

Artificial intelligence in cutaneous lesions: Where do we stand and what is next?

Edited by

Mara Giavina-Bianchi and Justin Ko

Published in

Frontiers in Medicine



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-4911-7
DOI 10.3389/978-2-8325-4911-7

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Artificial intelligence in cutaneous lesions: Where do we stand and what is next?

Topic editors

Mara Giavina-Bianchi — Albert Einstein Israelite Hospital, Brazil
Justin Ko — Stanford University, United States

Citation

Giavina-Bianchi, M., Ko, J., eds. (2024). *Artificial intelligence in cutaneous lesions: Where do we stand and what is next?* Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-4911-7

Table of contents

- 05 **Editorial: Artificial intelligence in cutaneous lesions: where do we stand and what is next?**
Mara Giavina-Bianchi and Justin Ko
- 08 **Explainability agreement between dermatologists and five visual explanations techniques in deep neural networks for melanoma AI classification**
Mara Giavina-Bianchi, William Gois Vitor, Victor Fornasiero de Paiva, Aline Lissa Okita, Raquel Machado Sousa and Birajara Machado
- 21 **Effectiveness of an image analyzing AI-based Digital Health Technology to identify Non-Melanoma Skin Cancer and other skin lesions: results of the DERM-003 study**
Helen Marsden, Caroline Morgan, Stephanie Austin, Claudia DeGiovanni, Marcello Venzi, Polychronis Kemos, Jack Greenhalgh, Dan Mullarkey and Ioullos Palamaras
- 32 **Development and validation of an artificial intelligence-powered acne grading system incorporating lesion identification**
Jiaqi Li, Dan Du, Jianwei Zhang, Wenjie Liu, Junyou Wang, Xin Wei, Li Xue, Xiaoxue Li, Ping Diao, Lei Zhang and Xian Jiang
- 40 **Principles, applications, and future of artificial intelligence in dermatology**
Jesutofunmi A. Omiye, Haiwen Gui, Roxana Daneshjou, Zhuo Ran Cai and Vijaytha Muralidharan
- 49 **Finetuning of GLIDE stable diffusion model for AI-based text-conditional image synthesis of dermoscopic images**
Veronika Shavlokhova, Andreas Vollmer, Christos C. Zouboulis, Michael Vollmer, Jakob Wollborn, Gernot Lang, Alexander Kübler, Stefan Hartmann, Christian Stoll, Elisabeth Roeder and Babak Saravi
- 57 **Real-world post-deployment performance of a novel machine learning-based digital health technology for skin lesion assessment and suggestions for post-market surveillance**
Lucy Thomas, Chris Hyde, Dan Mullarkey, Jack Greenhalgh, Dilraj Kalsi and Justin Ko
- 69 **Patient perspectives of artificial intelligence as a medical device in a skin cancer pathway**
Anusuya Kawsar, Khawar Hussain, Dilraj Kalsi, Polychronis Kemos, Helen Marsden and Lucy Thomas
- 74 **Artificial intelligence for skin cancer detection and classification for clinical environment: a systematic review**
Brunna C. R. S. Furriel, Bruno D. Oliveira, Renata Prôa, Joselisa Q. Paiva, Rafael M. Loureiro, Wesley P. Calixto, Márcio R. C. Reis and Mara Giavina-Bianchi

- 87 **Artificial intelligence and skin cancer**
Maria L. Wei, Mikio Tada, Alexandra So and Rodrigo Torres
- 97 **Accuracy of an artificial intelligence as a medical device as part of a UK-based skin cancer teledermatology service**
Helen Marsden, Polychronis Kemos, Marcello Venzi, Mariana Noy, Shameera Maheswaran, Nicholas Francis, Christopher Hyde, Daniel Mullarkey, Dilraj Kalsi and Lucy Thomas



OPEN ACCESS

EDITED AND REVIEWED BY
Robert Gniadecki,
University of Alberta, Canada

*CORRESPONDENCE
Mara Giavina-Bianchi
✉ marahgbianchi@gmail.com

[†]These authors have contributed equally to this work

RECEIVED 19 April 2024
ACCEPTED 30 April 2024
PUBLISHED 08 May 2024

CITATION
Giavina-Bianchi M and Ko J (2024) Editorial:
Artificial intelligence in cutaneous lesions:
where do we stand and what is next?
Front. Med. 11:1420152.
doi: 10.3389/fmed.2024.1420152

COPYRIGHT
© 2024 Giavina-Bianchi and Ko. This is an
open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Editorial: Artificial intelligence in cutaneous lesions: where do we stand and what is next?

Mara Giavina-Bianchi^{1*†} and Justin Ko^{2†}

¹Medical Image Research Department, Hospital Israelita Albert Einstein, São Paulo, Brazil, ²Clinical Dermatology, School of Medicine, Stanford University, Stanford, CA, United States

KEYWORDS

artificial intelligence (AI), cutaneous diseases, dermatology, skin cancer, teledermatology, acne, patient survey, systematic literature review

Editorial on the Research Topic

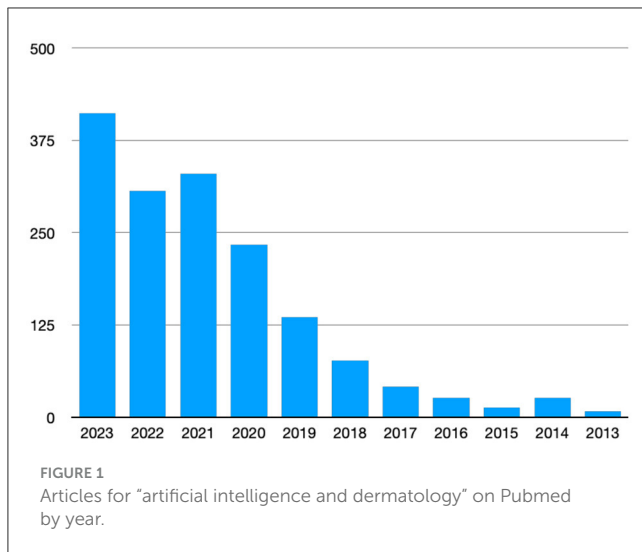
[Artificial intelligence in cutaneous lesions: where do we stand and what is next?](#)

We have seen, with great interest and enthusiasm, the continued growth in research output detailing the use of Artificial Intelligence (AI) in cutaneous diseases as can be seen in [Figure 1](#), as well as the maturation of content of the research bridging the gap from hype to reality; from pixels to practice (1).

The body of work spans a broad range, from skin cancer detection (2–4), inflammatory skin diseases (5, 6) surveys with dermatologists (7), patients perspectives (8), among others [(9); [Giavina-Bianchi et al.](#)]. While we are starting to see the initial glimpses of what clinical practice augmented and supported by AI capabilities might look like we do not yet have tools used regularly by dermatologists, other clinicians, or patients in daily practice. Why is this? Where do we stand now? What is next in this field? To try answer these questions, this special Research Topic solicited articles and resulted in 10 manuscripts from teams diverse in geographic representation as well as topic were accepted and published to shed light on these questions.

In setting the stage to answer the question “where are we now?”, [Furriel et al.](#) provided a systematic review of papers specifically on AI as applied to the detection, classification, and assessment of skin cancer images in the clinical setting. Their rigorous methodology identified 18 studies that encompassed a diversity of approaches in skin cancer detection, as well as significant differences in dataset size. They highlight the areas of convergence and divergence in the work and approaches to this topic, including more focused binary tools vs. broader approaches with multiclass output.

Two papers provide additional reflections on the state of the art as well as starting to answer the question “where are we going?”. [Omiye et al.](#) provided a broad overview of artificial intelligence (AI), as applied to dermatology with a primary focus on methodology, AI applications for various skin diseases, limitations, and future opportunities. They reviewed the current image-based models, highlighted the challenges facing widespread adoption and the future of AI in evolving the paradigm of large language, and multi-modal models.



Wei et al. discuss clinical applications including novel areas outside of visual assessment, as well as new methodological approaches like federated learning, multimodal learning, and new model architectures like vision transformers. The confluence of technological breakthroughs along with the breadth of clinical applications means that there will be opportunity for Research Topic on AI applied to dermatology for many years to come!

A set of four articles in the topic series focused on and highlight AI in real-world practice. They cover different aspects of pioneering endeavor in UK that is bringing these AI tools and capabilities into clinical practice with measurable benefit: from the model development to patient perceptions around the use of technology in aiding clinical decision-making. First, Marsden et al. had a goal to help improve the triage and management of suspicious skin lesions, using AI-based Digital Health Technology (DERM-003). This was a prospective, multi-center study that aimed to demonstrate the effectiveness of an AI as a Medical Device (AIaMD) to identify Squamous Cell Carcinoma, Basal Cell Carcinoma, pre-malignant and benign lesions from dermoscopic images of suspicious skin lesions. They found that the AIaMD AUROC varied from 0.85 to 0.89, demonstrating the potential to support the timely diagnosis of malignant and premalignant skin lesions.

Second, they aimed to implement the above AI solution, and safely reduce referral rates. Their objective was to demonstrate that the AIaMD had a higher rate of correctly classifying lesions that did not need to be referred for biopsy or urgent face-to-face dermatologist review, compared to teledermatology standard of care (SoC), maintaining the same sensitivity to detect malignancy. Their results showed a potential to reduce the burden of unnecessary referrals when used as part of a teledermatology service Marsden et al..

Third, patients recruited in this study were asked to complete an online questionnaire to evaluate their views regarding use of AIaMD in the skin cancer pathway by Kawsar et al. The majority of respondents felt confident in computers being used to help doctors diagnose and formulate management plans and as a support tool for general practitioners when assessing skin lesions and had no issues on their photographs being taken with a mobile phone device.

Lastly, Thomas et al. analyzed the real-world performance of the above medical device (AIaMD) tool for skin lesion assessment. They assessed the DERM deployment within skin cancer pathways at two National Health Service hospitals (UK) in 2 versions, which demonstrated very high sensitivity for detecting melanoma or malignancy, in-line with sensitivity targets and pre-marketing authorization research, reducing the caseload for hospital specialists.

The work of MB and team highlights an emerging important aspect of bringing AI capabilities into the real world—that of explainability and interaction with the clinician. This demonstrated the current state and variability between different models of saliency visualization that impacted clinician acceptance and preference. There is much to be done in the real of human/computer interface, and this work shows the nuance and importance of evaluating seemingly simple concepts like how we visualize and show data and information to clinicians (Giavina-Bianchi et al.).

Two additional papers represent progress and innovative approaches—Shavlokhova et al. explore the feasibility of leveraging advances in text-to-image generation capabilities in service of generating synthetic dermoscopic images of disease. While the results show that there is promise in preliminary aspects to this approach, it remains to be seen whether current state gaps in realism can be closed, and whether synthetic data may hold utility in supplementing or augmenting real data (Shavlokhova et al.).

Li et al. tackle a real world clinical use case of training and validating the ability of an algorithm to replicate human acne severity grading, demonstrating the utility of AI capabilities to use cases outside of skin lesion assessment and beyond classification/diagnosis tasks. The potential role for these efforts in creating efficiencies and fostering improved consistency in clinical assessment is on display, though begs the question of whether at this point clinician labeling as gold standard is the true gold standard (Li et al.).

This set of articles makes clear that we have traversed a significant distance from the initial hype around AI in dermatology toward an intimate understanding of what it takes to translate possibility to practice and patient impact.

Author contributions

MG-B: Writing—original draft, Writing—review & editing. JK: Writing—original draft, Writing—review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Narla A, Kuprel B, Sarin K, Novoa R, Ko J. Automated classification of skin lesions: from pixels to practice. *J Invest Dermatol.* (2018) 138:2108–10. doi: 10.1016/j.jid.2018.06.175
2. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human-computer collaboration for skin cancer recognition. *Nat Med.* (2020) 26:1229–34. doi: 10.1038/s41591-020-0942-0
3. 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) 542:115–8. doi: 10.1038/nature21056
4. 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) 29:1836–42. doi: 10.1093/annonc/mdy166
5. Han SS, Park I, Lim W, Kim MS, Park GH, Chae JB, et al. Augment intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *J Invest Dermatol.* (2020) 140:1753–61. doi: 10.1016/j.jid.2020.01.019
6. Han SS, Park GH, Lim W, Kim MS, Na JI, Park I, et al. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network. *PLoS ONE.* (2018) 13:e0191493. doi: 10.1371/journal.pone.0191493
7. Polesie S, Gillstedt M, Kittler H, Lallas A, Tschandl P, Zalaudek I, et al. Attitudes towards artificial intelligence within dermatology: an international online survey. *Br J Dermatol.* (2020) 183:159–61. doi: 10.1111/bjd.18875
8. Nelson CA, Pérez-Chada LM, Creadore A, Li SJ, Lo K, Manjaly P, et al. Patient perspectives on the use of artificial intelligence for skin cancer screening: a qualitative study. *JAMA Dermatol.* (2020). doi: 10.1001/jamadermatol.2019.5014
9. Wolf JA, Moreau JF, Akilov O, Patton T, English JC, Ho J, et al. Diagnostic inaccuracy of smartphone applications for melanoma detection. *JAMA Dermatol.* (2013) 149:422–6. doi: 10.1001/jamadermatol.2013.2382



OPEN ACCESS

EDITED BY

Andreas Recke,
University of Lübeck, Germany

REVIEWED BY

Jacob Furst,
DePaul University, United States
Antonio Neme,
National Autonomous University of Mexico,
Mexico

*CORRESPONDENCE

Mara Giavina-Bianchi
✉ marahgbianchi@gmail.com

RECEIVED 16 June 2023

ACCEPTED 14 August 2023

PUBLISHED 31 August 2023

CITATION

Giavina-Bianchi M, Vitor WG, Fornasiero de Paiva V, Okita AL, Sousa RM and Machado B (2023) Explainability agreement between dermatologists and five visual explanations techniques in deep neural networks for melanoma AI classification. *Front. Med.* 10:1241484. doi: 10.3389/fmed.2023.1241484

COPYRIGHT

© 2023 Giavina-Bianchi, Vitor, Fornasiero de Paiva, Okita, Sousa and Machado. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Explainability agreement between dermatologists and five visual explanations techniques in deep neural networks for melanoma AI classification

Mara Giavina-Bianchi*, William Gois Vitor,
Victor Fornasiero de Paiva, Aline Lissa Okita,
Raquel Machado Sousa and Birajara Machado

Department of Big Data, Hospital Israelita Albert Einstein, São Paulo, Brazil

Introduction: The use of deep convolutional neural networks for analyzing skin lesion images has shown promising results. The identification of skin cancer by faster and less expensive means can lead to an early diagnosis, saving lives and avoiding treatment costs. However, to implement this technology in a clinical context, it is important for specialists to understand why a certain model makes a prediction; it must be explainable. Explainability techniques can be used to highlight the patterns of interest for a prediction.

Methods: Our goal was to test five different techniques: Grad-CAM, Grad-CAM++, Score-CAM, Eigen-CAM, and LIME, to analyze the agreement rate between features highlighted by the visual explanation maps to 3 important clinical criteria for melanoma classification: asymmetry, border irregularity, and color heterogeneity (ABC rule) in 100 melanoma images. Two dermatologists scored the visual maps and the clinical images using a semi-quantitative scale, and the results were compared. They also ranked their preferable techniques.

Results: We found that the techniques had different agreement rates and acceptance. In the overall analysis, Grad-CAM showed the best total+partial agreement rate (93.6%), followed by LIME (89.8%), Grad-CAM++ (88.0%), Eigen-CAM (86.4%), and Score-CAM (84.6%). Dermatologists ranked their favorite options: Grad-CAM and Grad-CAM++, followed by Score-CAM, LIME, and Eigen-CAM.

Discussion: Saliency maps are one of the few methods that can be used for visual explanations. The evaluation of explainability with humans is ideal to assess the understanding and applicability of these methods. Our results demonstrated that there is a significant agreement between clinical features used by dermatologists to diagnose melanomas and visual explanation techniques, especially Grad-Cam.

KEYWORDS

melanoma, Grad-CAM, Grad-CAM++, Eigen-CAM, Score-CAM, LIME, explainability

1. Introduction

Melanoma is a skin cancer that is more lethal than all the other skin cancers combined, even though it accounts for less than 5% of all cases (1). The global incidence of melanoma rose from 11.8 to 17.5/100,000 inhabitants from 2003–2006 to 2011–2014 (2, 3). In Australia, one of the countries with the highest incidence of this pathology in the world, the number of deaths from melanoma of the skin increased from 596 in 1982 to 1,405 in 2019 (4). In 2021, in the U.S.A., 106,110 cases were diagnosed and 7,180 deaths by melanoma were estimated (5).

Melanoma represents a high cost to society. Loss of productivity due to morbidity or premature death, as well as the cost of treatments, are a considerable burden for health systems and have multiple implications in the life of such individuals (6). It is ranked as one of the most expensive cancers, with a significant decrease in cost when diagnosed in the early stages (7, 8). The average cost per patient with melanoma ranges from € 149 for disease stage 0 to € 66,950 for stage IV (9). When melanoma is diagnosed early, it can be treated effectively and with a high probability of survival (5). Therefore, it is essential to promote prevention programs with periodic examination of the skin for the early detection of suspicious lesions to reduce the costs and mortality of melanoma (6). The ABCDE rule is a widely used method to recognize characteristics often associated with melanoma. It was developed for both physicians and patients. It includes: Asymmetry, Border irregularity, Color heterogeneity, Diameter larger than 6mm, and Evolution or transformation of the lesion over time (10).

Since the detection of melanomas at an early stage is essential for a good prognosis, and the distinction between melanomas and harmless pigmented lesions is often not trivial, AI-based classification systems may bring important contributions to this field. Artificial intelligence algorithms have performed *in silico* at least as well as expert dermatologists in detecting melanoma lesions (11–13). Results have been encouraging, but there are only a few recent studies trying to use AI in the real world to detect melanoma lesions (14–16). There is still some controversy about the use of AI for diagnoses in “real-life” clinical settings. Concerns include the possibility of biases, the lack of transparency and explainability, scalability, data integration and interoperability, reliability, safety, privacy, and the ethics of aggregated digital data (17, 18). As with any other innovation, especially in healthcare, AI must prove to be efficient, reliable, reproducible, and friendly enough to be accepted by those who are actually going to use it; in this case, physicians (or perhaps other health professionals) and patients. As for physicians, a recent study in Korea has shown that, in general, physicians have a positive attitude toward AI in medicine (19). Another study has presented similar results in a large international survey among dermatologists, indicating that AI is well-accepted in the dermatology field and that AI should be a part of medical training (20). As for patients, one article concluded that they expressed a high level of confidence in decision-making by AI and that AI can contribute to improving diagnostic accuracy, but should not replace the dermatologist (21). Another survey has shown that patients and physicians are willing to use AI in the detection of melanoma lesions. Patients appear to be receptive to the use of AI for skin cancer screening if implemented in a manner that preserves the integrity of the human physician-patient relationship (22).

To satisfy the requirement for transparent and comprehensible treatment decisions, it will be necessary to work on strategies that allow AI results to be interpreted and verified (at least in part). Due to the high complexity of the algorithms, complete transparency of AI will probably not be possible. Still, it may be possible to explain the decisive influencing factors on individual decision steps within the algorithms. Explainable artificial intelligence (XAI) is an initiative that aims to “produce more explainable models while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially

intelligent partners” (23). The aim of enabling explainability in ML, as stated by FAT (fairness, accountability, and transparency) (24), “is to ensure that algorithmic decisions, as well as any data driving those decisions, can be explained to end-users and other stakeholders in non-technical terms”.

For deep learning models, the challenge of ensuring explicability is due to the trade-off in terms of powerful results and predictions (25) and the inherent opacity of black box models. This represents a serious disadvantage, as it prevents a human being from being able to verify, interpret and understand the system’s reasoning and how decisions are made (26). It is a common approach to understand the decisions of image classification systems by finding regions of an image that were particularly influential to the final classification. They are called sensitivity maps, saliency maps, or pixel attribution maps (27). These approaches use occlusion techniques or calculations with gradients to assign an “importance” value to individual pixels which are meant to reflect their influence on the final classification.

Gradient-weighted Class Activation Mapping (Grad-CAM) uses the gradients of any target concept flowing into the final convolutional layer to produce a coarse localization map highlighting important regions in the image for predicting the concept. It highlights pixels that the trained network deems relevant for the final classification (28). Grad-CAM computes the gradient of the class-score (called logit) with respect to the feature map of the final convolutional layer (28). Despite the difficulty of evaluating interpretability methods, some proposals have been made in this direction (29, 30). Grad-CAM is one method of local interpretability being used for deep learning models and was one of the few methods that passed the recommended sanity checks (29). There is also an improved version of the original Grad-CAM and CAM method, called Grad-CAM++. This method is based on the same principles as the original Grad-CAM method, but it uses a different weighted combination (31). Two other CAM techniques can be used: Eigen-CAM (32) and Score-CAM (33) which differ from the Grad-CAM by not relying on the backpropagation of gradients. A totally different approach can also be made using Local Interpretable Model-agnostic Explanations (LIME) technique, where the image is segmented into superpixels interconnected with similar colors (34).

To elucidate more about the explainability of deep neural network classification in melanoma lesions, we performed an exploratory experiment with 2 objectives. First, to assess the agreement rate between the features highlighted by 5 different techniques of visual saliency maps to the three most used clinical dermatological criteria for melanoma lesions: asymmetry, border irregularity, and color heterogeneity (ABC rule). Second, to subjectively evaluate the preferable techniques ranked by the dermatologists, the reasons for it and the degree of agreement between the two dermatologists about the five techniques.

2. Methodology

In this section, we will introduce the dataset used to build the classification model for evaluating the visual explanations, the Convolutional Neural Network (CNN) models used for the segmentation and classification tasks, the explainability methods

used for the visual explanations, and the experiment performed. The development of the algorithm and its performance were described in detail in a previous article (35).

This study was approved by Hospital Israelita Albert Einstein Ethics Committees under the identification CAAE:32903120.40000.0071.1 and it is in accordance with the ethical standards on human experimentation and with the Declaration of Helsinki. Dermatologists that took part in the experiment signed consent forms agreeing to participate. This research was performed at Hospital Albert Einstein, São Paulo, Brazil, from January–March 2023.

2.1. Melanoma dataset

For this study, we used the following datasets: HAM10000 Dataset (36), MSK Dataset (37), Dataset BCN20000 (38), and Derm7pt (39), all publicly available. The first three datasets compose the dermoscopic image data available by ISIC (37–39), an international competition for the identification of skin diseases. Derm7pt is composed of clinical and dermoscopic images categorized by the 7-point technique for the identification of melanoma, with more than 2000 images of melanoma and non-melanoma. In this study, we selected only dermoscopic images. The total dataset consists of 26,342 images. Only two different classes were established for our dataset: melanoma (18%) and non-melanoma (82%).

2.2. Convolutional neural networks models (CNN)

The classification model for melanoma lesions was constructed using two steps: image segmentation and image classification. For the segmentation, we used the MaskR-CNN architecture (40). The lesions in the dermoscopy images were segmented and then used in the classification model in a way that the latter could focus only on the patterns closely related to the lesion itself, excluding most of the background information that could impair its classification capabilities. To train the segmentation model, we used 2000 images previously annotated by specialists with the regions of interest. Using transfer learning with a Resnet50 backbone and 20 epochs, the trained model reached a 99.69% mAP for our test set.

