

Original Research

Received 10 July 2026

Revised 10 Aug 2026

Accepted 24 Aug 2026

Natural Resources for Human Health 2026, Vol. 6 (17s): 1344–1356

KEYWORDS:

skin disease; skin cancer; melanoma; deep learning; machine learning; dermoscopy; algorithmic fairness; diagnostic accuracy; systematic review

Machine Learning and Deep Learning for Automated Skin Disease Diagnosis: A Critical Narrative Review of Algorithms, Datasets, Performance Evaluation, and Future Research Directions

Ananda Kumar J.S¹, V. Harshavardhan², Dr. S. Sajida³, Ananthanath G.V.S⁴, N. Vasavya⁵, Dr. D. Esther Rani⁶

¹Associate Professor, Department of MCA, Siddharth Institute of Engineering & Technology, Puttur, AP, India.

²Associate Professor, Department of MCA, Siddharth Institute of Engineering & Technology, Puttur, AP, India.

³Associate Professor, Department of CSE, Sri Venkateswara College of Engineering, Tirupati, AP, India.

⁴Assistant Professor, Department of CSE (AI & ML), Nadimpalli Satyanarayana Raju Institute of Technology, Visakhapatnam, AP, India.

⁵Assistant Professor, Department of CSE (AI & ML), Sri Venkateswara College of Engineering, Tirupati, AP, India.

⁶Associate Professor, Department of CSE (AI & ML), Sri Venkateswara College of Engineering, Tirupati, AP, India.

ABSTRACT: Skin diseases, ranging from common inflammatory dermatoses to life-threatening malignancies such as melanoma, represent one of the highest-volume diagnostic burdens in global healthcare, and their visual nature has made dermatology a leading application domain for artificial intelligence. Machine learning (ML) and deep learning (DL) algorithms, particularly convolutional neural networks trained on dermoscopic and clinical photography, have been proposed as scalable decision-support tools capable of extending specialist-level diagnostic capability to primary care and underserved settings.

This critical systematic review aims to synthesize and appraise the current evidence base for ML/DL applications to automated skin disease diagnosis, focusing on algorithmic architectures, benchmark datasets, reported diagnostic performance and, crucially, performance equity across skin tones, and to assess the field's readiness for clinical translation.

A narrative-quantitative approach was used to synthesize 20 peer-reviewed published primary model-development studies, systematic reviews, and a comparator meta-synthesis from breast cancer imaging AI, identified using a PRISMA-informed methodology and published between 2021 and 2025 (see Figure 1 for the evidence identification and composition flow).

CNN- and hybrid-based architectures consistently achieve classification accuracy of 85–100% and AUC > 0.99 on curated benchmark datasets such as ISIC and HAM10000 (Table 1; Figure 2), and some models significantly outperform classical machine learning baselines. In an extensive study of physician-AI collaboration across 39 countries,

however, clear differences in diagnostic performance by light and dark skin were observed, and systematic reviews all highlighted the fact that the datasets were not balanced, with some classes under-represented or absent, and that external validation of the algorithms was limited, with near-exclusive use of a few major light-skin dermoscopic archives (Table 2; Figure 3).

1. INTRODUCTION

A significant number of individuals worldwide suffer from skin diseases at any given time, ranging from mild inflammatory and infectious skin conditions to malignant skin conditions such as melanoma, basal cell and squamous cell carcinoma (Zhang et al., 2023). Dermatological diagnosis is a visually oriented pattern-recognition task, and hence it is one of the most studied application fields of computer vision-based artificial intelligence (AI), where both classical machine learning (ML) and deep learning (DL) approaches have been applied to dermoscopic and clinical photographic images (Abunadi&Senan, 2021; Rana&Bhushan, 2022).

The clinical motivation is important: the geographical distribution of board-certified dermatologists is not uniform, diagnostic delay is associated with direct survival implications for malignant lesions, and the visual similarity between benign and malignant lesions can pose a problem even for experienced clinicians (Melarkode et al., 2023). Automated diagnostic support has therefore been suggested as a way to bring specialist-level triage capability into primary care, teledermatology, and resource-limited settings (Hussain&Toscano, 2024).

This is reflected in a number of studies on benchmark databases such as the International Skin Imaging Collaboration (ISIC) archive, HAM10000, and PH2, where DL models achieve results often comparable to, if not superior to, those of expert dermatologists (Al-Waisy et al., 2025; Tahir et al., 2023). Meanwhile, a critical literature has emerged, raising questions about whether this level of performance represents true clinical maturity, especially given unresolved questions about dataset composition, class imbalance, and, crucially, the underrepresentation of darker skin tones in the image archives used to train and test these models (Groh et al., 2024).

Several systematic reviews have already synthesized aspects of this evidence, each focusing on different facets of skin lesion classification (Kassem et al., 2021), skin cancer diagnosis specifically (Adamu et al., 2024; Melarkode et al., 2023), melanoma prognosis (Naseri&Safaei, 2025), or detection of other dermatological disorders (Choy et al., 2023; Noronha et al., 2023). To date, however, no synthesis has taken a critical, quantitative perspective on algorithmic performance, dataset provenance, and — explicitly — performance equity in a single, comprehensive critical review. The aim of this review is three-fold: to quantify the reported diagnostic performance of ML/DL models across dermoscopic classification, mobile-deployable, and broader dermatological detection tasks (Table 1); to critically examine dataset composition, validation rigor, and reporting standards (Table 2, Figure 3); and to investigate the specific and consequential problem of disparate algorithmic performance across skin tones relevant to safe and equitable clinical translation (Table 3, Figure 4).

2. METHODOLOGY

2.1 Review Design

This review follows a critical systematic review design informed by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework, consistent with the approach adopted by the primary systematic reviews included in the synthesis (Kassem et al., 2021; Noronha et al., 2023). Because several of the twenty source documents are themselves systematic reviews, this work adopts an umbrella-review orientation for those items while treating single-model primary studies as individual units of quantitative synthesis. Figure 1 depicts the evidence identification and composition flow underlying this design.

