**Electronic Journal of General Medicine**
2026, 23(6), em761
e-ISSN: 2516-3507
https://www.ejgm.co.uk/

MODESTUM

**Review Article**

**OPEN ACCESS**

# Artificial intelligence for non-invasive imaging diagnosis of basal cell carcinoma: A systematic review

Ibrahim Al Souleiman [1]*, Žanna Al Suleimane [2], Murat Abdiev [1], Habib Al Souleiman [1]

[1] Research Department, Swiss International University, Zurich, SWITZERLAND
[2] Faculty of Medicine, University of Latvia, Riga, LATVIA
***Corresponding Author:** dr.i@swiss.edu.eu

**Citation:** Al Souleiman I, Al Suleimane Z, Abdiev M, Al Souleiman H. Artificial intelligence for non-invasive imaging diagnosis of basal cell carcinoma: A systematic review. Electron J Gen Med. 2026;23(6):em761. https://doi.org/10.29333/ejgm/19771

**ARTICLE INFO**

Received: 01 Sep. 2026

Accepted: 16 Oct. 2026

**ABSTRACT**

**Background/objectives:** Basal cell carcinoma is the most common human cancer; it rarely spreads but destroys tissue locally, and its growth pattern determines treatment. We reviewed the accuracy of artificial intelligence (AI) applied to non-invasive skin imaging, against expert interpretation.

**Materials and methods:** Five databases were searched until 31 May 2026. Eligible studies applied AI to dermoscopy, confocal microscopy, optical coherence tomography, ultrasound or hyperspectral imaging with a separable result for this cancer. Studies were grouped by reference standard, appraised for risk of bias and rated for certainty; findings were synthesized narratively.

**Results:** Twenty-eight studies were included: fifteen histopathologically verified studies formed the primary evidence, and thirteen studies with archive-, consensus or undocumented reference standards were analyzed as supplementary evidence of technical performance. Accuracy varied widely within every technique and fell whenever a model was tested outside its development data. Only four studies compared stand-alone AI directly with human readers on the same lesions; the algorithm was level with or below the readers in three and better on specificity alone in one. Off-the-shelf general-purpose chatbots, in the versions and configurations evaluated, had the lowest sensitivities. Certainty was low or very low throughout.

**Conclusions:** Current evidence does not demonstrate superiority of AI over expert interpretation; direct comparisons are few, and most comparisons rely on expert benchmarks from other populations. Supervised decision support, not autonomous diagnosis, is the defensible position.

**Keywords:** carcinoma, basal cell, artificial intelligence, dermoscopy, diagnostic imaging, sensitivity and specificity, skin neoplasms

## INTRODUCTION

Basal cell carcinoma is the most common cancer in humans, and its incidence continues to rise in fair-skinned and ageing populations [1]. It almost never spreads in a Danish cohort of 126,627 people with a history of the tumor, only five metastatic cases were confirmed, a 14-year cumulative incidence of 0.0039% [2]. It matters for other reasons. It destroys tissue locally, it arises most often on the face, and its microscopic growth pattern determines how it must be treated.

Most suspicious skin lesions are first assessed outside specialist dermatology, where the practical question is not what the lesion is but who needs a biopsy and who needs referral. The reference standard is histopathology. European and British guidelines allow specialist clinical and dermoscopic diagnosis in typical lesions, require histopathological confirmation where the diagnosis is equivocal, and rely on histopathology in every case to assign the growth pattern and stratify risk [3, 4]. Biopsy is accurate but invasive, costly, and poorly matched to the volume of suspicious lesions seen in routine practice.

The tumor is a family of growth patterns with different behavior. Superficial disease is confined to the upper dermis and can be treated topically. Nodular disease is well circumscribed and is reliably cleared by standard excision. Infiltrative, micronodular and sclerosing patterns, together with basosquamous differentiation, are classified as high risk in both guidelines: they extend further and less predictably than their visible margins suggest, and both recommend margin-controlled surgery [3, 4]. Mixed tumors are managed according to their most aggressive component. Growth pattern thus determines aggressiveness, aggressiveness determines treatment, and non-invasive imaging is clinically valuable to the extent that it recovers the growth pattern without a biopsy.

Several bedside techniques can recover part of that information without a biopsy: dermoscopy [5, 6], reflectance confocal microscopy, optical coherence tomography and its line-field confocal variant [7], high-frequency ultrasound [8] and hyperspectral imaging. Their accuracy for growth pattern is set out later, and a review of regulatory-approved devices

*Copyright © 2026 by Author/s and Licensed by Modestum. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

concluded that the confocal tomography techniques perform best [9]. Every one demands considerable interpretive skill, and variation between observers remains the main barrier to wider use [10, 11].

Artificial intelligence (AI) has been proposed as a way to standardize this interpretation and extend it to less experienced clinicians. Early work showed dermatologist-level classification of skin cancer from photographs [12], and collaboration between clinician and algorithm can outperform either alone [13]. The landscape has since changed in a way that matters directly to general practice: general-purpose conversational models are now free and installed on ordinary telephones, and patients and non-specialist clinicians are already using them to interpret skin lesions. The question is therefore no longer only what a purpose-built system achieves in research conditions, but what a clinician, or a patient, obtains from a tool already in their pocket. Throughout this review, AI is treated as decision support for a trained clinician who can overrule it, not as a stand-alone test; no image-based dermatological algorithm is currently indicated for independent use by non-specialists [14].

Two earlier reviews cover parts of this ground: one of handcrafted and deep learning models across three techniques [15], and one confined to deep learning on dermoscopy [16]. The present review extends them in four ways. It covers the full range of non-invasive techniques. It appraises every study with a validated risk-of-bias instrument and rates certainty formally. It covers the three subsequent years, in which almost all of the confocal tomography evidence and all of the conversational-model evidence appeared. And it sets every estimate against what expert interpretation of the same technique already achieves, distinguishing direct comparisons within the same study from indirect comparisons with published benchmarks, without which no claim of added value can be judged. Our question was: in patients with lesions suspicious for, or confirmed as, basal cell carcinoma, how accurate is AI applied to non-invasive imaging when measured against histopathology, and how does it compare with expert interpretation of the same images?

## MATERIALS AND METHODS

### Protocol and Reporting

This review follows the preferred reporting items for systematic reviews and meta-analyses 2020 [17] and its diagnostic-test-accuracy extension [18]; completed checklists are provided in the **Supplementary Material.** The objectives, eligibility criteria, search strategy and analytic framework were agreed among the investigators before searching but were not documented as a written protocol until after data extraction, when they were deposited in a public repository as a read-only record of the methods as applied (link withheld for blinded review). The deposit post-dates the search close; the review was therefore not prospectively registered, and we make no claim of pre-specification. Because the decisions most vulnerable to knowledge of the retrieved literature are also those that determine which studies were included, a sensitivity analysis relaxing them is reported later, and every difference between the deposited protocol and the review as conducted is listed and justified in the **Supplementary Material**.

### Eligibility Criteria

The population comprises patients with lesions suspicious for, or histologically confirmed as, basal cell carcinoma. The index test was any AI or machine-learning system applied to dermoscopy, reflectance confocal microscopy, optical coherence tomography including its line-field variant, high-frequency ultrasound, hyperspectral imaging, or combinations of these. No restriction was placed on the type of model: purpose-built classifiers, general-purpose conversational models and dermatology foundation models were all eligible. Ordinary clinical photography was eligible only alongside one of these techniques. The comparator, where reported, was expert clinician assessment with or without dermoscopy; expert assessment was not accepted as a reference standard, and a single reader's interpretation of the index images was not accepted at all. Eligible designs were prospective and retrospective accuracy studies, reader studies and model development-and-validation studies, including research letters with complete primary data. To be included, a report had to provide a separable result specific to basal cell carcinoma: sensitivity, specificity, accuracy, F1 score or area under the curve (AUC). Multiclass studies reporting only aggregate performance were treated as background; studies without an AI index test were cited for context only; preprints and existing systematic reviews were not treated as evidence. Comparison against healthy skin or a single mimicking lesion was eligible, with the resulting indirectness recorded in the appraisal.

The reference standard required is histopathological examination. Applied strictly, that requirement would remove most of the dermoscopy evidence, much of which was developed on curated public image archives, without appraising it at all. Every report included was therefore assigned to one of two tiers. Tier 1 (histopathologically verified) means the extracted result derives from a cohort in which the contributing lesions were verified by histopathology, and that verification is documented in the primary report; where the cancers were histologically confirmed but benign comparators were verified by consensus or follow-up, the study was assigned to tier 1 with the partial verification recorded in the appraisal. Tier 2 (archive-, consensus, or undocumented labels) means the result rests on labels inherited from a curated public archive of mixed or unstated provenance, or on expert consensus reading, or on a reference standard that the primary report does not document for the lesions contributing the estimate; it contains archive-labelled evidence (6 studies), consensus-labelled evidence (2), evaluations of pre-trained models on archive-labelled benchmarks (2) and studies whose primary report does not document histopathological verification of the evaluated lesions (3), and the source of every label is stated in **Table 1**.

**Table 1.** Characteristics and principal results of the twenty-eight included studies, by tier and imaging technique

| Study | Imaging technique | Cohort and validation | Principal result for basal cell carcinoma |
|---|---|---|---|
| A. Tier 1 - Histopathologically verified; primary diagnostic-accuracy evidence (n = 15) | | | |
| Dermoscopy (n = 7) | | | |
| [19] | Dermoscopy | 7,192 images, prospective consecutive; split by patient; reader set 70 | Reader set 0.800 (0.442-0.965)/1.000 vs. 164 dermatologists 0.770/0.962 |

