Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AI Skin Analysis in Real-World Practice: A Multi-Source Evidence Assessment

Viktor Shpudeiko

v.shpudeiko@gmail.com

Skinive BV https://orcid.org/0000-0002-8612-3436

Kirill Sokolov

Skinive BV

Research Article

Keywords: Artificial Intelligence, Deep Learning, Dermatology, Skin Neoplasms, Clinical Decision Support Systems, Real-World Evidence, Triage

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11727541/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare potential competing interests as follows: The authors declare the following financial and non-financial competing interests: Kirill Sokolov is the founder and Chief Executive Officer of Skinive BV, and Viktor Shpudeiko is affiliated with Skinive BV. Both authors are directly involved in the research, development, validation, and commercialization of the Skinive AI software evaluated in this report. To mitigate potential bias, the clinical stress-test data were collected and reference diagnoses were established independently by clinicians at the participating tertiary referral center, which retained full autonomy over clinical interpretation. The authors declare no other competing interests.

Title: AI Skin Analysis in Real-World Practice: A Multi-Source Evidence Assessment

Authors: Kirill Sokolov[1], Viktor Shpudeiko[1]*
Affiliation: [1]Skinive BV, Jacob Van Lennepkade 155 H, 1054 ZL, Amsterdam, The Netherlands.
*Corresponding Author: Viktor Shpudeiko (v.shpudeiko@gmail.com)

Keywords: Artificial Intelligence; Deep Learning; Dermatology; Skin Neoplasms; Clinical Decision Support Systems; Real-World Evidence; Triage

Abstract

Background: Artificial intelligence (AI) is increasingly utilized to support the early assessment of skin lesions. While numerous studies report high diagnostic accuracy under controlled experimental conditions, considerably fewer integrate technical validation, independent clinical stress-testing, and large-scale real-world evidence (RWE) within a single transparent framework.

Objective: To evaluate the capabilities, clinical utility, and operational limitations of a mobile AI-assisted skin analysis system by synthesizing three complementary streams of evidence.

Methods: This study integrates: (1) an analysis of longitudinal technical performance metrics and big-data usage patterns from over 6.57 million real-world smartphone assessments; (2) an independent clinical evaluation ((n = 160) analyzed cases) conducted at a specialized tertiary referral dermatology center comparing AI predictions and triage routing against histopathological, cytological, or expert clinical reference standards; and (3) a comparative synthesis with contemporary peer-reviewed literature, including external validation on skin of color.

Results: Longitudinal technical validation demonstrated consistent, balanced improvements across sensitivity (97.4%), specificity (93.1%), and overall accuracy

(94.2%) by 2026. In real-world consumer usage, common inflammatory conditions (e.g., acne/rosacea, (n = 708,356)) accounted for the largest share of assessments, while preventable image-acquisition errors (poor lighting: 42.0%; poor focus: 28.0%; motion blur: 18.0%) represented the primary modifiable determinants of failure. In the independent tertiary referral cohort - heavily enriched with malignant and diagnostically challenging lesions - exact Top-1 and Top-3 diagnostic agreements were 54.4% and 70.6%, respectively. However, combining Top-3 agreement, correct broader disease-group assignment, and appropriate risk-based routing yielded a clinically useful triage outcome in 80.0% of cases.

Conclusions: AI-assisted skin analysis provides clinically meaningful support for early risk stratification and patient referral when deployed as a clinical decision-support tool rather than an autonomous diagnostic system. Comprehensive evaluation of medical AI requires pairing conventional accuracy metrics with triage-oriented clinical endpoints and transparent RWE reporting.

1. Introduction

Artificial intelligence (AI) has rapidly evolved within dermatology, particularly in the evaluation of skin lesions using clinical and smartphone-captured photography. Over the past decade, advances in deep learning have enabled convolutional neural networks and related architectures to achieve diagnostic performance comparable to that of experienced dermatologists under controlled experimental conditions.

These developments have stimulated substantial interest in AI-assisted decision support for skin cancer screening, teledermatology, and patient self-assessment. Nevertheless, translating algorithmic performance from curated benchmarks into routine clinical workflows remains a significant scientific and clinical challenge.

Most published validation studies evaluate AI models using conventional statistical indicators such as sensitivity, specificity, overall accuracy, or the area under the receiver operating characteristic curve (AUROC). While indispensable for

model optimization, these metrics capture only one dimension of clinical reliability. Performance measured on standardized validation datasets may not accurately reflect algorithmic behavior when confronted with heterogeneous lighting, variable smartphone camera pipelines, rare dermatoses, and the broad spectrum of skin phenotypes encountered in everyday practice.