For the classification task, we divided the total dataset as 80% for training, 10% for validation, and 10% for testing the classification model. To train the model, we used the EfficientNetB6 convolutional neural network (41). This family of architectures achieved some of the best precision and efficiency in the literature (41), performing better than previous CNN (42, 43). Through transfer learning with pre-trained weights from the ImageNet (44), the model was fine-tuned for 50 epochs using the Adam optimization (45) with a 0.001 starting learning rate and a batch size equal to 32. The learning rate was scheduled to be reduced by a factor of 30% if the model failed to improve with a stagnant validation loss for 5 epochs. Finally, we used early stopping, also based on a validation loss of 10 epochs.

To address the imbalance in the two target classes, we trained the model using the focal loss function (46) to avoid bias for the most dominant class. We also weighted the classes according to their inverse frequency, in order to balance model attention in the loss function. All images were resized to 220×220 . In addition, we applied data augmentation using common image processing operations (rotation, shear, horizontal flip, zoom). The sigmoid function was used to deliver the prediction result. In the tests, our model has achieved an average ACC of 0.81, AUC of 0.94, sensitivity of 0.93 and specificity of 0.79, considering the threshold of 0.5. More details of the model can be found in our study previously reported (35).

2.3. Explainability methods adopted

2.3.1. Gradient-weighted class activation mapping (Grad-CAM)

Grad-CAM was proposed to produce visual explanations for decision-making in comprehensive classes of convolutional neural networks (28). The idea was to make AI models transparent and explainable, giving the possibility to identify flaws in the systems, mainly of deep learning models that were considered difficult to interpret. Some proposals have used Grad-CAM in an attempt to explain possible decisions of the model (47) in the medical field (48–51).

Since Grad-CAM does not require any particular CNN architecture, it can be used with fixed weights (after being trained), and it is able to explore the spatial information of the last convolutional layers through feature maps that are weighted and calculated, based on gradients. The positive values, which are the most “relevant” information for the classification result, can be obtained through a ReLU operation, defined as,

$$L_{Grad-CAM}^c = ReLU \left(\sum_k \alpha_k^c A^k \right) \quad (1)$$

$$\text{where } \alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}.$$

2.3.2. Grad-CAM++

Grad-CAM++ technique is an improved version of the original Grad-CAM and CAM method. The Grad-CAM++ method is based on the same principles as the original Grad-CAM method, but it uses a weighted combination of the positive partial derivatives of the last convolutional layer feature maps with respect to a specific class score as weights to generate a visual explanation for the class label under consideration (Equation2) (31).

$$L_{Grad-CAM}^c = ReLU \left(\sum_k \alpha_k^c A^k \right) \quad (2)$$

The class-discriminative saliency map generated by Grad-CAM++ is a high-resolution heatmap that indicates the regions of the input image that are most relevant to the specific prediction made by the network. For a given image, L_c is calculated as a linear

combination of the forward activation maps, followed by a relu layer (Equation 3) (31).

$$L_{ij}^c = \text{ReLU} \left(\sum_k w_k^c A_{ij}^k \right) \quad (3)$$

2.3.3. Eigen-CAM

The Eigen-CAM technique leverages the principal components on the activation maps of the convolutional layers (32). It does not rely on the backpropagation of gradients. For the last convolutional layer:

1. Singular value decomposition (SVD) is used to factorize the combined activation map A for input X as $A = U \sum V^t$;
2. The activation map is then projected on the first eigenvector of the V matrix;
3. The projection highlights the principal components of the activation map.

In this method, there is no use of a ReLU activation function. Conceptually, the Eigen-CAM can be defined as,

$$L_{\text{Eigen-CAM}} = AV_1 \quad (4)$$

where V_1 denotes the first the eigenvector at the first position in the V matrix.

2.3.4. Score-CAM

Like Eigen-CAM, Score-CAM does not rely on the backpropagation of gradients. It borrows from the Grad-CAM technique in the sense that it is also non-dependent on a particular architecture; where they differentiate, however, is in the way they deal with the flow of gradient information. Instead of using the gradient from the last convolutional layer to build on the importance of each region of input X toward class C , the Score-CAM technique assimilates the importance of each region as an increase of confidence in the overall prediction (33). For a specific convolutional layer:

1. Each activation map is upsampled, normalized, and then used as a mask for input X , highlighting the most activated regions;
2. The masked input image is passed through the CNN resulting in a logit for each class;
3. All logits and activation maps are linearly combined;
4. A ReLU activation function is applied to the combined product, resulting in the Score-CAM output.

Because gradients can be noisy, explode, and/or vanish (52), these characteristics can also be present in the layer activations (53), thus resulting in suboptimal CAM visualizations. The Score-CAM technique, however, is not dependent on the model gradient.

Conceptually, the Score-CAM can be defined as,

$$L_{\text{Score-CAM}}^k = \text{ReLU} \left(\sum_k \alpha_k^c A_l^k \right) \quad (5)$$

where $\alpha_k^c = C \left(A_l^k \right)$, and $C \left(A_l^k \right) = f \left(X \cdot H_l^k \right) - f(X_b)$.

2.3.5. Local interpretable model-agnostic explanations (LIME)

LIME is model agnostic, which allows it to be utilized across a wide range of machine learning models. The locally weighted square loss (\mathcal{L}) as the metric choice by authors (Equation 6). This loss function takes into account the exponential kernel $\kappa(z)$, which is defined as $\exp(-D(x, z)^2/\sigma^2)$, where D represents a distance function, such as the cosine distance for text or the L_2 distance for images, and σ is the width of the kernel (54).

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) (f(z) - g(z'))^2 \quad (6)$$

How LIME is used for image:

1. The image is segmented into superpixels. Superpixels are interconnected pixels with similar colors;
2. The surrogate model highlights the superpixels of the image that are the most active in predicting a certain class;
3. The image is transformed into a binary vector where 1 indicates the original superpixel and 0 indicates a grayed-out super-pixel.

The complexity depends on the time required to compute the prediction of the relevant class and the number of samples N . Due to this complexity, LIME may take longer than other methods, especially when applied to image data (34, 54). In the present publication, the LIME is used to highlight superpixels that have the maximum positive and negative influence on the model's prediction.

2.4. The experiment

In order to analyze the impact of the five different explainability techniques on humans, we defined two major questions to be addressed experimentally. They are:

1. Is there a quantitative agreement between dermatologists ABC rule and the visual explanation techniques for melanoma?
2. Do dermatologists qualitatively agree with the visual explanation techniques for melanoma?

In the next sections, we will explore each question in further detail.

2.4.1. Is there a quantitative agreement between dermatologists ABC rule and the visual explanation techniques for melanoma?

In this experiment, we aimed to apply an explainability method visual analysis by human experts, such as dermatologists, comparing the highlighted areas in the saliency maps with the areas of the lesion that show asymmetry, border irregularity, and color heterogeneity (ABC rule), three of the main features evaluated in a melanoma lesion.

From the dataset, we selected 100 lesions correctly classified by the model as melanoma. These 100 dermoscopy images were analyzed by two experienced and Board-Certified dermatologists (MGB and ALO). They first assessed only the dermoscopy image and graded three of the five most frequently melanoma criteria (ABCDE) used in clinical practice: asymmetry (A), border

irregularity (B), and color heterogeneity (C). They did not grade diameter (D) because most of the dermoscopy images had no scale measure and evolution in time (E) due to the fact that the clinical photographs in the dataset were taken at one point in time and no follow-up images were available.

Both dermatologists had to reach a consensus to use a semi-quantitative scale from 0 to 2 to grade the ABC features in the lesions, as shown in Figure 1. To assess asymmetry, the lesion was divided into 4 quadrants, and its shape and color distribution was analyzed. If all 4 quadrants had regular shapes and colors, there was no asymmetry (0); if 2 or 3 quadrants were similar, there was mild asymmetry (1); and if all four quadrants were different, there was severe asymmetry (2). For borders, they evaluated the shape and regularity. If the aspect was smooth and regular in color, the borders were considered benign (0). If $\leq 50\%$ of the border area presented irregular borders or signs of color abnormality, it was called partial involvement (1), and if $> 50\%$, severe involvement (2). If $> 50\%$ of the lesion's limits could not be evaluated, they were designed as non-available (N/A). For color, we assessed the degree of color heterogeneity by the number of colors present in the lesion: one color present, no heterogeneity (0); two colors present, mild heterogeneity (1); three or more colors present, severe heterogeneity (2).

Next, they analyzed each visual explanation technique (Grad-CAM, Grad-CAM ++, Eigen-CAM, Score-CAM, and LIME) in conjunction with its dermoscopy image, separately, in pairs, and blindly to the techniques name. For each of them, they assessed the features highlighted by the saliency map, using the following criteria (Figure 2). For asymmetry, it was the same criteria as for clinical features. The visual explanation map was divided into 4 quadrants and shape and color distribution were analyzed. If all 4 quadrants showed the same color and format, there is no asymmetry (0); if 2 or 3 quadrants are similar, there was mild asymmetry (1); and if all four quadrants were different, there is severe asymmetry (2). The clinical border area was compared to the highlighted visual map for borders. If the visual technique showed no highlight or $\leq 50\%$ of the border area highlighted with cold colors for the clinical borders, it was classified as no highlight (0). If $\leq 50\%$ of the area was highlighted with heat colors or $> 50\%$ with cold colors, it was called partial border highlight (1). If $> 50\%$ of the area were highlighted with heat colors, it was designated as total border highlight (2) or non-available (N/A), and if $> 50\%$ of lesion's limits could not be evaluated clinically.

For color assessment, we had to pursue a different strategy, mainly because visual heat maps, by definition, ought to display multiple colors, leaving all the maps to be rated as showing severe heterogeneity of colors (2), which would not be meaningful to the dermatologists understanding. Thus, dermatologists decided to compare the most significant color abnormalities presented in the dermoscopy image (as if they had a saliency map in their minds) to the heat colors of the visual map, considering its location and intensity, and grading the match between them. If the clinical color abnormalities presented an agreement area was $\leq 75\%$ for heat colors, it was called total agreement (0). If the matched area was 25–75% for heat colors or $> 75\%$ for cold colors, it was designated as partial agreement (1). If the matched area for heat colors was $< 25\%$ or 25–75% for cold colors, it was considered total disagreement (2). For grading the highlight colors, we established blue/purple as cold

colors and orange/red for heat colors. Examples of high and low agreement cases can be seen in Figure 3.

To calculate the agreement rate between the clinical criteria and visual techniques, we used the following criteria: if the difference between their grade scales was zero, they were in total agreement. If the difference was one, they had a partial agreement and if the difference was two, they had no agreement. For example, if dermatologists graded the heterogeneity of colors as 0 in the clinical image and as 0 in the visual technique, the difference was zero, so they were in total agreement. On the other hand, if dermatologists graded border irregularity as 2 for the clinical image and as 0 for the visual explanation technique, the difference was 2, and therefore there was no agreement. At last, if the asymmetry was rated as 0 for the clinical image and as 1 for the explanation technique, the difference was 1, so that corresponded to a partial agreement.

2.4.2. Do dermatologists qualitatively agree with the visual explanation techniques for melanoma?

The rationale for this part of the qualitative study was to capture the overall characteristics perceived by the experts about each explainability technique, making comments about each of them and ranking their preferable techniques. For this purpose, after grading ABC, we showed all the images again, with the respective label for each technique to both dermatologists and asked them to make comments about each technique and how they would rank the techniques in order of the most preferable to the least (1–5). After that, they were also asked to read the comments and determine if they agree or not with the other experts observations, according to the following criteria: total agreement; partial agreement; no agreement nor disagreement; partial disagreement; and total disagreement. Examples of clinical melanoma images and their respective visual maps using Score-CAM, Eigen-CAM, LIME, Grad-CAM, and Grad-CAM ++ can be seen in Figure 4.

3. Results

3.1. Quantitative results

To assess the AB clinical criteria for melanoma in our study, a confusion matrix was constructed after grading melanoma images, as depicted in Figure 5. The diagonal of the matrix signifies instances where the reference and dermatologists concurred, indicating total agreement. The off-diagonal elements, displaced either one or two columns away from the main diagonal, denote partial agreement or disagreement, respectively. The generated confusion matrix was used to construct (Table 1), presenting a comprehensive overview of the inter-rater reliability of the AB clinical criteria for melanoma in our study.

Table 1 shows the results of total, partial, and no agreement rates to ABC melanoma rule. Asymmetry was the criterium of the highest agreement rate among the three. LIME, Grad-CAM, and Grad-CAM++ were the top techniques for asymmetry, all of them showing $> 50\%$ of total agreement rates. 40–50% of all techniques showed a partial agreement rate in this criterium. Eigen-CAM had the poorest performance, with $> 25\%$ of no agreement rate, while

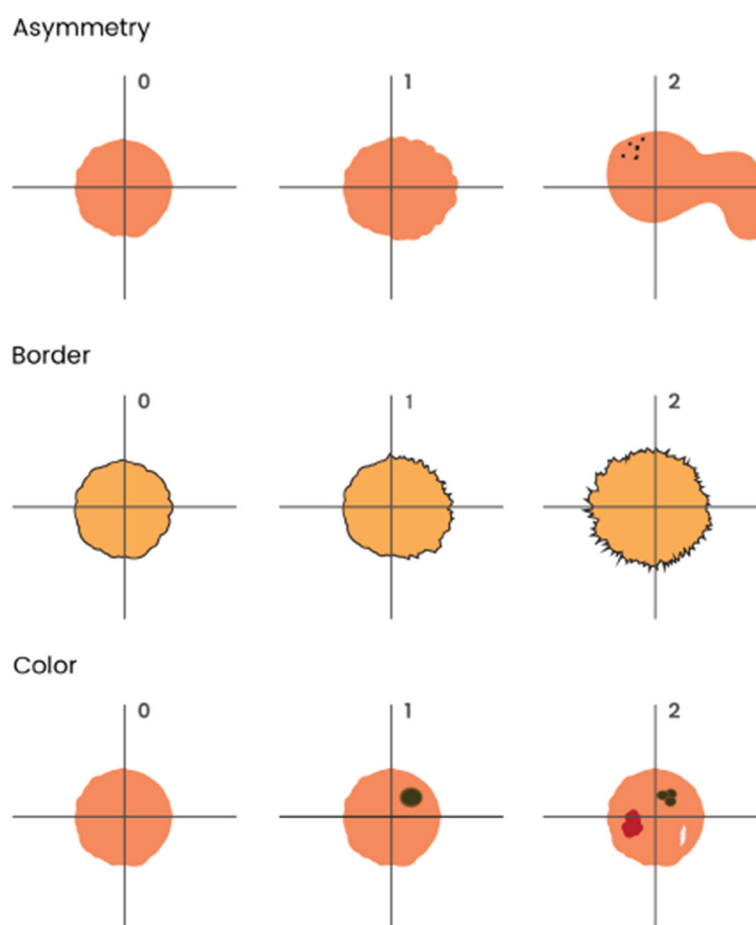


FIGURE 1

Graphical representation of ABC melanoma criteria used in clinical images: asymmetry, border irregularity, and color heterogeneity. To assess asymmetry, the lesion was divided into 4 quadrants, and its shape and color distribution were analyzed. If all 4 quadrants had regular shapes and colors, there was no asymmetry (0); if 2 or 3 quadrants were similar, there was mild asymmetry (1); and if all four quadrants were different, there was severe asymmetry (2). For borders, they evaluated shape and regularity. If the aspect was smooth and regular in color, the borders were considered benign (0); if $\leq 50\%$ of the border area presented irregular borders or signs of color abnormality, it was considered as partial involvement (1), and if $> 50\%$, severe involvement (2). Finally, if $> 50\%$ of the lesion's limits could not be evaluated, it was considered non-available (N/A). For color, we assessed the degree of color heterogeneity by the number of colors present in the lesion: presence of one color was considered as no heterogeneity (0); presence of two colors was considered as mild heterogeneity (1); presence of three or more colors was considered as severe heterogeneity (2).

Grad-CAM++ and LIME showed only around 3% of no agreement. Thus, Grad-CAM++ seems to be the best technique for asymmetry detection in melanoma cases.

Regarding border evaluation, all visual explanation techniques showed similar total agreement rates, between 32 and 39%, but Score-CAM and Grad-CAM++ showed no agreement in $\geq 20\%$ of the cases. For partial agreement, Grad-CAM and Eigen-CAM showed the best numbers. Taking all into account, it looks like Grad-CAM is the most reliable technique to identify border abnormalities by visual maps.

As for the color match, Grad-CAM presented the top performance, with 40% of total agreement, followed by Grad-CAM++ and LIME. For partial agreement, all techniques showed similar results. As Grad-CAM had only 6% of no agreement, it was considered the best technique for this aspect.

Analyzing the three criteria together, Grad-CAM was the best visual explanation technique in agreement with the ABC rule of melanoma cases. In second and third places,

respectively, are LIME and Grad-CAM++, which performed very similarly in this experiment. Eigen-CAM and Score-CAM finalized in the fourth and fifth places, respectively, Eigen-CAM presenting a little better result for total and no agreement rates.

3.2. Qualitative results

Comments of both dermatologists about the five different visual explanation methods can be seen in Table 2, as well as their preferable choices, and their inter-expert agreement rates. Grad-CAM and Grad-CAM++ were in the top position for both. Score-CAM was unanimous the third place in choice and the worst positions were occupied by LIME and Eigen-CAM techniques. The overall inter-expert agreement rates was 60% total and 40% partial, although they were not coincident for each explainability method. There were no disagreements.

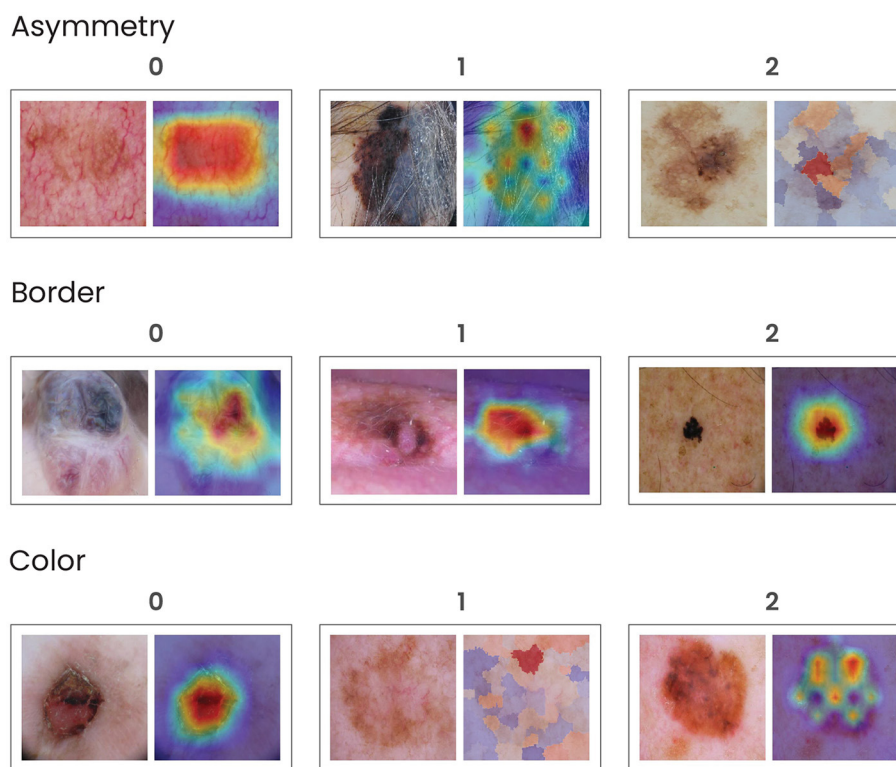


FIGURE 2

Grading examples of the visual map explanation techniques. For asymmetry, the visual explanation map was divided into 4 quadrants and shape and color distribution were analyzed. If all four quadrants showed the same color and format, there was no asymmetry (0); if 2 or 3 quadrants were similar, there was mild asymmetry (1); and if all four quadrants were different, there was severe asymmetry (2). For borders, the clinical border area was compared to the highlighted visual map. If the visual technique showed no highlight or $\leq 50\%$ of the border area highlighted with cold colors, it was considered as no highlight (0). If $\leq 50\%$ of the area was highlighted with warm colors or $> 50\%$ with cold colors, it was considered partial border highlight (1); if $> 50\%$ of the areas was highlighted with warm colors, it was considered total border highlight (2). Finally, if $> 50\%$ of the lesion's limits could not be evaluated clinically, it was considered non-available (N/A). For color abnormality, dermatologists decided to compare the most significant color abnormalities in the dermatoscopy image as if they had a saliency map in their minds, comparing the imaginary heatmaps to the ones in the visual techniques. If the clinical color abnormalities presented an agreement area of $\leq 75\%$ for warm colors, it was considered total agreement (0); if it was 25–75% for warm colors or $> 75\%$ for cold colors, it was considered as partial agreement (1); if it was $< 25\%$ for warm colors or 25–75% for cold colors, it was considered total disagreement (2). For grading the highlight colors, we established blue/purple as cold colors and orange/red as warm colors.

