



International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

Operating Publisher
SciFormat Publishing Inc.
ISNI: 0000 0005 1449 8214

2734 17 Avenue SW,
Calgary, Alberta, T3E0A7,
Canada
+15878858911
editorial-office@sciformat.ca

ARTICLE TITLE	ARTIFICIAL INTELLIGENCE IN DERMATOLOGICAL IMAGE CLASSIFICATION: DIAGNOSTIC ACCURACY, SKIN TONE DISPARITIES, AND CLINICAL GOVERNANCE
----------------------	---

DOI	https://doi.org/10.31435/ijitss.4(52).2026.6380
------------	---

RECEIVED	11 August 2026
-----------------	----------------

ACCEPTED	04 October 2026
-----------------	-----------------

PUBLISHED	15 October 2026
------------------	-----------------

LICENSE



The article is licensed under a **Creative Commons Attribution 4.0 International License**.

© The author(s) 2026.

This article is published as open access under the Creative Commons Attribution 4.0 International License (CC BY 4.0), allowing the author to retain copyright. The CC BY 4.0 License permits the content to be copied, adapted, displayed, distributed, republished, or reused for any purpose, including adaptation and commercial use, as long as proper attribution is provided.

ARTIFICIAL INTELLIGENCE IN DERMATOLOGICAL IMAGE CLASSIFICATION: DIAGNOSTIC ACCURACY, SKIN TONE DISPARITIES, AND CLINICAL GOVERNANCE

Faustyna Magdalena Kołodziej (Corresponding Author, Email: kolodziej.faustyna.fk@gmail.com)
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0002-8778-7100

Mateusz Lis
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0006-8441-8296

Zuzanna Julia Komin
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0001-6284-6410

Maryan Liasota
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0006-0266-4849

Eliza Zofia Kabat
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0003-6531-967X

Michalina Justyna Kozłowska
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0003-9561-2957

Krzysztof Konopka
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0003-4055-0074

Maciej Kamil Majchrzak
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0007-4739-0092

Jakub Juskowiak
Poznań University of Medical Sciences, Poznań, Poland
ORCID ID: 0009-0001-0372-400X

Alicja Kwiatkowska
Pomeranian Medical University, Szczecin, Poland
ORCID ID: 0009-0003-1027-2114

ABSTRACT

Background: Artificial intelligence (AI) is increasingly used to classify skin lesions from dermoscopic and clinical images, yet high performance in curated datasets may not translate into safe or equitable clinical use.

Objective: This structured narrative review evaluates diagnostic accuracy, skin-tone-related performance disparities, dataset representation, and governance requirements for AI-based skin-lesion classification.

Methods: PubMed/MEDLINE was used as the primary database, supplemented by PubMed Central, Google Scholar, Semantic Scholar, journal websites, and reference-list screening. English-language, peer-reviewed literature published mainly between 2021 and 2026 was synthesised thematically across diagnostic performance, external validity, skin-tone representation, clinical implementation, and governance.

Results: Meta-analytic evidence indicates high sensitivity for AI-based skin-lesion classification, including pooled sensitivity of 0.91 and an HSROC-AUC of 0.88, but lower specificity and substantial heterogeneity limit direct clinical translation. Performance declined in darker skin-tone groups in several external evaluations; in selected models, sensitivity fell between Fitzpatrick types I–II and V–VI. Limited reporting of skin tone, reliance on internally tested or curated datasets, and sparse prospective evidence further constrain generalisability.

Conclusion: AI should be implemented as clinician-supervised decision support rather than as an autonomous diagnostic system. Equitable deployment requires representative datasets, external and subgroup-stratified validation, transparent reporting, defined clinical accountability, regulatory oversight, and continuous post-deployment monitoring.

KEYWORDS

Artificial Intelligence, Dermatology, Skin Cancer, Skin-Lesion Classification, Algorithmic Bias, Clinical Governance

CITATION

Kołodziej, F. M., Lis, M., Komin, Z. J., Liasota, M., Kabat, E. Z., Kozłowska, M. J., Konopka, K., Majchrzak, M. K., Juskowiak, J., & Kwiatkowska, A. (2026). Artificial intelligence in dermatological image classification: Diagnostic accuracy, skin tone disparities, and clinical governance. *International Journal of Innovative Technologies in Social Science*, 4(52). [https://doi.org/10.31435/ijitss.4\(52\).2026.6380](https://doi.org/10.31435/ijitss.4(52).2026.6380)

COPYRIGHT

© The author(s) 2026. This article is published as open access under the **Creative Commons Attribution 4.0 International License (CC BY 4.0)**, allowing the author to retain copyright. The CC BY 4.0 License permits the content to be copied, adapted, displayed, distributed, republished, or reused for any purpose, including adaptation and commercial use, as long as proper attribution is provided.

1. Introduction

1.1 Background and Clinical Context

Skin cancer remains a central concern in dermatological diagnosis, as early detection can guide treatment decisions and influence morbidity and long-term clinical outcomes (Liopyris et al., 2022). This is particularly important in melanoma, where delayed diagnosis may result in progression to invasive or metastatic disease, while earlier identification is associated with more effective management and a better prognosis (Liopyris et al., 2022). Non-melanoma skin cancers, such as basal cell carcinoma and squamous cell carcinoma, are also clinically relevant because of their high prevalence, potential for local tissue invasion, risk of recurrence, and, in some cases, metastatic potential (Sanchez et al., 2023). Rising demand for dermatology services has increased pressure on referral pathways, primary care triage, and teledermatology systems, creating interest in technologies that may support earlier recognition of suspicious lesions and improve referral prioritisation (Escalé-Besa et al., 2023; Marsden et al., 2023; Sangers et al., 2023; Smak Gregoor et al., 2023).

Dermatology is a highly visual specialty, making it particularly suitable for the computational analysis of dermoscopic images, clinical photographs, mobile-acquired images, and multimodal data (Li et al., 2022; Smak Gregoor et al., 2023; Tschandl et al., 2026). However, accurate clinical evaluation also requires consideration of lesion progression, anatomical site, symptoms, patient history, individual risk factors, and the broader clinical context (Liopyris et al., 2022; Sangers et al., 2023). Accordingly, performance on isolated images should not be equated with comprehensive clinical reasoning.