Consequently, increasing attention is directed toward Real-World Evidence (RWE) as a vital complement to traditional validation frameworks. For decision-support technologies in dermatology, clinically meaningful outcomes encompass not only exact diagnostic agreement, but also appropriate risk stratification, patient triage, and specialist referral prioritization.

The objective of this report is to provide a balanced, transparent assessment of AI-assisted skin analysis by integrating three complementary sources of evidence: (1) longitudinal technical validation and large-scale real-world usage data from the Skinive platform; (2) an independent clinical stress-test conducted in a specialized tertiary referral dermatology center managing diagnostically complex lesions; and (3) a comparative synthesis of contemporary peer-reviewed literature, regulatory guidance, and external clinical validations [1 - 7, 10 - 17]. Rather than claiming diagnostic autonomy, this framework examines how mobile AI functions as an adjunct for early risk awareness and clinical triage.

2. Methods

2.1. Longitudinal Technical Validation and Real-World Big-Data Analysis

To evaluate algorithmic stability outside controlled clinical trials, aggregated, anonymized metadata from over 6.57 million real-world smartphone skin assessments (accumulated through Q3 2026) were analyzed. These examinations were initiated voluntarily by users in non-clinical environments across diverse geographical regions - predominantly Europe (43.1%), Asia (25.1%), South America (15.2%), Oceania and Australia (7.2%), North America (6.3%), and Africa

(3.0%) - utilizing varied consumer mobile devices (approximately 60.0% Android and 40.0% iOS).

Technical performance across successive software iterations (2021 - 2026) was evaluated against a standardized, clinically annotated internal validation benchmark comprising approximately 250,000 images across 55 dermatological conditions. Unsuccessful image submissions were audited to categorize user-modifiable acquisition errors.

2.2. Independent Clinical Stress-Test in a Tertiary Referral Center

In 2026, an independent clinical evaluation was performed at a specialized tertiary referral dermatology center whose institutional identity remains confidential under a non-disclosure agreement. Because this institution manages diagnostically uncertain lesions referred from primary care and general dermatology practices, the study functioned as a clinical stress-test rather than a population-representative screening trial.

A total of 241 lesion examinations were conducted; 160 cases possessed complete AI outputs and definitive reference standards, forming the primary analytical cohort. Reference diagnoses were established via histopathology (n = 117), cytology (n = 22), or extended specialist clinical assessment (n = 21). Treating dermatologists evaluated patients independently of the AI output. Smartphone photographs were processed by the algorithm, which returned three ranked diagnostic suggestions alongside an associated risk tier.

Performance was stratified across five clinically distinct endpoints:

Top-1 Diagnostic Agreement: Exact match between the highest-ranked AI prediction and the reference diagnosis.

Top-3 Diagnostic Agreement: Reference diagnosis appearing within any of the top three ranked AI predictions.

Correct Disease-Group Classification: Exact diagnostic label mismatch, but correct assignment to the broader pathological category (e.g., identifying a malignant lesion as an alternative malignant tumor type).

Correct Risk-Based Routing: Incorrect precise label and disease group, but accurate triage recommendation prompting necessary specialist evaluation (e.g., predicting actinic keratosis for a lesion confirmed as basal cell carcinoma, where both trigger clinical referral).

Clinically Useful Composite Outcome: The aggregate proportion of cases achieving Top-3 agreement, correct disease-group classification, or correct risk-based routing.

Differences between basal cell carcinoma (BCC) and squamous cell carcinoma (SCC), as well as between invasive melanoma and lentigo maligna melanoma, were treated as concordant within broader oncological management pathways due to visual overlap requiring histological confirmation.

2.3. Comparative Literature Synthesis and External Validation Context

To situate these findings within the broader scientific landscape, results were compared against systematic reviews, international regulatory guidelines, and independent clinical trials [1 - 7, 10 - 17]. Particular emphasis was placed on an external, cross-sectional clinical validation conducted in Egypt by Zaki et al. (2025), which evaluated the algorithm alongside two comparator AI applications against histopathology in patients with darker skin phototypes (Fitzpatrick types I - IV).

3. Results

3.1. Real-World Usage Patterns, Image Quality, and Model Evolution

Analysis of the 6.57 million real-world assessments revealed that users employ mobile dermatology tools across a wide spectrum of conditions rather than exclusively for pigmented lesion screening. Inflammatory dermatoses represented the largest volume of identified findings, as detailed in Table 1.