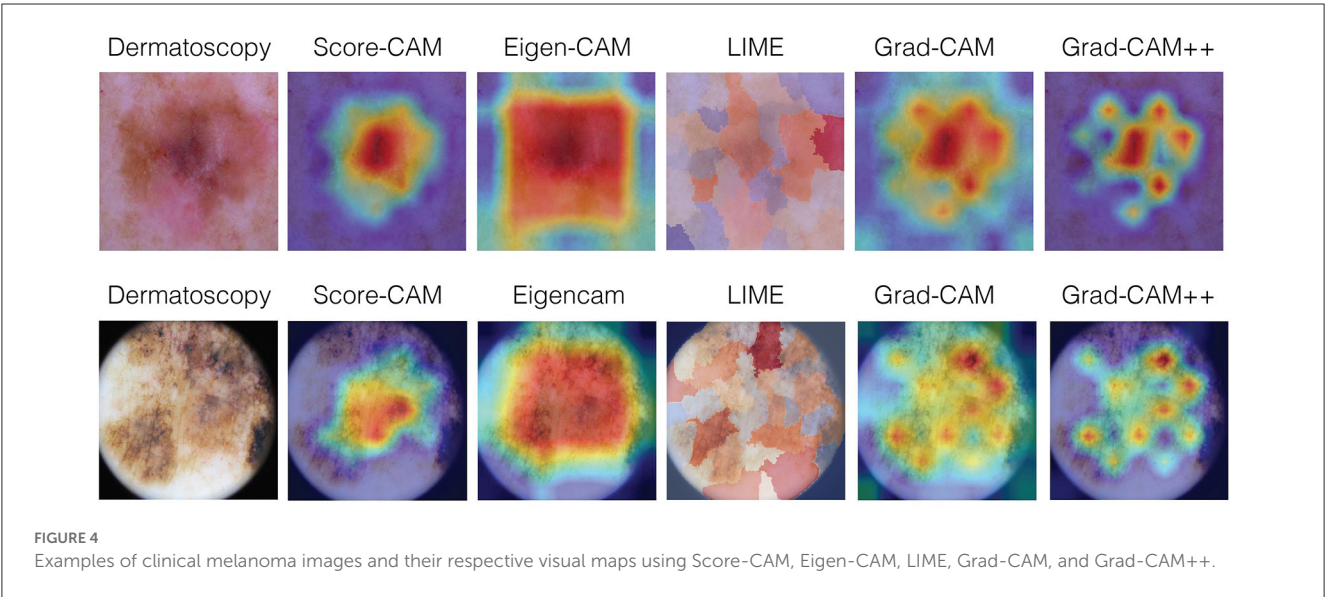
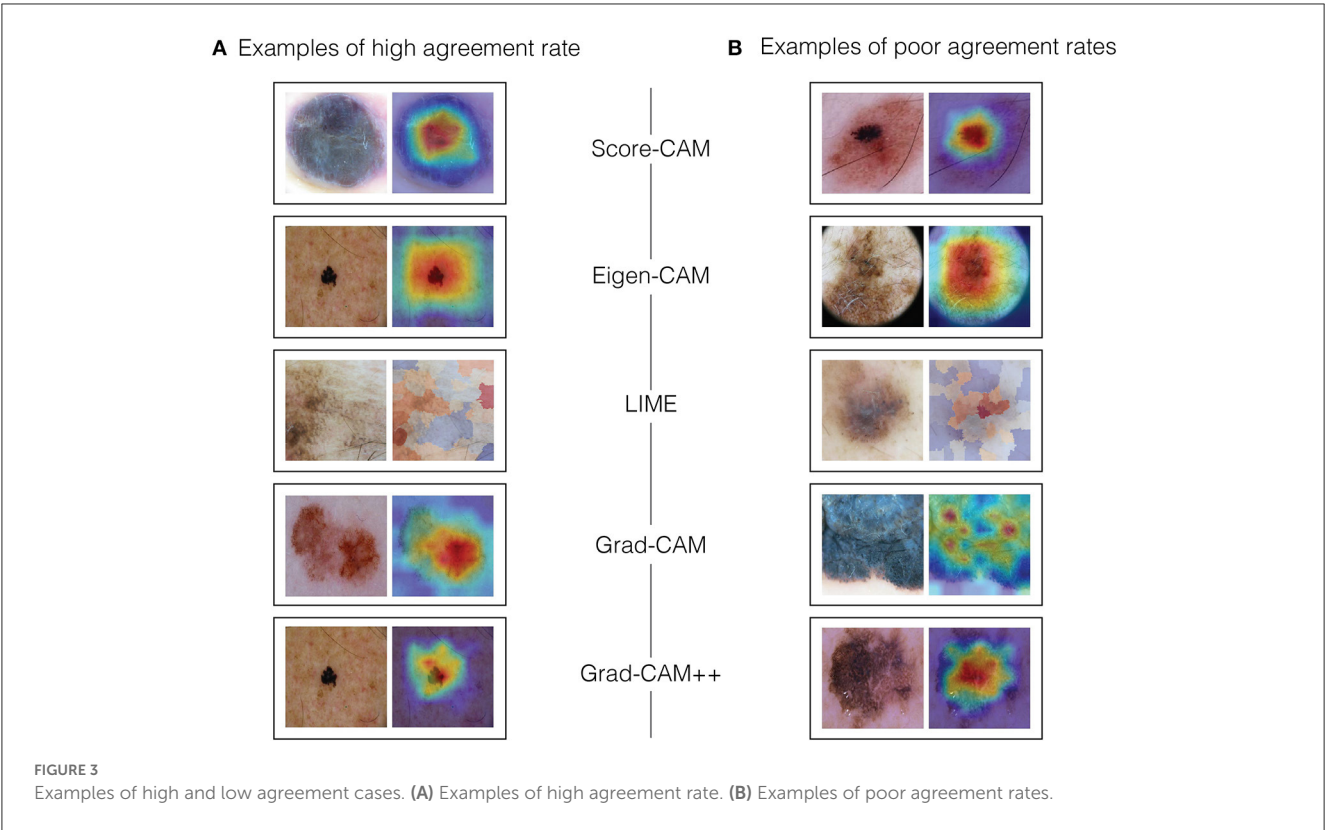
4. Discussion

Due to the difficulty of interpreting deep learning models and giving a plausible explanation for a prediction, this theme has been increasingly addressed in the literature through proposed methods, taxonomies, and benchmarks (29, 30, 55, 56). However, there is little consensus on what is interpretability/explainability in machine learning and how to evaluate it for benchmarking (55). Especially in the medical field, as physicians play a major role in endorsing (or not) the use of AI algorithms, it is important to reach out to them, understanding how and what they think about the explainability models. An adequate visual explanation should be able to identify details that help explain a particular classification (26). In this context, interpretability can be described as the degree to which a human can consistently predict the models result (25, 35).

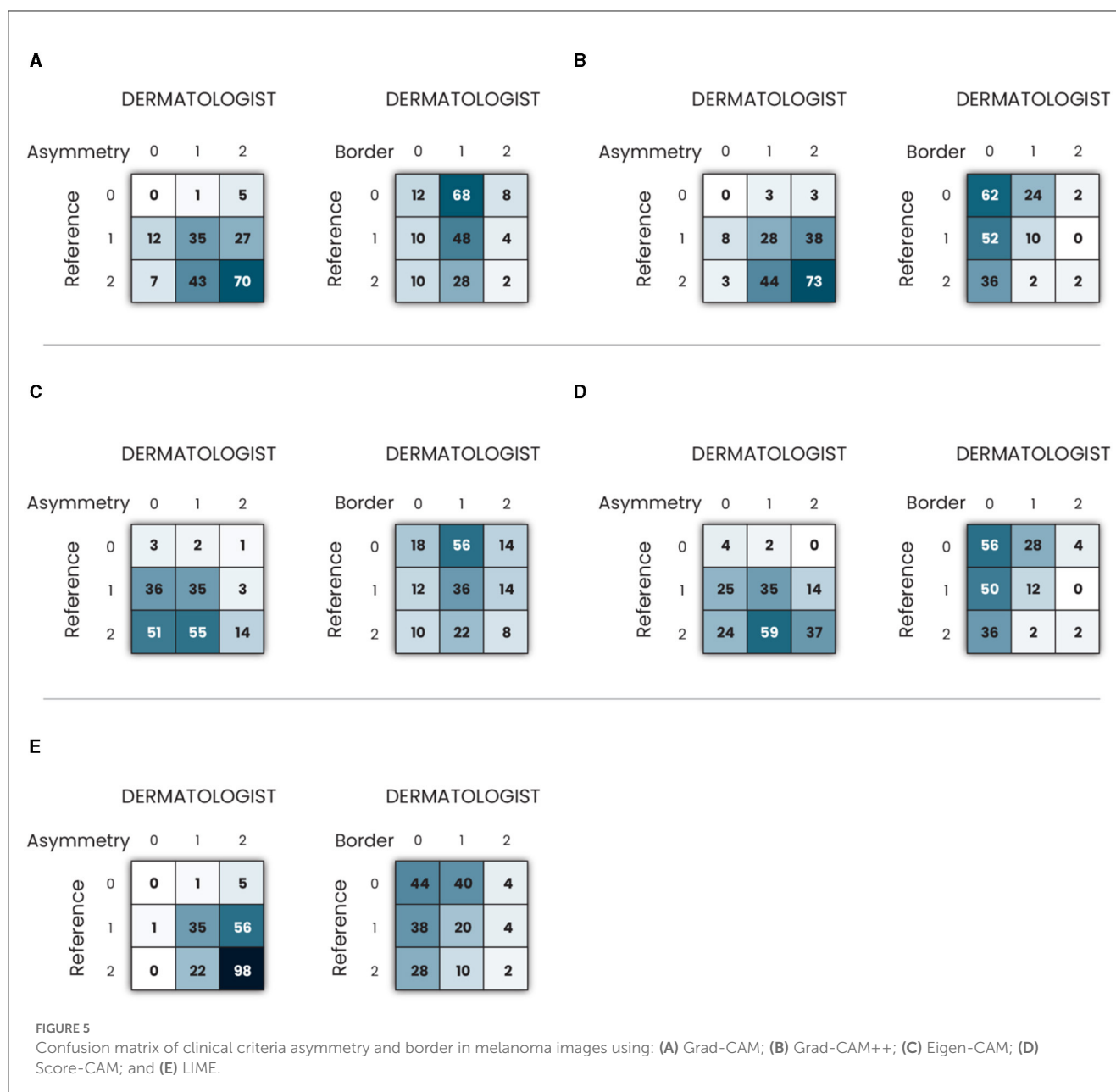
There are very few studies addressing this question in practice. Our work is likely one of the pioneers in this field, trying to bring light to the CNN black box, through practical

experiments using human experts in the field of Dermatology. Our methodology tested the discriminative visual explanation of five different techniques to support the understanding of the model's decision and our quantitative and qualitative results composed an interesting picture to compare the methods in a real-life situation.

Asymmetry was the criterium with the highest agreement rate, reaching 57.5% using LIME. This can be explained because the LIME technique is very geographical, dividing the maps lesion into several different areas and color tones, making it almost impossible to produce a symmetric visual map. As melanoma clinical lesions are often asymmetric themselves, the high agreement may be more of an expression of this fact rather than a true match with the dermatologists criterium. On the other hand, Eigen-CAM had the worst performance, justified by the fact that it often stamps a rectangle over the entire lesion, showing no asymmetry at all, poorly reflecting the reality of the clinical lesion. Grad-CAM and Grad-CAM++ also performed very well for asymmetry, with only $\leq 6\%$ of no agreement rate and excellent numbers for high and partial agreements rate.



Borders evaluation was the criterium with the lowest agreement rate. Grad-CAM showed the best results, with only 9% of no agreement rate, followed by Eigen-CAM. That corroborates the fact that Grad-CAM was the only technique cited as better limiting the border area. Eigen-CAM might have a good result in this assessment because, as said, the rectangle displayed in the visual map included, in most cases, the border area. As described above for LIME technique in asymmetry evaluation, Eigen-CAM may not reflect a true match with the border area, but only a coincidence dependent on the techniques visual map displayed. The worst performance techniques were Score-CAM and Grad-CAM++ was showing $\geq 20\%$ of no agreement rates, which was also pointed out by the dermatologists.



Color abnormalities assessment is probably the most relevant criterion when dermatologists evaluate lesions such as melanoma. Eigen-CAM and Grad-CAM presented the best results, over 30% of high agreement and $\leq 4\%$ of no agreement. As already mentioned, Eigen-CAM, as its visual map prints a big rectangle over the lesion, it did match the color abnormalities, but indiscriminately, as pointed out by the dermatologists. Thus, for this criterium, when the qualitative study is considered, Grad-CAM seemed to better match the relevant areas of color abnormalities of the lesions. LIME and Score-CAM did poorly in this evaluation, showing only around 15% of high agreement and 12–13% of no agreement.

Overall, Grad-CAM showed the best agreement rate with 40% of total agreement and only 6% of no agreement. This was also reflected by the dermatologists opinion, which ranked it in the top two techniques. The LIME technique ended

up in the second position in the quantitative study, probably because of the high performance for asymmetry, but was ranked very low by the dermatologists, in the last two spots. Grad-CAM++ turned up to be third in quantitative agreement, but it was highly ranked by the dermatologists (first and second places). Eigen-CAM performed fourth in the agreement experiment and it was disliked, as well, by the experts. Finally, Score-CAM showed the worst performance in the quantitative assessment, but it assumed a unanimous third place among the dermatologists, only after Grad-CAM and Grad-CAM++.

Another study, recently published, tested four Convolutional Neural Network models using five different interpretation techniques (saliency, guided backpropagation, integrated gradients, input gradients, and DeepLIFT)

TABLE 1 Agreement between clinical ABC melanoma features and each visual explanation.

Technique	Total agreement		Partial agreement		No agreement		Total
Assymetry							
Eigen-CAM	52	26.00%	96	48.00%	52	26.00%	200
Grad-CAM	105	52.50%	83	41.50%	12	6.00%	200
Grad-CAM++	101	50.50%	93	46.50%	6	3.00%	200
LIME	115	57.50%	80	40.00%	5	2.50%	200
Score-CAM	76	38.00%	100	50.00%	24	12.00%	200
Border							
Eigen-CAM	62	32.63%	104	54.74%	24	12.63%	190
Grad-CAM	62	32.63%	110	57.89%	18	9.47%	190
Grad-CAM++	74	38.95%	78	41.05%	38	20.00%	190
LIME	66	34.74%	92	48.42%	32	6.84%	190
Score-CAM	70	36.84%	80	42.11%	40	21.05%	190
Color							
Eigen-CAM	75	37.50%	121	60.50%	4	2.00%	200
Grad-CAM	69	34.50%	123	61.50%	8	4.00%	200
Grad-CAM++	41	20.50%	132	66.00%	27	13.50%	200
LIME	29	14.50%	148	74.00%	23	11.50%	200
Score-CAM	32	16.00%	141	70.50%	27	13.50%	200
TOTAL							
Eigen-CAM	189	32.03%	321	54.41%	80	13.56%	590
Grad-CAM	236	40.00%	316	53.56%	38	6.44%	590
Grad-CAM++	216	36.61%	303	51.36%	71	12.03%	590
LIME	210	35.59%	320	54.24%	60	10.17%	590
Score-CAM	178	30.17%	321	54.41%	91	15.42%	590

to compare their agreement with experts previous annotations of esophagus cancerous tissue, showing that saliency attributes match best with the manual experts delineations and that there was moderate to high correlation between the sensitivity of a model and the human-and-computeragreement (57).

Saliency maps are one of the few methods that can be used for visual explanations. As in our study, the evaluation of explainability with humans is ideal to assess the understanding and applicability of these methods (55). A large variety of methods have been applied for this aim. However, recent work has shown that many are, in fact, independent of the model weights and/or the class labels. In these cases, it is likely that the model architecture itself is constraining the saliency maps to look falsely meaningful: frequently, the maps just act as a variant of an edge detector. This is particularly dangerous in the context of skin cancer detection, as features at the borders of lesions are often considered diagnostic for melanoma: saliency maps that highlight the edges of a lesion may be misconstrued as clinically meaningful (51). Interestingly, our results in the experiment showed that most of the techniques fail to identify

the borders of the lesions, and only Grad-CAM showed a good performance.

Although human evaluation is essential to assess interpretability, the evaluation of the human subject is not an easy task (55). In our experiment, it is not possible to measure, in a concrete way, if the techniques are looking at the same features as the experts to confirm or not the agreement. Some studies claimed that people tend to disregard information that is inconsistent with their prior beliefs. This effect is called confirmation bias (25) and that is why our dermatologists assessed the dermoscopic images and Grad-CAM visual maps separately and blindly, trying to avoid it. Also, relying only on examples to explain the models behavior can lead to over-generalization and misunderstanding (58), and observing where the network is looking at the image does not tell the user what the CNN is actually doing with that part of the image (59).

Furthermore, when evaluating the most appropriate explanation, one must take into account the social environment of the ML system and the target audience. This means that the best explanation varies depending on the domain of the application and the use case (60). Despite the fact that a

TABLE 2 Qualitative results of each visual map technique showing the comments, ranking and inter-expert agreement.

Visual map technique	Dermatologist 1			Dermatologist 2		
	Comments	Preference ranking	Inter-expert agreement	Comments	Preference ranking	Inter-expert agreement
Score-CAM	Poor delimitation of the lesion, very specific, but very low sensitivity	3	Total	It points only to specific areas, but not necessarily the relevant ones	3	Partial
Eigen-CAM	It creates a rectangle over the central area; does not seem specific nor sensitive	4	Total	It maps a great area, without differentiation between relevant areas; it only points to the lesion	5	Total
LIME	It creates geographical areas, hard to interpret; it can delimitate the lesion very well, but does not seem specific or sensitive	5	Total	Maps do not explain why clinically similar areas of the skin show different patterns in the map; does not seem sensitive or specific	4	Total
Grad-CAM	It delimitates the lesion most accurately, and have better match to clinically relevant areas	1	Partial	It seems more specific, but not so much sensitivity; it points correctly to the whole lesion	2	Partial
Grad-CAM++	It does not delimitate the lesion; it highlights only the major relevant areas; high specificity and low sensitivity	2	Total	It also seems more specific, localizing the relevant areas but less sensitive; it points only to parts of the lesion, not delimitating the whole area	1	Total

saliency map located on the lesion cannot yet be viewed as justification that clinically meaningful correlations have been learned, a map that is clearly located on a clinically irrelevant region could be used to signal a prediction that should be ignored (51).

In our study, we encouraged experts to provide quantitative and qualitative analyses of the different explainability techniques to assess subjective matters related to how they visually interpreted melanoma lesions alongside the technique's results. By doing that, we touched unknown territory in terms of analyzing how useful these visual explainability techniques can be in clinical practice. In our study design, the experts gave important feedback that was statically detailed and explored. There was no adoption of a method described in the scientific literature because it was not possible to find one. In the future, it may be pertinent to carefully explore and propose study designs to address this issue, preferably exploring subjective matters objectively, minimizing model and expert biases, and focusing on the real-world gains of adopting AI algorithms in clinical practice.

5. Conclusion

Our work is likely one of the pioneers using experts to try to bring light to the CNN black box in the Dermatology area,

performing quantitative and qualitative studies on different visual explanation techniques for melanoma. Our results demonstrated that there is a significant agreement between clinical features used by dermatologists to diagnose melanomas and visual explanation techniques, especially Grad-Cam. The interpretation of black-box generalization in melanoma images based on visual maps showed up to be promising, presenting trustworthy outputs compared to experts interpretations and encouraging new studies.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Hospital Israelita Albert Einstein. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

MG-B, WV, and VF had the idea, designed the experiments, wrote, and reviewed the final manuscript. MG-B and AO performed the experiments and reviewed the final manuscript. RS reviewed the literature and developed the CNN. BM overviewed the entire process, was responsible for accessing the funding, and reviewed the final manuscript.

Funding

This study was supported by Brazilian Ministry of Health, process number: 25000.121674/2018-13.

References

- Najita JS, Swetter SM, Geller AC, Gershenwald JE, Zelen M, Lee SJ. Sex differences in age at primary melanoma diagnosis in a population-based analysis (US Surveillance, Epidemiology, and end results, 2005-2011). *J Invest Dermatol.* (2016) 136:1894. doi: 10.1016/j.jid.2016.03.044
- Steglich RB, Cardoso S, Gaertner MHdCN, Coelho KMdPA, Cestari TF, Franco SC. Differences in the diagnosis of primary cutaneous melanoma in the public and private healthcare systems in Joinville, Santa Catarina State, Brazil. *Anais brasileiros de dermatologia.* (2018) 93:507–12. doi: 10.1590/abd1806-4841.20185767
- Steglich RB, Coelho KMdPA, Cardoso S, Gaertner MHdCN, Cestari TF, Franco SC. Epidemiological and histopathological aspects of primary cutaneous melanoma in residents of Joinville, 2003-2014. *Anais brasileiros de dermatologia.* (2018) 93:45–53. doi: 10.1590/abd1806-4841.20185497
- Melanoma of the Skin Statistics. Available online at: <https://www.cancer australia.gov.au/cancer-types/melanoma/statistics> (accessed 29 May, 2023).
- Skin Cancer Facts and Statistics. Available online at: <https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/#melanoma> (accessed 19 May, 2023).
- Krensel M, Schäfer I, Augustin M. Cost-of-illness of melanoma in Europe: a modelling approach. *J Eur Acad Dermatol Venereol.* (2019) 33:34–45. doi: 10.1111/jdv.15308
- Alexandrescu DT. Melanoma costs: a dynamic model comparing estimated overall costs of various clinical stages. *Dermatol Online J.* (2009) 15:11. doi: 10.5070/D353F8Q915
- Guy Jr GP, Ekwueme DU, Tangka FK, Richardson LC. Melanoma treatment costs: a systematic review of the literature, 1990-2011. *Am J Prev Med.* (2012) 43:537–45. doi: 10.1016/j.amepre.2012.07.031
- Buja A, Sartor G, Scioni M, Vecchiato A, Bolzan M, Rebba V, et al. Estimation of direct melanoma-related costs by disease stage and by phase of diagnosis and treatment according to clinical guidelines. *Acta Derm Venereol.* (2018) 98:218–24. doi: 10.2340/00015555-2830
- Ward WH, Farma JM. *Cutaneous Melanoma: Etiology and Therapy*. Brisbane, QLD: Codon Publications (2017).
- Esteve 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) 542:115–8. doi: 10.1038/nature21056
- 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) 29:1836–42. doi: 10.1093/annonc/mdy166
- Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *Lancet Oncol.* (2019) 20:938–47. doi: 10.1016/S1470-2045(19)30333-X
- Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human-computer collaboration for skin cancer recognition. *Nat Med.* (2020) 26:1229–34. doi: 10.1038/s41591-020-0942-0
- Han SS, Park I, Chang SE, Lim W, Kim MS, Park GH, et al. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *J Investigat Dermatol.* (2020) 140:1753–61. doi: 10.1016/j.jid.2020.01.019
- Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med.* (2020) 26:900–8. doi: 10.1038/s41591-020-0842-3
- Lau AY, Staccini P, et al. Artificial intelligence in health: new opportunities, challenges, and practical implications. *Yearb Med Inform.* (2019) 28:174–8. doi: 10.1055/s-0039-1677935
- Cath C. Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philos Trans Royal Soc.* (2018) 376:20180080. doi: 10.1098/rsta.2018.0080
- Oh S, Kim JH, Choi SW, Lee HJ, Hong J, Kwon SH. Physician confidence in artificial intelligence: an online mobile survey. *J Med Internet Res.* (2019) 21:e12422. doi: 10.2196/12422
- Polesie S, Gillstedt M, Kittler H, Lallas A, Tschandl P, Zalaudek I, et al. Attitudes towards artificial intelligence within dermatology: an international online survey. *Br J Dermatol.* (2020) 183:159–61. doi: 10.1111/bjd.18875
- Jutzi TB, Kriehoff-Henning EI, Holland-Letz T, Utikal JS, Hauschild A, Schadendorf D, et al. Artificial intelligence in skin cancer diagnostics: the patients' perspective. *Front Med.* (2020) 7:233. doi: 10.3389/fmed.2020.00233
- Nelson CA, Pérez-Chada LM, Creadore A, Li SJ, Lo K, Manjaly P, et al. Patient perspectives on the use of artificial intelligence for skin cancer screening: a qualitative study. *JAMA Dermatol.* (2020) 156:501–12. doi: 10.1001/jamadermatol.2019.5014
- Explainable Artificial Intelligence. Available online at: <http://www.darpa.mil/program/explainable-artificialintelligence> (accessed 29 May, 2023).
- Fairness, Accountability, and Transparency in Machine Learning. Available online at: <https://www.fatml.org/> (accessed 19 May, 2023).
- Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: a survey on methods and metrics. *Electronics.* (2019) 8:832. doi: 10.3390/electronics8080832
- Montavon G, Lapuschkin S, Binder A, Samek W, Müller KR. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern recognition.* (2017) 65:211–22. doi: 10.1016/j.patcog.2016.11.008
- Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. Smoothgrad: removing noise by adding noise. *arXiv.* (2017). doi: 10.48550/arXiv.1706.03825
- Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*. Venice: IEEE. (2017) p. 618–626. doi: 10.1109/ICCV.2017.74
- Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. In: *Advances in Neural Information Processing Systems (NeurIPS 2018)*. Montréal, QC (2018). p. 31.
- Hooker S, Erhan D, Kindermans PJ, Kim B. A benchmark for interpretability methods in deep neural networks. In: *Advances in Neural Information Processing Systems (NeurIPS 2019)*. Vancouver, BC (2019). p. 32.
- Chattopadhyay A, Sarkar A, Howlader P, Balasubramanian VN. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Lake Tahoe, NV: IEEE. (2018) p. 839–847. doi: 10.1109/WACV.2018.00097