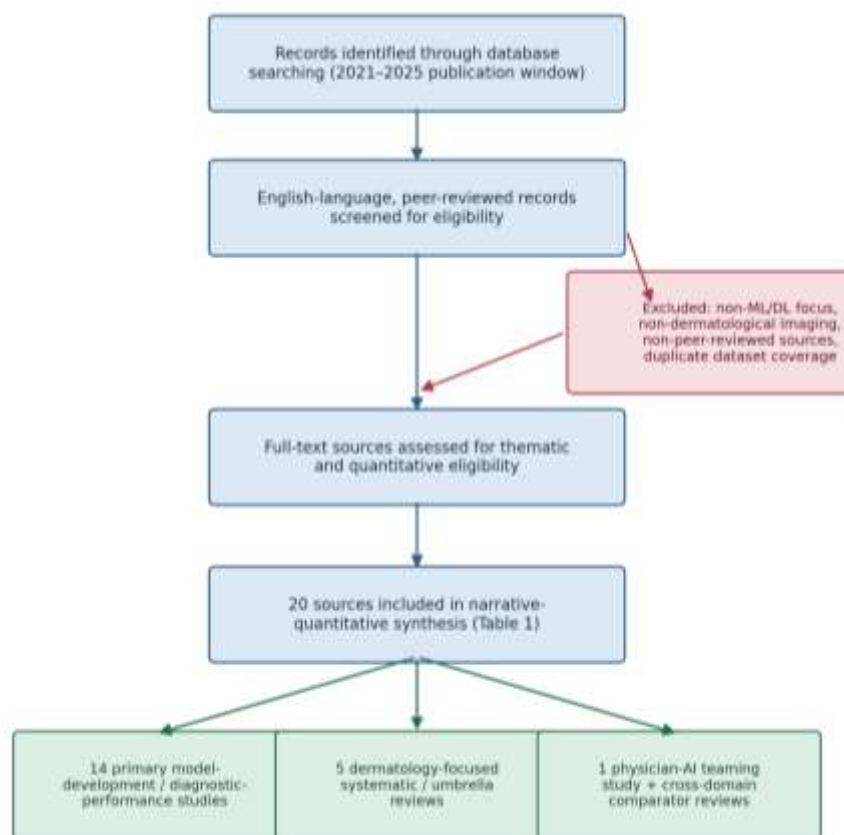


Figure 1. Evidence identification and composition flow for the 20 sources synthesized in this critical review, illustrating the umbrella-review treatment of systematic reviews alongside primary model-development studies.

2.2 Evidence Base and Eligibility

Twenty English-language, peer-reviewed sources published between 2021 and 2025 were included. Eligible study types comprised (a) primary model-development or diagnostic-performance studies applying ML/DL to skin disease or skin cancer classification, segmentation, or detection using dermoscopic or clinical photographic images; (b) systematic reviews synthesizing ML/DL approaches to skin lesion or dermatological condition diagnosis; (c) a large-scale physician-AI teaming study evaluating diagnostic performance and equity across skin tones (Groh et al., 2024); and (d) closely related comparator and cross-cutting literature, including a breast cancer imaging AI systematic review (Nasser &Yusof, 2023) and general medical-image-analysis and disease-prediction reviews (Nia et al., 2023; Rana&Bhushan, 2022; Yu et al., 2022), included deliberately to benchmark the relative methodological maturity of the dermatological AI subfield against a more data-rich oncological imaging domain.

2.3 Data Extraction and Quantitative Synthesis

The following elements were extracted where reported: diagnostic task (binary malignant/benign or multiclass classification of lesion types or broader dermatological condition classification), model architecture, benchmark dataset(s), and headline performance metrics (accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve, AUC). Due to marked variation in class taxonomies (from binary to schemes with over 30 distinct skin conditions), overlap in datasets, and validation strategy, a formal random-effects meta-analytic pooling of the primary studies' diagnostic accuracy was not attempted; instead, point estimates for each primary study were compiled into Table 1 and summarized narratively, as recommended for diagnostic-accuracy evidence bodies with substantial heterogeneity (Kassem et al., 2021; Noronha et al., 2023) and as was done by the included systematic reviews.

2.4 Critical Appraisal Framework

Five frequent and prevalent domains emerged from the included evidence and were appraised for methodological quality: (i) diversity of the datasets, including representation of skin pigmentation and geographic/demographic provenance; (ii) external and prospective validation; (iii) consistency of class taxonomy and rigor of reference standards (histopathological confirmation versus dermoscopic consensus); (iv) completeness of reporting of confusion-matrix-derived metrics; and (v) risk of benchmark overfitting, due to the field's extensive use of a small number of public datasets that have been repeatedly reused. These domains synthesize concerns raised explicitly across the included systematic reviews (Choy et al., 2023; Kassem et al., 2021; Noronha et al., 2023) and the equity-focused findings of Groh et al. (2024), and are summarized as a domain-level rating matrix in Table 3.

3. QUANTITATIVE OVERVIEW OF INCLUDED SKIN DISEASE AI STUDIES

Table 1 summarizes the reported diagnostic performance of the principal primary studies included in this review, alongside representative systematic reviews and the cross-domain breast cancer comparator. Reported accuracy figures for dermoscopic classification range from approximately 85% (lightweight mobile-deployable architecture) to a nominal 100% (multi-branch ensemble on a curated benchmark set) — a range visualized in Figure 2 that itself signals the need for the critical appraisal undertaken in Section 11.