**Table 1 (Continued).**

| Study | Imaging technique | Cohort and validation | Principal result for basal cell carcinoma |
|---|---|---|---|
| **A. Tier 1 – Histopathologically verified; primary diagnostic-accuracy evidence (n = 15)** | | | |
| **Dermoscopy (n = 7)** | | | |
| [26] | Dermoscopy + clinical, fine-tuned | 2,675 and 594 images; internal validation only | Near-perfect internal separation; growth-pattern accuracy 0.899, weaker for aggressive patterns |
| [27] | Dermoscopy + photography, three chatbots | 772 images from 290 confirmed cancers + mimicker cohort | Accuracy 0.507-0.750 photography, 0.509-0.698 dermoscopy; sensitivity 0.231-0.634 and 0.304-0.554 |
| [30] | Dermoscopy | 13,603 images/2,538 patients; reader set 200 cases | Reader set 0.880/0.994 vs. 280 dermatologists 0.805/0.967; internal figures unusable |
| [32] | Dermoscopy | 5,572 archive records + prospective cohort of 707 patients | AUC 0.79 (0.75-0.83) test, 0.71 (0.66-0.76) independent cohort |
| [44] | Dermoscopy + photography | 90 patients, one lesion each; off-the-shelf chatbot | Accuracy 0.796 photography, 0.678 dermoscopy; biopsy advised for every lesion |
| [45] | Dermoscopy, task-adapted model | 550 archival images: 300 cancers + 250 controls | Sensitivity 0.940/specificity 0.767; expert comparator 0.983/0.920 on the same images |
| **Reflectance confocal microscopy (n = 3)** | | | |
| [47] | Confocal microscopy | 52 consecutive patients, 66 lesions; external 34 lesions | AUC 0.900, external 0.861; blinded experts 0.774/0.652 |
| [48] | Confocal microscopy | 230 lesions (82 cancers) + external cohorts (n = 115) | Internal 0.984 (0.902/0.953); external 0.735, 0.864 combined |
| [49] | Confocal microscopy, intraoperative | 62 patients stated, 14 analyzed | Pixel-level 0.46/0.85; cancer vs. healthy skin |
| **Optical coherence tomography (n = 2; a third, multimodal, listed under ultrasound)** | | | |
| [51] | Tomography, research prototype | 34 consecutive patients, 78 lesions; biopsy same day | 81% correctly classified; comparison against a single mimic |
| [52] | Line-field confocal tomography | 200 equivocal lesions/175 patients; 43 readers | Readers 0.842/0.894 unaided vs. 0.947/0.919 assisted; not stand-alone |
| **High-frequency ultrasound, multimodal (n = 1)** | | | |
| [28]† | Ultrasound + tomography + spectroscopy | 37 tumors, prospective, single phototype | 0.891/0.951; only study reporting by clinical form |
| **Hyperspectral imaging (n = 2)** | | | |
| [57]† | Hyperspectral | 11 tumors; training and test within the same image | Median per-lesion 0.892; margin rather than diagnostic task |
| [58] | Hyperspectral | 26 lesions (10 pigmented cancers) | 0.90 (0.60-0.98)/1.00 |
| **B. Tier 2 – Archive-, consensus or undocumented labels; supplementary evidence of technical performance (n = 13)** | | | |
| **Archive-labelled (n = 6)** | | | |
| [33] | Dermoscopy | Public archive; split by image; test 5,066 images | Derived from confusion matrix: 0.884/0.979 |
| [34] | Dermoscopy | Public archive, 10,015 images; test set doubles as validation | F1 score 0.72 only |
| [35] | Dermoscopy, color-transformed | 1,854 archive images; split after augmentation | Recall 0.747/0.791; transformed images worse on every measure |
| [36, 37] | Dermoscopy | 2,000 images; 395-image holdout; split by image | Accuracy 0.972-0.974; 0.972-0.979/0.965-0.978 |
| [38] | Dermoscopy | 904 images; internal test set contains all training images | 1.000 training, 0.634-0.818 internal, 0.634 external |
| [39] | Dermoscopy, color-transformed | 954 archive images; no split reported | Recall 0.95, precision 0.87; no specificity |
| **Consensus-labelled (n = 2)** | | | |
| [40] | Dermoscopy | Public archive; 479 cancer images | Readers without vs. with assistance: 0.84 to 0.91 and 0.78 to 0.89 |
| [41] | Dermoscopy, teledermatology | 1,559 consecutive images, 60 primary-care centers | Accuracy 0.90, recall 0.89; no specificity |
| **Pre-trained models on archive-labelled benchmarks (n = 2)** | | | |
| [42] | Dermoscopy, general-purpose chatbot | 950 archive images (200 cancers); no training step | 0.070-0.081/0.959-0.987; F1 0.125-0.132 |
| [43] | Dermoscopy, foundation model | Pretrained on >2 million images; 1,511-image test pool | Recall 0.860, reported incidentally; no specificity |
| **Reference standard not documented in the primary report (n = 3)** | | | |
| [31] | Dermoscopy | 100 biopsy-proven cancers + 254 mimics; separate 1,024-image test set, verification not described | AUC 0.878; no operating point |
| [55] | Ultrasound (5-18 MHz) | 310 lesions (130 cancers); folds by image; patients and reference standard not reported | 0.840-0.858 vs. other lesions; 0.735-0.783 vs. melanoma |
| [56] | Hyperspectral | 992 signatures from 41 patients; split by signature; reference standard not described | Accuracy 0.91, 0.89/0.92; cancer vs. healthy skin |

Note. Accuracy by growth pattern was reported only where stated; Tier 1 denotes a result verified against histopathology; Tier 2 one resting on curated archive labels or, expert consensus, or a reference standard that the primary report does not document; Tier 1 is the primary diagnostic-accuracy evidence; Tier 2 is supplementary evidence of technical performance and does not measure diagnostic accuracy; †Tier 1 with an outcome-level verification limit: the cancers were histologically confirmed, but the reported outcome was scored against expert pixel annotation [49], oncologist-assigned clinical form [28] or dermoscopic margins [57]; Full extracted variables, reference standards and model training details are given in the **Supplementary Material**

At the second round of peer review the documentation requirement was re-applied to every included study against its primary report: three studies previously assigned to tier 1 on the strength of their clinical setting or of a histologically verified development set, whose reports do not state how the evaluated lesions were verified, were reassigned to tier 2; no risk-of-bias judgement changed. Because tier 2 results are not anchored to documented histopathology, they do not answer the same diagnostic-accuracy question as tier 1. Tier 1 is therefore the primary diagnostic-accuracy evidence of this

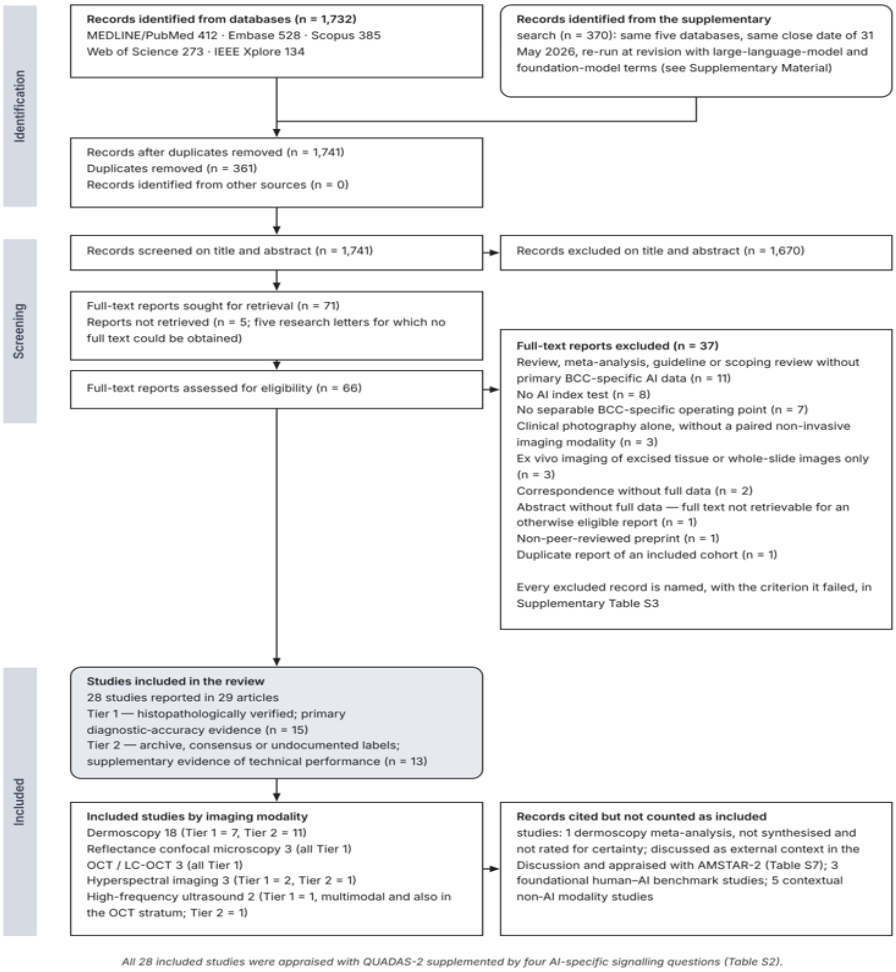



**Figure 1.** Study identification, screening, and inclusion, with reasons for all full-text exclusions (28 studies reported in 29 articles were included [15 histopathologically verified, forming the primary diagnostic-accuracy evidence, and 13 with archive, consensus, or undocumented labels, analyzed as supplementary evidence of technical performance]; Record-by-record exclusion reasons are given in the **Supplementary Material**) (the authors' own elaboration)

review, and every conclusion about diagnostic accuracy and about comparison with expert interpretation rests on it; tier 2 is reported as supplementary evidence of technical performance, that is, of how well a model reproduces archive, consensus or undocumented labels, and is used only to show how far such figures depart from the verified evidence. Within tier 1, three studies carry an outcome-level verification limit, marked † in the tables and figures and identified: the cancers were histologically confirmed, but the reported outcome was scored against unblinded expert pixel annotation of the same confocal images rather than spatially matched histology, against clinical forms assigned by the treating oncologist rather than histological growth patterns, or, for the tumor margins measured, against dermoscopy. These studies remain in tier 1 because their lesions were histologically verified, but their estimates are interpreted as evidence of segmentation or classification performance rather than of histopathologically verified diagnostic accuracy, and a sensitivity analysis excluding them is reported. Both tiers were screened by the same criterion, appraised with the same instrument and reported in full, but were rated separately and never pooled, because the reference standard is the dominant source of heterogeneity in this literature [16]. The two-tier structure is a post-hoc device and is declared as such.

Studies were excluded if they imaged excised tissue or histological slides only, evaluated photography without a paired imaging technique, evaluated other tumors without separable data, reported no extractable accuracy measure, were reviews, editorials or abstracts without full data, or were non-English. Two reports of one experiment were counted as one study.

### Information Sources and Search Strategy

We searched MEDLINE/PubMed, Embase, Scopus, Web of Science and IEEE Xplore from inception to 31 May 2026, supplemented by hand-searching the reference lists of included studies and relevant reviews. The search combined controlled vocabulary and free-text terms for the disease, the imaging techniques and the analytic approach. As originally run it contained no terms for conversational or foundation models; on review of the analytic strategy it was re-run across the same databases to the same close date with those terms added. Database-specific strings for both runs are provided in the **Supplementary Material**.

### Study Selection, Data Extraction, and Appraisal

Records were de-duplicated and screened in two stages, title and abstract then full text, independently by two reviewers, with disagreements adjudicated by a third; agreement at the full-text stage was almost perfect across the 66 records assessed (**Figure 1**).