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

32. Muhammad MB, Yeasin M. Eigen-cam: Class activation map using principal components. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. Glasgow: IEEE. (2020). p. 1–7. doi: 10.1109/IJCNN48605.2020.9206626
33. Wang H, Wang Z, Du M, Yang F, Zhang Z, Ding S, et al. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. Seattle, WA: IEEE. (2020). p. 2425. doi: 10.1109/CVPRW50498.2020.00020
34. Ribeiro MT, Singh S, Guestrin C. “Why should I trust you?” Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY: Association for Computing Machinery (2016). p. 1135–1144. doi: 10.1145/2939672.2939778
35. Giavina-Bianchi M, de Sousa RM, Paciello VZdA, Vitor WG, Okita AL, Prôa R, et al. Implementation of artificial intelligence algorithms for melanoma screening in a primary care setting. *PLoS ONE*. (2021) 16:e0257006. doi: 10.1371/journal.pone.0257006
36. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*. (2018) 5:1–9. doi: 10.1038/sdata.2018.161
37. Codella NC, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza SW, et al. Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. Washington, DC: IEEE. (2018) p. 168–172. doi: 10.1109/ISBI.2018.8363547
38. Combalia M, Codella N, Rotemberg V, Helba B, Vilaplana V, Reiter O, et al. BCN20000: Dermoscopic lesions in the wild. *arXiv*. (2019). doi: 10.48550/arXiv.1908.02288
39. Kawahara J, Daneshvar S, Argenziano G, Hamarneh G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE*. (2018) 23:538–46. doi: 10.1109/JBHI.2018.2824327
40. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: *Proceedings of the IEEE International Conference on Computer Vision*. Venice: IEEE. (2017) p. 2961–2969. doi: 10.1109/ICCV.2017.322
41. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: *International Conference on Machine Learning*. New York: PMLR (2019). p. 6105–6114.
42. Iandola F, Moskewicz M, Karayev S, Girshick R, Darrell T, Keutzer K. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv*. (2014). doi: 10.48550/arXiv.1404.1869
43. Xia X, Xu C, Nan B. Inception-v3 for flower classification. In: *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*. Chengdu: IEEE. (2017) p. 783–787.
44. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. *Int J Comput Vis*. (2015) 115:211–52. doi: 10.1007/s11263-015-0816-y
45. Kingma D, Ba J. Adam: A method for stochastic optimization. Published as a conference paper at ICLR (2015). *arXiv*. (2015). doi: 10.48550/arXiv.1412.6980
46. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*. Venice: IEEE. (2017) p. 2980–2988. doi: 10.1109/ICCV.2017.324
47. Woo S, Park J, Lee JY, Kweon IS. Cbam: Convolutional block attention module. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. (2018). p. 3–19. doi: 10.48550/arXiv.1807.06521
48. Kim I, Rajaraman S, Antani S. Visual interpretation of convolutional neural network predictions in classifying medical image modalities. *Diagnostics*. (2019) 9:38. doi: 10.3390/diagnostics9020038
49. Yang C, Rangarajan A, Ranka S. Visual explanations from deep 3D convolutional neural networks for Alzheimers disease classification. In: *AMIA Annual Symposium Proceedings*. Bethesda, MD: American Medical Informatics Association. (2018) p. 1571.
50. Iizuka T, Fukasawa M, Kameyama M. Deep-learning-based imaging-classification identified cingulate island sign in dementia with Lewy bodies. *Sci Rep*. (2019) 9:8944. doi: 10.1038/s41598-019-45415-5
51. Young K, Booth G, Simpson B, Dutton R, Shrapnel S. Deep neural network or dermatologist? In: *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support: Second International Workshop, iMIMIC 2019 and 9th International Workshop, ML-CDS 2019 Held in Conjunction with MICCAI 2019, China, October 17, 2019 Proceedings 9*. Cham: Springer. (2019) p. 4855.
52. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans Neural Netw*. (1994) 5:157–66. doi: 10.1109/72.279181
53. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv*. (2013). doi: 10.48550/arXiv.1312.6034
54. Garreau D, Luxburg U. Explaining the explainer: A first theoretical analysis of LIME. In: *International Conference on Artificial Intelligence and Statistics*. PMLR (2020). p. 1287–1296.
55. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv*. (2017).
56. Nie W, Zhang Y, Patel A. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In: *International Conference on Machine Learning*. New York: PMLR. (2018). p. 3809–3818.
57. de Souza Jr LA, Mendel R, Strasser S, Ebigbo A, Probst A, Messmann H, et al. Convolutional Neural Networks for the evaluation of cancer in Barrett’s esophagus: Explainable AI to lighten up the black-box. *Comput Biol Med*. (2021) 135:104578. doi: 10.1016/j.combiomed.2021.104578
58. Kim B, Khanna R, Koyejo OO. Examples are not enough, learn to criticize! criticism for interpretability. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R, editors. *Advances in Neural Information Processing Systems (NIPS 2016)*. (2016). p. 29.
59. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intel*. (2019) 1:206–15. doi: 10.1038/s42256-019-0048-x
60. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV: IEEE. (2016) p. 770778.



OPEN ACCESS

EDITED BY

Mara Giavina-Bianchi,
Albert Einstein Israelite Hospital, Brazil

REVIEWED BY

Giusto Trevisan,
University of Trieste, Italy
Darius Mehregan,
Wayne State University, United States

*CORRESPONDENCE

Helen Marsden
✉ helen@skinanalytics.co.uk

RECEIVED 04 September 2023

ACCEPTED 18 September 2023

PUBLISHED 06 October 2023

CITATION

Marsden H, Morgan C, Austin S, DeGiovanni C,
Venzi M, Kemos P, Greenhalgh J,
Mullarkey D and Palamaras I (2023)
Effectiveness of an image analyzing AI-based
Digital Health Technology to identify
Non-Melanoma Skin Cancer and other skin
lesions: results of the DERM-003 study.
Front. Med. 10:1288521.
doi: 10.3389/fmed.2023.1288521

COPYRIGHT

© 2023 Marsden, Morgan, Austin, DeGiovanni,
Venzi, Kemos, Greenhalgh, Mullarkey and
Palamaras. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in this
journal is cited, in accordance with accepted
academic practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Effectiveness of an image analyzing AI-based Digital Health Technology to identify Non-Melanoma Skin Cancer and other skin lesions: results of the DERM-003 study

Helen Marsden^{1*}, Caroline Morgan², Stephanie Austin²,
Claudia DeGiovanni³, Marcello Venzi¹, Polychronis Kemos¹,
Jack Greenhalgh¹, Dan Mullarkey¹ and Ioulia Palamaras⁴

¹Skin Analytics Ltd., London, United Kingdom, ²Dermatology Unit, University Hospitals Dorset, Poole Hospital, Poole, United Kingdom, ³Dermatology Unit, University Hospitals Sussex NHS Foundation Trust, Brighton, United Kingdom, ⁴Department of Dermatology, Barnet and Chase Farm Hospitals, Royal Free London NHS Foundation Trust, London, United Kingdom

Introduction: Identification of skin cancer by an Artificial Intelligence (AI)-based Digital Health Technology could help improve the triage and management of suspicious skin lesions.

Methods: The DERM-003 study (NCT04116983) was a prospective, multi-center, single-arm, masked study that aimed to demonstrate the effectiveness of an AI as a Medical Device (AlaMD) to identify Squamous Cell Carcinoma (SCC), Basal Cell Carcinoma (BCC), pre-malignant and benign lesions from dermoscopic images of suspicious skin lesions. Suspicious skin lesions that were suitable for photography were photographed with 3 smartphone cameras (iPhone 6S, iPhone 11, Samsung 10) with a DL1 dermoscopic lens attachment. Dermatologists provided clinical diagnoses and histopathology results were obtained for biopsied lesions. Each image was assessed by the AlaMD and the output compared to the ground truth diagnosis.

Results: 572 patients (49.5% female, mean age 68.5 years, 96.9% Fitzpatrick skin types I–III) were recruited from 4 UK NHS Trusts, providing images of 611 suspicious lesions. 395 (64.6%) lesions were biopsied; 47 (11%) were diagnosed as SCC and 184 (44%) as BCC. The AlaMD AUROC on images taken by iPhone 6S was 0.88 (95% CI: 0.83–0.93) for SCC and 0.87 (95% CI: 0.84–0.91) for BCC. For Samsung 10 the AUROCs were 0.85 (95% CI: 0.79–0.90) and 0.87 (95% CI: 0.83–0.90), and for the iPhone 11 they were 0.88 (95% CI: 0.84–0.93) and 0.89 (95% CI: 0.86–0.92) for SCC and BCC, respectively. Using pre-determined diagnostic thresholds on images taken on the iPhone 6S the AlaMD achieved a sensitivity and specificity of 98% (95% CI, 88–100%) and 38% (95% CI, 33–44%) for SCC; and 94% (95% CI, 90–97%) and 28% (95% CI, 21–35%) for BCC. All 16 lesions diagnosed as melanoma in the study were correctly classified by the AlaMD.

Discussion: The AlaMD has the potential to support the timely diagnosis of malignant and premalignant skin lesions.

KEYWORDS

skin cancer, Artificial Intelligence, Digital Health Technology, skin lesions, smartphone cameras

1. Introduction

Non-Melanoma Skin Cancer (NMSC) is the fifth most common form of all types of cancer worldwide, with the most common NMSC types being Basal Cell Carcinoma (BCC), accounting for 75% of cases, and Squamous Cell Carcinoma (SCC), accounting for 23% of NMSC cases (1). In the UK, there are around 156,000 NMSC cases diagnosed, resulting in 920 deaths, *per annum*. The actual incidence of NMSC may be higher however, as it is known to be under-reported due to the number of multiple diagnoses per patient. Incidence rates of skin cancer have increased by over 2.5-fold (169%) since the early 1990s and are projected to rise by 14% in the UK between 2023 and 2025 (2). While NMSCs make up most of skin cancer diagnoses, melanoma has a much higher mortality rate due to high risk of metastasis, and early diagnosis is critical. When melanoma is caught early, the chances of survival are greatly improved (3).

Currently, diagnosis of NMSC is usually clinical, with subsequent histological confirmation following excision and specialist interpretation (4). To facilitate early diagnosis, alongside managing patient concern, a high proportion of 'suspicious moles' are referred from primary care on the two-week wait pathway, which has seen an increase from 332-thousand referrals in 2015/16 to 509-thousand referral in 2019/20 (5). However, a high proportion of these lesions are benign (6) with the main diagnoses being melanocytic naevi or seborrheic keratosis. Due to the nature of these referrals, they are awarded an inappropriate priority at the expense of more serious disorders. As a result, healthcare services are under pressure with the number of patients being referred for specialist evaluation, onward biopsies and subsequent management of suspicious skin lesions, such that a decreasing percentage of patients referred on a two-week wait pathway are seen within 14 days (5). There is a need to improve diagnostic accuracy of skin lesions earlier on in this process, in order to minimize unnecessary referrals and skin biopsies.

Deep Ensemble for the Recognition of Malignancy (DERM) is a Digital Health Technology that includes an Artificial Intelligence as a Medical Device (AIaMD) algorithm that is able to analyze dermoscopic images of a skin lesion and determine the presence of melanoma in pigmented lesions, with a similar accuracy to clinicians specialized in skin cancer detection (7). The AIaMD has been trained and tested on dermoscopic images of skin lesions with confirmed diagnoses of a range of malignant and non-malignant lesions and sub-types. This helps ensure that, for example, melanoma lesions with different clinical appearance like amelanotic melanoma (8), would be classified as melanoma. However, the AIaMD would not be expected to identify skin cancer from different image types, such as that from reflectance confocal microscopy. The AIaMD is also able to detect BCC and SCC, premalignant and selected benign lesions [such as Intraepidermal Carcinoma (IEC/SCC *in situ*), actinic keratosis, seborrheic keratosis, and benign melanocytic nevi] providing additional information to aid the clinician in differentiating skin cancers, including melanoma, from benign conditions. The AIaMD provides a high degree of accuracy in the diagnosis of NMSC using historical dermoscopic images, but clinical validation is necessary to demonstrate its utility in clinical practice. DERM is a Class IIa UKCA marked medical device and has been deployed in clinical pathways within the UK since 2020.

2. Materials and methods

The DERM-003 study was a prospective, multi-center, single-arm, cross-sectional, blinded study (NCT04116983), designed to demonstrate the effectiveness of the AIaMD to identify SCC and BCC. Secondary objectives included demonstrating the effectiveness of the AIaMD to identify premalignant and benign conditions, comparing the AIaMD performance to dermatologists, and demonstrating the feasibility of image capture in a clinic setting. Ethical approval for the study was granted by the Leicester South National Research Ethics committee.

Eligible participants were patients attending dermatology clinics with at least one suspicious skin lesion that was suitable for photographing. Lesions were defined as suspicious by a dermatologist, with no requirement on lesions being of a particular type or pigmentation. Patients provided written informed consent for the study. Recruitment was on a consecutive, competitive recruitment basis in 4 UK hospitals between June 2020 and February 2022. Lesions needed to be less than 15 mm in diameter, not located on an anatomical site unsuitable for photographing (genitals, hair-bearing areas, under nails) or in an area of visible scarring or tattooing, and not previously biopsied, excized or otherwise traumatized. Suitable lesions were photographed by three smartphones (iPhone 6S, iPhone 11 Apple Inc., Samsung Galaxy S10) with (dermoscopic image) or without (macroscopic image) a DermLite DL1 Basic (DermLite LLC) lens attached, providing a 10x magnification. In addition, one dermoscopic image of healthy skin was also taken by each camera. The AIaMD assessment was not shared with the investigator, who managed the patient in accordance with standard of care. The patient had completed the protocol-defined procedures once the photographs had been taken. For each lesion included in the study, a clinical diagnosis and the clinician's assessment of the likelihood of skin cancer, using a four-point Likert scale (unlikely, equivocal, likely, highly likely), was collected. Where a biopsy was taken, the histopathology-confirmed diagnosis was collected and categorized as melanoma, SCC, BCC, IEC, Actinic Keratosis (AK), Atypical, Benign or other. When there was histopathological uncertainty in the diagnosis, investigators reported the most likely diagnosis. 'Other' diagnoses were reviewed by the Chief Investigator.

Images of skin lesions were captured electronically and securely transferred to DERM for analysis by the AIaMD. All images were analyzed by DERM v3 after the completion of the study. The AIaMD generates a numeric output (continuous scale) for each of the examined classes, which reflects its confidence that the lesion is that condition. The sum of the numeric output of all classes is always 1. Threshold settings are defined for each lesion type, above which a lesion is classified as that lesion type. The AIaMD returns the most serious lesion type where the confidence score is above the threshold setting.

2.1. Statistical aspects

Patients and lesions that did not meet the inclusion criteria were excluded from the Intention To Treat population (ITT), as were those lesions without a final diagnosis available. Lesions with no AIaMD result available (missing dermoscopic images, and/or where these failed the DERM v3 image quality assessment) were excluded from the

Per Protocol (PP) population. The primary analyses were conducted on biopsied lesions in the PP population only.

Area Under the Receiver Operator Characteristic (AUROC) curves were used to examine the association of the algorithm's confidence scale with the histopathology-confirmed diagnosis (biopsied lesions) or clinical diagnosis (non-biopsied lesions). The co-primary outcome measures of the study were the one-against-all AUROC for both SCC and BCC. The iPhone 6S camera was used as the reference device. The study aimed to demonstrate both co-primary endpoints were above 0.9.

Assuming the true AUROC curve of the AIaMD is 0.98 and an incidence rate of 11% for SCC and 43% for BCC, a sample size of 45 SCC and 50 BCC lesions was required to demonstrate the AUROCs were superior to 0.9 at $\alpha=0.05$, with 90% power. A sample size of 543 patients, with an average of 1.2 lesions per patient, was expected to provide sufficient numbers of lesions diagnosed as SCC and BCC, but recruitment remained open until 45 SCC lesions had been included in the study.

Diagnostic accuracy indices (sensitivity, specificity, predictive values, false-positive rates, and false-negative rates) were calculated using decision thresholds determined prior to the image analysis, and applying the hierarchy within the AIaMD. The hierarchy means that, if the AIaMD identifies a lesion as potentially either a BCC or melanoma, it will return the classification of melanoma. Therefore, for a lesion diagnosed as SCC, an output from the AIaMD of "suspected melanoma" is considered a true positive, whereas for a lesion diagnosed as melanoma, an output from the AIaMD of "suspected SCC" is a false negative. The definition of true positive will therefore vary depending on the lesion type being assessed. The likelihood assessment scale was used to calculate a clinician AUROC that could be compared to the AIaMD.

The influence of patient and lesion variables that may affect the AIaMD's accuracy were investigated. The following co-variables were examined: age, sex, Fitzpatrick skin type, skin cancer risk factors including past medical history of skin cancer, lesion body location, experience of reviewing clinician, lesion change, patient's level of concern, clinician's assessment of likelihood of skin cancer, malignancy sub-type and staging.

A p -value of <0.05 was regarded as statistically significant, and all tests were two-tailed. Statistical estimates of accuracy are reported with 95% Confidence Intervals (CIs). Statistical analysis was conducted using R language version 4.1.3 (The R Project for Statistical Computing).

3. Results

A total of 572 patients consented to the study, providing 611 suspicious lesions. Nine patients (6 lesions) were withdrawn / excluded from the study. Eighteen lesions were excluded from the ITT population due to failing to meet eligibility criteria, resulting in 18 patients being excluded due to no eligible lesions. Two further lesions were excluded from the PP population due to missing AIaMD results, resulting in 1 further patient being excluded from the PP population (Figure 1). Of the lesions included in the PP population, 96.7% had images available from all three combinations of hardware, 2.9% had 2 images available, and 2 lesions had just one image available. Nine images failed image quality checks.

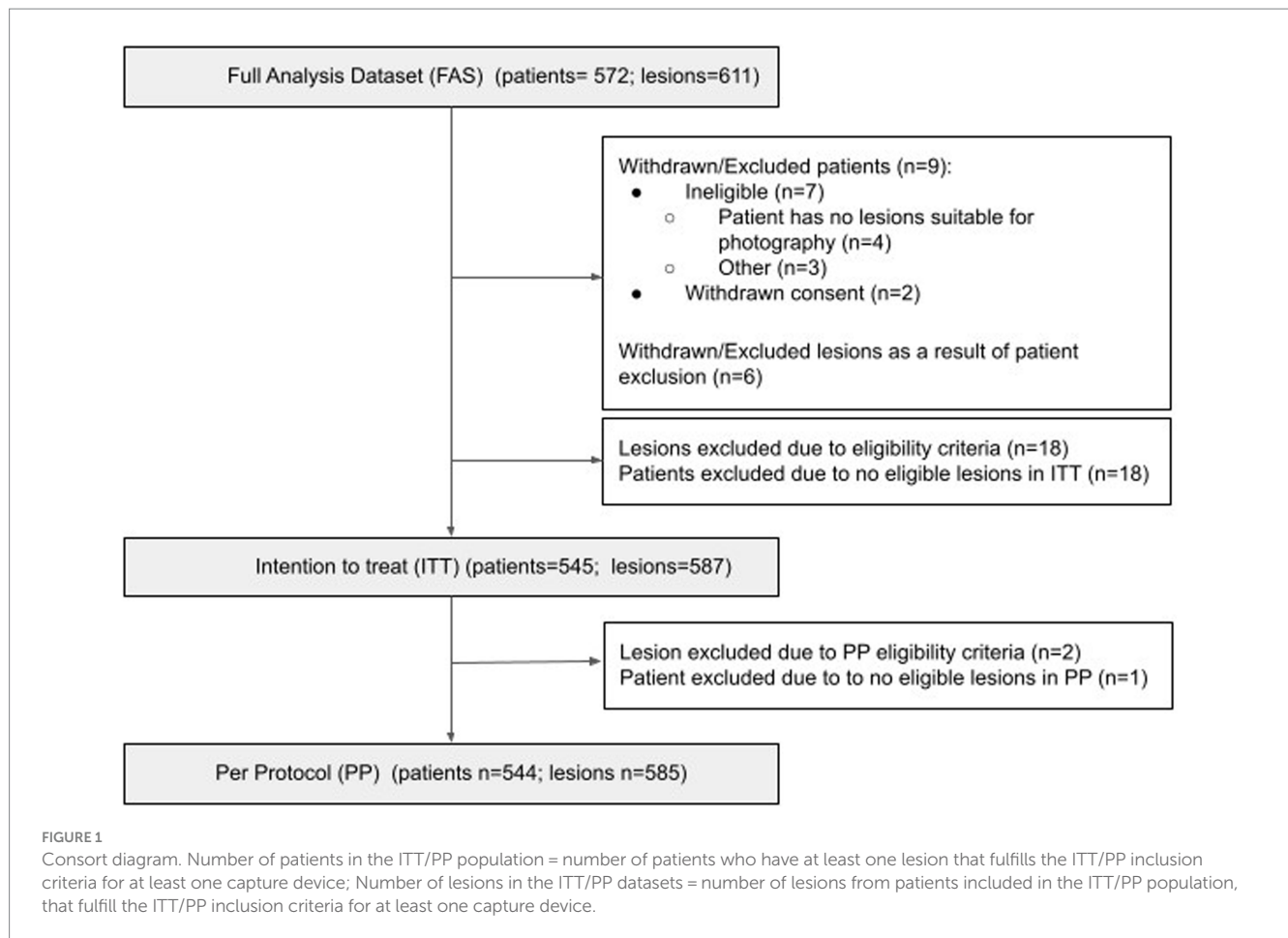
The PP population was equally distributed between females and males, mostly White (94%) and ranged in age from 18 to 97 years (median 73). Most patients (97.8%) had Fitzpatrick skin type I-III, with over half (56.8%) the patients reporting having Fitzpatrick skin type II (Table 1). Most lesions were located on the face and scalp (46.3%), posterior chest and back (14.5%), arms (13.5%), and legs (12.3%). On average, lesions were 8.9 (± 3.5 standard deviation) mm in size, ranging from 0.8 to 15 mm (Table 2).