1.2 Artificial Intelligence in Skin Lesion Classification

AI-based skin lesion classification commonly uses supervised machine learning or deep learning models trained to distinguish benign from malignant dermatological lesions (Li et al., 2022; Liopyris et al., 2022). Convolutional neural networks are widely used in dermatological image analysis because they can automatically learn relevant visual features from dermoscopic or clinical images without requiring manually designed feature extraction (Li et al., 2022; Rezk et al., 2022). These models can be applied to binary classification tasks, such as distinguishing benign from malignant lesions, as well as to multiclass classification involving melanoma, basal cell carcinoma, squamous cell carcinoma, benign melanocytic lesions, seborrheic keratoses, and other diagnostic categories (Li et al., 2022; Tschandl et al., 2026). Multimodal approaches combining dermoscopy, clinical photography, metadata, or contextual information may more closely approximate clinical decision-making, although they require careful validation across datasets and clinical settings (Li et al., 2022; Tschandl et al., 2026).

Potential applications include highlighting suspicious lesions, prioritising referrals, supporting teledermatology triage, and providing second-reader support. Their clinical value, however, depends on safety, workflow compatibility, accountability, real-world performance, and equitable outcomes across patient groups. (Escalé-Besa et al., 2023; Sangers et al., 2023; Thomas et al., 2023).

1.3 Diagnostic Performance and Clinical Applicability

Research on AI in dermatology has frequently reported promising diagnostic results, which are usually assessed using sensitivity, specificity, and the area under the receiver operating characteristic curve (AUC; Li et al., 2022; Rezk et al., 2022; Salinas et al., 2024). These measures are clinically important because missed malignant lesions may delay diagnosis and treatment, whereas false-positive results may lead to avoidable biopsies, greater patient anxiety, and increased pressure on healthcare services (Li et al., 2022; Liopyris et al., 2022).

Nevertheless, high overall accuracy in benchmark datasets does not automatically indicate that a model will perform safely, consistently, or fairly in routine clinical care (Daneshjou et al., 2022; Salinas et al., 2024; Tjiu & Lu, 2025). Model performance may vary according to dataset characteristics, image quality, the range and prevalence of lesions, the reference standard used for diagnosis, and whether the model has been evaluated using internal or external data (Daneshjou et al., 2022; Salinas et al., 2024; Tjiu & Lu, 2025). Generalisability is therefore essential when considering the clinical usefulness of AI-based dermatology tools. Claims of clinical utility therefore require external validation across independent datasets, devices, healthcare settings, and patient populations.

1.4 Skin Tone Bias and Dataset Underrepresentation

Underrepresentation of darker skin tones is one of the most clinically and ethically significant limitations of dermatology AI (Ang & Oh, 2025; Daneshjou et al., 2022; Liu et al., 2023; Rezk et al., 2022; Tjiu & Lu, 2025). Many dermatological image datasets predominantly contain images of lighter skin tones, whereas individuals with Fitzpatrick skin types IV–VI and other populations with skin of colour are less consistently represented or insufficiently annotated (Ang & Oh, 2025; Daneshjou et al., 2022; Liu et al., 2023; Rezk et al., 2022). This imbalance may affect diagnostic safety because lesion morphology, contrast against the surrounding skin, background pigmentation, and clinical presentation can vary across skin tones (Daneshjou et al., 2022; Liu et al., 2023; Tjiu & Lu, 2025).

The Fitzpatrick Skin Type scale is an imperfect proxy for skin tone because it was originally developed to classify tanning and burning responses to ultraviolet exposure rather than to provide an objective measurement of skin pigmentation (Weir et al., 2024).

1.5 Governance, Safety, and Aim of the Review

Governance should be central to AI-based skin lesion classification. Depending on their intended use, level of risk, and the relevant jurisdiction, AI systems used to support diagnosis, triage, or referral may be regulated as medical device software or classified as high-risk medical AI (Medical Device Coordination Group & Artificial Intelligence Board, 2025; Medicines and Healthcare products Regulatory Agency [MHRA], 2023; World Health Organization [WHO], 2023).

Although AI-based skin lesion classification has advanced rapidly, substantial gaps remain in translating technical performance into equitable clinical benefit (Daneshjou et al., 2022; Tjiu & Lu, 2025; Tschandl et al., 2026). Existing literature frequently emphasises aggregate diagnostic accuracy but less consistently integrates subgroup performance, skin-tone representation, dataset transparency, implementation setting, and regulatory governance (Daneshjou et al., 2022; Liu et al., 2023; Tjiu & Lu, 2025). This structured narrative review therefore synthesises evidence on diagnostic accuracy, performance variation across skin tones and dataset populations, and the governance measures required for safe and equitable implementation.

2. Methods

2.1 Research Design

This article was designed as a structured narrative literature review examining artificial intelligence-based skin-lesion classification in dermatology, with particular emphasis on diagnostic accuracy, skin-tone-related performance disparities, dataset representation, external validity, clinical implementation, social and institutional implications, and regulatory governance.

The review did not involve the collection or analysis of original patient data, experimental model development, prospective patient recruitment, or an independent quantitative meta-analysis. Numerical estimates reported in the review were derived from previously published studies, systematic reviews, and meta-analyses.

A structured narrative approach was selected because the topic spans several interconnected fields, including clinical dermatology, oncology, medical imaging, machine learning, health equity, digital health, ethics, implementation science, and medical-device regulation. This approach allowed evidence concerning diagnostic performance to be interpreted alongside subgroup disparities, dataset limitations, patient safety, clinician accountability, institutional consequences, and post-deployment oversight.

2.2 Data Sources and Search Strategy

PubMed, including MEDLINE-indexed records, was used as the primary bibliographic database because of its relevance to biomedical, dermatological, oncological, and clinical implementation literature. Google Scholar and Semantic Scholar were used for supplementary literature discovery, while PubMed Central was used to retrieve full-text articles and identify related publications. Backward citation tracking was performed using the reference lists of eligible publications, while forward citation tracking was conducted using citation records available through Google Scholar and Semantic Scholar. Official sources, including World Health Organization publications, MHRA guidance, GOV.UK policy documents, and European Union legislation available through EUR-Lex, were searched separately for regulatory, governance, and policy documents relevant to artificial intelligence in healthcare and medical-device software.