**Figure 2.** Study-level sensitivity and specificity for every included study reporting an extractable estimate, by tier and imaging technique (Stand-alone algorithms, AI-assisted readers and human readers evaluated in the same study [direct comparators] are distinguished by marker; Horizontal lines are 95% confidence intervals (CIs) as reported by the source or computed by the Wilson method from derived two-by-two counts; Estimates are shown for transparency only; they are not pooled and are not comparable across units of analysis; †Tier 1 with an outcome-level verification limit) (the authors' own elaboration)

A piloted, standardized form was used to extract, in duplicate, the setting and design; the imaging technique and device; the type of model and how it was trained; the numbers of patients, lesions and images and how the data were divided for training and testing; the reference standard; whether validation was internal or external; the unit of analysis; the comparison class; how the operating threshold was chosen; skin phototype; whether accuracy was reported by growth pattern; and every accuracy measure with intervals and two-by-two counts where available. Where a source does not report a variable, no value was imputed. Every reference was checked against the primary source record, and every extracted number against the full text of its source rather than its abstract, with one documented exception in which two tables could not be rendered from the publisher's file [19].

Quality and applicability were appraised with the quality assessment of diagnostic accuracy studies tool [20] across patient selection, index test, reference standard, and flow and timing, by two independent reviewers with third-reviewer adjudication. Because the index tests are algorithms, four additional signaling questions were applied and reported separately in the **Supplementary Material**: overlap between the data used to build and to test the model; the unit of analysis; selection of the threshold or of the model itself on the test data; and the presence of an independent external test set. Reporting completeness was assessed against the checklist for AI in medical imaging [21] and the artificial-intelligence extension of the standards for reporting Diagnostic accuracy studies [22]: seven reporting domains common to both instruments were scored as reported, partially reported or not reported, and the results are summarized later and given study by study in **Table S4**. These assessments describe reporting quality; they informed interpretation but did not alter any risk-of-bias judgement or certainty rating. A dedicated appraisal instrument for such studies has been announced but was not usable at the time of this review [23].

## Data Synthesis

Meta-analysis was considered and rejected for specific reasons rather than because of heterogeneity in general. Bivariate or hierarchical summary receiver operating characteristic pooling requires two-by-two counts at a defined operating point from several studies estimating the same quantity. Two-by-two counts were reported or derivable in only nine of the twenty-eight studies (six at tier 1 and three at tier 2), and no imaging technique contributed to more than three tier 1 studies with such counts. Even within those, the studies differed in the unit of analysis (image, case, lesion, pixel, spectral signature or reader decision), in the reference standard (histopathology, archive labels or consensus), in the clinical task (basal cell carcinoma against all other lesions, against a single named mimic or against healthy skin) and in the comparator class (multiclass, binary or reader-assisted), so that no two studies with derivable counts estimated the same parameter in the same population. Pooling across these differences would have produced a summary estimate without a clinical referent. A structured narrative synthesis organized by imaging technique is therefore presented, with tier 1 as the primary evidence and tier 2 reported separately as supplementary evidence of technical performance. To make the spread of results transparent without pooling, every extractable study-level sensitivity and specificity is plotted in **Figure 2**, with 95% CIs taken from the source or computed by the Wilson method from derived two-by-two counts.

**Table 2.** Certainty of the evidence by imaging technique and tier, rated separately for sensitivity and specificity

| Technique (tier) | Studies (n) | Illustrative accuracy | Main reasons for downgrading | Sensitivity | Specificity |
|---|---|---|---|---|---|
| Dermoscopy, tier 1 | 7 | Reader-set 0.800-0.880 with 0.994-1.000; 0.71 in the only independent prospective cohort; chatbot accuracy 0.507-0.796 | Risk of bias (all high for patient selection); inconsistency (sensitivity 0.231 to 0.940); imprecision (reader sets of 70 and 200) | Very low | Low |
| Dermoscopy, tier 2 (supplementary: technical performance) | 11 | Classifiers 0.90-0.974; the one study with external data fell from 1.000 to 0.634; chatbot sensitivity 0.070-0.081; color-texture AUC 0.878 (test-set verification undocumented) | Risk of bias (archive, consensus or undocumented labels in all eleven; split by image or none stated in nine); indirectness; inconsistency; publication bias plausible | Very low | Very low |
| Reflectance confocal microscopy, tier 1 | 3 | Internal 0.900-0.984; external 0.735-0.864; intraoperative sensitivity 0.46 | Inconsistency (0.984 internal vs. 0.735 external); risk of bias (all high for patient selection and index test); imprecision (14 analyzed cases); indirectness (pixel-level expert annotation in one†) | Very low | Low |
| Optical coherence tomography, tier 1 | 3 | 81% correct vs. a single mimic; assisted readers 0.947/0.919; multimodal 0.891/0.951 | Risk of bias (all high for patient selection; test-set independence unestablished in the largest study); indirectness (reader improvement, not stand-alone accuracy); imprecision | Low | Low |
| Hyperspectral imaging, tier 1 | 2 | Median per-lesion 0.892 (11 tumors)†; 0.90 (0.60-0.98) in 26 lesions | Imprecision (11 and 26 lesions); risk of bias (training and test within the same image in both; no external validation); indirectness (margins rather than diagnosis in one†) | Very low | Very low |
| Hyperspectral imaging, tier 2 (supplementary: technical performance) | 1 | Signature-level accuracy 0.91, 0.89/0.92 (41 patients) | Risk of bias (reference standard not described; split by signature; architecture selected on the test metrics); indirectness (cancer vs. healthy skin); imprecision (no intervals) | Very low | Very low |
| High-frequency ultrasound, tier 1 | 1 | Multimodal 0.891/0.951 (37 tumors)† | Imprecision (one small single-center cohort); risk of bias (tuning on the evaluation data; no external validation); indirectness (ultrasound contribution not separable; clinical-form labels†) | Very low | Very low |
| High-frequency ultrasound, tier 2 (supplementary: technical performance) | 1 | 0.840-0.858 vs. other lesions; 0.735-0.783 vs melanoma (310 lesions) | Risk of bias (reference standard not stated; folds by image; no patient denominator; no external validation); indirectness (AUC only, no operating point); imprecision (single study, no intervals) | Very low | Very low |

Note. Certainty was rated with GRADE guidance for test accuracy [24, 25], separately for sensitivity and specificity because these diverged systematically; The groups come to twenty-nine because one multimodal study [28] contributes to two; No group reached moderate certainty; Each row is one body of evidence, defined by imaging technique, tier and outcome (composition in **Table S12**); a rating expresses certainty that accuracy lies within the range shown, not in a single point estimate; Tier 2 ratings refer to technical performance against archive, consensus, or undocumented labels; †Outcome-level verification limit; The complete evidence profile is given in the **Supplementary Material**

Comparisons with human interpretation were classified as direct, where stand-alone or assisted AI and human readers were evaluated on the same lesions within one study, or indirect, where an artificial-intelligence estimate is set against a published expert benchmark derived from a different population. Indirect comparisons are affected by differences in case mix, prevalence, reference standard, threshold and unit of analysis, and are presented as context, not as evidence of superiority or inferiority.

Certainty was rated with the grading of recommendations assessment, development, and evaluation guidance for test accuracy [24, 25], separately for each tier and separately for sensitivity and specificity, because these diverged systematically (**Table 2**). Each body of evidence was defined by four elements: the target condition (basal cell carcinoma as a class), the imaging technique, the tier of reference standard and the outcome. Within a body, different model types, diagnostic tasks, comparators and units of analysis were retained rather than separated, and their differences were rated down under inconsistency and indirectness. A single rating for such a body therefore expresses certainty that the accuracy of AI for that technique lies within the range shown, not certainty in any single point estimate. Splitting further, for example by model type, would have produced bodies of one to three studies; because every subgroup would retain serious or very serious risk of bias and serious imprecision, we verified for each technique that no subdivision would have raised any rating, so the ratings reported are an upper bound on the certainty of any narrower question. The composition of each body of evidence is listed in **Table S12**.

**Table 3** expresses the principal estimates as expected false-negative and false-positive results per 1,000 lesions at two prevalences derived from Cochrane reviews of the same imaging techniques, one representing equivocal lesions in which imaging is used to decide on biopsy and one representing lesions referred with clinical suspicion of the tumor; the sources are given beneath **Table 3**.

**Table 3.** AI and human interpretation compared by type of comparison, expressed as expected false negatives and false positives per 1,000 lesions

| Test (source) | Sensitivity | Specificity | FN-1 | FP-1 | FN-2 | FP-2 |
|---|---|---|---|---|---|---|
| **A. Direct comparisons: AI and human readers evaluated on the same lesions within one study** | | | | | | |
| Dermoscopy, AI model on reader set [19] | 0.800 (0.442-0.965) | 1.000 (0.925-1.000) | 30 | 0 | 120 | 0 |
| Dermoscopy, 164 dermatologists, same reader set [19] | 0.770 | 0.962 | 35 | 32 | 138 | 15 |
| Dermoscopy, AI model on reader set [30] | 0.880 | 0.994 | 18 | 5 | 72 | 2 |
| Dermoscopy, 280 dermatologists, same reader set [30] | 0.805 | 0.967 | 29 | 28 | 117 | 13 |
| Dermoscopy, task-adapted conversational model [45] | 0.940 | 0.767 | 9 | 198 | 36 | 93 |
| Dermoscopy, expert reading of the same images [45] | 0.983 | 0.920 | 3 | 68 | 10 | 32 |
| Confocal microscopy, blinded experts (model reported AUC only) [47] | 0.774 (0.673-0.873) | 0.652 (0.537-0.766) | 34 | 296 | 136 | 139 |
| Line-field tomography, AI-assisted readers [52] | 0.947 (0.939-0.954) | 0.919 (0.912-0.926) | 8 | 69 | 32 | 32 |
| Line-field tomography, unaided readers, same study [52] | 0.842 (0.829-0.854) | 0.894 (0.886-0.902) | 24 | 90 | 95 | 42 |
| **B. AI estimates without a same-study human comparator (comparison with experts only indirect)** | | | | | | |
| Dermoscopy, off-the-shelf chatbot on a public archive (tier 2) [42] | 0.081 | 0.959 | 138 | 35 | 551 | 16 |
| Confocal microscopy, best internal model [48] | 0.902 | 0.953 | 15 | 40 | 59 | 19 |
| Confocal microscopy, intraoperative model, against expert pixel annotation [49]† | 0.46 | 0.85 | 81 | 128 | 324 | 60 |
| Multimodal ultrasound and tomography, clinical-form labels [28]† | 0.891 | 0.951 | 16 | 42 | 65 | 20 |
| Hyperspectral, pigmented lesions [58] | 0.90 (0.60-0.98) | 1.00 | 15 | 0 | 60 | 0 |
| **C. Published expert benchmarks for indirect comparison (not included studies of this review)** | | | | | | |
| Confocal microscopy, pooled expert benchmark [50] | 0.92 (0.87-0.95) | 0.93 (0.85-0.97) | 12 | 60 | 48 | 28 |
| Line-field tomography, expert benchmark [10] | 0.98 | 0.80 | 3 | 170 | 12 | 80 |
| Tomography, expert benchmark with dermoscopy [53] | 0.95 (0.91-0.97) | 0.77 (0.69-0.83) | 8 | 196 | 30 | 92 |