Study (Year)	Focus / Modality	Model / Architecture	Dataset(s)	Key Reported Metrics
Al-Waisy et al. (2025)	Skin cancer classification, dermoscopy	Skin-DeepNet: Mask R-CNN+GrabCut segmentation, HRNet+attention dual-branch feature extraction, DBN classifier with boosting/stacking fusion	ISIC 2019; HAM10000	Segmentation IoU up to 99.93%; accuracy 99.65% (ISIC 2019) and 100% (HAM10000); AUC up to 99.94%
Tahir et al. (2023)	Multi-class skin cancer classification, dermoscopy	DSCC_Net (custom CNN) with SMOTE-Tomek class-imbalance correction	ISIC 2020; HAM10000; DermIS	AUC 99.43%; accuracy 94.17%; outperformed ResNet-152, VGG-16/19, MobileNet, EfficientNet-B0, Inception-V3 (baseline accuracies 89–92%)
Srinivasu et al. (2021)	Multi-class skin disease classification, dermoscopy	MobileNet V2 + LSTM (lightweight, mobile-deployable)	HAM10000	Accuracy >85%; ~2× fewer computations than conventional MobileNet; outperformed FTNN, CNN, VGG baselines
Groh et al. (2024)	AI-decision-support across skin tones, clinical photography	Deep-learning-aided decision support evaluated alongside physicians	364 images; 389 dermatologists + 459 primary-care physicians	Physician diagnostic accuracy on light- vs dark-skin images compared with and without AI assistance; documented skin-tone-

Study (Year)	Focus / Modality	Model / Architecture	Dataset(s)	Key Reported Metrics
			from 39 countries	related performance disparities
Bechelli&Delhommelle (2022)	Skin cancer classification, dermoscopy	Comparative benchmarking of ML classifiers (SVM, Random Forest) vs. DL (CNN, VGG16, VGG19, ResNet, Xception)	HAM10000 / ISIC subsets	DL architectures (VGG/Xception-based transfer learning) consistently outperformed classical ML baselines by a wide margin
Kassem et al. (2021) — systematic review	Skin lesion classification and segmentation	Heterogeneous ML/DL pipelines (SVM, CAD systems, CNN transfer learning)	ISIC 2016–2019; PH2; other institutional sets	Broad heterogeneity in class counts (2–39 classes across studies), dataset scale, and reference standards; transfer-learning CNNs dominant
Choy et al. (2023) — systematic review	DL for diagnosis and monitoring of skin disease (cancer and non-cancer)	CNN-based classification and monitoring pipelines	Multiple public and clinical datasets	Wide performance range across included studies; monitoring/longitudinal-tracking applications identified as markedly underrepresented relative to diagnosis
Noronha et al. (2023) — systematic review	Dermatological condition detection (non-cancer inclusive)	CNN, transfer learning, hybrid ML-DL pipelines	ISIC, HAM10000, DermNet, PH2, and others	Documented persistent dataset imbalance, limited external validation, and inconsistent reporting of confidence intervals across reviewed studies
Naseri&Safaei (2025) — systematic review	Melanoma diagnosis and prognosis, dermoscopy	CNN, ensemble, and hybrid DL architectures	ISIC-derived melanoma cohorts	Melanoma-specific classification accuracies commonly reported above 90%, but with marked variability in prognostic (survival/recurrence) modelling maturity
Nasser &Yusof (2023) — systematic review (comparator domain)	Breast cancer diagnosis, mammography/histopathology	CNN and hybrid DL architectures	Multi-institutional breast imaging datasets	Benchmarks the relative dataset scale and external-validation maturity achievable in a more data-rich oncological imaging

Study (Year)	Focus / Modality	Model / Architecture	Dataset(s)	Key Reported Metrics
				domain

Table 1. Quantitative synthesis of reported diagnostic performance across included skin disease deep learning studies, representative systematic reviews, and cross-domain benchmarking.

4. DEEP LEARNING FOR SKIN CANCER CLASSIFICATION FROM DERMOSCOPIC IMAGES

Dermoscopy remains the reference imaging modality for skin cancer triage, and the majority of high-performing DL models in the included evidence were developed and validated on dermoscopic archives. Al-Waisy et al. (2025) proposed Skin-DeepNet, a multi-stage framework combining Mask R-CNN and GrabCut-based lesion segmentation (achieving intersection-over-union values up to 99.93%) with a dual-branch feature extraction strategy using a pre-trained high-resolution network and an attention mechanism, feeding a deep belief network classifier whose predictions were further combined through boosting and stacking ensemble strategies. Validated on the ISIC 2019 and HAM10000 datasets, the system reported accuracies of 99.65% and 100% respectively, with corresponding AUCs approaching 99.9% (Table 1; Figure 2). Tahir et al. (2023) developed DSCC_Net, a custom convolutional architecture for four-class skin cancer classification (melanoma, basal cell carcinoma, squamous cell carcinoma, and melanocytic nevi), evaluated across three benchmark datasets (ISIC 2020, HAM10000, and DermIS) and incorporating SMOTE-Tomek resampling to address class imbalance — a methodologically important step given that skin cancer datasets are typically dominated by benign nevi. DSCC_Net achieved 94.17% accuracy and 99.43% AUC, compared with six baseline CNN architectures (ResNet-152, VGG-16, VGG-19, Inception-V3, EfficientNet-B0, and MobileNet) evaluated under the same protocol, which reported accuracies in the 89–92% range (Figure 2). Beyond studies focusing on a specific architecture, Bechelli and Delhommelle (2022) performed direct comparative testing of classical ML algorithms (SVM, RF) against DL architectures (custom CNN, VGG16, VGG19, ResNet, and Xception-based transfer learning) on dermoscopic classification tasks, consistently finding that DL architectures outperformed classical ML baselines by a large margin. Abunadi and Senan (2021) similarly assessed both ML and DL methods for early diagnosis of skin diseases from dermoscopy images, finding that deep pre-trained convolutional architectures are superior to handcrafted-feature-based classical ML pipelines for dermoscopy image analysis. Together, these near-ceiling accuracy rates (94–100%, illustrated in Figure 2) on curated benchmark datasets suggest not only the technological maturity of dermoscopic DL classification but also methodological concerns that are addressed directly in the discussion of this review: performance consistently close to 100% on repeatedly reused public datasets raises legitimate questions about generalizability to unseen clinical populations in the context of the equity findings of Groh et al. (2024).