The literature search covered publications from 2019 to 2026, with priority given to recent evidence published between 2021 and 2026. Publications from 2019 to 2020 were eligible when they provided clinically, methodologically, or conceptually important background.

Free-text keywords and Boolean operators were used, with database-specific subject headings incorporated where applicable. Terms describing similar concepts were combined using the operator OR, while relevant concept groups were combined using AND. Representative search terms included “artificial intelligence”, “machine learning”, “deep learning”, “convolutional neural network”, “skin lesion”, “skin cancer”, “melanoma”, “non-melanoma skin cancer”, “basal cell carcinoma”, “squamous cell carcinoma”, “dermoscopy”, “dermatoscopy”, “clinical image”, “teledermatology”, “diagnostic accuracy”, “sensitivity”, “specificity”, “AUC”, “external validation”, “skin tone”, “Fitzpatrick skin type”, “skin of colour”, “darker skin”, “bias”, “equity”, “fairness”, “dataset diversity”, “governance”, “ethics”, “regulation”, “transparency”, “explainability”, “post-market surveillance”, and “medical device”.

Because this was a structured narrative review, the searches were iterative and adapted to the thematic domain under investigation. The objective was broad and balanced coverage of the principal clinical, equity-related, implementation, and governance issues rather than exhaustive retrieval of every publication concerning dermatology AI.

The search was centred primarily on biomedical literature and supplemented by multidisciplinary discovery tools and official regulatory sources. Consequently, some relevant engineering, legal, social-science, or policy literature not indexed in the selected sources may not have been identified.

2.3 Eligibility Criteria

The eligibility criteria used to screen and select studies are summarized in Table 1.

Table 1. Eligibility Criteria for Study Selection

Criterion	Included	Excluded
Source type	Peer-reviewed original studies, validation studies, systematic or scoping reviews, meta-analyses, dataset studies, digital health evaluations, implementation studies, and governance or ethics papers	Unverifiable publications, conference abstracts without accessible full text, opinion pieces without direct relevance, and publications lacking sufficient methodological information
Contextual and regulatory sources	Authoritative legislation, regulatory guidance, institutional reports	Commercial materials, non-authoritative regulatory commentary, and unverified online content
Population and setting	Human dermatology, skin cancer detection, lesion triage, primary care, specialist dermatology, and tele dermatology	Animal, laboratory-only, or non-clinical studies without relevance to dermatological diagnosis or triage
Technology and focus	AI, machine learning, deep learning, CNNs, multimodal AI, dermoscopy-based or clinical image-based tools, mobile health, and AI as a medical device for skin lesion or skin cancer classification	Technologies unrelated to image-based dermatology AI or studies unrelated to skin lesion classification
Outcomes	Diagnostic accuracy, sensitivity, specificity, AUC, external validation, subgroup performance, dataset diversity, skin tone representation, clinical safety, implementation, or governance	Purely technical model-development papers without clinical, equity, validation, or implementation relevance
Language and date	English-language full-text publications published between 2019 and 2026, with priority given to evidence published between 2021 and 2026.	Non-English publications, publications without accessible full text, and publications published before 2019.

Note. AI = artificial intelligence; AUC = area under the receiver operating characteristic curve; CNN = convolutional neural network.

2.4 Study Selection and Data Extraction

The core scholarly evidence base comprised 19 publications included in the thematic synthesis. Additional authoritative regulatory, policy, and institutional documents were used to contextualise governance, medical-device regulation, and post-deployment oversight. These contextual sources were analysed separately and were not included in the count of 19 core publications.

Data were extracted and organised using a structured thematic framework. The extracted domains included publication type, clinical setting, imaging modality, AI method, lesion categories, diagnostic reference standard, validation design, performance metrics, dataset characteristics, skin-tone or Fitzpatrick skin-type reporting, population diversity, external validation strategy, implementation context, and governance implications.

Particular attention was given to whether studies used internal validation, independent external validation, prospective evaluation, or real-world post-deployment assessment. Information concerning image sources, acquisition modalities, histopathological confirmation, clinical settings, and the representation of different skin-tone groups was considered where reported. Regulatory and institutional documents were analysed separately in relation to intended use, human oversight, transparency, accountability, risk management, post-market monitoring, and lifecycle governance.

Greater interpretive weight was given to systematic reviews and meta-analyses, independent external validation studies, prospective or real-world evaluations, multicentre datasets, and studies using histopathological confirmation. Findings from retrospective single-dataset evaluations, synthetic augmentation studies, and early multimodal or vision-language models were interpreted more cautiously because their reported performance may not directly reflect routine clinical practice.

3. Results

3.1 Structural Overview of the Evidence

The evidence base included systematic reviews and meta-analyses, external validation studies, prospective and real-world evaluations, dataset studies, qualitative implementation research, and regulatory documents. Quantitative evidence was most developed for diagnostic performance, whereas evidence concerning skin-tone equity, real-world implementation, and governance was more limited and heterogeneous. AI systems were most commonly evaluated using dermoscopic images, clinical photographs, or datasets combining both modalities (Salinas et al., 2024; Thomas et al., 2023; Tjiu & Lu, 2025).

3.2 Performance of AI-Based Models in Skin Lesion Classification

In a meta-analysis of 18 studies including more than 70,000 test images, AI-based skin lesion classification showed strong overall diagnostic performance. The pooled sensitivity was 0.91, indicating a high ability to identify lesions classified as positive, whereas the pooled specificity was 0.64, indicating a comparatively higher false-positive rate. Both the HSROC-AUC (hierarchical summary receiver operating characteristic area under the curve) and pooled AUROC (area under the receiver operating characteristic curve) reached 0.88, indicating good overall discrimination between malignant and benign lesions (Tjiu & Lu, 2025).

Salinas et al. (2024) reported a pooled sensitivity of 87.0% and specificity of 77.1% for AI systems. In the comparative meta-analysis, clinicians overall achieved a pooled sensitivity of 79.8% and specificity of 73.6%. In a separate subgroup analysis, the performance of AI was clinically comparable to that of expert dermatologists. However, such comparisons should be interpreted cautiously, as clinicians are often assessed using isolated images without access to the broader clinical context. Evidence indicates that the performance of some AI systems may be lower for darker skin tones and less common diseases when models are evaluated using diverse clinical image sets (Daneshjou et al., 2022; Patel Housley et al., 2025; Tjiu & Lu, 2025). These comparisons were conducted primarily under image-based evaluation conditions and did not consistently reproduce the full clinical context available during routine clinical assessment.