Note. Expected frequencies are computed from the point estimates at two prevalences taken from Cochrane reviews of the same imaging techniques: 15%, the median prevalence of basal cell carcinoma observed in studies of equivocal lesions in the review of reflectance confocal microscopy [59], representing lesions in which imaging is used to decide on biopsy; and 60%, the mean prevalence observed in studies of lesions referred with clinical suspicion in the review of optical coherence tomography [53], representing a specialist or surgical-referral setting; False negatives are missed cancers; false positives are lesions without basal cell carcinoma classified as positive, a proxy for potentially avoidable biopsies or excisions; The frequencies are illustrative and not adjusted for heterogeneity; CIs are shown only where the source reports them; Section A rows can be compared with each other within the same source; rows in section B and section C can be compared only indirectly, across different populations, case mixes and reference standards; Rows are not comparable where the unit of analysis differs; †Outcome-level verification limit: the estimate is scored against expert pixel annotation [49] or clinical-form labels [28] rather than histopathology; FN-1: False negatives (missed cancers) per 1,000 at 15%; FP-1: False positives per 1,000 at 15%; FN-2: False negatives (missed cancers) per 1,000 at 60%; FP-2: False positives per 1,000 at 60%

## RESULTS

### Study Selection and Characteristics

The original search identified 1,732 records and the supplementary search a further 370. After de-duplication, 1,741 records were screened, 71 full texts were sought, 66 assessed and 37 excluded (**Figure 1**). Twenty-eight studies, reported in twenty-nine articles, met the inclusion criteria: fifteen at tier 1 and thirteen at tier 2. These comprise eighteen dermoscopy studies (seven tier 1, eleven tier 2), three reflectance confocal microscopy studies, and three optical coherence tomography studies (all tier 1), three hyperspectral studies (two tier 1, one tier 2) and two of high-frequency ultrasound (one tier 1, one tier 2), the tier 1 ultrasound study being multimodal and also contributing to the tomography group. Five research letters could not be obtained in full text. Study characteristics and principal results are given in **Table 1** and study-level sensitivity and specificity in **Figure 2**; three studies report accuracy below the level of the tumor as a single class, two by histological growth pattern [26, 27] and one by clinical form assigned by the treating oncologist [28].

### Dermoscopy

Dermoscopy resolves surface and superficial dermal structures to a depth of about 100 micrometers. The tumor is recognized by the absence of a pigment network together with branching telangiectasia, blue-grey ovoid nests and globules, leaf-like and spoke-wheel areas, and ulceration, a criterion set reaching 97% sensitivity with 92-93% specificity in expert hands [29]. Because dermoscopy resolves neither cellular detail nor depth, it separates superficial from other growth patterns only moderately well and cannot show deep extension [5]. AI is therefore asked here to standardize pattern recognition and extend it to non-experts, not to overcome the physical limit of the technique.

Primary evidence (tier 1). Seven studies are histopathologically verified (**Table 1**), and the two with direct comparisons against dermatologists reading the same reader sets place the algorithm level with, not above, its comparators. A classifier developed on 7,192 prospectively and consecutively collected images, the only dermoscopy study to separate training and test data by patient, reached a reader-set sensitivity of 0.800 (95% CI 0.442-0.965) and specificity of 1.000 (0.925-1.000), against 164 dermatologists at 0.770 and 0.962 [19]. A second reached 0.880 and 0.994, against 280 dermatologists at 0.805 and 0.967 [30]. In both, only the reader sets are usable, because the reports do not provide separable internal test results. A vision-language model tested on an independently and prospectively collected 707-patient cohort reached 0.71 (0.66-0.76) against a single mimicking diagnosis [32].

Supplementary evidence of technical performance (tier 2). Eleven further studies report results that do not rest on documented histopathology: six inherit labels from curated public archives [33-39], two rest on expert consensus [40, 41], two evaluate pre-trained models on archive-labelled benchmarks [42, 43], and one does not document the reference standard for the lesions it evaluates [31]. In that study a color-and-texture model developed on biopsy-proven cancers

achieved an AUC of 0.878 against competitive benign mimics, on a separate 1,024-image test set whose composition and verification are not described. The biases differ. Archive labels import unknown verification and documented duplication of the same lesion; consensus labels are circular to the extent that the readers see the same images as the algorithm; and pre-trained models add undisclosed training corpora whose overlap with the evaluation archive cannot be excluded. Two features determine how much weight this evidence can carry. First, the archive- and consensus-labelled classifiers reach higher figures than any verified estimate, accuracies of 0.90 to 0.974 against verified reader-set sensitivities of 0.80 to 0.88, yet every one either divides its data by image in archives known to hold several images of the same lesion or reports no division at all; one reported an internal AUC of 0.818 on a test set containing its entire training data [38]. Second, a curated archive does not guarantee a high figure: a conversational model evaluated on the same archives reached a sensitivity of only 0.070-0.081 [42]. What a curated archive guarantees is only that the number cannot be read as clinical accuracy.

Seven included studies evaluate conversational, vision-language or dermatology foundation models [26, 27, 32, 42-45]; the five built on general-purpose conversational or multimodal models [26, 27, 42, 44, 45] divide sharply on whether the model was adapted to the task. Used off the shelf, in the versions, prompts and image sets evaluated, these models had the lowest sensitivities in this review. One widely used assistant, tested on 90 confirmed lesions, reached an accuracy of 0.796 on ordinary photographs, falling to 0.678 with dermoscopy, and recommended biopsy for every lesion [44]. Three web-based assistants tested on 772 images from 290 confirmed cancers and an independent cohort of mimicking lesions reached accuracies of 0.507-0.750 on photographs and 0.509-0.698 on dermoscopy, with sensitivities of 0.231-0.634 and 0.304-0.554; the authors concluded that such models are not clinically suitable [27]. A fourth, on 950 archival images, gave a sensitivity of 0.070-0.081 [42]. Adapted to the task, the same technology performs far better: a configuration instructed in the dermoscopic criteria reached 0.940 and 0.767 on 550 archival images, against 0.983 and 0.920 for expert reading of the same images [45], and a model fine-tuned on paired clinical and dermoscopic images predicted the histological growth pattern at an accuracy of 0.899, weaker for micronodular, infiltrative and basosquamous disease [26]. Neither adapted model has been externally validated. In these studies, therefore, the difference between off-the-shelf and adapted use was larger than the differences between imaging techniques. Because commercial multimodal models are updated frequently and their output depends on version, prompt and image set, these findings describe the configurations evaluated, from GPT-4 Vision to ChatGPT-5 (**Table S5**) and should not be generalized to the technology as a class.

On the verified evidence, superiority of AI over expert dermoscopic interpretation is not demonstrated: in the three direct comparisons the algorithm was level with dermatologists [19, 30] or below expert reading of the same images [45], and higher figures appear only in the archive-labelled estimates, which measure technical performance rather than diagnostic accuracy. What the evidence supports is narrower and still useful: assisted reading raises less experienced readers towards expert performance [40].

### Reflectance Confocal Microscopy

Reflectance confocal microscopy images the living skin at near-cellular resolution to a depth of 200-300 micrometers, with contrast derived mainly from melanin. The tumor is recognized by nests with peripheral palisading and clefting, dark silhouettes, elongated monomorphic nuclei and increased dermal vasculature, reaching 82.9% sensitivity and 95.7% specificity when four or more criteria are present [46]. The limits follow from the physics: depth restricts assessment of deeper tumor, so infiltrative disease is systematically harder to capture, and contrast weakens in non-pigmented lesions [9]. AI is therefore asked to do two different things here: to standardize interpretation, which is within reach, and to extract signal from images in which the diagnostic features are incompletely represented, which is not.

Three studies met the criteria, all histopathologically verified, and their spread is one of the most informative results in this review. A model developed on 52 consecutively enrolled patients with equivocal lesions reached a lesion-level AUC of 0.900, falling to 0.861 on an independent multinational set; the blinded expert comparator in the same study achieved a sensitivity of 0.774 (95% CI 0.673-0.873) and specificity of 0.652 (0.537-0.766) [47]. A second reached 0.984 internally but fell to 0.735 externally [48]. A third, applied during surgery, achieved a pixel-level sensitivity of 0.46 with specificity 0.85; its analyzed cohort is 14 cases, not the 62 patients its abstract states, and its reference standard is unblinded expert annotation of the same images rather than matched histology [49]; this estimate therefore measures agreement with expert annotation, not histopathologically verified accuracy, and is marked accordingly.

The evidence that AI improves conventional interpretation here is partial and confined to the one direct comparison. Against blinded expert readers in the same study, the algorithm was better on specificity [47]. Against the pooled expert benchmark for this technique, sensitivity 0.92 (0.87-0.95) and specificity 0.93 (0.85-0.97) across 15 studies and 4,163 lesions [50], an indirect comparison across different populations, no externally validated estimate here can be shown to be higher. During surgery, where the diagnostic features are physically under-represented in the image, the algorithm agreed poorly with expert annotation [49], which suggests that it inherits that limitation rather than correcting it, although without matched histology this cannot be measured directly.

### Optical Coherence Tomography

Conventional optical coherence tomography and its line-field confocal variant are distinct techniques and should not be conflated. The conventional technique resolves tissue architecture rather than cells, to a depth of up to 2 mm, and identifies the tumor by avoiding dark structures, a dark border beneath it, disruption of the dermo-epidermal junction and dilated vessels [9]. The line-field variant reaches near-histological detail at a reduced depth, agrees with the histological growth pattern in 90.4% of tumors [7] and reaches 100% sensitivity for infiltrative disease in pooled analysis [9]. Its practical limits are depth, field of view and the interpretive expertise it demands, which is the constraint AI is best placed to relieve.

Three studies contribute, all histopathologically verified; two are tabulated here and a third, multimodal, is listed under ultrasound [28] and carries clinical-form rather than

histological labels. An early feature-based classifier applied to 34 consecutively enrolled patients, each biopsied at the same consultation, classified 81% of cancers and 73% of actinic keratoses correctly [51]. The principal evidence is a multicenter reader study of a real-time assistant, in which 43 dermatologists read 200 equivocal lesions from 175 patients in a double-round design with a two-week washout [52]. Pooled reader sensitivity and specificity were 0.689 (95% CI 0.677-0.700) and 0.751 (0.742-0.759) with clinical and dermoscopic assessment alone, 0.842 (0.829-0.854) and 0.894 (0.886-0.902) with unaided imaging, and 0.947 (0.939-0.954) and 0.919 (0.912-0.926) with assistance. The headline gains of 25.8 and 16.8 percentage points are measured against clinical and dermoscopic assessment and therefore conflate the value of the imaging with that of the algorithm. Measured against unaided imaging, the assistant added 10.5 points of sensitivity ($p < 0.001$) and 2.5 points of specificity, the latter not significant for experts or intermediates; gains were largest among novices (0.792 to 0.930) and smallest among experts (0.901 to 0.972). These are reader-improvement effects, not stand-alone operating points, and the report does not state that the 200 evaluation lesions were excluded from a training corpus of more than 685,000 images.