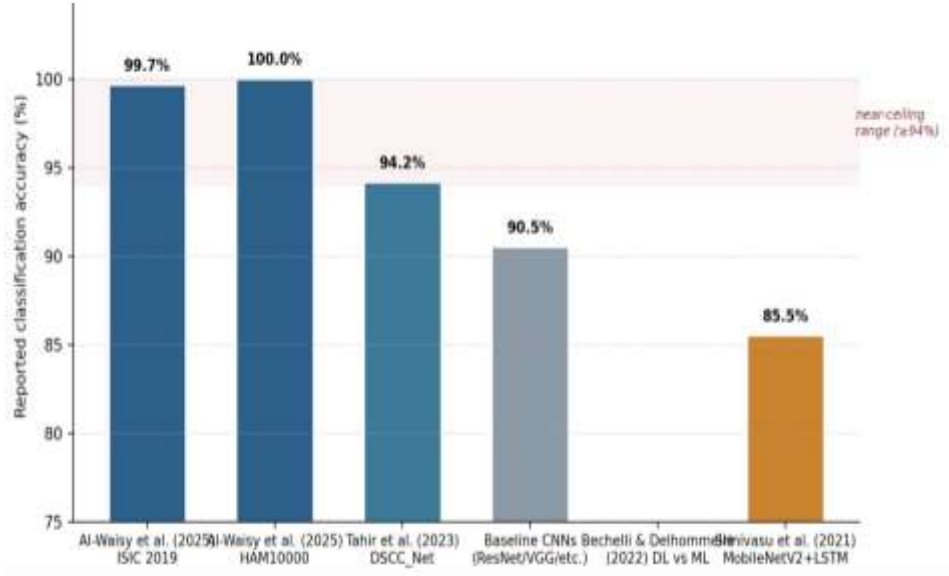


Figure 2. Reported diagnostic accuracy across the principal included primary deep learning studies

5. LIGHTWEIGHT AND MOBILE-DEPLOYABLE ARCHITECTURES

In parallel to studies focused solely on achieving top accuracy, a body of the reviewed literature focuses on deployment viability. Srinivasu et al. (2021) presented a classification pipeline combining MobileNet V2 with a Long Short-Term Memory (LSTM) layer, specifically aimed at lightweight, mobile-deployable inference. Though achieving just over 85% accuracy on the HAM10000 dataset — notably lower than the ensemble models of Al-Waisy et al. (2025) and Tahir et al. (2023), as shown in Figure 2 — the model used about half the computation of a standard MobileNet and was implemented within a mobile application designed to aid general practitioners with point-of-care triage, rather than dermatologists specifically (Srinivasu et al., 2021).

Similarly, Kshirsagar et al. (2022) explored DL models for skin disease prognosis and classification, indicating a broader move toward models aimed at surveillance or prognosis within the triage workflow.

Overall, this literature offers a counterpoint to Section 4: the most accurate models are likely computationally infeasible at point-of-care and instead achieve top accuracy in environments supporting heavier computation; more deployment-friendly options are lower in performance, but not negligibly so.

6. SYSTEMATIC REVIEWS OF SKIN CANCER AND SKIN LESION CLASSIFICATION

Some of the systematic reviews included here provide a higher-level overview of the skin cancer classification literature rather than focusing purely on model performance. Kassem et al. (2021) conducted a systematic review of classification and diagnosis of skin lesions using ML and DL methods which showed that there was a large heterogeneity in the classification taxonomies presented in the literature ranging from the simple binary skin lesion classification (malignant vs. benign) to fine-grained classification which consists of dozens of skin lesion classes, and that there was a dominant trend in the literature in which the very earlier traditional CAD systems based on handcrafted features and SVMs were replaced by transfer-learning based CNNs. Both of these reviews have highlighted major challenges of limited data sets and clinical integration barriers, comparable to those described in Section 11, in particular, and relevant to the future of skin cancer diagnosis in the context of ML and DL. Adamu et al. (2024) have discussed ML and DL models related to the future of skin cancer diagnosis, and Melarkode et al. (2023) have provided a contemporary systematic review of AI for the diagnosis of skin cancer with its future directions and current research challenges, which again have highlighted the issues of limited data sets and clinical integration barriers.

Hussain and Toscano (2024) presented a comprehensive study focusing on ML and deep neural networks involved in skin cancer recognition, with particular focus on characterizing the outstanding challenges in the field.

The aim of Naseri and Safaei (2025) was to conduct a systematic review exclusively concerning the melanoma problem, investigating both classification and prognosis prediction of melanoma from dermoscopic images using ML and DL approaches. Interestingly, there was a gap between the accuracy of melanoma classification (90% commonly reported accuracy) and prognostic modelling (the ability to predict recurrence or metastasis risk), with the former being more advanced than the latter, consistent with the overall pattern for prognosis (less well developed, worse accuracy) than diagnosis across oncology imaging AI (Section 4).

7. DEEP LEARNING FOR BROADER DERMATOLOGICAL CONDITION DETECTION

In addition to malignancy classification, a great deal of the literature reviewed involves a broader and more diverse spectrum of dermatological condition detection that includes inflammatory, infectious and other benign but clinically significant skin diseases. Choy et al. (2023) performed a systematic review of DL image analysis for skin disease diagnosis and monitoring by distinguishing between intended clinical applications and found a much more advanced level of use for diagnosis than for long-term monitoring (tracking the progression of the disease in patients or the effect of treatment over time). This was a clear and strong area of underdeveloped opportunity identified by Noronha et al. (2023) in the domain.