Daneshjou et al. (2022) reported substantial limitations in the performance of state-of-the-art dermatology AI models on the Diverse Dermatology Images dataset, which contained pathologically confirmed images representing a broader range of skin tones. Their findings indicated poorer model performance for darker skin tones and uncommon diseases, demonstrating that dataset composition is directly relevant to algorithmic fairness, external validity, and clinical reliability. Limited reporting of skin tone or Fitzpatrick skin type also restricts subgroup performance auditing and the assessment of algorithmic fairness (Daneshjou et al., 2022; Liu et al., 2023; Tjiu & Lu, 2025).

3.3 Social and Institutional Implications

The potential impact of dermatology AI extends beyond diagnostic accuracy. Its introduction may alter how work is distributed between dermatologists, primary care clinicians, teledermatology services, administrative staff, and patients. AI-assisted triage could improve referral prioritisation and reduce some unnecessary specialist assessments. However, a tool with high sensitivity but limited specificity may transfer rather than reduce workload by generating additional referrals, image reviews, biopsies, or follow-up assessments (Sangers et al., 2023; Smak Gregoor et al., 2023; Thomas et al., 2023).

Evidence from patient-facing applications reflects a similar trade-off between potential benefits and unintended effects on healthcare systems. Mobile tools may encourage earlier assessment of suspicious lesions, but they may also increase consultations for benign lesions, healthcare utilisation, and associated costs when outputs are not integrated into clearly defined referral pathways (Smak Gregoor et al., 2023). Conversely, false reassurance may delay professional assessment. The effects of AI should therefore be evaluated not only through diagnostic metrics but also through changes in referral volume, waiting times, clinical workload, biopsy rates, healthcare expenditure, and access to specialist care.

Clinician acceptance is also an important determinant of implementation. Dermatologists and general practitioners have identified potential benefits, including more efficient care pathways and improved support for clinical decision-making, while expressing concerns about diagnostic error, health inequalities, and the possibility that AI could replace rather than assist clinical judgement (Sangers et al., 2023).

3.4 Regulatory Oversight, Transparency, and Post-Market Monitoring

AI systems used to support diagnosis, triage, referral, or treatment decisions may fall within medical-device regulatory frameworks depending on their intended purpose, clinical claims, and jurisdiction. Regulatory assessment should therefore begin with a clearly defined intended use, target population, clinical setting, user group, and role within the decision-making pathway (Medicines and Healthcare products Regulatory Agency [MHRA], 2023; World Health Organization [WHO], 2023).

Relevant requirements include risk management, documentation of training and validation data, evaluation in representative populations, transparency regarding limitations, appropriate human oversight, quality-management procedures, and mechanisms for detecting changes in performance after deployment. The European Union Artificial Intelligence Act additionally establishes lifecycle requirements for certain high-risk AI systems, including human oversight and structured post-market monitoring (Regulation (EU) 2024/1689, 2024). Regulatory documents emphasised transparent reporting of intended use, validation populations, performance limitations, subgroup outcomes, and known failure modes..

Clinicians should receive appropriate training and retain the authority to question or override AI recommendations when they do not match the clinical findings. Responsibility for AI-supported decisions must also be clearly assigned among developers, manufacturers, healthcare organisations, clinicians, and service commissioners. The use of an algorithm should never create uncertainty about who is accountable for referral, discharge, biopsy, follow-up, or communication with the patient.

Post-market monitoring is particularly important because performance may change when a system is applied to new populations, devices, image-acquisition conditions, disease prevalence, or clinical workflows. Monitoring should include false-negative and false-positive results, subgroup performance, technical failures, inappropriate reassurance, referral outcomes, histopathological confirmation, user complaints, and changes associated with software updates. Model versions should be traceable so that changes in performance can be linked to specific algorithm updates.

Thomas et al. (2023) found that an AI-based system for assessing skin lesions maintained very high sensitivity for melanoma and other skin cancers when used in skin cancer referral pathways at two NHS hospitals. The newer version also classified a larger proportion of cases as potentially eligible for discharge within the evaluated clinical pathways, which included additional review procedures during the study period. These findings show that algorithm performance may change after software updates and should therefore be continuously evaluated in clinical practice rather than assessed only before deployment.

3.5 Quality and Limitations of the Included Evidence

The quality of the evidence varied substantially according to study design and research question. The strongest quantitative evidence was provided by systematic reviews and meta-analyses that synthesised diagnostic performance across multiple studies. Nevertheless, pooled estimates were affected by heterogeneity in datasets, lesion categories, imaging modalities, comparator groups, decision thresholds, and reference standards. Aggregate performance values may therefore conceal important differences between clinical settings and patient subgroups (Salinas et al., 2024; Tjiu & Lu, 2025).

A major limitation was the predominance of retrospective evaluations using curated or publicly available datasets. Salinas et al. (2024) reported that most included studies used internal test sets and that only 5.7% were prospective. This limits confidence that reported performance will be maintained when systems encounter new populations, different cameras, variable image quality, realistic disease prevalence, uncommon diagnoses, and routine clinical workflows. Studies using histopathological confirmation provide stronger diagnostic reference standards, but even these may be affected by spectrum bias when datasets contain a higher proportion of malignant or diagnostically challenging lesions than would be expected in primary care.

Evidence concerning skin tone-related performance was clinically important but remained fragmented. Several studies demonstrated poorer performance in darker skin tones, yet subgroup analyses were often limited by small sample sizes, inconsistent skin-tone classification, incomplete demographic reporting, or the absence of malignant cases within particular subgroups. These limitations make it difficult to estimate the magnitude of disparities with precision or to determine whether performance differences are attributable to pigmentation, image quality, disease distribution, dataset composition, or interacting factors (Daneshjou et al., 2022; Liu et al., 2023; Tjiu & Lu, 2025).

The evidence concerning regulatory governance was derived mainly from authoritative policy and regulatory documents together with a limited number of real-world implementation studies. These sources clarify expected principles, such as intended-use specification, risk management, human oversight, transparency, and post-market monitoring, but they do not establish that currently available dermatology AI systems consistently satisfy these requirements.