For the line-field technique, the direct comparison within the reader study shows that assistance improved on unaided imaging, unevenly and as decision support. Indirectly, published expert reading without assistance already achieves 0.98 sensitivity with 0.80 specificity in equivocal lesions [10] and 0.98 with 0.90 against mimicking lesions [11]; the assisted figures of 0.947 and 0.919 are comparable with these, not superior, although differences in case mix and design limit the comparison. For the conventional technique, the only estimate, 81% correct classification against a single mimic [51], is below the pooled expert estimate of 0.95 (0.91-0.97) and 0.77 (0.69-0.83) [53], again an indirect comparison.

### High-Frequency Ultrasound and Hyperspectral Imaging

High-frequency ultrasound resolves tumor geometry, margins and thickness, not the cellular features on which diagnosis rests. A Cochrane review found the evidence for diagnosis too variable to pool [54]; the established role of the technique is in depth and extent and, guided by dermoscopy, in separating aggressive from non-aggressive growth patterns [8]. Two included studies nevertheless show that useful signal can be extracted from ultrasound texture and shape. A support-vector model applied to 310 lesions separated the cancer from other lesions at an AUC of 0.840-0.858 but from melanoma at only 0.735-0.783, with no patient denominator reported [55]; because the report does not state how the lesion diagnoses were established, the study is assigned to tier 2 and read as supplementary evidence of technical performance. A prospective study combining ultrasound shape features with tomographic texture and diffuse reflectance spectroscopy in 37 tumors reported sensitivity 0.891 and specificity 0.951 [28]; it is the only study reporting accuracy by clinical form, and sensitivity was lowest in pigmented and infiltrative-ulcerative lesions, with pigmented tumors misclassified as superficial in 35% of cases. Its labels, however, are clinical forms assigned by the treating oncologist rather than histological growth patterns, and all 37 participants were of a single skin phototype. It is therefore marked as tier 1 with an outcome-level verification limit, and its form-specific results are not interpreted as accuracy for histological growth pattern.

Three studies used hyperspectral acquisition. The largest analyzed 992 spectral signatures from 41 patients and reported accuracy 0.91 with sensitivity 0.89 and specificity 0.92 but divided its data by signature rather than by patient [56]; its report does not describe the reference standard, and its healthy-skin labels rest on distance from the tumor, so it is assigned to tier 2. A margin-assessment study of 11 tumors, marked as tier 1 with an outcome-level verification limit, reported a median per-lesion AUC of 0.892 but trained and tested on different halves of the same image and never verified the margins histologically [57]. A pilot in 26 lesions correctly classified 9 of 10 pigmented cancers, a sensitivity of 90% (95% CI 60-98); it is the only included study to account fully for excluded images, but it too trained on one half of each test lesion's own image and excluded the benign mimics that make pigmented disease a diagnostic problem [58]. All three compare the tumor against healthy skin or a single mimicking class, and none has been externally validated.

For ultrasound the question of added value cannot yet be answered, because no expert diagnostic benchmark exists [54]; what can be said is that AI extracts signal conventional reading does not use, and that the one model reporting by clinical form fails where the stakes are highest. For hyperspectral imaging, the prior question is whether the technique carries a diagnostic signal at all, and on three small studies, only two of them histopathologically verified, that remains unresolved.

### Risk of Bias

Domain-level results are shown in **Figure 3** and justified study by study in the **Supplementary Material**. Three findings carry the certainty ratings. First, no included study is at low risk of patient-selection bias: all twenty-eight either assemble images retrospectively from a curated archive, restrict the comparison to the cancer versus healthy skin or a single named mimic [32, 49, 51, 56], or apply selective curation to a consecutively enrolled cohort [19, 27, 47]. Second, the index-test domain is high risk in twenty-four studies; the recurring mechanism is not a tuned threshold but selection of the model itself on the test data, including in one case the best of 207 combinations [36-38, 56]. Overlap between the data used to build and to test the model is high or unclear in twenty-six studies, almost always because the data were divided by image rather than by patient in datasets holding many images of the same lesion; only three studies divided by patient [19, 32, 51] and one by tumor [28]. In the three evaluations of off-the-shelf conversational models the training corpora are undisclosed and overlap with the archives used to evaluate them can be neither excluded nor quantified [27, 42, 44]. Third, flow and timing is low risk in one study alone [51]; the interval between imaging and histology is reported in only three studies, and exactly one accounts for every excluded image [58]. No included study reports blinding the pathologist to the algorithm's output.

### Reporting Completeness

Reporting completeness against CLAIM and STARD-AI is given study by study in **Table S4** and summarized here. Reporting was most complete for data source and provenance, reported in full in 14 of the 28 studies and partially in 13, and least complete for the elements on which re-analysis and appraisal depend. The level at which data were partitioned was stated explicitly in only six studies and was partial or absent in nineteen; the remaining three involved no training. The

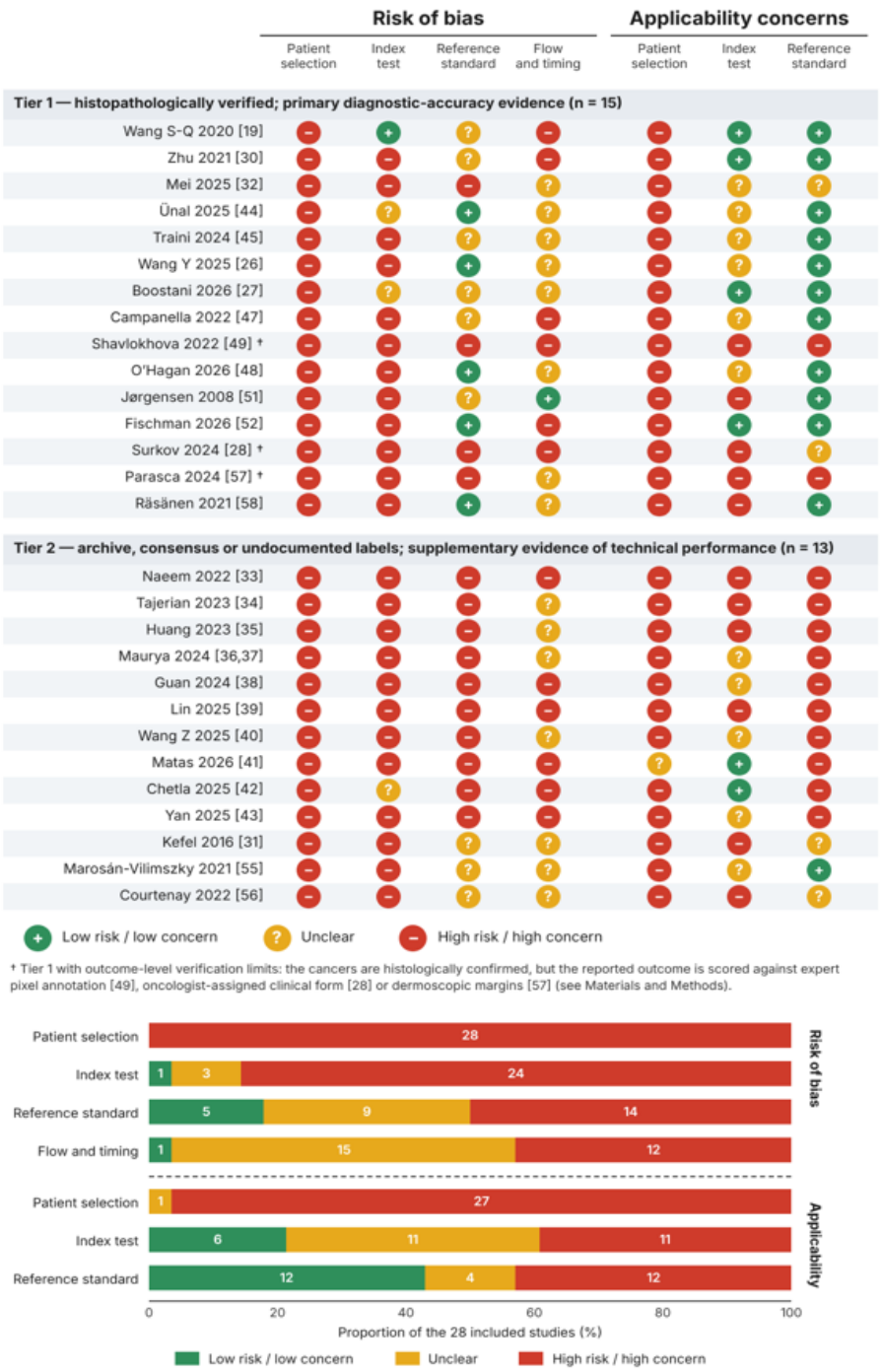



**Figure 3.** Risk-of-bias and applicability assessment for all 28 included studies (The upper panel gives study-level judgements, the two tiers as separate blocks, with the three studies whose reports do not document histopathological verification in the tier 2 block [31, 55, 56]; the lower panel summarizes them by domain; Study labels carry the reference numbers used in **Table 1** and the reference list; †Tier 1 with an outcome-level verification limit) (the authors' own elaboration)

operating threshold and how it was chosen were reported in five studies and not at all in seventeen; accuracy metrics with CIs were fully reported in two and absent in fifteen; participant flow and exclusions were fully reported in four and absent in eighteen; and skin phototype and subtype were reported in three and absent in twenty. Model, framework and version were fully reported in eight studies, a gap that matters most for the general-purpose conversational models, whose version and configuration were fully reported in only two of the four such evaluations [27, 42]. These deficits are the main reason so few estimates could be re-derived or plotted with intervals (**Figure 2**).

### Sensitivity Analysis of Eligibility Decisions

Because the protocol was not documented before data extraction, the eligibility decisions cannot be independently verified as pre-specified. We therefore repeated the synthesis under six alternative rules: as applied; histopathologically verified studies only; restricted to comparisons against clinical mimics; restricted to studies with external validation; restricted to studies qualifying on sensitivity, specificity or AUC alone; and, added at the previous revision, histopathologically verified studies without an outcome-level verification limit, which removes the intraoperative confocal, multimodal and

hyperspectral margin studies [28, 49, 57]. These leave 28, 15, 21, 4, 26, and 12 studies, respectively, the second, fifth and sixth having been recalculated after the reassignment of three studies to tier 2. Under all six the direction of the conclusions is unchanged, no group reaches moderate certainty, and the fall in performance from internal to external testing remains the dominant finding. The rules are set out in full in the **Supplementary Material**.