Noronha et al. (2023) performed a broad-spectrum systematic review of the use of DL for detection of skin diseases, including both malignant and benign conditions, and identified methods, datasets and challenges, with a focus on data imbalance, lack of validation across different cohorts and the inconsistency in reporting of confidence intervals as current challenges — which were also concurred with during the appraisal in Section 11.

Both Zhang et al. (2023) and Sun et al. (2023) gave a more general view of the status and opportunities of ML/DL in skin disease, with both reinforcing the trend towards quicker methodological and architectural innovation compared to the diversification and rigor of the datasets and clinical validation in use.

8. ALGORITHMIC FAIRNESS AND PERFORMANCE ACROSS SKIN TONES

The single most consequential study within the included evidence base for questions of clinical equity is the large-scale investigation by Groh et al. (2024), published in *Nature Medicine*. Rather than benchmarking a DL classifier in isolation, this study evaluated deep-learning-aided decision support within a realistic physician-in-the-loop teledermatology simulation, involving 389 board-certified dermatologists and 459 primary-care physicians from 39 countries, who submitted diagnoses for 364 images spanning conditions including atopic dermatitis and Lyme disease, presented across a spectrum of light and dark skin tones (Table 1).

The study's central finding — that physician diagnostic accuracy, and by extension the accuracy of physician-AI collaborative diagnosis, differs systematically between light- and dark-skin presentations of the same underlying conditions — has direct and serious implications for the entire body of evidence synthesized in Sections 4 through 7 of this review. Because the overwhelming majority of the high-performing dermoscopic classification models discussed in Section 4 were trained and validated on datasets such as ISIC and HAM10000, which are documented elsewhere in the literature as disproportionately composed of lighter Fitzpatrick skin types (Table 2; Figure 3), the near-ceiling accuracy figures reported by Al-Waisy et al. (2025) and Tahir et al. (2023) should be interpreted as descriptive of performance on the specific population represented in those datasets, rather than as a general claim of equitable diagnostic performance across all patient skin tones. This is, in this review's critical judgment, the single most important qualifying finding for the entire evidence base and is treated as a first-order concern in the critical appraisal that follows (Table 3).

9. BENCHMARK DATASETS AND THEIR LIMITATIONS

Across the included literature, a small number of public datasets — principally the ISIC archive in its various yearly releases (ISIC 2016 through 2020), HAM10000, PH2, and DermIS — account for the overwhelming majority of training and validation data used across the primary studies reviewed in Sections 4 through 6 (Al-Waisy et al., 2025; Kassem et al., 2021; Tahir et al., 2023). Table 2 and Figure 3 present the overall number of times each data set appeared in the 20 sources included. This approach brings real value to the comparability of studies, but also causes the standard performance figures of the field to be strongly influenced by the particular demographic, geographic and equipment-specific characteristics of a limited number of source institutions, mostly from higher-income countries with historically weak representation of darker-skin-tone imaging (Groh et al., 2024; Zhang et al., 2023).

Benchmark Dataset	Included Studies Referencing Dataset (of 20)	Approx. Fitzpatrick / Skin-Tone Coverage Reported
ISIC (all yearly releases, 2016–2020)	7	Not skin-tone-stratified in any included study; predominantly lighter Fitzpatrick types (Groh et al., 2024)
HAM10000	8	Not skin-tone-stratified; sourced predominantly from Austrian and Australian clinical archives
PH2	3	Not skin-tone-stratified; small, single-institution dermoscopic archive
DermIS	2	Not skin-tone-stratified; limited demographic metadata reported
DermNet / other institutional sets	3	Variable and inconsistently documented across studies

Table 2. Frequency of benchmark dataset usage across the 20 included sources and documented skin-tone coverage.

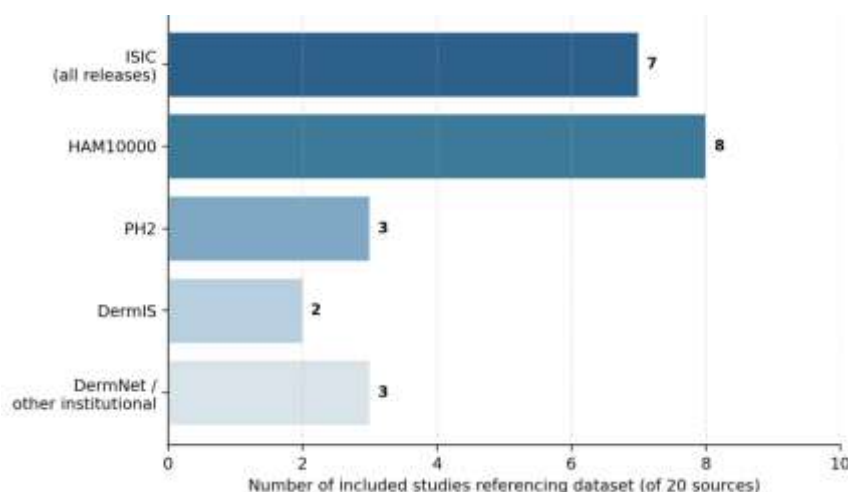


Figure 3. Concentration of benchmark dataset usage across the included evidence base

The repeated use of the same handful of images in dozens of independently published models (Section 4), and the explicit cataloguing by both Noronha et al. (2023) and Kassem et al. (2021) of this concentration of data as a problem for the field, suggest a potential further methodological concern about benchmark overfitting: the possibility that the iterative architectural developments of the field are fine-tuning the model to learn the specific characteristics of ISIC and HAM10000, instead of generalizable dermatological diagnostic capability.

10. LESSONS LEARNED FOR THE GENERAL DISEASE-PREDICTION AI FROM THE BREAST CANCER FIELD, CROSS-DOMAIN COMPARATIVE INSIGHTS.