Across the complete dataset, 14.68% of assessments flagged clinically relevant risk indicators warranting professional medical consultation, while 85.32% indicated no immediate high-risk features.

Table 1. Distribution of Major Dermatological Categories Identified in Real-World Mobile Assessments (2026) [8]

Dermatological Category	Number of AI Assessments (n)
Acne / Rosacea	708,356
Suspected Skin Cancer Risks	198,004
HPV-Related Lesions	105,706
Fungal Skin Conditions	1,867

Audit of unsuccessful or low-quality image submissions demonstrated that image degradation was predominantly driven by preventable user-end factors: insufficient illumination accounted for 42.0% of failures, poor camera focus for 28.0%, motion blur for 18.0%, and miscellaneous hardware or framing issues for 12.0%.

Longitudinal benchmarking across software iterations (2021 - 2026) demonstrated sustained improvements across all core diagnostic metrics without sacrificing specificity for sensitivity (Table 2).

Table 2. Longitudinal Technical Validation Metrics Across Algorithm Iterations (2021 - 2026) [8]

Performance Metric	2021	2022	2024	2026
Sensitivity	94.6%	98.2%	95.9%	97.4%
Specificity	93.0%	91.7%	91.5%	93.1%
Overall Accuracy	93.5%	92.9%	93.1%	94.2%
F1-Score	86.4%	83.5%	84.2%	89.1%
Precision	80.6%	72.7%	75.1%	82.2%

3.2. Clinical Stress-Test Outcomes in the Tertiary Referral Cohort

The referral cohort (n = 160) exhibited a high prevalence of atypical, premalignant, and malignant lesions - including suspected melanoma, BCC, SCC, Bowen disease, and dysplastic nevi - with minimal representation of straightforward benign nevi.

Top-1 and Top-3 Diagnostic Agreement: The highest-ranked AI prediction matched the reference diagnosis in 87 of 160 cases (54.4%). An additional 19 cases (11.9%) matched on the second prediction, and 7 cases (4.4%) matched on the third, yielding a cumulative Top-3 diagnostic agreement of 70.6% (113/160).

Disease-Group and Routing Concordance: Six additional cases (3.75%) were assigned to the correct broader disease category despite an exact label mismatch, and 9 cases (5.6%) generated an appropriate specialist referral recommendation despite diagnostic disagreement.

Clinically Useful Composite Outcome: Combining Top-3 agreement (n = 113), correct disease-group classification (n = 6), and correct risk-based routing (n = 9) yielded a clinically actionable outcome in 80.0% of cases (128/160).

Discordant Cases and Subgroup Analysis: Unrecognized rare pathologies accounted for 3.75% (n = 6), while completely discordant outputs (disagreement across diagnosis, disease group, and routing) occurred in 16.25% (n = 26), primarily involving atypical presentations or rare entities such as Spitz nevus, juvenile xanthogranuloma, and pyogenic granuloma.

False-positive risk flags outnumbered false-negative assessments. In a separate subgroup of 23 cases lacking definitive histology or conclusive cytology, the algorithm's Top-3 output agreed with the specialist's clinical diagnosis in 19 cases (82.6%).

3.3. External Validation on Skin of Color and Literature Synthesis

In the independent Egyptian outpatient study by Zaki et al. (2025), 200 histopathologically confirmed lesions (100 benign: 48 seborrheic keratoses, 40 melanocytic nevi, 12 actinic keratoses; 100 malignant: 51 BCC, 37 SCC, 12 melanoma) were evaluated in patients with Fitzpatrick skin types I - IV.

As summarized in Table 3, the evaluated system achieved an overall accuracy of 93.0% (sensitivity: 94.0%; specificity: 92.0%), performing comparably to board-certified dermatologists (92.5% accuracy) and outperforming comparative

commercial tools in Rank-1 detection of SCC (81.1%) and seborrheic keratosis (85.4%), though exhibiting a higher failure rate in actinic keratosis (25.0%).