## DISCUSSION

### Principal Findings

Five findings stand out. First, accuracy varies more within a single imaging technique than between techniques, and it varies systematically. Within confocal microscopy alone, one model falls from an internal AUC of 0.984 to 0.735 externally [48], another from 0.900 to 0.861 [47], and a model used during surgery agrees with expert pixel annotation at a sensitivity of only 0.46 [49]—at the 15% prevalence used in **Table 3**, 81 missed cancers per 1,000 lesions. Within dermoscopy the spread runs from 0.070 for an unadapted conversational model [42] to 0.979 on a curated archive [36, 37].

Second, the highest estimate in every group comes from a curated, internally validated evaluation. In all four studies reporting both, the estimate obtained under external, prospective or in vivo conditions is no higher than the same model's internal figure [32, 38, 47, 48]. The converse does not hold: several archive-labelled studies report internal estimates that are themselves low [34, 38, 42].

Third, the methodological problem is not threshold optimization but the construction of the evidence itself: no study at low risk of patient-selection bias, three dividing their data by patient, and selection of the model on the test data so routine that it is rarely acknowledged as bias.

Fourth, within the dermoscopy evidence, how a general-purpose model was used mattered as much as which images it received. Off-the-shelf conversational models, in the versions and with the prompts evaluated, had the lowest sensitivities here [27, 42, 44], whereas the same technology adapted to the task performed among the best, on internal evaluation only [26, 45]. Because these products change rapidly and their output depends on configuration, this contrast describes the evaluations available, not a stable property of the model class.

Fifth, the three studies reporting accuracy below the level of the tumor as a single class, two by histological growth pattern [26, 27] and one by oncologist-assigned clinical form [28], are all internally validated, and each performs worst where the therapeutic consequence is greatest.

### Comparison with Expert Interpretation

Two kinds of comparison are available, and they carry different weight. Direct comparisons, in which AI and human readers were evaluated on the same lesions in the same study, are few: four studies compared a stand-alone algorithm with human readers [19, 30, 45, 47] and two measured the effect of assistance on readers [40, 52], one of them at tier 2. All other comparisons in this section are indirect: they set an artificial-intelligence estimate from one population against an expert benchmark from another. Differences in case mix, prevalence, reference standard, threshold and unit of analysis mean that indirect comparisons can show whether estimates are broadly comparable but cannot establish superiority or inferiority.

For indirect comparison, the published expert benchmarks are as follows (**Table 3**, section C); none is an included study, and each was derived in a different population. For conventional optical coherence tomography, the Cochrane review reported sensitivity 95% (95% CI 91-97) and specificity 77% (69-83) alongside naked-eye and dermoscopic assessment [53]. For reflectance confocal microscopy, it reported 94% (79-98) and 85% (72-92) in difficult lesions [59], and a meta-analysis of 15 studies and 4,163 lesions reported 0.92 (0.87-0.95) and 0.93 (0.85-0.97) [50]. For the line-field technique, prospective expert reading without assistance achieves 0.98 with 0.80 in equivocal lesions [10] and 0.97 (0.91-1.00) with 0.95 (0.90-0.98) in a histology-only analysis [11, 60]. For high-frequency ultrasound the evidence was too variable to pool, so no benchmark exists [54].

A published meta-analysis of deep learning on dermoscopy supplies a further comparison. It is not an evidence stream of this review, but it is the largest pooled estimate available: sensitivity 0.96 (0.93-0.98) and specificity 0.98 (0.96-0.99) on internal validation, against a dermatologist comparator of 0.75 (0.66-0.82) and 0.97 (0.95-0.98) in the same source [16]. The apparent advantage therefore lies almost entirely in sensitivity, and the comparison was not head-to-head: only one of the fifteen pooled studies contributed an external validation set, of roughly 200 images against 32,069 internal. We appraised the source formally and rated confidence in it as low [61].

In the direct comparisons, the two verified dermoscopy studies show the algorithm level with dermatologists reading the same reader sets [19, 30]; the one study comparing a task-adapted conversational model with expert reading of the same images finds the model lower on both measures [45]; the confocal study finds the algorithm better than blinded experts on specificity [47]; and in the line-field reader study assistance improved readers over unaided imaging, most among the least experienced [52]. In the indirect comparisons, the best externally validated confocal result, an AUC of 0.864 [48], cannot be shown to exceed the pooled expert benchmark, and the assisted line-field figures are comparable with, not superior to, published unaided expert reading [10, 11, 52]. The intraoperative sensitivity of 0.46 [49] is not comparable with any benchmark, because it measures agreement with expert annotation at pixel level. The defensible statement is therefore that current evidence does not demonstrate superiority of AI over expert interpretation: performance is comparable under favorable conditions in the few direct comparisons, with a larger and less predictable deterioration under unfavorable ones. This does not make the technology valueless: standardization, scalability and the extension of expert-level reading to less experienced clinicians are real benefits. But headline accuracy figures should not be presented as evidence that algorithms outperform clinicians using the same images.

### Implications for Clinical Practice

The most plausible near-term application is decision support, not autonomous diagnosis. In the line-field reader study the assistant added 10.5 points of sensitivity over unaided imaging, with the gains concentrated among readers less experienced in the technique [52]. The same signature appears in dermoscopy, where assisted reading raised reader sensitivity from 0.84 to 0.91 and specificity from 0.78 to 0.89 [40]. This is a tool that addresses variability in reader expertise,

not a physical limitation of the imaging. Used this way, assisted imaging could plausibly reduce unnecessary biopsies of equivocal lesions on the face and extend specialist-level assessment to less experienced clinicians, which is precisely the setting in which most of these lesions are first seen.

Automated assignment of growth pattern is no longer wholly unevidenced, but the evidence is thin, internal only and, in one of the three studies, based on clinical rather than histological labels [26-28]. The benchmark it must meet is already set by expert imaging: dermoscopy separates superficial from other patterns at 81.9% sensitivity and 81.8% specificity [5], line-field tomography agrees with histology in 90.4% of tumors [7], and dermoscopy-guided ultrasound separates aggressive from non-aggressive disease at 82.4% and 91.3% [8]. No model has been shown to reach these prospectively, and this is the application with the clearest therapeutic consequence [3, 4].

Two qualifications follow. First, none of this supports unsupervised use by non-specialists or by patients. The models most accessible to a non-expert are the general-purpose conversational assistants, and in the versions evaluated those had the lowest sensitivities here; the largest study of them concludes that they should not be used to guide care for this tumor at all [27], and the consensus position is that safety depends on a clinician being able to overrule the algorithm [14]. In the evaluations assembled here, off-the-shelf chatbots missed a substantial proportion of cancers, in one archive-based evaluation more than nine in ten [42], but this figure is specific to the model version, prompts and images tested and should not be read as a general rate for current or future products. Clinicians should be aware that such a consultation may have taken place before the patient arrives and may wish to ask about it. Second, adoption will turn on more than accuracy: in an international survey of 300 dermatologists across 13 countries, the leading concern was the absence of regulation and standardization rather than any technical limitation, and 75% considered formal training in these tools a necessary part of medical education [62].

### Barriers to Translation

Several obstacles remain. Datasets are small, single-center, and either enriched for equivocal lesions or assembled from curated public archives, both of which inflate apparent performance, and external validation in diverse cohorts is scarce. Skin phototype is reported in only three of the twenty-eight studies, and where reported the cohorts are confined to lighter phototypes, with no representation of the darkest anywhere in this literature [27]. Two-by-two counts are reported or derivable in only nine studies. Explainability remains unresolved: clinicians are reluctant to act on predictions whose reasoning they cannot inspect, and while explanation expressed in the diagnostic terms clinicians already use raises confidence more than conventional heat maps [63], it has been argued that current methods cannot deliver the transparency claimed for them [64]. Prospective studies measuring downstream outcomes such as biopsy rates are largely absent.

A final barrier became apparent during our own verification: reporting errors are common enough to compromise secondary analysis. Among the studies we checked we found transposed lesion counts between abstract and results, a test set containing the entire training set, and a pair of near-identical publications reporting one experiment. Anyone pooling this literature should verify every figure against the full text rather than the abstract.

### Strengths and Limitations

Strengths include a focused clinical question covering all non-invasive techniques; duplicate screening and extraction with third-reviewer adjudication; verification of every included report against its primary source; risk-of-bias appraisal with additional algorithm-specific questions; certainty rated separately for sensitivity and specificity; explicit comparison against expert interpretation, distinguishing direct from indirect comparisons; and a sensitivity analysis of the eligibility decisions.

The limitations are these. The review was not prospectively registered and the protocol was not documented in writing until after data extraction; the methods are fully reported and were agreed before searching, but cannot be independently verified as pre-specified, and partial mitigation is offered. The original search contained no terms for conversational or foundation models; the omission was identified during synthesis and corrected to the same close date, but a search that had to be corrected once may be incomplete in ways not yet identified. The two-tier structure is a post-hoc device; it changed no study's appraisal, and the sensitivity analysis reports what the review would conclude without it. Tier assignment depends on what the primary reports document: three studies whose reports do not state how the evaluated lesions were verified are assigned to tier 2 although some of their lesions may in fact have been examined histologically, which makes the tier 1 evidence conservative rather than inflated. Tier 2 is internally heterogeneous, and its four label sources are rated within a single body of evidence for each technique because each alone contains too few studies for a stable rating. The evidence base remains heterogeneous, precluding meta-analysis, and publication bias favoring positive results is likely. Most included studies report neither CIs nor two-by-two counts; five research letters could not be obtained in full text; eligibility was restricted to English-language reports; and a dedicated appraisal instrument for these studies was not available. Direct comparisons between AI and human readers were available in only six studies, so most comparisons with expert interpretation are indirect. Evidence on general-purpose conversational models is specific to the model versions, prompts, image sets and evaluation dates studied; these products change frequently, and newer versions or other configurations may perform differently. Finally, accuracy figures are not comparable across techniques or units of analysis.

### Key Messages

- Across twenty-eight studies and six imaging techniques, current evidence does not demonstrate superiority of AI over expert interpretation. Direct same-study comparisons are few, and most comparisons rely on expert benchmarks from other populations.
- Histopathologically verified studies are the primary evidence; results resting on archive-, consensus or undocumented labels measure technical performance against labels, not diagnostic accuracy.
- The fall in performance from internal to external testing is the most reproducible finding in this literature, and it

is larger than any difference between imaging techniques.