4. Discussion

4.1 AI-Based Skin Lesion Classification as a Socio-Technical Intervention

The findings of this structured narrative review indicate that AI-based skin lesion classification should be understood not only as an image-recognition task but also as a socio-technical intervention embedded within clinical workflows, referral systems, healthcare institutions, datasets, regulatory frameworks, and patient–clinician relationships. Although deep learning and convolutional neural network systems can support the classification of dermoscopic and clinical images, their clinical value depends on how their outputs are interpreted and incorporated into real diagnostic pathways (Escalé-Besa et al., 2023; Liopyris et al., 2022; Sangers et al., 2023; Thomas et al., 2023).

This distinction is clinically important because the diagnosis of skin cancer involves substantially more information than the visual appearance of an isolated lesion. Comparisons between AI systems and clinicians may therefore be misleading when clinicians are evaluated using decontextualised images without access to the information normally available during clinical assessment (Salinas et al., 2024).

AI systems should consequently be positioned primarily as clinical decision-support tools rather than independent substitutes for professional judgement.

The effects of AI implementation also extend beyond the individual diagnostic decision. Introducing an AI tool may change referral volumes, consultation patterns, clinician workload, biopsy rates, waiting times, healthcare expenditure, and the distribution of responsibility between primary care, specialist dermatology, teledermatology providers, and technology developers. An evaluation limited to technical accuracy may therefore overlook clinically and institutionally important consequences.

4.2 Why Overall Accuracy Can Mask Clinical Risk

A major implication of the reviewed literature is that aggregate diagnostic performance can conceal clinically meaningful risks when outcomes are not assessed across relevant patient subgroups, lesion categories, imaging conditions, and implementation settings. Sensitivity, specificity, area under the receiver operating characteristic curve, and classification accuracy are important measures, but they do not independently demonstrate clinical safety, fairness, or real-world effectiveness (Daneshjou et al., 2022; Liopyris et al., 2022; Salinas et al., 2024; Tjiu & Lu, 2025).

False-negative outputs may delay referral, biopsy, or treatment, whereas false-positive outputs may increase unnecessary interventions, patient anxiety, and pressure on specialist services. These risks may be concentrated among uncommon malignancies, atypical presentations, lower-quality images, or underrepresented populations. Moreover, a high-sensitivity system may generate substantial service burden when specificity is limited, whereas threshold changes intended to improve specificity may increase the risk of missed malignancies.

Benchmark datasets may also fail to reproduce the complexity of routine clinical practice. Curated dermoscopic images generally differ from smartphone photographs or standard clinical images with respect to lighting, focus, magnification, background, compression, framing, and device characteristics. Benchmark datasets may also contain a higher prevalence of malignant or diagnostically challenging lesions than would normally be encountered in primary care or patient-facing applications. Diagnostic accuracy reported under controlled conditions should therefore be interpreted as evidence of technical potential rather than proof of real-world clinical effectiveness (Escalé-Besa et al., 2023; Salinas et al., 2024; Thomas et al., 2023).

4.3 Skin Tone Bias as a Patient Safety and Equity Problem

Skin tone-related performance disparities represent one of the most important clinical and ethical concerns in dermatology AI. The appearance and visual contrast of a lesion may vary with background pigmentation, inflammatory response, lesion morphology, and the conditions under which an image is acquired. When training and validation datasets mainly include lighter skin tones, AI models may perform less reliably in patients with darker skin because the features they learn may not apply equally well across different skin tones (Daneshjou et al., 2022; Liu et al., 2023; Tjiu & Lu, 2025).

In their 2022 study, Daneshjou et al. found that leading dermatology AI models performed considerably worse when tested on the Diverse Dermatology Images dataset, which included patients with a broader range of skin tones, both common and rarer skin conditions, and diagnoses confirmed by histopathological examination. The models showed lower performance than in their original evaluations, with poorer results for

darker skin tones and less common diagnoses. The narrower performance gap observed after the model was further trained on more diverse images suggests that the range of data used for training directly affects how fairly and reliably it performs.

Similar concerns have been identified in evaluations of newer multimodal and vision-language systems. Differences in diagnostic performance across Fitzpatrick skin types suggest that advances in model architecture do not automatically eliminate biases arising from unrepresentative training data, inappropriate skin-tone measurement, or limited subgroup validation (Cirone et al., 2024; Patel Housley et al., 2025).

Skin tone bias should therefore not be treated as a marginal technical limitation. Lower sensitivity for malignant lesions in patients with darker skin could contribute to delayed referral, biopsy, or diagnosis. Lower specificity could lead to unnecessary interventions, anxiety, and additional healthcare use. These harms may disproportionately affect populations that are already underrepresented in dermatology education, clinical image datasets, and research, and that also face limited access to specialist care.

Equity concerns extend beyond algorithmic performance to the accessibility of the digital pathway in which a system is deployed. Patient-facing AI and teledermatology tools may improve access where specialist services are limited, but their benefits may be distributed unevenly. At the model level, the underrepresentation of darker skin tones and uncommon diagnoses in development and validation datasets may reduce diagnostic reliability for particular patient groups (Daneshjou et al., 2022; Liu et al., 2023). At the pathway level, age, previous computer or smartphone-application use, and confidence in digital communication may influence patients' willingness and ability to use remote dermatological care, while disparities in teledermatology uptake have been observed across demographic and socioeconomic groups (Choi et al., 2022; Jallow et al., 2023; Khoong et al., 2021). Limited access to reliable internet or mobile data may create further barriers, and the variable quality and clinical usefulness of patient-submitted photographs indicate that access to a device alone does not ensure effective participation (Jiang et al., 2022; Khoong et al., 2021). Equity assessment should therefore consider both subgroup diagnostic performance and whether patients can realistically access, navigate, and use the technology required to obtain a clinically interpretable assessment.

4.4 From Dataset Diversity to Clinically Meaningful Validation

The reviewed evidence indicates that dataset diversity is necessary but not sufficient for safe clinical implementation. Including more images from underrepresented populations may improve diversity in the dataset, but reliable clinical validation also requires testing the model in independent patient groups, different healthcare settings, and across a range of imaging devices, lesion types, and conditions in which images are taken (Salinas et al., 2024; Tjiu & Lu, 2025; Tschandl et al., 2026).