Table 3. Comparative Diagnostic Performance in an Independent Histopathology-Confirmed Cohort (Zaki et al., 2025) [15]

Evaluation Metric / Endpoint	Skinive AI	Model Dermatology	Tibot AI	Dermatologists
Sensitivity	94.0%	94.0%	95.0%	95.0%
Specificity	92.0%	95.0%	69.0%	90.0%
Positive Predictive Value (PPV)	92.2%	94.9%	75.4%	90.5%
Negative Predictive Value (NPV)	93.9%	94.1%	93.2%	94.7%
Overall Accuracy	93.0%	94.5%	82.0%	92.5%
Rank-1 Accuracy: BCC	84.3%	82.4%	79.2%	92.2%
Rank-1 Accuracy: SCC	81.1%	73.0%	40.5%	40.5%
Rank-1 Accuracy: Melanoma	83.3%	66.7%	83.3%	83.3%
Rank-1 Accuracy: Seborrheic Keratosis	85.4%	64.6%	52.1%	52.1%
Failure Rate: Actinic Keratosis	25.0%	0.0%	0.0%	16.7%

These external findings align with prior clinical evaluations reporting 95.3% sensitivity and 93.5% specificity across skin neoplasms, as well as broader literature demonstrating that deep learning systems achieve robust diagnostic parity when paired with clinical metadata, whereas earlier single-center evaluations of alternative consumer apps showed lower specificity or benign-lesion accuracy.

4. Discussion

A central insight emerging from this multi-source assessment is that evaluating medical AI solely through single-label accuracy metrics under controlled conditions provides an incomplete measure of clinical value [5 - 7]. In the tertiary referral stress-test, exact Top-1 agreement was 54.4% - a predictable consequence of evaluating smartphone photographs of diagnostically ambiguous lesions that required histopathology for definitive resolution. However, because the system preserved clinically useful differential diagnoses, disease-group categorization, or

appropriate risk routing in 80.0% of cases, it demonstrated substantial utility as a triage adjunct.

From an oncological safety perspective, the algorithm's tendency toward false-positive rather than false-negative classifications in complex lesions reduces the risk of missed malignancies during early screening. Nonetheless, false-positive alerts carry non-trivial clinical trade-offs, including patient anxiety and increased healthcare utilization, underscoring the necessity of positioning AI strictly as a decision-support tool rather than an autonomous diagnostic replacement [10 - 14]. Furthermore, the real-world analysis of 6.57 million checks demonstrates that image quality - specifically lighting and focus - remains the primary modifiable barrier to reliable mobile assessment, highlighting the need for automated on-device image quality gating prior to inference.

Finally, cross-referencing internal longitudinal data with external cohorts such as the Egyptian study [15] addresses a critical gap in dermatological AI: generalizability across non-European skin phototypes (Fitzpatrick types I - IV). While differences in study design and disease prevalence preclude direct head-to-head superiority claims across platforms, the convergence of internal RWE and independent clinical testing supports the responsible integration of validated AI tools into primary care triage and teledermatology workflows.

Study Limitations

Several methodological limitations must be acknowledged when interpreting these findings:

Selection and Spectrum Bias: The independent clinical stress-test was conducted in a tertiary referral center enriched with malignant and atypical lesions (n = 160). Consequently, these metrics cannot be extrapolated to unselected primary care or general consumer populations where benign lesions predominate.

Reference Standard Heterogeneity: While histopathology confirmed the majority of referral cases (73.1%), cytology and expert clinical consensus served as

reference standards for the remaining cohort, introducing minor diagnostic certainty variance.

Proprietary Dataset Constraints: Because the underlying 250,000-image training repository and 6.57-million-check RWE logs constitute proprietary clinical and commercial assets, full external pixel-level reproducibility is restricted, although validation protocols and aggregated metrics are reported transparently.

Phototype Representation: External validation in darker skin tones excluded Fitzpatrick type V - VI individuals, necessitating further prospective evaluation across the full spectrum of skin pigmentation.

5. Conclusion

Integrating longitudinal technical validation, real-world usage analytics, and independent clinical stress-testing offers a transparent framework for assessing AI-assisted dermatology. While mobile deep learning models cannot replace dermoscopy, clinical examination, or histopathology, they demonstrate strong potential to support early risk awareness, patient self-screening, and specialist referral prioritization. Continued prospective, multicenter trials in unselected patient populations remain essential to refine clinical thresholds and ensure equitable performance across all healthcare settings.

Declarations

Ethics approval and consent to participate: This study synthesized retrospectively aggregated, fully anonymized usage metadata and secondary summary data from an independent institutional evaluation. The independent clinical assessment was conducted by clinicians at a specialized tertiary referral dermatology center in accordance with routine clinical practice and local institutional governance; no experimental interventions were performed, and no identifiable patient data were accessible to the authors. For the aggregated mobile usage metrics, all data were de-identified prior to analysis in compliance with the European Union General Data Protection Regulation (GDPR) and platform privacy

terms. Consequently, formal ethics committee approval and individual informed consent for this secondary analysis were waived/not applicable.

Consent for publication: Not applicable. The manuscript contains no individual person's data in any form (including individual details, images, or videos).