- Off-the-shelf conversational models, the tools most accessible to non-specialists and patients, had the lowest sensitivities in the versions and configurations evaluated; adapted to the task, the same technology performed far better on internal evaluation. These findings are version- and configuration-specific.
- These systems relieve variability in reader expertise, not the physical limits of the imaging. Where the diagnostic features are under-represented in the image, the algorithm inherits the limitation rather than correcting it.
- Accuracy below the level of the tumor as a single class is reported by three studies, one using clinical rather than histological labels; all are internally validated, and all perform worst in the aggressive patterns that determine surgical management.
- Certainty is low or very low for every technique, every tier and both sensitivity and specificity, and no included study is at low risk of patient-selection bias.
- The defensible clinical position is supervised decision support, not autonomous or unsupervised diagnosis.

## CONCLUSIONS

AI applied to non-invasive skin imaging for basal cell carcinoma shows wide and largely unexplained variation in diagnostic accuracy. Within every technique the highest estimates come from curated, internally validated evaluations, and no model evaluated externally, prospectively or in living skin performs better there than on its own development data. Certainty is low or very low throughout, and no included study is at low risk of patient-selection bias. Current evidence does not demonstrate superiority of AI over expert interpretation. In the four direct comparisons of a stand-alone algorithm with human readers on the same lesions, the algorithm was level with or below the readers in three and better on specificity alone in one, and the two reader studies show that assistance improves readers, most among the less experienced. All other comparisons are indirect, against expert benchmarks such as pooled confocal sensitivity 0.92 with specificity 0.93 and prospective line-field sensitivity 0.98, which no externally validated algorithm here can be shown to exceed. The most reproducible signal in this literature is not any accuracy figure but the fall in performance from internal to external testing, and accuracy by growth pattern is lowest in exactly the aggressive patterns that determine surgical management. The priorities for future work are external and prospective validation in the populations in which these tools would be used, across diverse skin phototypes; division of data by patient; complete reporting of two-by-two counts; and trials measuring what changes for the patient, above all the number of biopsies avoided and the number of cancers missed. Until then, the defensible clinical position is supervised decision support for a trained clinician, not autonomous or unsupervised diagnosis.

**Author contributions: IAS:** conceptualization, data curation, formal analysis, investigation, methodology, validation, writing – original draft, writing – review & editing; **ŽAS:** conceptualization, data curation, formal analysis, investigation, supervision, validation, writing – review & editing; **MA:** data curation, methodology, validation, writing – review & editing; **HAS:** formal analysis, methodology, validation, writing – review & editing. All authors have agreed with the results and conclusions.

**Funding:** No funding source is reported for this study.

**Ethical statement:** This is a systematic review of previously published data and did not involve new studies on humans or animals. Hence, ethics approval is not required.

**AI statement:** A large language model was used for language editing of text already drafted by the authors. It was not used for searching, screening, extraction, appraisal, certainty rating, or the generation of data, results, citations or conclusions, all of which were carried out manually and in duplicate by the authors. No AI tool is listed as an author, and the authors take full responsibility for the content of this publication.

**Declaration of interest:** No conflict of interest is declared by the authors.

**Data sharing statement:** Data supporting the findings and conclusions are available upon request from the corresponding author.

## REFERENCES

1. Rogers HW, Weinstock MA, Feldman SR, Coldiron BM. Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the U.S. population, 2012. JAMA Dermatol. 2015;151(10):1081-6. https://doi.org/10.1001/jamadermatol.2015.1187 PMid:25928283
2. Nguyen-Nielsen M, Wang L, Pedersen L, et al. The incidence of metastatic basal cell carcinoma (mBCC) in Denmark, 1997-2010. Eur J Dermatol. 2015;25(5):463-8. https://doi.org/10.1684/ejd.2015.2546 PMid:26105129
3. Peris K, Fargnoli MC, Kaufmann R, et al. European consensus-based interdisciplinary guideline for diagnosis and treatment of basal cell carcinoma-update 2023. Eur J Cancer. 2023;192:113254. https://doi.org/10.1016/j.ejca.2023.113254 PMid:37604067
4. Nasr I, McGrath EJ, Harwood CA, et al. British Association of Dermatologists guidelines for the management of adults with basal cell carcinoma 2021. Br J Dermatol. 2021;185(5):899-920. https://doi.org/10.1111/bjd.20524 PMid:34050920
5. Lallas A, Tzellos T, Kyrgidis A, et al. Accuracy of dermoscopic criteria for discriminating superficial from other subtypes of basal cell carcinoma. J Am Acad Dermatol. 2014;70(2):303-11. https://doi.org/10.1016/j.jaad.2013.10.003 PMid:24268311
6. Longo C, Lallas A, Kyrgidis A, et al. Classifying distinct basal cell carcinoma subtype by means of dermatoscopy and reflectance confocal microscopy. J Am Acad Dermatol. 2014;71(4):716-24.e1. https://doi.org/10.1016/j.jaad.2014.04.067 PMid:24928707
7. Ruini C, Schuh S, Gust C, et al. Line-field optical coherence tomography: In vivo diagnosis of basal cell carcinoma subtypes compared with histopathology. Clin Exp Dermatol. 2021;46(8):1471-81. https://doi.org/10.1111/ced.14762 PMid:34047380
8. Bozsanyi S, Boostani M, Farkas K, et al. Optically guided high-frequency ultrasound to differentiate high-risk basal cell carcinoma subtypes: A single-centre prospective study. J Clin Med. 2023;12(21):6910. https://doi.org/10.3390/jcm12216910 PMid:37959375 PMCid:PMC10648659
9. Boostani M, Pellacani G, Wortsman X, et al. FDA and EMA-approved noninvasive imaging techniques for basal cell carcinoma subtyping: A systematic review. JAAD Int. 2025;21:73-86. https://doi.org/10.1016/j.jdin.2025.05.006 PMid:40689050 PMCid:PMC12270798

10. Gust C, Schuh S, Welzel J, et al. Line-field confocal optical coherence tomography increases the diagnostic accuracy and confidence for basal cell carcinoma in equivocal lesions: A prospective study. Cancers (Basel). 2022;14(4):1082. https://doi.org/10.3390/cancers14041082 PMid:35205830 PMCid:PMC8870684
11. Cano CO, Sanak D, Lenoir C, et al. Diagnostic performance of line-field confocal optical coherence tomography for basal cell carcinoma: A prospective study. Cancers (Basel). 2026;18(1):153. https://doi.org/10.3390/cancers18010153 PMid:41514661 PMCid:PMC12784862
12. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-8. https://doi.org/10.1038/nature21056 PMid:28117445
13. Tschandl P, Rinner C, Apalla Z, et al. Human-computer collaboration for skin cancer recognition. Nat Med. 2020;26(8):1229-34. https://doi.org/10.1038/s41591-020-0942-0 PMid:32572267 PMCid:PMC7341032
14. Daneshjou R, Barata C, Betz-Stablein B, et al. Checklist for evaluation of image-based artificial intelligence reports in dermatology: CLEAR Derm consensus guidelines from the International Skin Imaging Collaboration Artificial Intelligence Working Group. JAMA Dermatol. 2022;158(1):90-6. https://doi.org/10.1001/jamadermatol.2021.4915 PMid:34851366 PMCid:PMC9845064
15. Widaatalla Y, Wolswijk T, Adan F, et al. The application of artificial intelligence in the detection of basal cell carcinoma: A systematic review. J Eur Acad Dermatol Venereol. 2023;37(6):1160-7. https://doi.org/10.1111/jdv.18963 PMid:36785993
16. Liu H, Shang G, Shan Q. Deep learning algorithms in the diagnosis of basal cell carcinoma using dermatoscopy: Systematic review and meta-analysis. J Med Internet Res. 2025;27:e73541. https://doi.org/10.2196/73541 PMid:41043135 PMCid:PMC12534767
17. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. https://doi.org/10.1136/bmj.n71 PMid:33782057
18. McInnes MDF, Moher D, Thombs BD, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: The PRISMA-DTA statement. JAMA. 2018;319(4):388-96. https://doi.org/10.1001/jama.2017.19163 PMid:29362800
19. Wang S-Q, Zhang X-Y, Liu J, et al. Deep learning-based, computer-aided classifier developed with dermoscopic images shows comparable performance to 164 dermatologists in cutaneous disease diagnosis in the Chinese population. Chin Med J (Engl). 2020;133(17):2027-36. https://doi.org/10.1097/CM9.0000000000001023 PMid:32826613 PMCid:PMC7478660
20. Whiting PF, Rutjes AWS, Westwood ME, et al. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529-36. https://doi.org/10.7326/0003-4819-155-8-201110180-00009 PMid:22007046
21. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. Radiol Artif Intell. 2024;6(4):e240300. https://doi.org/10.1148/ryai.240300 PMid:38809149 PMCid:PMC11304031
22. Sounderajah V, Guni A, Liu X, et al. The STARD-AI reporting guideline for diagnostic accuracy studies using artificial intelligence. Nat Med. 2025;31(10):3283-9. https://doi.org/10.1038/s41591-025-03953-8 PMid:40954311
23. Sounderajah V, Ashrafian H, Rose S, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med. 2021;27(10):1663-5. https://doi.org/10.1038/s41591-021-01517-0 PMid:34635854
24. Schunemann HJ, Mustafa RA, Brozek J, et al. GRADE guidelines: 21 part 1. Study design, risk of bias, and indirectness in rating the certainty across a body of evidence for test accuracy. J Clin Epidemiol. 2020;122:129-41. https://doi.org/10.1016/j.jclinepi.2019.12.020 PMid:32060007
25. Schunemann HJ, Mustafa RA, Brozek J, et al. GRADE guidelines: 21 part 2. Test accuracy: Inconsistency, imprecision, publication bias, and other domains for rating the certainty of evidence and presenting it in evidence profiles and summary of findings tables. J Clin Epidemiol. 2020;122:142-52. https://doi.org/10.1016/j.jclinepi.2019.12.021 PMid:32058069
26. Wang Y, Zhang X, Wu H, Xiao L, Liu J. Fine-tuning a large multimodal neural network with clinical and dermoscopic images for accurate detection and histopathological subtype prediction of basal cell carcinoma in long-tailed distributions. Chin Med J (Engl). 2025;138(10):1251-3. https://doi.org/10.1097/CM9.0000000000003498 PMid:40090971 PMCid:PMC12091604
27. Boostani M, Zouboulis CC, Pellacani G, et al. ChatGPT, Gemini, and Claude in clinical and dermoscopic image analysis of basal cell carcinoma and its common mimickers: A comparative performance analysis. JID Innov. 2026;6(1):100463. https://doi.org/10.1016/j.xjidi.2026.100463 PMid:41952838 PMCid:PMC13054413
28. Surkov YI, Serebryakova IA, Kuzinova YK, et al. Multimodal method for differentiating various clinical forms of basal cell carcinoma and benign neoplasms in vivo. Diagnostics (Basel). 2024;14(2):202. https://doi.org/10.3390/diagnostics14020202 PMid:38248078 PMCid:PMC10814941
29. Menzies SW, Westerhoff K, Rabinovitz H, Kopf AW, McCarthy WH, Katz B. Surface microscopy of pigmented basal cell carcinoma. Arch Dermatol. 2000;136(8):1012-6. https://doi.org/10.1001/archderm.136.8.1012 PMid:10926737
30. Zhu CY, Wang YK, Chen HP, et al. A deep learning based framework for diagnosing multiple skin diseases in a clinical environment. Front Med (Lausanne). 2021;8:626369. https://doi.org/10.3389/fmed.2021.626369 PMid:33937279 PMCid:PMC8085301
31. Kefel S, Pelin Kefel S, LeAnder RW, et al. Adaptable texture-based segmentation by variance and intensity for automatic detection of semitranslucent and pink blush areas in basal cell carcinoma. Skin Res Technol. 2016;22(4):412-22. https://doi.org/10.1111/srt.12281 PMid:26991418
32. Mei L-H, Cao M-K, Li J, Ye X-G, Liu X-D, Yang G. Deep learning in assisting dermatologists in classifying basal cell carcinoma from seborrheic keratosis. Front Oncol. 2025;15:1507322. https://doi.org/10.3389/fonc.2025.1507322 PMid:40342818 PMCid:PMC12058839