The MILK10k dataset represents an important advance in dermatology AI research, as it includes paired clinical and dermoscopic images of more than 5,000 skin lesions collected across multiple centres, covering a broad range of diagnoses, skin tones, and clinical presentations (Tschandl et al., 2026). Nevertheless, the existence of a more diverse dataset does not by itself demonstrate that a model will perform safely within primary care, teledermatology, specialist pathways, or patient-facing applications. External, prospective, and workflow-specific evaluations remain necessary.

The findings of Escalé-Besa et al. (2023) illustrate why retrospective or theoretical performance should not be assumed to translate directly into primary care. In their prospective real-world evaluation, machine learning performance was lower than that of general practitioners and teledermatology-based dermatologists. The study also demonstrated that model performance depended on whether the assessed condition was represented in the system's training categories.

Dataset transparency is central to assessing generalisability. Relevant information includes image source, clinical setting, imaging modality, device type, diagnostic category, reference standard, histopathological confirmation, anatomical site, patient demographics, skin-tone distribution, and whether the data were used for training, internal testing, or external validation. Without this information, clinicians and regulators cannot determine whether aggregate performance results are applicable to the population in which a system is intended to be used.

Fitzpatrick skin type also has important limitations as a measure of image-based skin colour because it was developed to describe skin responses to ultraviolet exposure rather than to provide a direct objective measurement of pigmentation. Research has shown that Fitzpatrick categories may be poorly predictive of objectively measured skin tone, particularly when categories are assigned retrospectively from uncontrolled images (Howard et al., 2021). Newer approaches, including the Monk Skin Tone Scale and computational skin-colour estimation, may provide more granular skin-tone annotation. However, their application remains sensitive to lighting, imaging conditions, annotator subjectivity, and the need for transparent validation in dermatological image datasets (Heldreth et al., 2024; Schumann et al., 2023; Weir et al., 2024).

4.5 Governance Responsibilities and Institutional Accountability

Governance responsibilities extend beyond regulatory approval and involve developers, manufacturers, healthcare institutions, clinicians, service commissioners, regulators, and patients. AI systems that influence lesion triage, referral, diagnosis, discharge, or biopsy decisions require evidence appropriate to their intended use and the level of clinical risk they introduce.

From a regulatory perspective, the intended purpose of the system should determine the population, setting, user group, performance requirements, and evidence needed for evaluation. A tool intended to prioritise referrals should not automatically be considered appropriate for autonomous diagnosis or patient reassurance. Similarly, evidence generated from retrospective dermoscopic datasets may not be sufficient to support deployment in primary care, smartphone-based assessment, or teledermatology pathways. Clinician concerns regarding accuracy, inequality, accessibility, and professional liability further support the need for supervised and accountable implementation (Sangers et al., 2023).

Institutional readiness involves more than purchasing or technically integrating a system. Healthcare organisations require clear procedures defining who reviews an AI output, who communicates the result, when a clinician must override the system, how uncertain results are escalated, and who remains responsible for the final clinical decision. Training should include not only how to operate the technology but also its intended use, limitations, target population, known failure modes, and the meaning of its performance metrics.

Governance should also address automation bias, in which clinicians may place excessive confidence in algorithmic recommendations, and algorithm aversion, in which visible errors lead users to disregard potentially helpful outputs. AI-supported decisions should remain reviewable, and clinicians should retain the authority to question or override recommendations that are inconsistent with the clinical presentation.

Safe use requires transparency regarding the system's intended purpose, training data, validation populations, performance metrics, subgroup performance, known failure modes, clinical oversight, and post-market monitoring (Sangers et al., 2023; Thomas et al., 2023; WHO, 2023). Explainability methods may support the interpretation of model outputs, but heatmaps and similar visualisation tools should not be treated as substitutes for clinical validation or as complete explanations of model reasoning (Liopyris et al., 2022; WHO, 2021).

Thomas et al. (2023) demonstrated the value of post-deployment evaluation by assessing changes in real-world performance between algorithm versions within NHS pathways. Their findings show that model performance and potential effects on patient discharge may evolve after implementation. Pre-deployment validation should therefore be supplemented by continuous surveillance and predefined corrective actions, such as recalibration, additional validation, restricted use, or withdrawal when safety requirements are not maintained.

4.6 Practical Recommendations for Future Research and Implementation

Future research should prioritise prospective, multicentre, and external validation in realistic clinical environments, including primary care, teledermatology, specialist dermatology, and patient-acquired mobile imaging. Validation should be conducted in populations that differ meaningfully from the data used for model development (Escalé-Besa et al., 2023; Salinas et al., 2024; Tjiu & Lu, 2025).

Studies should report sensitivity, specificity, area under the receiver operating characteristic curve, predictive values, calibration, and clinically relevant error patterns. Aggregate results should be accompanied by subgroup analyses according to skin tone, image modality, lesion category, device, and clinical setting when sample sizes are sufficient. Researchers should report subgroup denominators and confidence intervals to avoid overinterpreting unstable estimates derived from very small samples.

Dataset development should intentionally include darker skin tones, uncommon diseases, non-pigmented lesions, benign mimics, varied imaging conditions, and clinically representative disease spectra. Synthetic data augmentation may supplement dataset development but should not replace the collection of real, clinically validated images from diverse patient populations (Rezk et al., 2022).

Diagnostic reference standards should be reported clearly, including whether diagnoses were established through histopathology, expert consensus, clinical follow-up, or image-derived annotation. Histopathological confirmation generally provides a stronger reference standard for malignant lesions, although datasets composed predominantly of biopsied lesions may not reflect the disease spectrum encountered in routine triage.

Implementation studies should assess more than diagnostic discrimination. Relevant outcomes include referral volume, waiting times, biopsy and excision rates, clinician workload, patient anxiety, costs, access to

care, and unintended inequalities. These outcomes are necessary to determine whether an AI tool improves the overall care pathway rather than only performing well on selected image sets.

AI implementation should initially occur through clinician-supervised workflows with clear escalation procedures for suspicious, uncertain, technically inadequate, or discordant cases. Patient-facing applications require particular caution because false reassurance may delay professional medical assessment, while excessive warnings may increase anxiety and unnecessary healthcare use (Cirone et al., 2024; Smak Gregoor et al., 2023). To ensure fair evaluation, studies should report performance across different patient groups, examine both false-negative and false-positive results, clearly describe the distribution of skin tones in the dataset, and include sufficient numbers of malignant and benign lesions within each subgroup.