Availability of data and materials: The complete training image repository and raw consumer assessment logs are not publicly available due to patient privacy protections, contractual confidentiality agreements with the clinical referral institution, and proprietary intellectual property restrictions held by Skinive BV. Aggregated validation metrics and methodological summaries are presented within the article and the cited Skinive Accuracy Report 2026. Additional de-identified, aggregated statistical summaries may be made available from the corresponding author upon reasonable scientific request.

Competing interests: The authors declare the following financial and non-financial competing interests: Kirill Sokolov is the founder and Chief Executive Officer of Skinive BV, and Viktor Shpudeiko is affiliated with Skinive BV. Both authors are directly involved in the research, development, validation, and commercialization of the Skinive AI software evaluated in this report. To mitigate potential bias, the clinical stress-test data were collected and reference diagnoses were established independently by clinicians at the participating tertiary referral center, which retained full autonomy over clinical interpretation. The authors declare no other competing interests.

Funding: This research was supported internally by Skinive BV. No external public, commercial, or not-for-profit grants were received for the preparation of this manuscript.

Authors' contributions: VS (MD, oncologist) led the clinical interpretation of the findings, curated the tertiary referral center evaluation data, performed the comparative literature synthesis, and drafted the manuscript. KS (technical lead and

director) conceptualized the multi-source evaluation framework, oversaw the deep learning algorithm architecture and technical validation metrics, and critically revised the manuscript. Both authors read and approved the final manuscript.

Acknowledgments: None.

References

1. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115 - 118. doi:10.1038/nature21056.
2. Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol. 2018;29(8):1836 - 1842. doi:10.1093/annonc/mdy166.
3. Han SS, Moon IJ, Kim SH, Na JI, Kim MS, Park GH, et al. Assessment of deep neural networks for the diagnosis of benign and malignant skin neoplasms in comparison with dermatologists: A retrospective validation study. PLoS Med. 2020;17(11):e1003381. doi:10.1371/journal.pmed.1003381.
4. Freeman K, Dinnes J, Chuchu N, Takwoingi Y, Bayliss SE, Matin RN, et al. Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. BMJ. 2020;368:m127. doi:10.1136/bmj.m127.
5. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44 - 56. doi:10.1038/s41591-018-0300-7.
6. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against healthcare professionals in detecting diseases from medical imaging: a systematic review and meta-

analysis. Lancet Digit Health. 2019;1(6):e271 - e297. doi:10.1016/S2589-7500(19)30123-2.

7. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. doi:10.1186/s12916-019-1426-2.
8. Skinive. Skinive Accuracy Report 2026. Amsterdam: Skinive BV; 2026.
9. Skinive. Clinical Assessment of Skinive AI in a Specialized Tertiary Referral Dermatology Center. Internal Clinical Evaluation Report. 2026.
10. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347 - 1358. doi:10.1056/NEJMra1814259.
11. Chen JH, Asch SM. Machine learning and prediction in medicine - Beyond the peak of inflated expectations. N Engl J Med. 2017;376(26):2507 - 2509. doi:10.1056/NEJMp1702071.
12. Mathews SC, McShea MJ, Hanley CL, Ravitz A, Labrique AB, Cohen AB. Digital health: a path to validation. NPJ Digit Med. 2019;2:38. doi:10.1038/s41746-019-0111-3.
13. World Health Organization. Ethics and Governance of Artificial Intelligence for Health. Geneva: World Health Organization; 2021.
14. U.S. Food and Drug Administration, Health Canada, Medicines and Healthcare products Regulatory Agency. Good Machine Learning Practice for Medical Device Development: Guiding Principles. 2021.
15. Zaki AM, Abd-Alaziz HM, Abd-Alaziz MAS. Evaluation of an Artificial Intelligence Technology for the Diagnosis of Pigmented Skin Lesions in Egyptian Patients. Al-Azhar Int Med J. 2025;2025(5):40 - 47.
16. Sokolov K, Shpudeiko V. Dynamics of the neural network accuracy in the context of modernization of the algorithms of skin pathology recognition. Indian J Dermatol. 2022;67(4):386 - 392. doi:10.4103/ijd.ijd_1155_21.
17. Patil S, Rao ND, Patil A, Basar F, Bate S. Assessment of Tibot® Artificial Intelligence Application in Prediction of Diagnosis in Dermatological

Conditions: Results of a Single Centre Study. Indian Dermatol Online J. 2020;11(6):910 - 914. doi:10.4103/idoj.IDOJ_271_20.