Forty-three lesions in the PP population were diagnosed as SCC and 176 as BCC (Table 3) by histopathology. A further 22 lesions were diagnosed as SCC or BCC by clinical diagnosis only, which were excluded from the primary analysis. These lesions did not undergo a biopsy because either the dermatologist chose to treat the lesion ($n=10$), the patient refused biopsy ($n=3$) or other reason ($n=9$), including the biopsy occurred outside the study window. The PP population also included 16 lesions diagnosed as melanoma, and two lesions diagnosed as other malignancies [one Neuroendocrine, and one Spitzoid tumor of uncertain malignant potential (STUMP)] (Supplementary Table 1). Most malignancies were at an early stage.

The AUROC for SCC and BCC produced on images of biopsied lesions captured on each camera were: iPhone 6S 88.5% (95% CI: 83.9–93.1%) and 89.6% (95% CI: 86.5–92.7%) respectively; iPhone 11 88.9% (95% CI: 83.8–94.0%) and 89.5% (95% CI: 86.4–92.6%) respectively; and Samsung S10 84.9% (95% CI: 79.1–90.7%) and 87.2% (95% CI: 83.8–90.7%) respectively (Figure 2 and Table 4). The AUROCs for BCC and SCC, when calculated on all lesions, were $>90\%$ except for SCC in images captured on the Samsung 10 camera, where the AUROC was 87% (Figure 3). The AUROC for benign lesions produced by the AIaMD when assessing biopsied lesions only was between 74.9–76.8%, while the AUROC for benign lesions when all lesions were assessed, ranged between 79.8–80.9%. The AUROC for melanoma was $\geq 91.8\%$ for all cameras when the AIaMD assessed both biopsied lesions and all lesions. Moderate concordance (72.9% percentage agreement) was found between the AIaMD output label using images from the two iPhones; between iPhone 6S and Samsung 10 the percentage agreement of the AIaMD output label was 60.3%, and between the iPhone 11 and Samsung 10, it was 61.7%.

The AUROC for SCC and BCC produced by clinicians were 74.0% (95% CI: 66.4–81.6%) and 85.6% (95% CI: 81.8–89.3%) for biopsied lesions, and 76.9% (95% CI: 69.6–84.3%) and 90.0% (95% CI: 87.3–92.7%) for all lesions, respectively (Table 5). The AUROCs for SCC lesions were significantly lower than those produced by the AIaMD ($p<0.026$ for each camera). The clinician AUROCs were also significantly lower than those produced by the AIaMD ($p\leq 0.04$) for lesions diagnosed as IEC, AK and benign by histopathology. A weak to moderate level of agreement between clinical and histopathology diagnosis labels was found (percentage agreement 66.4%; Cohen's kappa = 0.52, $p<0.001$).

When pre-set threshold settings were applied, the sensitivity of the AIaMD to identify malignant lesions was above 90%, and the specificity of the AIaMD for malignant lesions was above 41.5% for each individual malignant lesion type and for all malignant lesions (Table 6). Both "other malignant" lesions were classified as malignant by the AIaMD using images from all cameras. The sensitivity and specificity of the AIaMD was more variable for other lesion types, particularly atypical lesions where the sensitivity varied between 38.1% for the Samsung and 86.4% for the iPhone 6S. In comparison, when considering the suspected diagnosis documented by the



clinician at the time of their assessment, they labeled fewer melanoma and SCC lesions accurately compared to the AIaMD (melanoma sensitivity of 81.2% compared to >93% by the AIaMD, SCC sensitivity of 63.6% compared to >90%), and more BCC lesions (sensitivity of 97.5% compared to <96%). Conversely, clinicians achieved a much higher specificity for malignant lesions and were more accurate at identifying benign lesions than the AIaMD.

Univariate analyses and multiple logistic regression analyses were performed on the FA population, filtered for those images with a final diagnosis available, to identify patient and lesion characteristics that might have influenced the accuracy of the AIaMD results and clinical diagnosis. Age above 60 was associated with a non-significant reduction in the accuracy of both dermatologists and the AIaMD to identify malignant lesions in images from the iPhones (Odds Ratio (OR) = 0.37–0.88, $p > 0.16$) and minor improvement in images from the Samsung 10 (OR = 1.07–1.18, $p > 0.7$). The impact only reached significance ($p = 0.034$) for the AIaMD with images from the iPhone 11, in patients aged 74–82. No significant impact was seen for either the AIaMD assessment or clinicians to accurately identify malignant lesions due to the Fitzpatrick skin type, however no cancers were detected in patients with Fitzpatrick skin types V and VI. Indeed, the only factor associated with a significant improvement on the accuracy of dermatologists to identify malignant lesions was a likely or high likelihood of skin cancer (OR > 7, $p < 0.018$), and on the AIaMD was a high level of patient concern (OR = 1.95, $p = 0.008$).

4. Discussion

The DERM-003 study is the first prospective, powered, clinical validation study that specifically evaluates the ability of the AIaMD to identify NMSC. Previously, the performance of the AIaMD to identify melanoma was evaluated (7), though this was on an earlier version of the software which focused solely on the identification of melanoma. DERM v3 is designed to identify SCC and BCC, alongside melanoma, as well as a range of premalignant, atypical and benign lesions often mistaken for skin cancer. The study recruited patients in dermatology clinics across the UK, such that the population reflects the aging, primarily Caucasian, population seen in these clinics. Although patients with Fitzpatrick Skin types V and VI were recruited, no skin cancers were diagnosed in these patients. Indeed, only 2.2% of the study population had Fitzpatrick skin type IV–VI, limiting the generalizability of these results for patients with darker skin tones. However, this reflects the trend seen in other clinical studies, and in the real world, where few patients with Fitzpatrick skin types IV–VI are seen in dermatology clinics with suspicious skin lesions (7, 9) and as such the study population can be seen as representative of the population that DERM would be used on. Robust performance evaluation of technologies, such as DERM, in patients with darker skin types may only be possible through post-market surveillance analyses, where more patients with these skin types can be evaluated (10). Similarly, the study included lesions across a good distribution of body locations, including those with higher sun exposure (head,

TABLE 1 Patient demographics by analysis population.

		FA (N)	ITT (N)	PP (N)
Total		572	545	544
Sex	Female	283	273	272
	Male	286	272	272
	Missing	3	0	0
Age	Mean	68.5	68.4	68.4
	SD	17.3	17.4	17.3
	Median	73	73	73
	Minimum	18	18	18
	Maximum	97	97	97
Ethnicity	White	534	512	511
	Asian	9	8	8
	Black	3	2	2
	Mixed	1	1	1
	Other	1	1	1
	Missing/Not stated	24	21	21
Fitzpatrick skin type	I	115	113	113
	II	327	309	309
	III	112	110	110
	IV	8	8	7
	V&VI	7	5	5
	Missing	3	0	0
Past medical history	Melanoma	38	37	37
	SCC	54	51	51
	BCC	127	126	126
	Other skin cancer	6	6	6
	None	332	313	312
	Unknown	15	12	12
Family medical history	Melanoma	27	27	26
	SCC	4	4	4
	BCC	23	23	23
	Other skin cancer	30	27	27
	None	439	418	418
	Unknown	49	46	46

FA, Full Analysis; ITT, Intention-to-Treat; PP, Per Protocol; SD, Standard Deviation. Family history of skin cancer is defined as first degree family only.

neck upper body) and lower limbs, where lesions can look different, and a range of skin cancer sub-types and stages that are seen in dermatology clinics. The study also included two “other malignant” lesions, which were diagnosed as STUMP and neuroendocrine, and a range of benign lesions.

When the study was designed, the calculations used to determine the success criteria and sample size were based on *in silico* performance data, which provided an assumption that the true AUROC for both SCC & BCC was 98%. The clinical performance of AI-based devices has

frequently been shown to be lower than that of laboratory-based data (11–13), and as such an expectation that the true AUROC achieved by the AIaMD on fresh clinical data would be comparable to laboratory results was perhaps unrealistic. Although the study failed to meet either of the co-primary endpoints, the AUROCs achieved by the AIaMD for SCC and BCC were still high and at least comparable to dermatologists. Indeed, the AUROCs of the clinical diagnosis for SCC and BCC lesions do not achieve a 90% AUROC either, indicating that even between clinician and histology there is a huge amount of diagnostic variability. This may be a reflection of clinical practice, where uncertainty of diagnosis drives a conservative view and decision to biopsy. Reassuringly, the AUROC produced by the AIaMD for melanoma was higher than that previously reported (7), demonstrating an improved performance of the AIaMD over the earlier version of the algorithm.

It should be noted that for non-biopsied lesions, the clinical diagnosis was used as the ground truth against which both the AIaMD and clinical diagnosis were compared. Clinical diagnosis therefore will appear more accurate in an all-lesion population, compared to a biopsy-only population, for those lesions where a high proportion do not have a histopathology diagnosis, specifically BCC, AK, and benign lesions. Despite this, the AUROCs achieved by the AIaMD for non-malignant lesions are comparable to those achieved by dermatologists in an all-lesion population, and indeed are notably higher than dermatologists in a biopsy only population.

The study assessed the performance of the AIaMD on images captured by three smartphone cameras available in the UK market at the time of the study. They were chosen to demonstrate performance of the AIaMD across different physical hardware devices (camera specification), operating systems, and price points and included a reference combination (iPhone 6S/DL1) which Skin Analytics has used in a previous study (7). Across the three cameras, the AUROCs for melanoma, SCC and BCC were very similar, indicating a good generalizability of the algorithm across the image capture hardware used. Although a greater variability across the cameras is seen for non-malignant lesions, the AUROCs achieved by the AIaMD from all cameras are still high.

The thresholds used to determine the sensitivity and specificity of the AIaMD were defined to be suitable for use in a secondary care setting at the beginning of the study. The sensitivity achieved by the AIaMD for melanoma, SCC and all malignant lesions were higher than achieved by clinical diagnosis alone, though clinicians referred these lesions for biopsy, so their management decision ensured a sensitivity of 100%. Even for BCC, sensitivity achieved by the AIaMD was around 95% using images from all cameras, and the sensitivity and specificity of the AIaMD to identify premalignant and atypical lesions are at a level that are clinically useful. Additionally, the specificity and NPV values for malignant lesions indicate that the AIaMD could aid the appropriate management of benign lesions. The threshold settings used in live deployments of the AIaMD are different than used in this study, and the sensitivity across all malignant lesions achieved in the real world have been demonstrated to be even higher (10), demonstrating the value in optimizing the settings within the AIaMD for the population it is being used to assess. The sensitivities achieved by the AIaMD for non-malignant lesions are more variable across the cameras than seen for malignant lesions, specifically atypical and benign lesions. Similarly, there was only a moderate concordance between the outputs produced by the AIaMD when analyzing images captured by the different image capture hardware.

TABLE 2 Lesion characteristics by analysis population.

		FA (N)	ITT (N)	PP (N)
Total		611	587	585
No. of Lesions assessed (count = number of participants)	1	532	505	504
	2	38	38	38
	3	2	2	2
Lesion size (mm)	Mean	9	8.6	8.6
	SD	4.9	3.5	3.5
	Median	8	8	8
	Minimum	0.8	0.8	0.8
	Maximum	64	20	20
Lesion location	Face and scalp	281	271	271
	Neck	21	21	21
	Anterior chest and abdomen	56	55	54
	Posterior chest and back	90	85	85
	Arms, excluding palms	80	79	79
	Palms	1	1	1
	Legs, excluding soles	80	73	72
	Soles	2	2	2
Patient level of concern	Not concerned	144	138	138
	A little concerned	307	299	299
	Very concerned	135	126	124
	Unknown	25	24	24
Experience of reviewing clinician	Foundation doctor	55	54	54
	Specialty registrar	20	19	18
	Consultant	455	440	439
	Other/GPwSI	81	74	74
	Missing	0	0	0
Lesion change	None	110	104	104
	Changed color	20	20	20
	Symptomatic	179	172	172
	Grown a bit	112	109	108
	New lesion	160	154	154
	Grown a lot	30	28	27
Clinician assessment of likelihood of skin cancer	Unlikely	224	216	215
	Equivocal	61	59	59
	Likely	211	203	202
	Highly likely	115	109	109
Biopsy taken	Lesion not referred for biopsy	167	163	162
	Further clinical review determined no biopsy needed	7	7	7
	Biopsy taken	418	398	397
	Patient refused biopsy	5	5	5
	Other	14	14	14

FA, Full Analysis; ITT, Intention-to-Treat; PP, Per Protocol; SD, Standard Deviation; GPwSI, General Practitioner with Special Interest. Number of lesions equates to number of lesion records created in the study database, the lesion count is based on clinician provided information on the number of lesions they assessed for each patient.

This may be due to variances in the hardware and post-processing software, or a factor of the threshold settings used by the AIaMD to assign the output label. If the confidence scores produced by the AIaMD on images of the same lesion taken on two different cameras were similar, but fell either side of the threshold set, the AIaMD output label from each image could be different. Since the AUROCs for these lesions were similar, this suggests that the thresholds applied could

TABLE 3 Breakdown of lesion diagnoses in the PP population.

Diagnosis	Subtype/stage	Clinical diagnosis	Histopathology
Melanoma	All	0	16
	Superficial spreading		9
	Lentigo maligna		1
	Other		1
	Not given/ambiguous		5
	<i>In situ</i>		2
	<1.0 mm		7
	1.01–2.0 mm		2
	2.01–4.0 mm		4
	>4 mm		0
	Not available		1
SCC	All	1	43
	Poorly differentiated		4
	Moderately differentiated		15
	Well differentiated		16
	Other/unknown		8
	Tis		1
	T1		38
	T2		0
	T4		3
	Not available		1
BCC	All	21	176
	Superficial		13
	Nodular		94
	Infiltrative		17
	Morphoeic		0
	Micronodular		2
	Basosquamous		1
	Other/unknown		49

(Continued)

TABLE 3 (Continued)

Diagnosis	Subtype/ stage	Clinical diagnosis	Histopathology
	Tis		3
	T1		141
	T2		2
	T4		0
	Not available		30
Other malignant		0	2
IEC		0	11
Actinic keratosis		40	21
Dysplastic nevus	All	2	20
	Mild atypia		9
	Moderate atypia		4
	Severe atypia		2
	Unknown severity		5
Seborrheic keratosis		59	12
Dermatofibroma		8	7
Vascular lesion		3	0
Lentigo		0	1
Benign melanocytic nevi		10	12
Other (benign)		43	75
Unknown/missing		1	1
Total lesions		188	397

SCC, Squamous Cell Carcinoma; BCC, Basal Cell Carcinoma; IEC, Intraepidermal Carcinoma.

be optimized for the image capture hardware being used, to achieve the best sensitivity.

The multivariate analysis identified a different impact of patient factors on the accuracy of malignant lesion detection by the AIaMD compared with previously reported analyses (7). This may reflect a change in how the AIaMD works between the two versions assessed. However, since the impact of patient factors on the accuracy of dermatologists is also different, it may be more a reflection that the previous study focused on melanoma detection, whereas this analysis considered all malignant lesions included in the study population. Further analyses are needed to understand whether these translate into a clinically relevant reduction in sensitivity and/or specificity of the AIaMD in different patient groups.

The main limitation to the DERM-003 study is the clinical setting in which it was conducted, and therefore the population studied. The study was conducted in UK secondary care dermatology clinics in order to include sufficient numbers of SCC and BCC lesions in the study population, and to easily capture the histopathology confirmed diagnosis of biopsied lesions and a dermatologist's clinical assessment of the lesion. This means the study population was made up of patients and lesions that dermatologists determined were suitable for inclusion in the study, which may not be representative of all patients and lesions that would be assessed by DERM. For example, lesions that were clearly benign may have been excluded by a study dermatologist, but on which a less experienced clinician may use DERM to support their patient management decision. That said, the study recruited a broader spectrum of lesions in the study population compared to a previous study (7), where the study population was limited to patients with a pigmented lesion that was due for biopsy. The results of this study are therefore more generalizable to the population of patients seen in secondary care in the UK. Indeed, data from ongoing post-market surveillance monitoring indicates that DERM can be deployed safely as an adjuvant tool in live clinical services accessible to patients with eligible skin lesions (i.e., excluding those under nails, on genitalia or on hairy areas of skin), from a broad range of age groups and most representative skin types with suspicious skin lesions, with sensitivity

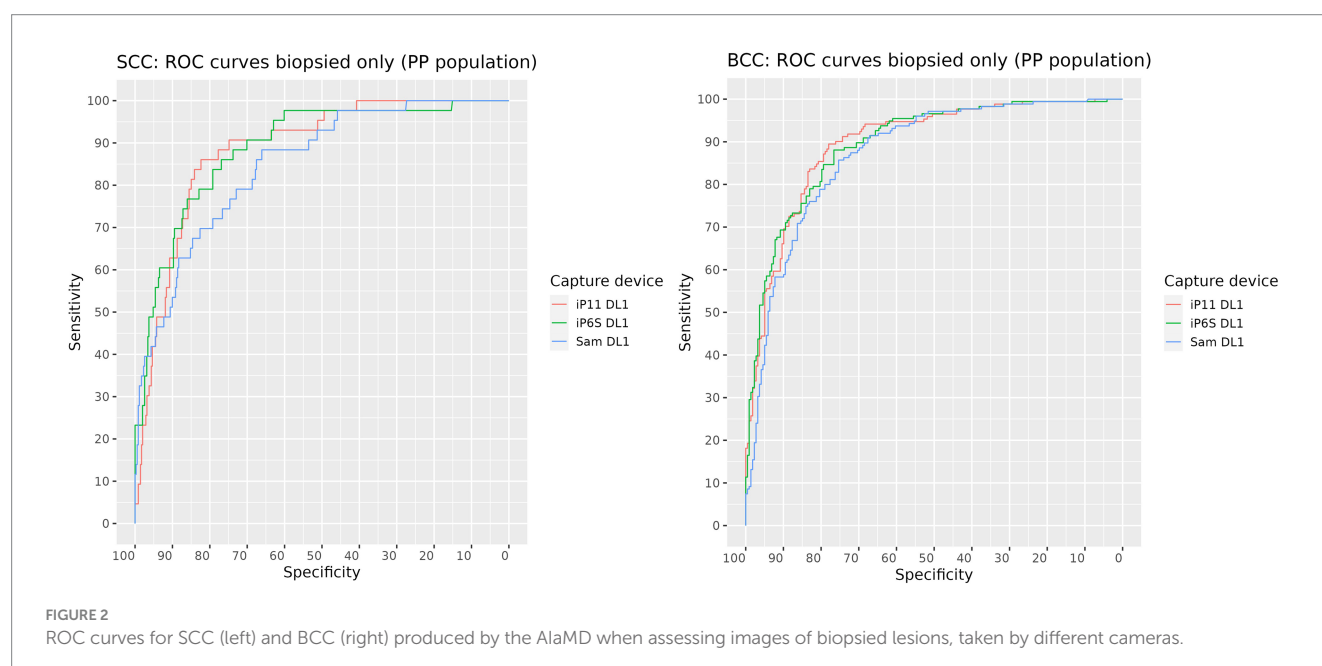


TABLE 4 AUROCs produced by DERM, using images taken on each camera.

Lesions	iPhone 11 (95% CI)		iPhone 6S (95% CI)		Samsung 10 (95% CI)	
	Biopsied	All	Biopsied	All	Biopsied	All
Melanoma	91.8% (82.9–100%)	92.6% (84.3–100%)	97.5% (94.8–100%)	97.5% (94.8–100%)	94.4% (89.2–99.6%)	94.6% (89.9–99.3%)
SCC	88.5% (83.9–93.1%)	90.1% (86.1–94.0%)	88.9% (83.8–94.0%)	90.0% (85.3–94.7%)	84.9% (79.1–90.7%)	87.0% (82.1–91.9%)
BCC	89.6% (86.5–92.7%)	92.0% (89.7–94.3%)	89.5% (86.4–92.6%)	92.3% (90.1–94.6%)	87.2% (83.8–90.7%)	90.9% (88.4–93.3%)
IEC	87.7% (82.0–93.4%)	89.0% (84.2–93.8%)	81.2% (73.3–89.2%)	83.3% (76.6–90.1%)	78.2% (67.8–88.6%)	80.2% (71.1–89.3%)
AK	77.3% (66.7–87.9%)	81.1% (75.0–87.2%)	86.1% (78.5–93.7%)	82.8% (77.0–88.7%)	77.8% (68.4–87.3%)	76.4% (69.6–83.3%)
Atypical	91.5% (85.4–97.5%)	89.4% (82.7–96.2%)	93.9% (87.0–100%)	93.0% (86.1–99.9%)	80.2% (68.3–92.1%)	80.9% (70.6–91.3%)
Benign	75.2% (69.9–80.6%)	80.9% (77.3–84.5%)	76.8% (71.6–81.9%)	80.4% (76.8–83.9%)	74.9% (69.3–80.4%)	79.8% (76.1–83.5%)

AUROC, Area Under the Receiver Operator Characteristic Curve; SCC, Squamous Cell Carcinoma; BCC, Basal Cell Carcinoma; IEC, Intraepidermal Carcinoma; AK, Actinic keratosis; CI, Confidence Intervals. Because of the necessity for a dermoscopic image of the lesion to be available for assessment by DERM, the number of lesions included was different for each camera.

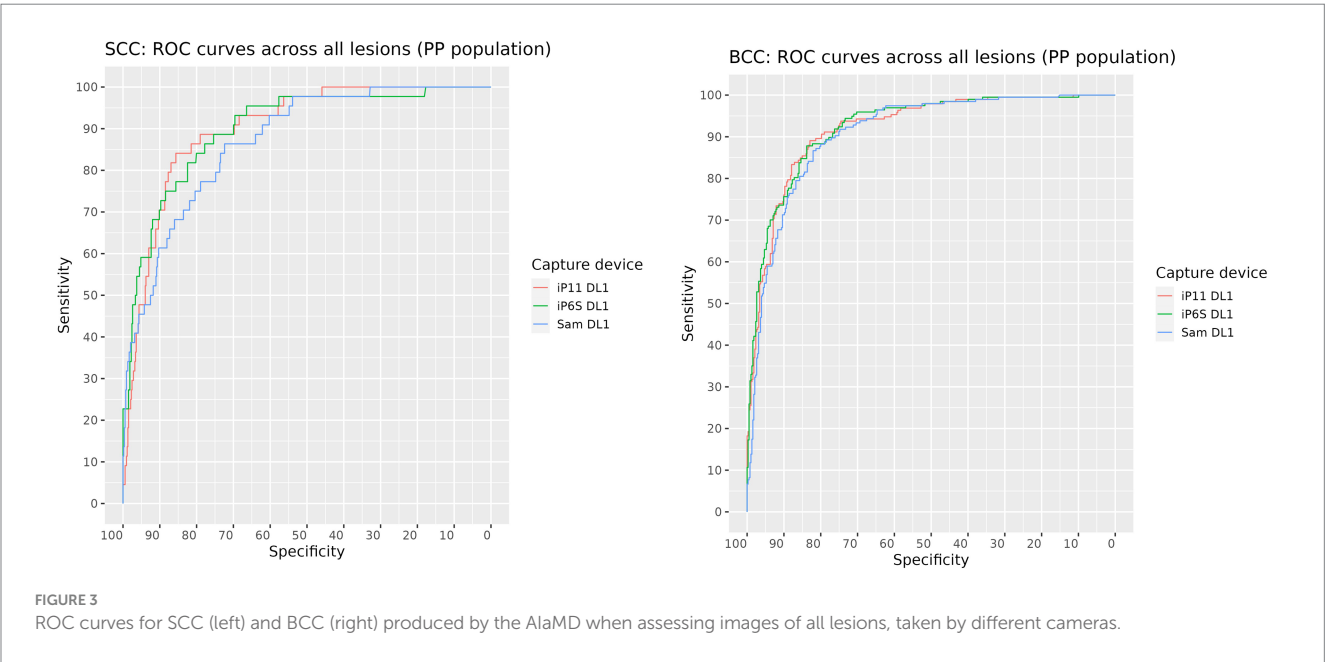


TABLE 5 AUROC of clinician assessment of likelihood of skin cancer.