33. Naeem A, Anees T, Fiza M, Naqvi RA, Lee S-W. SCDNet: A deep learning-based framework for the multiclassification of skin cancer using dermoscopy images. Sensors (Basel). 2022;22(15):5652. https://doi.org/10.3390/s22155652 PMid:35957209 PMCid:PMC9371071
34. Tajerian A, Kazemian M, Tajerian M, Malayeri AA. Design and validation of a new machine-learning-based diagnostic tool for the differentiation of dermatoscopic skin cancer images. PLoS One. 2023;18(4):e0284437. https://doi.org/10.1371/journal.pone.0284437 PMid:37058446 PMCid:PMC10104315
35. Huang H-Y, Hsiao Y-P, Mukundan A, Tsao Y-M, Chang W-Y, Wang H-C. Classification of skin cancer using novel hyperspectral imaging engineering via YOLOv5. J Clin Med. 2023;12(3):1134. https://doi.org/10.3390/jcm12031134 PMid:36769781 PMCid:PMC9918106
36. Maurya A, Stanley RJ, Lama N, et al. Hybrid topological data analysis and deep learning for basal cell carcinoma diagnosis. J Imaging Inform Med. 2024;37(1):92-106. https://doi.org/10.1007/s10278-023-00924-8 PMid:38343238 PMCid:PMC10976904
37. Maurya A, Stanley RJ, Aradhyula HY, et al. Basal cell carcinoma diagnosis with fusion of deep learning and telangiectasia features. J Imaging Inform Med. 2024;37(3):1137-50. https://doi.org/10.1007/s10278-024-00969-3 PMid:38332404 PMCid:PMC11169204
38. Guan H, Yuan Q, Lv K, et al. Dermoscopy-based radiomics help distinguish basal cell carcinoma and actinic keratosis: A large-scale real-world study based on a 207-combination machine learning computational framework. J Cancer. 2024;15(11):3350-61. https://doi.org/10.7150/jca.94759 PMid:38817855 PMCid:PMC11134443
39. Lin T-L, Mukundan A, Karmakar R, Avala P, Chang W-Y, Wang H-C. Hyperspectral imaging for enhanced skin cancer classification using machine learning. Bioengineering (Basel). 2025;12(7):755. https://doi.org/10.3390/bioengineering12070755 PMid:40722447 PMCid:PMC12292285
40. Wang Z, Hu H, Liu Z, et al. Rethinking BCC diagnosis: Automating concept-specific detection of BCC in dermatoscopic images. J Dtsch Dermatol Ges. 2025;23(2):184-93. https://doi.org/10.1111/ddg.15608 PMid:39737619
41. Matas I, Serrano C, Silva-Claveria F, Serrano A, Toledo-Pastrana T, Acha B. MultiTask learning AI system to assist BCC diagnosis with dual explanation. Sci Rep. 2026;16:40229. https://doi.org/10.1038/s41598-026-40229-8 PMid:42156415 PMCid:PMC13187323
42. Chetla N, Chen M, Chang J, et al. Assessing the diagnostic accuracy of ChatGPT-4 in identifying diverse skin lesions against squamous and basal cell carcinoma. JMIR Dermatol. 2025;8:e67299. https://doi.org/10.2196/67299 PMid:40116774 PMCid:PMC11951809
43. Yan S, Yu Z, Primiero C, et al. A multimodal vision foundation model for clinical dermatology. Nat Med. 2025;31(8):2691-702. https://doi.org/10.1038/s41591-025-03747-y PMid:40481209 PMCid:PMC12353815
44. Ünal E, Yücel MB, Gedikli SN, Akay BN. ChatGPT-4o as a diagnostic tool for skin cancer: Diagnostic accuracy in melanoma and non-melanoma detection. Cutan Ocul Toxicol. 2026;45(1):43-50. https://doi.org/10.1080/15569527.2025.2598573 PMid:41355746
45. Traini DO, Palmisano G, Peris K. Large language models and dermoscopy: Assessing the potential of task-specific GPT-4 vision in diagnosing basal cell carcinoma. J Eur Acad Dermatol Venereol. 2024;38(12):2320-2. https://doi.org/10.1111/jdv.20333 PMid:39258894
46. Nori S, Rius-Diaz F, Cuevas J, et al. Sensitivity and specificity of reflectance-mode confocal microscopy for in vivo diagnosis of basal cell carcinoma: A multicenter study. J Am Acad Dermatol. 2004;51(6):923-30. https://doi.org/10.1016/j.jaad.2004.06.028 PMid:15583584
47. Campanella G, Navarrete-Dechent C, Liopyris K, et al. Deep learning for basal cell carcinoma detection for reflectance confocal microscopy. J Invest Dermatol. 2022;142(1):97-103. https://doi.org/10.1016/j.jid.2021.06.015 PMid:34265329 PMCid:PMC9338423
48. O'Hagan R, Acciardi A, Rodriguez MT, et al. Deep learning analysis of reflectance confocal microscopy for basal cell carcinoma diagnosis. JAAD Int. 2026;27:21-2. https://doi.org/10.1016/j.jdin.2026.04.010 PMid:42256604 PMCid:PMC13240722
49. Shavlokhova V, Vollmer M, Gholam P, et al. Deep learning on basal cell carcinoma in vivo reflectance confocal microscopy data. J Pers Med. 2022;12(9):1471. https://doi.org/10.3390/jpm12091471 PMid:36143256 PMCid:PMC9506260
50. Lupu M, Popa IM, Voiculescu VM, Caruntu A, Caruntu C. A systematic review and meta-analysis of the accuracy of in vivo reflectance confocal microscopy for the diagnosis of primary basal cell carcinoma. J Clin Med. 2019;8(9):1462. https://doi.org/10.3390/jcm8091462 PMid:31540342 PMCid:PMC6780971
51. Jørgensen TM, Tycho A, Mogensen M, Bjerring P, Jemec GBE. Machine-learning classification of non-melanoma skin cancers from image features obtained by optical coherence tomography. Skin Res Technol. 2008;14(3):364-9. https://doi.org/10.1111/j.1600-0846.2008.00304.x PMid:19159385
52. Fischman S, Viel T, Perrot J-L, et al. AI-assisted basal cell carcinoma diagnosis with LC-OCT: A multicentric retrospective study. J Eur Acad Dermatol Venereol. 2026;40(5):1059-68. https://doi.org/10.1111/jdv.70099 PMid:41109970 PMCid:PMC13206264
53. Ferrante di Ruffano L, Dinnes J, Deeks JJ, et al. Optical coherence tomography for diagnosing skin cancer in adults. Cochrane Database Syst Rev. 2018;12:CD013189. https://doi.org/10.1002/14651858.CD013189 PMid:30521690 PMCid:PMC6516952
54. Dinnes J, Bamber J, Chuchu N, et al. High-frequency ultrasound for diagnosing skin cancer in adults. Cochrane Database Syst Rev. 2018;12:CD013188. https://doi.org/10.1002/14651858.CD013188 PMid:30521683 PMCid:PMC6516989
55. Marosán-Vilimszky P, Szalai K, Horváth A, et al. Automated skin lesion classification on ultrasound images. Diagnostics (Basel). 2021;11(7):1207. https://doi.org/10.3390/diagnostics11071207 PMid:34359290 PMCid:PMC8303815
56. Courtenay LA, Gonzalez-Aguilera D, Lagüela S, et al. Deep convolutional neural support vector machines for the classification of basal cell carcinoma hyperspectral signatures. J Clin Med. 2022;11(9):2315. https://doi.org/10.3390/jcm11092315 PMid:35566440 PMCid:PMC9102335

57. Parasca SV, Calin MA, Manea D, Radvan R. Hyperspectral imaging with machine learning for in vivo skin carcinoma margin assessment: A preliminary study. Phys Eng Sci Med. 2024;47(3):1141-52. https://doi.org/10.1007/s13246-024-01435-8 PMid:38771442 PMCid:PMC11408400
58. Räsänen J, Salmivuori M, Pölönen I, Grönroos M, Neittaanmäki N. Hyperspectral imaging reveals spectral differences and can distinguish malignant melanoma from pigmented basal cell carcinomas: A pilot study. Acta Derm Venereol. 2021;101(2):adv00405. https://doi.org/10.2340/00015555-3755 PMid:33521835 PMCid:PMC9366698
59. Dinnes J, Deeks JJ, Chuchu N, et al. Reflectance confocal microscopy for diagnosing keratinocyte skin cancers in adults. Cochrane Database Syst Rev. 2018;12:CD013191. https://doi.org/10.1002/14651858.CD013191 PMid:30521687 PMCid:PMC6516892
60. Cinotti E, Brunetti T, Cartocci A, et al. Diagnostic accuracy of line-field confocal optical coherence tomography for the diagnosis of skin carcinomas. Diagnostics (Basel). 2023;13(3):361. https://doi.org/10.3390/diagnostics13030361 PMid:36766466 PMCid:PMC9914674
61. Shea BJ, Reeves BC, Wells G, et al. AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358:j4008. https://doi.org/10.1136/bmj.j4008 PMid:28935701
62. Karampinis E, Zoumpourli C-M, Lallas A, et al. Perspectives on artificial intelligence in dermatology: An international cross-sectional study. Medicina (Kaunas). 2026;62(4):759. https://doi.org/10.3390/medicina62040759 PMid:42075630 PMCid:PMC13118061
63. Chanda T, Hauser K, Hobelsberger S, et al. Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. Nat Commun. 2024;15:524. https://doi.org/10.1038/s41467-023-43095-4 PMid:38225244
64. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health. 2021;3(11):e745-50. https://doi.org/10.1016/S2589-7500(21)00208-9 PMid:34711379