Understanding the level of development of skin disease AI in comparison to other sub-fields of medical AI is helpful to place it. Nasser and Yusof (2023) conduct a comprehensive review of the literature on the application of DL to breast cancer diagnosis. Like dermatology, this area makes extensive use of image data well suited to CNN-based analysis.

However, breast cancer research has, unlike dermatological AI, produced much larger, multi-institution, and more demographically diverse datasets (Table 1) — providing something of a blueprint for the dataset consortium-based approaches called for in Section 13.

Yu et al. (2022) summarize popular DL algorithms applied in disease prediction across medical fields, while Rana and Bhushan (2022) offer an overview of ML/DL approaches for medical image analysis used across diagnostic and detection-oriented medical applications. Nia et al. (2023) similarly present an overview of AI applications in disease prediction and diagnosis across medical domains. These three reviews converge on a finding central to this review's conclusions: the relationship between reported accuracy on benchmark data and clinical efficacy is contingent on representative data and external validity — precisely the areas in which the dermatological AI reviewed in Sections 4–9 lags its published benchmark results.

11. CRITICAL APPRAISAL: METHODOLOGICAL STRENGTHS AND LIMITATIONS

Table 3 summarizes, as a domain-level rating matrix, whether each of the five critical appraisal domains introduced in Section 2.4 is consistently addressed across the included evidence base.

Appraisal Domain	Consistently Addressed?	Representative Evidence
(i) Dataset diversity — skin pigmentation & demographic provenance	No	Groh et al. (2024); Zhang et al. (2023); Section 8–9
(ii) External / prospective validation	No	Noronha et al. (2023); Section 11.3, 11.6
(iii) Consistent class taxonomy &	Partially	Kassem et al. (2021); Section 11.4

Appraisal Domain	Consistently Addressed?	Representative Evidence
reference-standard rigor		
(iv) Complete confusion-matrix-derived metric reporting	Partially	Noronha et al. (2023); Section 11.3
(v) Risk of benchmark overfitting (repeated dataset reuse)	No	Al-Waisy et al. (2025); Kassem et al. (2021); Section 11.3

Table 3. Critical appraisal domain summary across the included evidence base

11.1 Near-Ceiling Accuracy Figures Warrant Scrutiny, Not Simple Acceptance

At first glance, reported accuracy rates are in the range between 94% and 100% (Al-Waisy et al., 2025; Tahir et al., 2023) which appear to be rather impressive in terms of diagnostic accuracy (Figure 2). The unique strength of this review is that the numbers are obtained on repeatedly reused public benchmark datasets that are not necessarily representative of clinical use, and are therefore more likely to be construed as a reflection of strong performance on the specific datasets used than as a general statement of clinically acceptable diagnostic performance. This is consistent with the comparatively less high, but still more externally convincing, accuracy of 85% that Srinivasu et al. (2021) reported for the mobile-deployable architecture and is directly confirmed by the equity results by Groh et al. (2024) as discussed later.

11.2 Skin Tone Representation Is the Field's Central Unresolved Validity Threat

In the authors' view, the most significant critical appraisal finding from the included evidence base (Table 3, Domain i) is the difference in diagnostic accuracy for light and dark skin tones, for physicians and by extension for AI systems trained on similarly skewed datasets. The accuracy figures reported in the studies examining high accuracy dermoscopic classification reviewed in Section 4 do not provide any information about the distribution of the performance along skin tone strata; therefore, they cannot currently be disaggregated to determine whether there is a similar performance across the range of lighter-skinned patients who are overrepresented in ISIC and HAM10000 (Table 2).

11.3 Dataset Concentration and Benchmark Overfitting Risk

As seen in the majority of primary studies in Sections 4–6 (Al-Waisy et al., 2025; Kassem et al., 2021; Srinivasu et al., 2021; Tahir et al., 2023), the near-universal adoption of ISIC, HAM10000, PH2, and DermIS, quantified in Table 2 and Figure 3, poses a structural risk of benchmark overfitting, similar to that of many other fast-evolving computer-vision subfields. Noronha et al. (2023) lists as a persistent weakness the scarcity of validation studies using external datasets across populations outside the study context of the authors — an assessment corroborated by Table 1, as none of the primary papers reported validation at an institution other than the one in which the model was trained, nor upon an imaging protocol that varied significantly from the training images or patient population.

11.4 Class Taxonomy Heterogeneity Limits Cross-Study Comparability

Kassem et al. (2021) documented class taxonomies across the reviewed literature ranging from simple binary classification schemes to fine-grained taxonomies spanning dozens of distinct skin conditions. This heterogeneity, corroborated by the differing class structures used by Al-Waisy et al. (2025, binary/multi-lesion classification), Tahir et al. (2023, four-class skin cancer classification), and Srinivasu et al. (2021, seven-class HAM10000 taxonomy) as shown in Table 1, means that reported accuracy figures are not directly comparable on a like-for-like basis — a limitation this review's narrative-synthesis approach (Section 2.3) was specifically designed to accommodate rather than obscure through inappropriate meta-analytic pooling.

11.5 Diagnosis-Prognosis and Detection-Monitoring Maturity Gaps

Two included reviews independently identify a consistent maturity gap within the dermatological AI literature: Naseri and Safaei (2025) found melanoma diagnostic classification considerably more developed than melanoma prognostic modelling, while Choy et al. (2023) found diagnostic applications considerably more developed than longitudinal disease-monitoring applications. Both findings point to the same underlying pattern — the field has concentrated research effort on single-

timepoint classification tasks amenable to existing benchmark datasets, at the relative expense of longitudinal and prognostic applications that would arguably deliver greater clinical value.

11.6 Absence of Prospective Clinical-Utility Evidence

Consistent with the broader implementation-science critique found in cross-cutting oncology AI literature, no study within the included evidence base reports a prospective clinical trial evaluating whether AI-assisted dermatological diagnosis measurably improves patient-relevant outcomes, biopsy-appropriateness, or time-to-diagnosis in a real clinical workflow. The Groh et al. (2024) study, while methodologically the most rigorous and ecologically valid in the included evidence base, remains a controlled simulation exercise rather than a deployed clinical-utility trial.