4.7 Limitations of This Review

Several limitations should be considered when interpreting this review. As a structured narrative review, it did not follow a registered systematic-review protocol or the full PRISMA framework. Its findings should therefore not be regarded as an exhaustive assessment of the literature or as a pooled estimate of diagnostic accuracy. Although the narrative approach allowed evidence from diagnostic studies, dataset analyses, implementation research, and regulatory sources to be examined together, it carries a greater risk of selection and interpretation bias than a systematic review based on duplicate screening and formal risk-of-bias assessment.

Interpretation was additionally constrained by heterogeneity across study designs, datasets, imaging modalities, reference standards, validation methods, and skin-tone reporting.

Evidence on multimodal models, synthetic data, clinical implementation, and governance remains limited or exploratory, while real-world post-deployment studies are less common than retrospective image-classification research. Conclusions regarding long-term safety, workflow effects, cost-effectiveness, clinician behaviour, and equity should therefore be interpreted cautiously. Regulatory frameworks are also continuing to evolve, meaning that the conclusions of this review may require revision as further prospective studies, external validations, regulatory guidance, and post-market surveillance data become available.

5. Conclusions

AI-based skin lesion classification has considerable potential to support dermatological care, particularly in lesion triage, teledermatology, primary care referral pathways, and settings where access to specialist dermatology is limited. Current evidence suggests that deep learning models, especially convolutional neural networks, can achieve promising diagnostic performance in the classification of dermoscopic and clinical images. However, performance reported in curated datasets or internally validated studies should not be assumed to reflect safe, equitable, or generalisable performance in routine clinical practice.

A central finding of this review is that skin tone-related bias represents both a patient safety concern and a health equity issue. The underrepresentation of darker skin tones, particularly Fitzpatrick skin types IV–VI, may reduce the reliability of AI systems for patients with skin of colour and contribute to delayed diagnosis, inappropriate reassurance, or unequal diagnostic pathways. High overall measures of performance, such as AUC, sensitivity, or specificity, may therefore conceal clinically important subgroup differences when validation datasets lack diversity or fail to report skin tone composition transparently.

Safe implementation of AI in dermatology requires governance that extends beyond technical accuracy. Regulatory oversight, transparent dataset reporting, external validation, subgroup performance analysis, explainability, clinician accountability, and post-deployment monitoring should be treated as essential requirements for systems that influence diagnostic or triage decisions. AI should be used to support clinical judgement, rather than replace medical assessment, dermoscopic expertise, or histopathological confirmation when malignancy is suspected.

Future research should prioritise diverse and transparently annotated datasets, prospective external validation, clinically realistic testing environments, robust methods for measuring skin tone, fairness metrics, and real-world monitoring of diagnostic outcomes across different patient populations. With equity-centred validation and accountable governance, AI-based skin lesion classification has the potential to contribute to dermatological care that is safer, more inclusive, and clinically meaningful.

Author Contributions:

Conceptualization, Investigation, Visualization: Faustyna Magdalena Kołodziej

Conceptualization, Resources, Visualization: Maciej Kamil Majchrzak

Methodology, Supervision: Krzysztof Konopka

Methodology, Supervision, Data Curation: Eliza Zofia Kabat

Software, Data Curation, Project Administration: Alicja Kwiatkowska

Validation, Writing – Original Draft, Project Administration: Jakub Juskowiak

Validation, Writing – Original Draft: Maryan Liasota

Formal Analysis, Writing – Review & Editing: Michalina Justyna Kozłowska

Formal Analysis, Writing – Review & Editing: Mateusz Lis

Investigation, Writing – Review & Editing: Zuzanna Julia Komin

All authors have read and agreed to the submitted version of the manuscript.

Funding statement: This research received no external funding.

No conflicts of interest to declare.

Declaration of the use of generative AI and AI-assisted technologies in the writing process: In preparing this work, the authors used ChatGPT for the purpose of improving language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

1. Ang, X. L., & Oh, C. C. (2025). The use of artificial intelligence for skin cancer detection in Asia: A systematic review. *Diagnostics*, 15(7), 939. <https://doi.org/10.3390/diagnostics15070939>
2. Cirone, K., Akrou, M., Abid, L., & Oakley, A. (2024). Assessing the utility of multimodal large language models (GPT-4 Vision and Large Language and Vision Assistant) in identifying melanoma across different skin tones. *JMIR Dermatology*, 7, e55508. <https://doi.org/10.2196/55508>
3. Daneshjou, R., Vodrahalli, K., Novoa, R. A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S. M., Bailey, E. E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Allerup, J. A. C., Okata-Karigane, U., Zou, J., & Chiou, A. S. (2022). Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science Advances*, 8(32), eabq6147. <https://doi.org/10.1126/sciadv.abq6147>
4. Escalé-Besa, A., Yélamos, O., Vidal-Alaball, J., Fuster-Casanovas, A., Miró Catalina, Q., Börve, A., Ander-Egg Aguilar, R., Fustà-Novell, X., Cubiró, X., Esquius Rafat, M., López-Sanchez, C., & Marin-Gomez, F. X. (2023). Exploring the potential of artificial intelligence in improving skin lesion diagnosis in primary care. *Scientific Reports*, 13(1), 4293. <https://doi.org/10.1038/s41598-023-31340-1>
5. Heldreth, C. M., Monk, E. P., Clark, A. T., Schumann, C., Eyee, X., & Ricco, S. (2024). Which skin tone measures are the most inclusive? An investigation of skin tone measures for artificial intelligence. *ACM Journal on Responsible Computing*, 1(1), Article 7, 1–21. <https://doi.org/10.1145/3632120>
6. Howard, J. J., Sirotin, Y. B., Tipton, J. L., & Vemury, A. R. (2021). Reliability and validity of image-based and self-reported skin phenotype metrics. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(4), 550–560. <https://doi.org/10.1109/TBIOM.2021.3123550>
7. Li, Z., Koban, K. C., Schenck, T. L., Giunta, R. E., Li, Q., & Sun, Y. (2022). Artificial intelligence in dermatology image analysis: Current developments and future trends. *Journal of Clinical Medicine*, 11(22), 6826. <https://doi.org/10.3390/jcm11226826>
8. Liopyris, K., Gregoriou, S., Dias, J., & Stratigos, A. J. (2022). Artificial intelligence in dermatology: Challenges and perspectives. *Dermatology and Therapy*, 12(12), 2637–2651. <https://doi.org/10.1007/s13555-022-00833-8>
9. Liu, Y., Primiero, C. A., Kulkarni, V., Soyer, H. P., & Betz-Stablein, B. (2023). Artificial intelligence for the classification of pigmented skin lesions in populations with skin of color: A systematic review. *Dermatology*, 239(4), 499–513. <https://doi.org/10.1159/000530225>
10. Marsden, H., Morgan, C., Austin, S., DeGiovanni, C., Venzi, M., Kemos, P., Greenhalgh, J., Mullarkey, D., & Palamaras, I. (2023). Effectiveness of an image analyzing AI-based digital health technology to identify non-melanoma skin cancer and other skin lesions: Results of the DERM-003 study. *Frontiers in Medicine*, 10, 1288521. <https://doi.org/10.3389/fmed.2023.1288521>
11. Medical Device Coordination Group, & Artificial Intelligence Board. (2025). *FAQ on interplay between the Medical Devices Regulation (MDR) & In vitro Diagnostic Medical Devices Regulation (IVDR) and the Artificial Intelligence Act (AIA)* (MDCG 2025-6/AIB 2025-1). European Commission. https://health.ec.europa.eu/document/download/b78a17d7-e3cd-4943-851d-e02a2f22bbb4_en