Class	Biopsied lesions (95% CI)	All lesions (95% CI)
Lesions (N)	396	583
Melanoma	90.2% (80.2–100%)	90.3% (80.4–100%)
SCC	74.0% (66.4–81.6%)	76.9% (69.6–84.3%)
BCC	85.6% (81.8–89.3%)	90.0% (87.3–92.7%)
IEC	63.6% (49.8–77.4%)	63.6% (49.8–77.4%)
AK	56.9% (49.2–64.6%)	85.0% (79.2–90.8%)
Atypical	83.2% (72.3–94%)	85.1% (75.1–95%)
Benign	67.1% (62.2–72%)	82.1% (78.8–85.5%)

AUROC, Area Under the Receiver Operator Characteristic Curve; SCC, Squamous Cell Carcinoma; BCC, Basal Cell Carcinoma; IEC, Intraepidermal Carcinoma; AK, Actinic keratosis; CI, Confidence Intervals.

and specificity in-line with target thresholds and performance demonstrated in clinical studies (10).

Finally, the reliance on clinical diagnosis as the ground truth for non-biopsied lesions not only artificially increases the performance

metrics for the dermatologists, as discussed above, but potentially impacts the apparent performance of the AIaMD on non-biopsied lesions. The clinical diagnosis of skin cancer by clinicians is based on the subjective interpretation of morphological features and as such variability in the clinical diagnoses given by dermatologists is known to exist (14). The reliance on one dermatologist to provide the clinical diagnosis used as the ground truth for non-biopsied lesions introduces a potential bias to the results for both the AIaMD and dermatologists. The use of a panel of dermatologists to provide a consensus diagnosis would have provided a greater confidence in the clinical diagnosis ground truth, and provided an independent diagnosis against which to compare the investigating dermatologist.

In conclusion, even though the study failed to meet its co-primary endpoints, the results from the DERM-003 study showed that the AIaMD can detect NMSC and premalignant lesions with a similar level of accuracy as dermatologists, and that taking the images was a quick and well tolerated process. DERM could provide dermatologist level assessment of suspicious skin lesions earlier in the patient pathway, potentially enabling the earlier diagnosis of malignant lesions and improvement of differentiation between harmless and potentially harmful lesions by non-specialists.

TABLE 6 Diagnostic performance metrics of clinicians and DERM, using images from each camera, for all lesions in the Per Protocol population.

	Device	Lesions (N)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	FNR (95% CI)	FPR (95% CI)
Melanoma	Clinicians	581	81.2% (53.7–95.0%)	98.9% (97.6–99.6%)	68.4% (43.5–86.4%)	99.5% (98.3–99.9%)	18.8% (5.0–46.3%)	1.1% (0.4–2.4%)
	iPhone 6S	578	100% (74.7–100%)	69.6% (65.6–73.4%)	8.1% (4.7–13.2%)	100% (98.8–100%)	0% (0–25.3%)	30.4% (26.6–34.4%)
	iPhone 11	571	93.3% (66.0–99.7%)	73.6% (69.6–77.1%)	8.7% (5.0–14.4%)	99.8% (98.4–100%)	6.7% (0.3–34.0%)	26.4% (22.9–30.4%)
	Samsung	578	100% (75.9–100%)	65.5% (61.4–69.4%)	7.6% (4.6–12.3%)	100% (98.7–100%)	0% (0–24.1%)	34.5% (30.6–38.6%)
SCC	Clinicians	565	63.6% (47.7–77.2%)	89.1% (86–91.5%)	32.9% (23.4–44.1%)	96.7% (94.5–98.0%)	36.4% (22.8–52.3%)	10.9% (8.5–14.0%)
	iPhone 6S	563	95.4% (83.3–99.2%)	44.7% (40.4–49.1%)	12.8% (9.5–17%)	99.2% (96.6–99.9%)	4.6% (0.8–16.7%)	55.3% (50.9–59.6%)
	iPhone 11	556	93.2% (80.3–98.2%)	45.7% (41.3–50.1%)	12.8% (9.5–17.1%)	98.7% (96–99.7%)	6.8% (1.8–19.7%)	54.3% (49.9–58.7%)
	Samsung	562	90.9% (77.4–97%)	50.6% (46.2–55%)	13.5% (9.9–18.1%)	98.5% (95.9–99.5%)	9.1% (3–22.6%)	49.4% (45–53.8%)
BCC	Clinicians	521	97.5% (93.9–99.1%)	77.4% (72.4–81.8%)	72.6% (66.7–77.7%)	98% (95.2–99.3%)	2.5% (0.9–6.1%)	22.6% (18.2–27.6%)
	iPhone 6S	519	94.9% (90.6–97.4%)	41.6% (36.2–47.2%)	49.9% (44.7–55%)	93.1% (87.3–96.4%)	5.1% (2.6–9.4%)	58.4% (52.8–63.8%)
	iPhone 11	512	95.8% (91.7–98%)	45% (39.5–50.6%)	51.1% (45.8–56.4%)	94.7% (89.5–97.5%)	4.2% (2–8.3%)	55% (49.4–60.5%)
	Samsung	518	94.4% (89.9–97%)	54.5% (48.9–60%)	55.6% (50.1–61%)	94.1% (89.4–96.9%)	5.6% (3.0–10.1%)	45.5% (40–51.1%)
Malignant	Clinicians	583	93.8% (90–96.3%)	77.4% (72.4–81.8%)	77% (71.9–81.4%)	94.3% (90.6–96.7%)	5.8% (3.4–9.5%)	22.6% (18.2–27.6%)
	iPhone 6S	580	95.7% (92.3–97.7%)	41.6% (36.2–47.2%)	56.8% (52–61.5%)	92.4% (86.5–96%)	4.3% (2.3–7.7%)	58.4% (52.8–63.8%)
	iPhone 11	573	96.0% (92.6–98%)	45% (39.5–50.6%)	58% (53.1–62.7%)	93.5% (88.1–96.7%)	4% (2–7.4%)	55% (49.4–60.5%)
	Samsung	580	94.9% (91.3–97.2%)	54.5% (48.9–60%)	62.4% (57.4–67.2%)	93.1% (88.3–96.1%)	5.1% (2.8–8.7%)	45.5% (40–51.1%)
IEC	Clinicians	323	90.9% (57.1–99.5%)	78.8% (73.8–83.2%)	13.2% (6.8–23.3%)	99.6% (97.4–100%)	9.1% (0.5–42.9%)	21.1% (16.8–26.2%)
	iPhone 6S	322	100% (67.9–100%)	43.1% (37.5–48.8%)	5.9% (3.1–10.5%)	100% (96.5–100%)	0% (0–32.1%)	56.9% (51.2–62.5%)
	iPhone 11	320	100% (67.9–100%)	46.6% (41–52.3%)	6.2% (3.3–11.2%)	100% (96.8–100%)	0% (0–32.1%)	53.4% (47.7–59%)
	Samsung	323	90.9% (57.1–99.5%)	56.1% (50.4–61.6%)	6.8% (3.5–12.5%)	99.4% (96.4–100%)	9.1% (0.5–42.9%)	43.9% (38.4–49.6%)
AK	Clinicians	312	96.7% (87.6–99.4%)	79.3% (73.6–84%)	53.1% (43.5–62.6%)	99% (96.1–99.8%)	3.3% (0.6–12.4%)	20.7% (16–26.4%)
	iPhone 6S	311	85.0% (72.9–92.5%)	43.4% (37.2–49.8%)	26.4% (20.5–33.3%)	92.4% (85.6–96.2%)	15% (7.5–27.1%)	56.6% (50.2–62.8%)
	iPhone 11	309	84.8% (72.5–92.4%)	47.2% (40.9–53.6%)	27.5% (21.3–34.7%)	92.9% (86.6–96.5%)	15.2% (7.6–27.5%)	52.8% (46.4–59.1%)
	Samsung	312	83.6% (71.5–91.4%)	51.4% (45–57.7%)	29.5% (22.9–37%)	92.8% (86.8–96.3%)	16.4% (8.6–28.5%)	48.6% (42.3–55%)

(Continued)

TABLE 6 (Continued)

	Device	Lesions (N)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	FNR (95% CI)	FPR (95% CI)
Atypical	Clinicians	251	76.2% (52.5–90.9%)	73.9% (67.6–79.4%)	21% (12.9–32.2%)	97.1% (93.1–98.9%)	23.8% (9.1–47.5%)	26.1% (20.6–32.4%)
	iPhone 6S	251	86.4% (64.0–96.4%)	39.3% (33.0–46.0%)	12% (7.6–18.4%)	96.8% (90.2–99.2%)	13.6% (3.6–36.0%)	60.7% (54.0–67.0%)
	iPhone 11	250	59.1% (36.7–78.5%)	43.9% (37.4–50.6%)	9.2% (5.2–15.6%)	91.7% (84.5–95.9%)	40.9% (21.5–63.3%)	56.1% (49.4–62.6%)
	Samsung	251	38.1% (19.0–61.3%)	48.3% (41.7–54.9%)	6.3% (3.0–12.4%)	89.5% (82.4–94.1%)	61.9% (38.7–81.0%)	51.7% (45.1–58.3%)
Premalignant	Clinicians	323	91.4% (83.3–95.9%)	73.9% (67.6–79.4%)	58.6% (50.1–66.6%)	95.5% (91.0–97.9%)	8.6% (4.1–16.7%)	26.1% (20.6–32.4%)
	iPhone 6S	322	87.1% (78.2–92.9%)	39.3% (33.0–46.0%)	36.8% (30.5–43.6%)	88.2% (80.0–93.5%)	12.9% (7.1–21.8%)	60.7% (54.0–67.0%)
	iPhone 11	320	80.4% (70.6–87.7%)	43.9% (37.4–50.6%)	36.6% (30.1–43.7%)	84.8% (76.7–90.5%)	19.6% (12.3–29.4%)	56.1% (49.4–62.6%)
	Samsung	323	75.3% (65.0–83.4%)	48.3% (41.7–54.9%)	37% (30.2–44.4%)	82.8% (75.1–88.6%)	24.7% (16.6–35.0%)	51.7% (45.1–58.3%)
Benign	Clinicians	581	73.9% (67.6–79.4%)	93.7% (90.5–95.9%)	88.5% (83.0–92.5%)	84.6% (80.5–87.9%)	26.1% (20.6–32.4%)	6.3% (4.1–9.5%)
	iPhone 6S	578	39.3% (33.0–46.0%)	94.3% (91.1–96.4%)	81.8% (73.1–88.3%)	70.3% (65.9–74.4%)	60.7% (54.0–67.0%)	5.7% (3.6–8.9%)
	iPhone 11	571	43.9% (37.4–50.6%)	93.3% (90.0–95.6%)	81.3% (73.1–87.5%)	71.4% (67.0–75.5%)	56.1% (49.4–62.6%)	6.7% (4.4–10.0%)
	Samsung	578	48.3% (41.7–54.9%)	91.4% (87.8–94.0%)	78.7% (70.9–85.0%)	72.8% (68.3–76.8%)	51.7% (45.1–58.3%)	8.6% (6–12.2%)

SCC, Squamous Cell Carcinoma; BCC, Basal Cell Carcinoma; IEC, Intraepidermal Carcinoma; AK, Actinic Keratosis; CI, Confidence Intervals Rate; PPV, Positive Predictive Value; NPV, Negative Predictive Value; FPR, False Positive Rate; FNR, False Negative.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Leicester South National Research Ethics Committee. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

HM: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Writing – original draft. CM: Data curation, Investigation, Writing – review & editing. SA: Data curation, Investigation, Writing – original draft. CD: Data curation, Investigation, Writing – review & editing. MV: Data curation, Formal analysis, Investigation, Methodology, Writing – review & editing. CK: Conceptualization, Formal analysis, Methodology, Writing – review & editing. JG: Methodology, Software, Writing – review & editing. DM: Investigation, Resources, Supervision, Writing – review & editing. IP:

Conceptualization, Investigation, Supervision, Writing – review & editing. SA: Writing - reviewing and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Skin Analytics, London, UK sponsored and funded this study, as part of an InnovateUK BioMedical Catalyst project, and was involved with the study design, data collection, statistical analysis and interpretation of the data.

Acknowledgments

The authors would like to thank all patients who consented to the study, and all research staff involved in the conduct and data collection for the study. In particular, Philip Hampton for recruiting additional patients at the Royal Victoria Infirmary, Alicja Raginis-Zborowska, and other Skin Analytics staff involved in the operationalization of the study.

Conflict of interest

SA has received a non-financial gift from Skin Analytics for presenting the results of this research. PK was an employee of Skin Analytics. DM, JG, HM, and MV are employees of Skin Analytics and

have received Skin Analytics shares or share options. JG is named as an inventor on patents (pending) relating to DERM.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2023.1288521/full#supplementary-material>

References

1. Cancer Research UK. Types of skin cancer. (2023). Available at: <https://www.cancerresearchuk.org/about-cancer/skin-cancer/types> (Accessed August 14, 2023).
2. Cancer Research UK. Non-melanoma skin cancer statistics. (2022). Available at: <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/non-melanoma-skin-cancer> (Accessed August 14, 2023).
3. Cancer Research UK. Survival and Incidence by Stage at Diagnosis. (2023). Available at: <https://crukancerintelligence.shinyapps.io/EarlyDiagnosis/> (Accessed August 14, 2023).
4. Newlands C, Currie R, Memon A, Whitaker S, Woolford T. Non-melanoma skin cancer: United Kingdom National Multidisciplinary Guidelines. *J Laryngol Otol.* (2016) 130:S125–32. doi: 10.1017/S0022215116000554
5. NHS England. Cancer Waiting Times. (2023). Available at: <https://www.england.nhs.uk/statistics/statistical-work-areas/cancer-waiting-times/> (Accessed September 1, 2023).
6. Webb JB, Khanna A. Can we rely on a general practitioner's referral letter to a skin lesion clinic to prioritize appointments and does it make a difference to the patient's prognosis? *Ann R Coll Surg Engl.* (2006) 88:40–5. doi: 10.1308/003588406X82970
7. Phillips M, Marsden H, Jaffe W, Matin N, Wali GN, Greenhalgh J, et al. Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. *JAMA Netw Open.* (2019) 2:e1913436. doi: 10.1001/jamanetworkopen.2019.13436
8. Pizzichetta MA, Talamini R, Stanganelli I, Puddu P, Bono R, Argenziano G, et al. Amelanotic/hypomelanotic melanoma: clinical and dermoscopic features. *Br J Dermatol.* (2004) 150:1117–24. doi: 10.1111/j.1365-2133.2004.05928.x
9. Marsden H, Kemos P, Venzi M, Noy M, Maheswaran S, Francis N, et al. Accuracy of an artificial intelligence as a medical device as part of a UK-based skin cancer teledermatology service. *Front. Med.* (2023).
10. Thomas L, Hyde C, Mullarkey D, Greenhalgh J, Kalsi D, Ko J. Real-world post-deployment performance of a novel machine learning-based digital health technology for skin lesion assessment and suggestions for post-market surveillance. *Front. Med.* (2023).
11. Li CX, Fei WM, Shen CB, Wang ZY, Jing Y, Meng RS, et al. Diagnostic capacity of skin tumor artificial intelligence-assisted decision-making software in real-world clinical settings. *Chin Med J.* (2020) 133:2020–6. doi: 10.1097/CM9.0000000000001002
12. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.* (2018) 15:e1002683. doi: 10.1371/journal.pmed.1002683
13. Lin D, Xiong J, Liu C, Zhao L, Li Z, Yu S. Application of comprehensive artificial intelligence retinal expert (CARE) system: a national real-world evidence study. *Lancet Digit Health.* (2021) 3:e486–95. doi: 10.1016/S2589-7500(21)00086-8
14. Polesie S, Sundback L, Gillstedt M, Ceder H, Dahlén Gyllencreutz J, Fougberg J, et al. Interobserver agreement on dermoscopic features and their associations with in situ and invasive cutaneous melanomas. *Acta Derm Venereol.* (2021) 101:adv00570. doi: 10.2340/actadv.v101.281



OPEN ACCESS

EDITED BY
Justin Ko,
Stanford University, United States

REVIEWED BY
Karolina Chilicka-Hebel,
Opole University, Poland
Hongxiang Chen,
Huazhong University of Science
and Technology, China

*CORRESPONDENCE
Xian Jiang
✉ jiangxian@scu.edu.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 09 July 2023
ACCEPTED 12 September 2023
PUBLISHED 06 October 2023

CITATION
Li J, Du D, Zhang J, Liu W, Wang J, Wei X,
Xue L, Li X, Diao P, Zhang L and Jiang X (2023)
Development and validation of an artificial
intelligence-powered acne grading system
incorporating lesion identification.
Front. Med. 10:1255704.
doi: 10.3389/fmed.2023.1255704

COPYRIGHT
© 2023 Li, Du, Zhang, Liu, Wang, Wei, Xue, Li,
Diao, Zhang and Jiang. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted which
does not comply with these terms.

Development and validation of an artificial intelligence-powered acne grading system incorporating lesion identification

Jiaqi Li^{1,2,3†}, Dan Du^{1,2,3†}, Jianwei Zhang^{3,4}, Wenjie Liu^{3,4},
Junyou Wang^{3,4}, Xin Wei^{3,4}, Li Xue^{1,2}, Xiaoxue Li^{1,2}, Ping Diao^{1,2},
Lei Zhang^{3,4} and Xian Jiang^{1,2,3*}

¹Department of Dermatology, West China Hospital, Sichuan University, Chengdu, China, ²Laboratory of Dermatology, Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Clinical Institute of Inflammation and Immunology, Sichuan University, Chengdu, China, ³Med-X Center for Informatics, Sichuan University, Chengdu, China, ⁴College of Computer Science, Sichuan University, Chengdu, Sichuan, China

Background: The management of acne requires the consideration of its severity; however, a universally adopted evaluation system for clinical practice is lacking. Artificial intelligence (AI) evaluation systems hold the promise of enhancing the efficiency and reproducibility of assessments. Artificial intelligence (AI) evaluation systems offer the potential to enhance the efficiency and reproducibility of assessments in this domain. While the identification of skin lesions represents a crucial component of acne evaluation, existing AI systems often overlook lesion identification or fail to integrate it with severity assessment. This study aimed to develop an AI-powered acne grading system and compare its performance with physician image-based scoring.

Methods: A total of 1,501 acne patients were included in the study, and standardized pictures were obtained using the VISIA system. The initial evaluation involved 40 stratified sampled frontal photos assessed by seven dermatologists. Subsequently, the three doctors with the highest inter-rater agreement annotated the remaining 1,461 images, which served as the dataset for the development of the AI system. The dataset was randomly divided into two groups: 276 images were allocated for training the acne lesion identification platform, and 1,185 images were used to assess the severity of acne.

Results: The average precision of our model for skin lesion identification was 0.507 and the average recall was 0.775. The AI severity grading system achieved good agreement with the true label (linear weighted kappa = 0.652). After integrating the lesion identification results into the severity assessment with fixed weights and learnable weights, the kappa rose to 0.737 and 0.696, respectively, and the entire evaluation on a Linux workstation with a Tesla K40m GPU took less than 0.1s per picture.

Conclusion: This study developed a system that detects various types of acne lesions and correlates them well with acne severity grading, and the good accuracy and efficiency make this approach potentially an effective clinical decision support tool.

KEYWORDS

dermatology, acne, artificial intelligence, acne lesions, grading system

Introduction

Acne vulgaris is the eighth most prevalent disease affecting 9.4% of the global population (1). Although acne can occur at all ages, adolescents are the most prevalent group of acne sufferers, and eighty-five percent of adolescents are affected by acne (2). As a condition that alters appearance, acne affects patients' physical and psychological well-being and causes a strong desire for treatment (3). The large patient population and the strong desire for treatment seriously burden healthcare resources (4, 5). Assessment of acne severity is essential for the patient's stepwise therapy. There are more than 20 published scales for evaluating acne, but none is adopted universally for clinical practice (6).

Most scales can be classified as lesion-counting scales or text description scales. Lesion counting scales correspond to the severity by measuring different types of acne lesions, such as the Global Acne Grading System (7, 8). Counting acne lesions is supposed to be a more objective method. However, it shows a high degree of variability between raters due to ambiguity between different categories of skin lesions and interevaluator differences in the definition of skin lesions (9). In addition, a single counting process ignores the degree of inflammation, postinflammatory hyperpigmentation, scarring, and other features that affect the severity. In contrast to quantitative scales, qualitative scales distinguish between different levels of severity through textual descriptions. Although qualitative scales require more clinical experience from the evaluator, they simplify the tedious counting process to a certain extent and take care of other acne characteristics beyond the number of lesions. For example, Investigator Global Assessment classifies acne into five levels through text descriptions (clear, almost clear, mild, moderate, severe, and very severe) (10). On this basis, a recent study found that replacing the qualitative labels with the corresponding treatment intensity labels effectively reduced the high interrater variability, although these labels are more unstable since treatment options may change depending on regional perceptions and disciplinary developments (11).

Artificial intelligence (AI) for acne grading has been considered a promising research direction to increase the consistency and efficiency of assessment. Some AI systems focus on identifying and counting different types of lesions, but as with lesion-counting scales, they ignore considerable information beyond the countable lesions (12, 13). Other AI systems analyze the image as a whole but leave the evaluation free from clinical interpretability (14, 15). We believe that the quantity of different types of lesions is an inadequate but crucial component of acne severity assessment. Therefore we sought to develop a novel AI system that could integrate the identification and counting of skin lesions into the overall facial evaluation process, thereby improving the predictive accuracy.