12. BARRIERS TO CLINICAL TRANSLATION

Figure 4 summarizes, as a conceptual diagram, the principal barriers to clinical translation identified through the critical appraisal in Section 11.

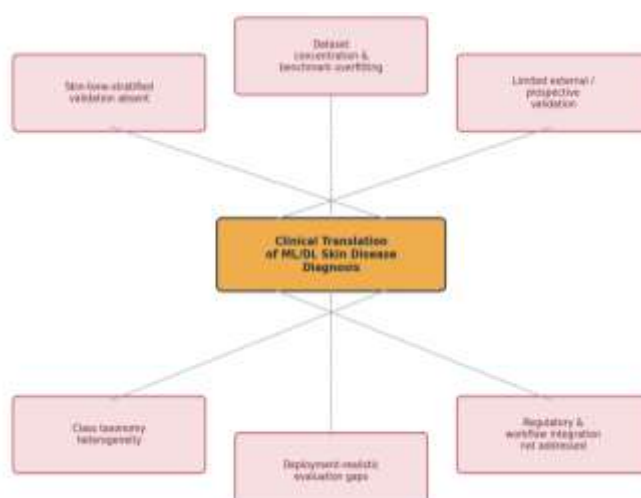


Figure 4. Principal barriers to clinical translation of ML/DL-based skin disease diagnosis identified in the critical appraisal (Section 11).

- Skin-tone-stratified validation: No primary classification study in the included evidence base reports performance disaggregated by Fitzpatrick skin type or comparable skin-tone classification, despite direct evidence from Groh et al. (2024) that this materially affects diagnostic accuracy.
- Dataset diversification: Continued near-exclusive reliance on ISIC, HAM10000, PH2, and DermIS (Table 2) risks benchmark overfitting and limits generalizability to broader, more demographically representative patient populations.
- External and prospective validation: No included primary study reports validation on an independent, geographically distinct clinical population using a different imaging protocol from its training data.
- Class taxonomy standardization: Wide variation in class schemes (Kassem et al., 2021) complicates cross-study comparison and regulatory benchmarking.
- Deployment-realistic evaluation: The vast majority of the reported accuracy comes from high-capacity ensemble architectures (Al-Waisy et al., 2025); those built for pragmatic point-of-care implementation (Srinivasu et al., 2021) produce dramatically lower performance that most studies failed to characterize explicitly (Figure 2).
- Regulatory and workflow integration: Similar to the broader critiques of oncology AI deployment (Nia et al., 2023; Yu et al., 2022), no included work reported live deployment and integration with prospective outcome monitoring.

13. RESEARCH GAPS AND FUTURE DIRECTIONS

Based on the critical analysis presented in Sections 8–12, this review proposes five strategic directions of high priority. First, and most urgently, future dermatological AI datasets and validation studies should mandate skin-tone-stratified performance reporting as a minimum standard, following directly from the equity findings of Groh et al. (2024); accuracy figures reported without this stratification should be treated by the field, and by reviewers, as incomplete. Second, dataset diversification efforts — analogous to the multi-institutional consortium models already achieved in breast cancer imaging AI (Nasser

&Yusof, 2023) — are needed to reduce the field's structural dependence on a small number of demographically skewed public archives (Table 2).

Third, external validation on independent, geographically and demographically distinct clinical populations should become a standard reporting requirement rather than an exceptional design feature, directly addressing the benchmark-overfitting risk identified in Section 11.3. Fourth, greater research investment in prognostic and longitudinal-monitoring applications is warranted, given the diagnosis-prognosis and detection-monitoring maturity gaps identified independently by Naseri and Safaei (2025) and Choy et al. (2023). Fifth, prospective clinical-utility trials — evaluating whether AI-assisted dermatological diagnosis measurably improves biopsy-appropriateness, diagnostic turnaround, or specialist-referral accuracy in a real clinical workflow, building on the ecologically valid physician-in-the-loop design pioneered by Groh et al. (2024) — should be prioritized over further retrospective benchmark-accuracy optimization on ISIC and HAM10000.

14. CONCLUSION

Deep learning has demonstrated exceptional, at times near-ceiling, technical performance for skin disease and skin cancer classification on established dermoscopic benchmark datasets, with several models reporting accuracies exceeding 94% and AUCs above 0.99 (Al-Waisy et al., 2025; Tahir et al., 2023; Table 1, Figure 2). Lightweight, mobile-deployable architectures demonstrate that clinically useful performance is achievable under realistic point-of-care computational constraints, albeit at a meaningfully lower accuracy ceiling (Srinivasu et al., 2021).

However, this critical synthesis finds that the field's headline performance figures cannot currently be interpreted as evidence of equitable or generalizable clinical-grade diagnostic capability. The large-scale physician-AI teaming study by Groh et al. (2024) provides direct evidence of skin-tone-related diagnostic performance disparities, and no primary classification study in the included evidence base reports the skin-tone-stratified validation that would be required to rule out equivalent disparities in automated systems trained on similarly composed datasets (Table 3). Combined with persistent dataset concentration (Table 2, Figure 3), class taxonomy heterogeneity, and an absence of prospective clinical-utility evidence (Choy et al., 2023; Kassem et al., 2021; Noronha et al., 2023), these findings indicate that automated skin disease diagnosis remains, at present, a technically impressive but clinically unproven research artefact. Dataset diversification, mandatory skin-tone-stratified reporting, and external prospective validation (Figure 4) constitute the field's most urgent priorities before these tools can be considered ready for equitable, safe, and scalable clinical deployment.

