina for the evaluation of AI tools in this population”. In: *Scientific Data* 10.1 (2023), p. 712.
- [22] María Agustina Ricci Lara et al. “Towards Unraveling Calibration Biases in Medical Image Analysis”. In: *Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging*. Ed. by Stefan Wesarg et al. Cham: Springer Nature Switzerland, 2023, pp. 132–141.
- [23] Veronica Rotemberg et al. *A Patient-Centric Dataset of Images and Metadata for Identifying Melanomas Using Clinical Context*. 2020. arXiv: 2008.07360 [eess.IV]. URL: <https://arxiv.org/abs/2008.07360>.
- [24] Yi Sheng et al. *Muffin: A Framework Toward Multi-Dimension AI Fairness by Uniting Off-the-Shelf Models*. 2023. arXiv: 2308.13730 [cs.LG]. URL: <https://arxiv.org/abs/2308.13730>.
- [25] Changjian Shui et al. “Mitigating Calibration Bias Without Fixed Attribute Grouping for Improved Fairness in Medical Imaging Analysis”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by Hayit Greenspan et al. Cham: Springer Nature Switzerland, 2023, pp. 189–198.
- [26] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions”. In: *Scientific Data* 5.1 (2018), p. 180161.