Materials and methods

Database

This study was conducted at sichuan university from January, 2020 to June, 2022, and was approved by the west china hospital institutional review board to use the patients' deidentified images

and records. This study followed the declaration of Helsinki and standards for reporting of diagnostic accuracy (STARD) reporting guidelines and the checklist for evaluation of image-based artificial intelligence algorithm reports in dermatology (CLEAR Derm) (16). We collected records of 3,098 visits to our dermatology specialist clinics with a diagnosis of acne without other inflammatory skin disease diagnoses. Of the 3,098 visits recorded, 1,501 had corresponding standardized pictures obtained via the VISIA system, including frontal, left and right profile photos, and information from these visits was included in the current study. To select labeling experts for the database and to evaluate the adequacy of the standardized frontal photo, 40 patients with acne (10 mild, 20 severe, 10 severe) were selected based on clinical records. seven experienced dermatologists first rated the frontal photos of the 40 patients, and the three evaluators with the highest average linear weighted Cohen's κ were selected to complete the severity marking of the 1,461 records. The median of their ratings was considered the true label. After disrupting the order of the 40 images, the 7 dermatologists again rated the combined photos (frontal and left and right side photos) of the 40 patients. To improve interrater agreement, in this study we used the Treatment Intensity label to distinguish between the severity of patients (11), and due to the low number of extremely severe cases, we combined Level 8 and Level 9 (Table 1).

Development of the skin lesion identification platform

For the acne detection module, we used a publicly available deep-learning method to detect acne lesions (17). We used a VISIA complexion analysis system to photograph 276 facial images as our samples, where each sample has a resolution from 3128×4171 to 3456×5184 pixels. All the samples were split 9:1 into training samples ($n = 248$) and test samples ($n = 28$). Six dermatologists participated in annotating all the samples. A total of 15,922 skin lesions with 10 lesion categories, i.e., *open comedone*, *closed comedone*, *papule*, *pustule*, *nodule/cyst*, *atrophic scar*, *hypertrophic scar*, *melasma* and *nevus* were generated. Next,

TABLE 1 Severity label and corresponding treatment intensity list.

Grading label	Severity description	Treatment intensity
1	Clear	No treatment necessary
2	Almost clear	BPO or a mild topical retinoid
3	Mild	BPO and a topical retinoid
4	Mild to moderate	BPO and a stronger topical retinoid or a topical retinoid and consideration of an oral antibiotics
5	Moderate	Topical treatment and an oral antibiotics
6	Moderate to less severe	Same as 5, but start considering isotretinoin
7	Less severe	Same as 5, but recommend isotretinoin
8	Severe or very severe	Should be on isotretinoin

BPO, benzoyl peroxide.

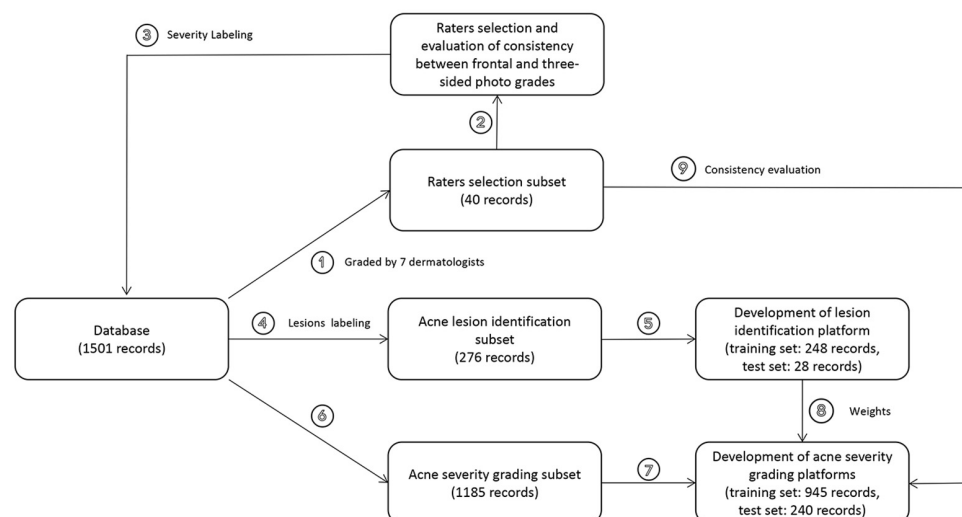


FIGURE 1

Overview of the development and validation of our AI systems. After confirming the adequacy of the frontal photo information, the doctors with the highest agreement with other peers were selected as true label raters for the remaining 1,461 frontal photos. Of the 1,461 photos, 276 were used to develop a skin lesion identification platform and 1,185 were used to develop an acne severity rating system. Then, we sought to incorporate skin lesion identification results into the severity evaluation and validated the feasibility in test set and rater selection subset.

the network is trained by an SGD optimizer with 15 epochs, where the learning rate, momentum, and weight decay were 0.002, 0.9, and 0.0001, respectively.

Development of acne grading systems

We used ResNet50 as the training network for the baseline results (18). This network contains four large blocks, each with 3, 4, 6, and 3 small blocks, and each small block consists of three convolutional layers. In addition, the network contains jump connections to alleviate the problem of gradient explosion and gradient disappearance during training, thereby allowing the model to extract deeper features. A total of 1,185 images were used for the grading experiments, of which, 945 were used for training and 240 for testing. For the training set, all images were first resized to 256×256 pixels and later randomly cropped to 224×224 pixels to meet the input size of the network. Furthermore, the images are randomly flipped horizontally (50% probability) and randomly rotated from -20° to $+20^\circ$ to expand the data to prevent training overfitting. The model was trained using cross entropy loss with a total of 200 epochs and a batch size of 32. The initial learning rate was 0.001, and it decayed to 0.0001 using a cosine annealing function. The optimizer was the Adam optimizer with a weight decay of 0.0001. The training was conducted on a Tesla K40m GPU. For the acne grading task, the number of acne lesions as well as the overall assessment are an important reference for acne grading. Therefore, we propose a method that combines dermatologists' *a priori* knowledge with a CNN to automatically grade pictures. The acne counts of all samples were semiautomatically labeled by the trained detection model and manually validated by an experienced dermatologist. The rule divides each image into a grading interval instead of a single grade to guide the network to better predict the image grading.

We propose two methods to integrate the proposed rules into the network, i.e., fixed weights and learnable weights, and the two methods are shown in Figure 1. For the fixed-weights approach, the probability weight of the interval is fixed. If the interval does not contain the grading, the weight is 0; otherwise, it is 1. Each input image is fed into the CNN first to learn the image features. The image features are average-pooled and mapped to an 8-dimensional vector to correspond to the probability of each classification. Then, the two vectors are multiplied by the corresponding position elements to obtain the predicted probability of each classification. Since the proposed rule reduces the weight of the intervals that do not belong to the image classification, only the predicted probability of the interval to which the image belongs is obtained. The classification corresponding to the highest probability is selected as the predicted class.

For the learnable weights approach, the network is given an initial value, after which the weights are fine-tuned through training. As shown in Figure 2, after training, the network outputs the graded probability values and the learned interval weights. The prediction probability of each classification is obtained by multiplying the classification probabilities with the corresponding interval weights. Again, the classification with the highest probability is the grading predicted by the model.

Statistical analysis

To determine the sample size of rater selection, assuming the interrater correlation coefficients were approximately 0.8, at least 7 raters and 40 subjects were needed. No formal sample size was calculated for validation of AI systems. Cohen's kappa with linear weights was used to evaluate the AI's performance against the true label or the 7 dermatologists on the rater selection dataset. A kappa value of less than 0.6 was considered unacceptably

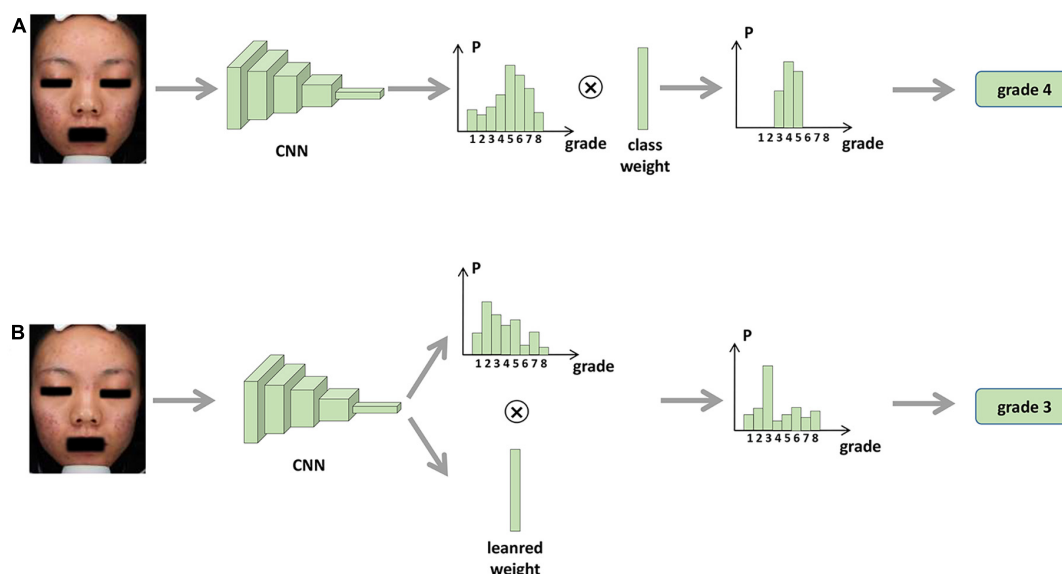


FIGURE 2

Procedure for integrating skin lesion identification with acne severity assessment based on AI. (A) Fixed weights approach. (B) Learnable weights approach.

low. The statistical analyses were performed using Prism software (GraphPad Prism 8.0) and R (version 4.2.1).

Results

The database was divided into three subsets, and the baseline characteristics are summarized in [Table 1](#). Forty records were enrolled to select the true label rater. The mean age of the 23 female and 17 male patients was 24.8 years, ranging from 16 to 39 years. Of the 560 assessments (7 raters, 40 patients and 2 rounds), each grading of severity was represented by at least 2 subjects. The evaluations obtained through the frontal photos are in good agreement with those obtained through the three-sided photos, indicating that the frontal photos are sufficiently informative as samples for the AI evaluation ([Table 2](#)). For interrater agreement of frontal photo assessment, the pairwise Cohen's kappa for each dermatologist ranked in descending order is shown in [Supplementary Figure 1](#), and the three raters with the greatest average kappa value were selected to rate all the photos in the database. For consistency of the assessment of frontal photographs and 3-side photographs, the overall ICC for frontal photo assessment and 3-side photograph was 0.878 (0.814, 0.916), which suggests that a frontal photograph taken with VISIA alone can yield a similar amount of information for acne as three-sided photos.

For the development of the acne lesion identification platform, 276 frontal photos were labeled by five doctors and reviewed by a senior doctor. In total, 3,060 closed pimples, 2,192 open pimples, 3,861 papules, 884 pustules, 113 nodules or cysts, 5,410 atrophic scars and 302 hypertrophic scars were marked in 276 images ([Figure 3](#)). The 276 images were divided into a training set and a test set at a ratio of 9:1. The average precision of our model for skin lesion identification was 0.507, and the average

recall was 0.775, which outperformed state-of-the-art one-stage and two-stage generic object detection methods. As previously anticipated, skin lesion counts are not sufficient for severity determination, and we were not able to build a decision tree model with good performance for acne severity evaluation, either based on the number of manually annotated lesions or the number of lesions identified by the algorithm (data not shown). However, different types of lesions have different distribution patterns on the face ([Supplementary Figure 2](#)). Inflammatory lesions (papules, pustules, nodules/cysts) are more evenly distributed, and non-inflammatory lesions and secondary lesions have unique distribution characteristics. Closed acne tends to be located on the forehead and midface, while open acne tends to cluster on the forehead. Atrophic scarring is concentrated on both cheeks, while hyperplastic scarring often occurs on the skin of the lower jaw.

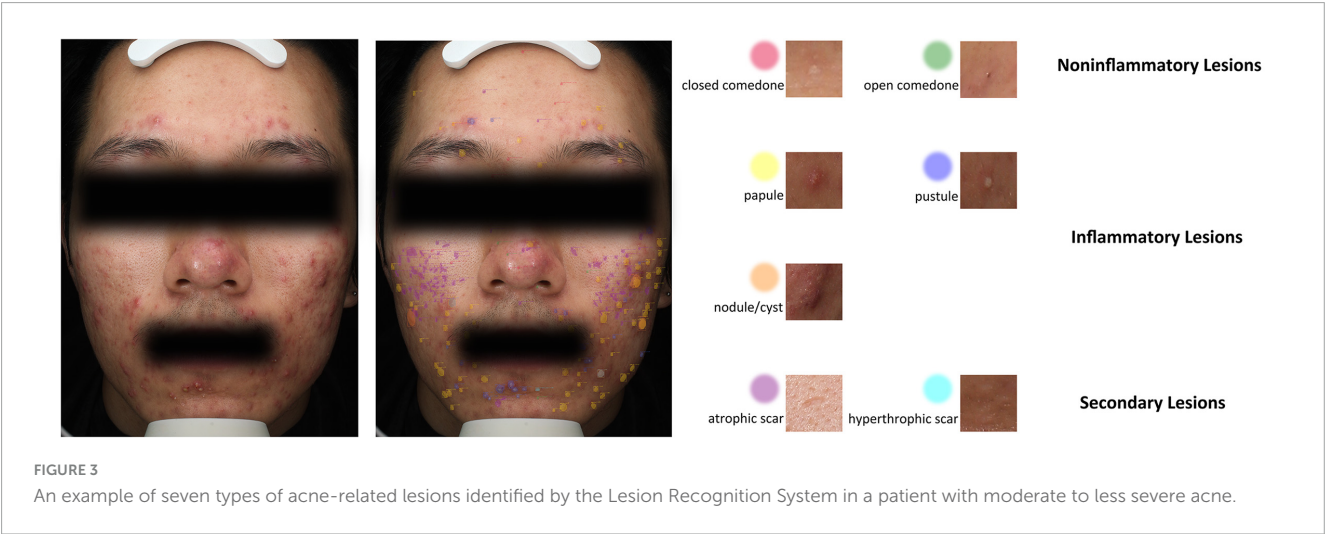
For the development and validation of the severity grading systems, totally 945 images were used for training and 240 for testing, and the kappa obtained by the AI system relative to the true label was 0.652 ([Figure 4A](#)). To further enhance the predictive power, we further constructed a fixed-weight model a learnable-weight model to integrate the lesion identification results of papule, pustule and nodule/cyst into the severity assessment based on lesion identification platform, which improved the kappa relative to the true label to 0.737 and 0.696, respectively ([Figures 4B, C](#)). The 40 images that were initially used to select database annotators were applied to the three models, and the mean pairwise kappa achieved by the three AI models ranked 7th, 2nd and 4th ([Figure 5](#)).

Discussion

In this study, we found that the artificial intelligence acne severity evaluation system we developed produced a reasonable

TABLE 2 Baseline characteristics.

	Rater selection subset (n = 40)	Lesion identification subset (n = 276)		Severity grading subset (n = 1185)	
		Training subset (n = 248)	Test subset (n = 28)	Training subset (n = 945)	Test subset (n = 240)
Age, years					
<20	4	43	3	190	45
20–29	29	164	23	633	161
30–39	5	39	2	109	30
40–49	2	2	0	13	4
Sex					
Female	23	165	19	621	142
Male	17	83	9	324	98
Severity (true label)					
Clear	/	5	1	34	11
Almost clear	/	28	3	178	37
Mild	/	87	7	321	94
Mild to moderate	/	57	5	172	40
Moderate	/	41	3	149	35
Moderate to less severe		20	1	68	15
Severe	/	7	1	14	9
Severe or very severe	/	3	0	9	3



evaluation of the frontal part of acne patients' photos, and its evaluation results were in good agreement with the true labels. Furthermore, we innovatively incorporated the lesion identification results into the severity evaluation with fixed weights and learnable weights, which improved the performance of the model. The AI system, whether weighted or not, can grade acne within the performance range of experienced dermatologists.

Artificial intelligence has powerful learning capabilities that enable it to capture the nuances of lesion images, including size, color and texture, etc (19). The morphological manifestation of the lesion is an important basis for diagnosing and evaluating dermatologic diseases, making AI even more distinctive in

dermatology (20). Currently, AI research in dermatology is focused on multiclassification tasks (21, 22) for disease diagnosis and binary classification (23, 24) for benign or malignant skin lesions, but the evaluation of the severity of a specific disease is also a research direction with great potential for application. The high prevalence and the lack of widely accepted evaluation criteria make acne a perfect fit for AI research. As the eighth most prevalent disease in the world, acne creates a medical need that cannot be met due to the current shortage and uneven distribution of dermatologists. AI can act as a decision aid for clinicians to improve the efficiency of evaluation, particularly in the identification and counting of acne lesions. In recent years,

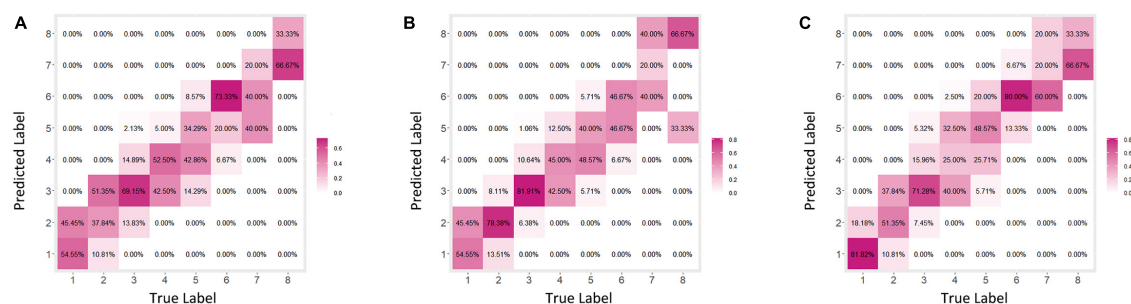


FIGURE 4

Confusion matrices for acne grading. (A) Original model. (B) Fixed weight model. (C) Learnable weight model.

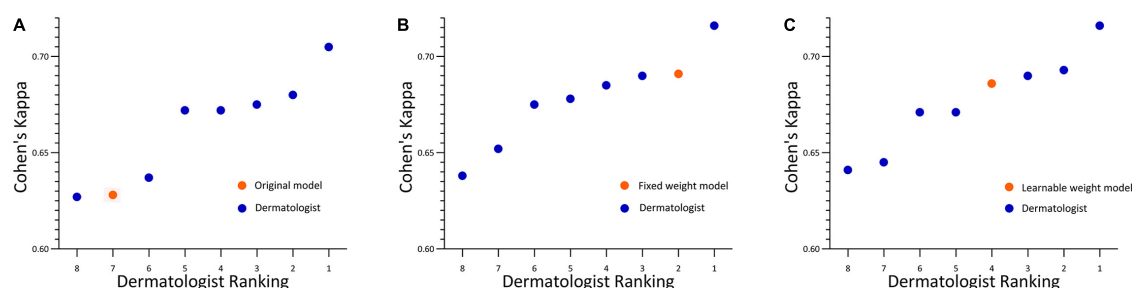


FIGURE 5

Acne grading performance on rater selection subsets. (A) Original model. (B) Fixed weight model. (C) Learnable weight model. Linear weighted Cohen's kappa for each pathologist ranked from lowest to the highest. Each kappa value is the average pairwise kappa for each of the dermatologists compared with the others. The AI is highlighted with an orange dot.

many advances have also been made in the evaluation of acne by AI. Sophie Seit  made several optimizations to their model to improve the recognition of inflammatory and non-inflammatory acne lesions, and their model achieves a GEA score similar to that of the dermatologists (13). Quan Thanh Huynh applied different models to complete the identification of acne lesions and the evaluation of severity with good accuracy, but their study did not incorporate the results of lesion identification into the severity evaluation (12). To the best of our knowledge, no previous studies have integrated skin lesion identification with severity assessment and consequently improved the accuracy of severity assessment. According to the principles of AI, skin lesion identification may no longer be important for severity evaluation when the sample size is sufficiently large, however, for more limited sample sizes, lesion identification can emphasize important information in the evaluation of severity and make the results more interpretable by doctors.

One of the major strengths of our study is that we have a much more detailed classification of severity (eight scales) than what is used by other common scales. One study found that the interobserver agreement using a crude acne severity scale was quite low (25). In order to improve interrater agreement, we referenced the treatment intensity label used by the Elena Bernardis's study to represent acne severity (11). The physicians in this study strongly endorsed the logic of this intensity label after discussion, although it differed slightly from the current Chinese Guidelines for the Management of Acne Vulgaris and medication habits of Chinese dermatologists. The use of treatment intensity for labeling, in

addition to increasing interrater consistency, provides doctors with an indication of the patient's treatment regimen. However, the doctors will need to take into account other information about the patient as well as the results prompted by the AI, because our model does not consider patient information outside of the image data, including but not limited to pregnancy and breastfeeding status, drug allergy history, financial situation, personal wishes etc. In addition we are more rigorous in testing of the models. Besides comparing the differences between the AI model predictions and the true labels, this study compared the AI predictions with the ratings of several experienced dermatologists. This step is important for grading systems that lack objective indicators such as acne severity.

Our study also suffered from a number of shortcomings. First, all of the patients we included were Chinese, and although there were different ethnic groups, all of the patients had skin types II to IV; thus, further validation of our model's ability to identify lesions and evaluate severity in patients with other skin types is needed. Second, our samples were sourced from hospital specialist clinics, and due to the low willingness of mild patients to seek treatment and the small proportion of patients with extremely severe illnesses, our sample is not evenly distributed at different levels. Finally, to obtain more reliable results, we included only patients with a diagnosis of acne and no other facial inflammatory diseases; however, in the real world acne is not exclusive to diseases such as rosacea and seborrheic dermatitis, and the AI evaluation for this group of patients requires a broader sample resource.

Conclusion

This study developed a system that detects various types of acne lesions and correlates them well with acne severity grading, and the good accuracy and efficiency make this approach potentially a very effective clinical decision support tool. However, further research is needed to validate the effectiveness of this AI system in real-world clinical settings.

Data availability statement

The original contributions presented in this study are included in this article/**Supplementary material**, further inquiries can be directed to the corresponding author.

Ethics statement

This study was approved by the West China Hospital Institutional Review Board to use the patients' de-identified images and records.

Author contributions

JL: Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Validation, Visualization, Writing – original draft. DD: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Validation, Writing – original draft. JZ: Methodology, Project administration, Software, Writing – original draft. WL: Methodology, Project administration, Software, Writing – original draft. JW: Methodology, Project administration, Software, Writing – original draft. XW: Methodology, Project administration, Software, Writing – original draft. LX:

Conceptualization, Investigation, Writing – original draft. XL: Investigation, Writing – original draft. PD: Investigation, Writing – original draft. LZ: Software, Supervision, Writing – review and editing. XJ: Funding acquisition, Resources, Supervision, Writing – review and editing.