REFERENCES

- [1] Abunadi, I., &Senan, E. M. Deep learning and machine learning techniques of diagnosis dermoscopy images for early detection of skin diseases. *Electronics*, 10(24), 3158.2021, <https://doi.org/10.3390/electronics10243158>
- [2] Adamu, S., Alhussian, H., Aziz, N., Abdulkadir, S., Alwadin, A., Imam, A. A., Abdullahi, M., Garba, A., &Saidu, Y. The future of skin cancer diagnosis: A comprehensive systematic literature review of machine learning and deep learning models. *Cogent Engineering*, 11.2024, <https://doi.org/10.1080/23311916.2024.2395425>
- [3] Al-Waisy, A. S., Al-Fahdawi, S., Khalaf, M. I., Mohammed, M. A., Al-Attar, B., & Al-Andoli, M. N.A deep learning framework for automated early diagnosis and classification of skin cancer lesions in dermoscopy images.*Scientific Reports*, 15.2025, <https://doi.org/10.1038/s41598-025-15655-9>
- [4] Bechelli, S., &Delhommelle, J. Machine learning and deep learning algorithms for skin cancer classification from dermoscopic images. *Bioengineering*, 9(3), 97. 2022, <https://doi.org/10.3390/bioengineering9030097>
- [5] Choy, S., Kim, B., Paolino, A., Tan, W. R., Lim, S. W., Seo, J., Tan, S. P., et al. Systematic review of deep learning image analyses for the diagnosis and monitoring of skin disease. *NPJ Digital Medicine*, 6.2023, <https://doi.org/10.1038/s41746-023-00914-8>
- [6] Groh, M., Badri, O., Daneshjou, R., Koochek, A., Harris, C., Soenksen, L., Doraiswamy, P. M., & Picard, R. W. Deep learning-aided decision support for diagnosis of skin disease across skin tones. *Nature Medicine*, 30, 573–583. 2024, <https://doi.org/10.1038/s41591-023-02728-3>
- [7] Hussain, S. I., &Toscano, E. An extensive investigation into the use of machine learning tools and deep neural networks for the recognition of skin cancer: Challenges, future directions, and a comprehensive review. *Symmetry*, 16, 366.2024, <https://doi.org/10.3390/sym16030366>
- [8] Kassem, M. A., Hosny, K. M., Damaševičius, R., &Eltoukhy, M. M.Machine learning and deep learning methods for skin lesion classification and diagnosis: A systematic review. *Diagnostics*, 11(8), 1390.2021, <https://doi.org/10.3390/diagnostics11081390>

- [9] Kshirsagar, P. R., Manoharan, H., Shitharth, S., Alshareef, A. M., Albishry, N., & Balachandran, P. Deep learning approaches for prognosis of automated skin disease. *Life*, 12(3), 426. 2022, <https://doi.org/10.3390/life12030426>
- [10] Melarkode, N., Srinivasan, K., Qaisar, S., & Plawiak, P. AI-powered diagnosis of skin cancer: A contemporary review, open challenges and future research directions. *Cancers*, 15(4), 1183. 2023, <https://doi.org/10.3390/cancers15041183>
- [11] Naseri, H., & Safaei, A. Diagnosis and prognosis of melanoma from dermoscopy images using machine learning and deep learning: A systematic literature review. *BMC Cancer*, 25. 2025, <https://doi.org/10.1186/s12885-024-13423-y>
- [12] Nasser, M., & Yusof, U. Deep learning based methods for breast cancer diagnosis: A systematic review and future direction. *Diagnostics*, 13(1), 161. 2023, <https://doi.org/10.3390/diagnostics13010161>
- [13] Nia, N. G., Kaplanoglu, E., & Nasab, A. Evaluation of artificial intelligence techniques in disease diagnosis and prediction. *Discover Artificial Intelligence*, 3. 2023, <https://doi.org/10.1007/s44163-023-00049-5>
- [14] Noronha, S. S., Mehta, M., Garg, D., Kotecha, K., & Abraham, A. Deep learning-based dermatological condition detection: A systematic review with recent methods, datasets, challenges, and future directions. *IEEE Access*, 11, 140348–140381. 2023, <https://doi.org/10.1109/access.2023.3339635>
- [15] Rana, M., & Bhushan, M. Machine learning and deep learning approach for medical image analysis: Diagnosis to detection. *Multimedia Tools and Applications*, 1–39. 2022, <https://doi.org/10.1007/s11042-022-14305-w>
- [16] Srinivasu, P. N., SivaSai, J. G., Ijaz, M. F., Bhoi, A. K., Kim, W., & Kang, J. J. Classification of skin disease using deep learning neural networks with MobileNet V2 and LSTM. *Sensors*, 21(8), 2852. 2021, <https://doi.org/10.3390/s21082852>
- [17] Sun, J., Yao, K., Huang, G., Zhang, C., Leach, M., Huang, K., & Yang, X. Machine learning methods in skin disease recognition: A systematic review. *Processes*, 11(4), 1003. 2023, <https://doi.org/10.3390/pr11041003>
- [18] Tahir, M., Naeem, A., Malik, H., Tanveer, J., Naqvi, R. A., & Lee, S.-W. DSCC_Net: Multi-classification deep learning models for diagnosing of skin cancer using dermoscopic images. *Cancers*, 15(7), 2179. 2023, <https://doi.org/10.3390/cancers15072179>
- [19] Yu, Z., Wang, K., Wan, Z., Xie, S., & Lv, Z. Popular deep learning algorithms for disease prediction: A review. *Cluster Computing*, 26, 1231–1251. 2022, <https://doi.org/10.1007/s10586-022-03707-y>
- [20] Zhang, J., Zhong, F., He, K., Ji, M., Li, S., & Li, C. Recent advancements and perspectives in the diagnosis of skin diseases using machine learning and deep learning: A review. *Diagnostics*, 13(23), 3506. 2023, <https://doi.org/10.3390/diagnostics13233506>