12. Medicines and Healthcare products Regulatory Agency. (2023, June 14). *Software and AI as a Medical Device Change Programme roadmap*. GOV.UK. <https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme/software-and-ai-as-a-medical-device-change-programme-roadmap>
13. Patel Housley, P., Samia, A. M., & Motaparathi, K. (2025). Performance evaluation of ChatGPT-4o in dermatological diagnoses across Fitzpatrick skin types. *Cureus*, 17(7), e87368. <https://doi.org/10.7759/cureus.87368>
14. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No. 300/2008, (EU) No. 167/2013, (EU) No. 168/2013, (EU) 2018/858, (EU) 2018/1139, and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797, and (EU) 2020/1828 (Artificial Intelligence Act), *Official Journal of the European Union*, L, 2024/1689 (2024). <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
15. Rezk, E., Eltorki, M., & El-Dakhakhni, W. (2022). Improving skin color diversity in cancer detection: Deep learning approach. *JMIR Dermatology*, 5(3), e39143. <https://doi.org/10.2196/39143>
16. Salinas, M. P., Sepúlveda, J., Hidalgo, L., Peirano, D., Morel, M., Uribe, P., Rotemberg, V., Briones, J., Mery, D., & Navarrete-Dechent, C. (2024). A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis. *npj Digital Medicine*, 7(1), 125. <https://doi.org/10.1038/s41746-024-01103-x>
17. Sanchez, K., Kamal, K., Manjaly, P., Ly, S., & Mostaghimi, A. (2023). Clinical application of artificial intelligence for non-melanoma skin cancer. *Current Treatment Options in Oncology*, 24(4), 373–379. <https://doi.org/10.1007/s11864-023-01065-4>
18. Sangers, T. E., Wakkee, M., Moolenburgh, F. J., Nijsten, T., & Lugtenberg, M. (2023). Towards successful implementation of artificial intelligence in skin cancer care: A qualitative study exploring the views of dermatologists and general practitioners. *Archives of Dermatological Research*, 315(5), 1187–1195. <https://doi.org/10.1007/s00403-022-02492-3>
19. Schumann, C., Olanubi, G. O., Wright, A., Monk, E., Jr., Heldreth, C., & Ricco, S. (2023). Consensus and subjectivity of skin tone annotation for ML fairness. *Advances in Neural Information Processing Systems*, 36, 30319–30348. https://proceedings.neurips.cc/paper_files/paper/2023/hash/60d25b3210c92f5ba2002a8e1f1adflc-Abstract-Datasets_and_Benchmarks.html
20. Smak Gregoor, A. M., Sangers, T. E., Bakker, L. J., Hollestein, L., Uyl-de Groot, C. A., Nijsten, T., & Wakkee, M. (2023). An artificial intelligence based app for skin cancer detection evaluated in a population based setting. *npj Digital Medicine*, 6(1), 90. <https://doi.org/10.1038/s41746-023-00831-w>
21. Thomas, L., Hyde, C., Mullarkey, D., Greenhalgh, J., Kalsi, D., & Ko, J. (2023). Real-world post-deployment performance of a novel machine learning-based digital health technology for skin lesion assessment and suggestions for post-market surveillance. *Frontiers in Medicine*, 10, 1264846. <https://doi.org/10.3389/fmed.2023.1264846>
22. Tjiu, J.-W., & Lu, C.-F. (2025). Equity and generalizability of artificial intelligence for skin-lesion diagnosis using clinical, dermoscopic, and smartphone images: A systematic review and meta-analysis. *Medicina*, 61(12), 2186. <https://doi.org/10.3390/medicina61122186>
23. Tschandl, P., Akay, B. N., Rosendahl, C., Rotemberg, V., Todorovska, V., Weber, J., Wolber, A. K., Müller, C., Kurtansky, N., Halpern, A., Weninger, W., & Kittler, H. (2026). MILK10k: A hierarchical multimodal imaging-learning toolkit for diagnosing pigmented and nonpigmented skin cancer and its simulators. *Journal of Investigative Dermatology*, 146(2), 357–364.e7. <https://doi.org/10.1016/j.jid.2025.06.1594>
24. Walker, B. N., Blalock, T. W., Leibowitz, R., Oron, Y., Dascalu, D., David, E. O., & Dascalu, A. (2025). Skin cancer detection in diverse skin tones by machine learning combining audio and visual convolutional neural networks. *Oncology*, 103(5), 413–420. <https://doi.org/10.1159/000541573>
25. Weir, V. R., Dempsey, K., Gichoya, J. W., Rotemberg, V., & Wong, A.-K. I. (2024). A survey of skin tone assessment in prospective research. *npj Digital Medicine*, 7(1), 191. <https://doi.org/10.1038/s41746-024-01176-8>
26. World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. <https://www.who.int/publications/i/item/9789240029200>
27. World Health Organization. (2023). *Regulatory considerations on artificial intelligence for health*. <https://www.who.int/publications/i/item/9789240078871>