Funding

This manuscript was supported by grants from the Med-X Center for Informatics Funding Project, Sichuan University (YGJC-003).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2023.1255704/full#supplementary-material>

References

1. Tan J, Bhate KA. Global perspective on the epidemiology of acne. *Br J Dermatol*. (2015) 172(Suppl. 1):3–12. doi: 10.1111/bjd.13462
2. Karimkhani C, Dellavalle R, Coffeng L, Flohr C, Hay R, Langan S, et al. Global skin disease morbidity and mortality: an update from the Global Burden of Disease Study 2013. *JAMA Dermatol*. (2017) 153:406–12. doi: 10.1001/jamadermatol.2016.5538
3. Bickers D, Lim H, Margolis D, Weinstock M, Goodman C, Faulkner E, et al. The burden of skin diseases: 2004 a joint project of the american academy of dermatology association and the society for investigative dermatology. *J Am Acad Dermatol*. (2006) 55:490–500. doi: 10.1016/j.jaad.2006.05.048
4. Layton A, Thiboutot D, Tan J. Reviewing the global burden of acne: how could we improve care to reduce the burden? *Br J Dermatol*. (2021) 184:219–25. doi: 10.1111/bjd.19477
5. Chilicka K, Rusztowicz M, Rogowska A, Szygula R, Nowicka D. Efficacy of oxyboration and cosmetic acids on selected skin parameters in the treatment with acne vulgaris. *Clin Cosmet Investig Dermatol*. (2023) 16:1309–17. doi: 10.2147/ccid.S407976
6. Eichenfield D, Sprague J, Eichenfield L. Management of acne vulgaris: a review. *JAMA*. (2021) 326:2055–67. doi: 10.1001/jama.2021.17633
7. Doshi A, Zaheer A, Stiller MJA. Comparison of current acne grading systems and proposal of a novel system. *Int J Dermatol*. (1997) 36:416–8. doi: 10.1046/j.1365-4362.1997.00099.x
8. Tan J, Jones E, Allen E, Pripotnev S, Raza A, Wolfe B. Evaluation of essential clinical components and features of current acne global grading scales. *J Am Acad Dermatol*. (2013) 69:754–61. doi: 10.1016/j.jaad.2013.07.029
9. Lucky A, Barber B, Girman C, Williams J, Ratterman J, Waldstreicher JA. Multirater validation study to assess the reliability of acne lesion counting. *J Am Acad Dermatol*. (1996) 35:559–65. doi: 10.1016/s0190-9622(96)90680-5
10. US Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research [CDER]. *Guidance for Industry: Acne Vulgaris: Developing Drugs for Treatment*. Washington, DC: US Department of Health and Human Services (2005)
11. Bernardis E, Shou H, Barbieri J, McMahon P, Perman M, Rola L, et al. Development and initial validation of a multidimensional acne global grading system integrating primary lesions and secondary changes. *JAMA Dermatol*. (2020) 156:296–302. doi: 10.1001/jamadermatol.2019.4668
12. Huynh Q, Nguyen P, Le H, Ngo L, Trinh N, Tran M, et al. Automatic acne object detection and acne severity grading using smartphone images and artificial intelligence. *Diagnostics*. (2022) 12:1879. doi: 10.3390/diagnostics12081879
13. Seite S, Khammari A, Benzaquen M, Moyal D, Dreno B. Development and accuracy of an artificial intelligence algorithm for acne grading from smartphone photographs. *Exp Dermatol*. (2019) 28:1252–7. doi: 10.1111/exd.14022

14. Yang Y, Guo L, Wu Q, Zhang M, Zeng R, Ding H, et al. Construction and evaluation of a deep learning model for assessing acne vulgaris using clinical images. *Dermatol Ther.* (2021) 11:1239–48. doi: 10.1007/s13555-021-00541-9
15. Lim Z, Akram F, Ngo C, Winarto A, Lee W, Liang K, et al. Automated grading of acne vulgaris by deep learning with convolutional neural networks. *Skin Res Technol.* (2020) 26:187–92. doi: 10.1111/srt.12794
16. Daneshjou R, Barata C, Betz-Stablein B, Celebi M, Codella N, Combalia M, 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) 158:90–6. doi: 10.1001/jamadermatol.2021.4915
17. Zhang J, Zhang L, Wang J, Wei X, Li J, Jiang X, et al. *Learning High-Quality Proposals for Acne Detection.* (2022). Available online at: <https://ui.adsabs.harvard.edu/abs/2022arXiv220703674Z> (accessed July 01, 2022).
18. He K, Zhang X, Ren S, Sun J editors. Deep residual learning for image recognition. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* London (2016).
19. Huang W, Tan K, Hu J, Zhang Z, Dong S. A review of fusion methods for omics and imaging data. *IEEE ACM Trans Comput Biol Bioinform.* (2022) 20:74–93. doi: 10.1109/tcbb.2022.3143900
20. Bajaj S, Marchetti M, Navarrete-Dechent C, Dusza S, Kose K, Marghoob A. The role of color and morphologic characteristics in dermoscopic diagnosis. *JAMA Dermatol.* (2016) 152:676–82. doi: 10.1001/jamadermatol.2016.0270
21. Liu Y, Jain A, Eng C, Way D, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med.* (2020) 26:900–8. doi: 10.1038/s41591-020-0842-3
22. Esteva A, Kuprel B, Novoa R, Ko J, Swetter S, Blau H, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* (2017) 542:115–8. doi: 10.1038/nature21056
23. Haenssle H, Fink C, Toberer F, Winkler J, Stolz W, Deinlein T, et al. Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. *Ann Oncol.* (2020) 31:137–43. doi: 10.1016/j.annonc.2019.10.013
24. Haenssle H, 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) 29:1836–42. doi: 10.1093/annonc/mdy166
25. Beylot C, Chivot M, Faure M, Pawin H, Poli F, Revuz J, et al. Inter-observer agreement on acne severity based on facial photographs. *J Eur Acad Dermatol Venereol.* (2010) 24:196–8. doi: 10.1111/j.1468-3083.2009.03278.x



OPEN ACCESS

EDITED BY

Mara Giavina-Bianchi,
Albert Einstein Israelite Hospital, Brazil

REVIEWED BY

Federica Veronese,
Azienda Ospedaliero Universitaria Maggiore
della Carità, Italy
Ionela Manole,
Colentina Clinical Hospital, Romania

*CORRESPONDENCE

Jesutofunmi A. Omiye
✉ tomiye@stanford.edu

[†]These authors have contributed equally to this work and share first authorship

[†]These authors have contributed equally to this work and share senior authorship

RECEIVED 16 August 2023

ACCEPTED 27 September 2023

PUBLISHED 12 October 2023

CITATION

Omiye JA, Gui H, Daneshjou R, Cai ZR and Muralidharan V (2023) Principles, applications, and future of artificial intelligence in dermatology.
Front. Med. 10:1278232.
doi: 10.3389/fmed.2023.1278232

COPYRIGHT

© 2023 Omiye, Gui, Daneshjou, Cai and Muralidharan. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Principles, applications, and future of artificial intelligence in dermatology

Jesutofunmi A. Omiye^{1*}, Haiwen Gui^{1†}, Roxana Daneshjou^{1,2},
Zhuo Ran Cai^{1†} and Vijaytha Muralidharan^{1†}

¹Department of Dermatology, Stanford University, Stanford, CA, United States, ²Department of Biomedical Data Science, Stanford University, Stanford, CA, United States

This paper provides an overview of artificial-intelligence (AI), as applied to dermatology. We focus our discussion on methodology, AI applications for various skin diseases, limitations, and future opportunities. We review how the current image-based models are being implemented in dermatology across disease subsets, and highlight the challenges facing widespread adoption. Additionally, we discuss how the future of AI in dermatology might evolve and the emerging paradigm of large language, and multi-modal models to emphasize the importance of developing responsible, fair, and equitable models in dermatology.

KEYWORDS

dermatology, artificial intelligence (AI), large language models (LLM), machine learning, melanoma, federated learning

1. Introduction

Recent advancements in artificial intelligence (AI) have fueled an interest in the utility of AI models in medicine (1). These models range from computer vision models that can interpret medical images (2) to large language models (LLM) that have capabilities for analyzing text data (3, 4) to multi-modal models that take both images and text as input (5). These AI models now have the capacity to analyze unstructured data such as clinical notes (3, 6), identify novel correlations in large datasets (7), and generate synthetic image data for improving model training (8, 9).

One medical specialty poised to benefit from these emerging AI technologies is dermatology. Its inherent visual diagnostic process, combined with an increasing volume of clinical photographs, dermoscopy images, and electronic health records (EHR) data (10) underscores its suitability for AI-augmented patient care. Moreover, the shortage of specialists—3.65 dermatologists per 100,000 people in the US (11, 12) and limited access to dermatological services in many regions (13, 14) provides a compelling case for augmented intelligent systems to help bridge this access gap (15). However, clinical integration of AI in dermatology workflow remains challenging. As novel medical applications arise, they also unveil problems that necessitate further research.

In this paper, we present a comprehensive overview of the fundamental principles of AI methodology as applied to dermatology, diving into categories and training approaches. Special emphasis is placed on the role of AI in the diagnosis and prognostication of an array of skin conditions. We also address the limitations of the AI models used in dermatology, notably issues of generalizability, bias, and explainability. Finally, we examine what the future might hold for dermatology-AI, while highlighting some research opportunities to help improve real-world utility of AI models. Our goal is to provide the readers with a panoramic view of AI's principles

and evolving role in dermatology, while equipping them with the knowledge to navigate this dynamic field.

2. Principles of artificial intelligence

AI is the ability of a computer system to mimic human cognitive functions and encompasses many computational subfields, including machine learning and natural language processing (Figure 1). Currently, major developments in AI are within the field of machine learning (ML), which are algorithms that make predictions about data without explicit programming. In other words, the machines are “learning” from the data and providing analyses without being explicitly told what features to prioritize. Examples from dermatology include identifying melanomas from clinical images (16), predicting efficacy of biologic therapies in psoriasis (17), and analyzing physician notes in electronic health records to determine focus of atopic dermatitis clinic visits (18).

Deep learning (DL) (19) is a subset of ML that uses algorithms modeled off human neurons that can model complex patterns and relationships in the data. ML techniques prior to the introduction of DL required domain expertise and human engineering to convert raw data into features that the algorithm can understand and detect patterns from. On the other hand, in DL, raw data can be inputted into the algorithm, and the machine is able to create its own representation needed for pattern recognition. These representations are typically arranged in sequential layers, where each layer is inputted into the next layer, increasing the abstraction of the data, collectively known as neural networks (6) (Figure 2). Within DL, there are multiple algorithms that are implemented, including convolutional neural networks (CNN) (20), traditionally used in image processing, and transformer models (21), which are neural networks that learn context and track relationships in sequential data.

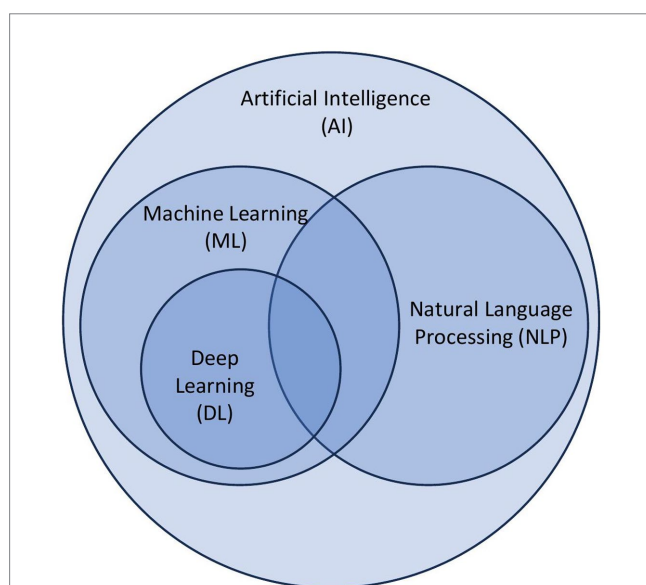


FIGURE 1
Overview of principles of artificial intelligence. Artificial intelligence (AI) is a broad categorization of algorithms that encompass subcategories including machine learning (ML), natural language processing (NLP), and deep learning (DL).

Within ML, there are different ways that algorithms can learn, including supervised learning, unsupervised learning, and reinforcement learning (Figure 2). Supervised learning, the most common form of machine learning, uses a labeled dataset to predict results. The algorithm learns to map the input data to the correct output, allowing it to make predictions on unseen data. The algorithm is given the data and the correct answers (ground truths) in a training set, which the algorithm uses to set its weights. Once the algorithm has learned from the training data, its performance is measured against a held-out test set that it has never encountered previously. This category of machine learning includes what most people are familiar with, such as logistic regression, linear regression, etc. Most of the image-based deep learning models in dermatology use supervised learning. Unsupervised learning is training a model on unlabeled datasets, meaning the data input does not have the ground truth. This algorithm aims to find patterns and relationships within the data, such as clustering similar data points together. Finally, reinforcement learning is when the agent (the algorithm) interacts with an environment to achieve specific goals. The agent receives feedback from the user (the human) in the form of rewards or penalties based on its actions, and it learns to optimize its behavior to maximize rewards. Compared to supervised and unsupervised learning, reinforcement learning has no predefined data input, but rather learns from the iterative feedback loops.

Natural language processing (NLP) is a branch of artificial intelligence that focuses on interpreting, analyzing, and generating human language. It combines linguistics with statistics, machine learning, and DL to process human language (22). NLP is generally divided into two subfields—natural language understanding (NLU), and natural language generation (NLG). NLU is focused on determining the understanding of the text, while NLG is focused on generating new text. Recent advancements in large language models, including OpenAI’s (San Francisco, United States) publicly-available ChatGenerative Pre-trained Transformer (ChatGPT) (23), fall under the subfield of NLG.

There are also recent emerging concepts of multimodal approaches, where algorithms are utilizing multiple data types to train their algorithms. Medicine is inherently a multimodal discipline, with clinicians interpreting lab values, clinical notes, radiology images, genomic data, etc. New development has been focused on utilizing the rich diversity of data to build more robust models and algorithms, including Med-PaLM Multimodal (Med-PaLM M) (24), LLaVa-Med (25), Med-Flamingo (5), and MiniGPT-4 (26). These new technologies are built on foundation models (FMs), which are models that are trained on a broad range of unlabeled data that are then adapted (fine-tuned) to specific downstream applications (27). These models can learn from the large amounts of data, and then transfer their learnings to a more specific application, like medicine.

3. Applications of AI in dermatology

There has been an abundance of work done to explore artificial intelligence use in all aspects of dermatology (28–30), ranging from skin malignancies to inflammatory skin conditions, to dermatopathology, to text-based analyses. The visual nature of Dermatology lends itself to many advancements that are image-based, though researchers are exploring other multimodal approaches that

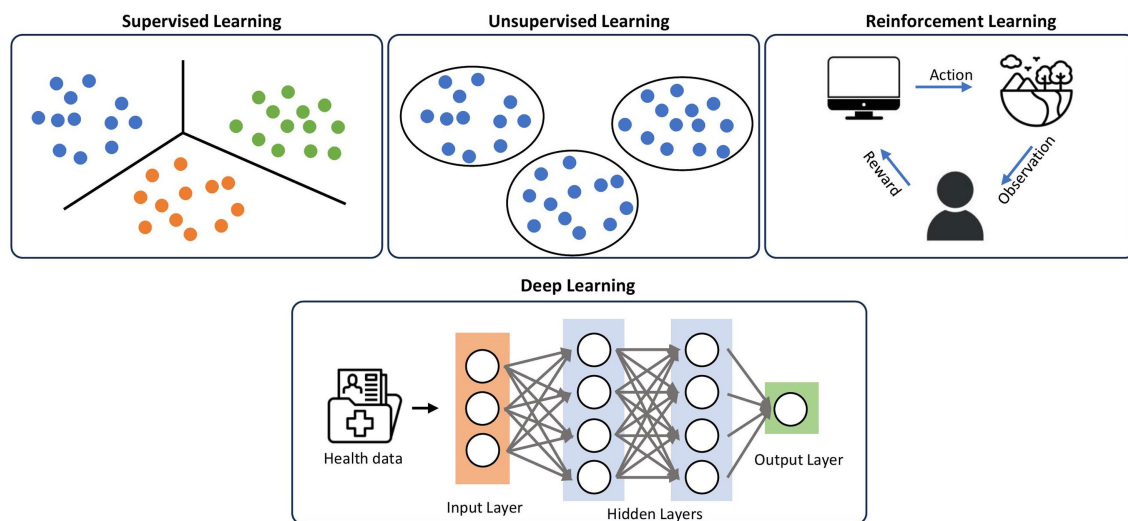


FIGURE 2

Classifications of machine learning. Supervised learning uses labeled datasets to categorize the data, while unsupervised learning does not have labeled datasets, using patterns and relationships in the data to create categories. Reinforcement learning uses iterative feedback loops to teach the algorithm. DL utilizes representation layers in a neural network to increase abstraction of the data, and employs techniques from supervised, unsupervised, and reinforcement learning.

use patient characteristics and clinical texts. Here, we will provide a broad overview of the different applications of AI in dermatology.

3.1. Skin malignancies

Applications of AI in dermatologic malignancies, which have been well described in the literature (31, 32), include identifying and distinguishing between benign nevi and melanoma. Researchers break down images of skin lesions to the pixel level for individual analysis and then utilize the techniques described above to predict and classify malignancies. There have been multiple landmark papers for AI applications in skin malignancies (16, 31, 33), resulting in high sensitivities and specificities when distinguishing malignant from benign lesions. Esteva et al. trained a CNN using a large dataset of over 100,000 biopsy-proven clinical images to determine keratinocyte carcinomas versus benign seborrheic keratoses, and malignant melanomas versus benign nevi (16). Han et al. fine-tuned a previously-built CNN model with clinical images to classify multiple malignancies, including basal cell carcinoma, squamous cell carcinoma, melanoma, etc. (34). There is also an annual international skin imaging competition, which provides publicly accessible dermatology images for researchers to build melanoma-classifying models (32, 35). Aside from identifying the primary lesion, there are also studies exploring metastases. Jansen et al. utilized histological tissue sections of sentinel lymph nodes in their convolutional neural network models to identify presence of metastases with high sensitivity and specificity (36).

3.2. Inflammatory skin diseases-psoriasis, dermatitis, and others

Aside from classification of melanomas and other malignant skin conditions, researchers are also exploring the identification and

management of inflammatory skin conditions, including psoriasis, dermatitis and acne. Similar to malignancy classification, a majority of work is focused on psoriasis identification and classification through images of skin (37–40), nails (41), and scalp (42) using CNN's and other DL techniques. In addition to diagnosing psoriasis via image recognition, researchers have utilized machine learning techniques to identify patients with increased risk of associated psoriatic conditions, including psoriatic arthritis (43). Work has also been done to determine the efficacy of psoriasis management by predicting outcomes of biologic therapies by using parameters such as patient demographics, clinical history of psoriasis, treatment history, and presence of other comorbidities (17, 44). These preliminary models could be used to eventually optimize therapy and management for patients. Finally, aside from determining outcomes of current biological treatments, AI techniques have been applied to genomic studies to help with drug target identification and drug repurposing (45), as well as screening for psoriasis biomarkers (46) and gene expression profiling (47).

Similar to diagnostic tasks with psoriasis, many researchers have explored using machine learning algorithms in dermatitis (48), ranging from image-based algorithms (49) to electronic health record text-based algorithms (50). Aside from determining diagnoses, researchers have developed proof-of-concept algorithms using self-reported eczema flare scores, patient demographics and treatment history to predict atopic dermatitis severity, resulting in a biologically interpretable model that focuses on patient's responsiveness to treatment (51). AI models have also been used to help prevent contact dermatitis by predicting skin sensitization potential and potency of substances (52).

In addition to psoriasis and dermatitis, researchers have developed acne lesion segmentation and evaluation tools (53–55) that can grade acne severity from easily-accessible smartphone images (56). There is also exploration in identifying lichen planus (41), and assessing the severity of hidradenitis suppurativa (57).

3.3. Ulcer assessment

One of the primary methods in identifying and classifying skin lesions is segmenting the lesion from the backdrop of normal skin. Multiple studies have explored determining and measuring ill-defined wound boundaries using techniques that simplify images down to the pixel level (58–61). Recent work has been done to apply these techniques into broader hospital systems to predict pressure ulcers (62), with the ultimate goal of pressure ulcer prevention (63, 64). Groups have even explored using body heat maps from pressure mats to identify poor in-bed position posture that could cause pressure ulcers (65). These proof-of-concept works, after validation in clinical trials, may ultimately translate into clinical-assist tools to aid clinicians in the management of ulcers.

3.4. Dermatopathology

Beyond identifying diagnoses via clinical images and electronic health record notes, machine learning techniques are being applied in dermatopathology (66, 67). Groups have developed models to classify basal cell carcinoma in digitized Mohs micrographic surgery histology slides to reduce the workload of manually examining these slides (68). Likewise, Hekler et al. used CNNs to aid in histopathologic melanoma diagnoses (69). There have also been studies done to interpret indirect immunofluorescence microscopies to classify bullous dermatoses (70).

3.5. Miscellaneous multiclass classification and text-based analysis

To better replicate real-world clinical scenarios of multiple differential diagnoses from a single skin lesion, technologies take a broader approach to solve multi-class classification problems. Many of the problems discussed above were binary classification, where algorithms strived to identify if a lesion was a specific disease or not; multi-class classification presents a more challenging problem with multiple possible diagnoses. Liu et al. created a DL system that provided a differential diagnosis for skin lesions, creating a ranked list of the most likely diagnoses for the skin lesion (71). Taking another multi-class approach, Sitaru et al. have worked to classify body parts from dermatology clinical images, creating body distribution maps for different diagnoses (72).

While dermatology is a visual specialty that focuses on using visual cues for diagnoses, there are aspects of written data that can be used to aid in better understanding questions posed by the research community. Frequently, this written data is unstructured and freeform, using natural human language; to understand and interpret this data, one needs to implement NLP techniques. Researchers have used NLP methods to examine dermatology discussion forums on social media to understand patient perceptions of the field (73). Others conducted analyses of clinical notes in the electronic health records to identify specific topics that providers and patients discuss during clinical visits (18). This analysis provided insights into the lack of documentation of the disease's impact on a patient's life, which may ultimately affect management and treatment options.

In addition to understanding natural language, there are also recent advances in technologies that generate new text, including ChatGPT. This technology could be utilized to guide patients, aid

clinicians with administrative tasks, educate trainees, etc. Groups are exploring ChatGPT's ability to generate responses to patient inquiries about melanoma (74), create patient education guides for acne (75), and even triage surgical management of cutaneous neoplasms (76).

3.6. Human-AI hybrid models

Given all the innovation that is occurring at the intersection of AI and dermatology, the logical next step is to evaluate the performance of these AI algorithms against clinicians (77, 78). Esteva et al.'s landmark study was the first to compare a DL algorithm against dermatologists, showing that their model was able to match the performance of 21 dermatologists in melanoma classification (16). Others have even shown that, in a group of 58 international dermatologists, many were outperformed by a CNN model (79). Because of the incredible ability of the technology to perform diagnostic tasks, many researchers are exploring ways to incorporate AI in a clinical workflow to help clinicians. There have now been multiple studies creating AI-based assistive tools to aid clinicians in interpreting clinical images. Groups have designed pipelines with the ultimate goal of real-time AI analysis of skin lesions in the clinics (77). Marchetti et al. prospectively assessed the diagnostic accuracy and utility of a melanoma AI algorithm used in real-world clinical settings to help determine the necessity of biopsying a suspicious lesion (33). Han et al. conducted a randomized trial and showed that AI can augment the accuracy of non-expert physicians in the real-world setting (80). With these smaller pilot studies showing promising results, the AI research community may be looking to increase prospective studies and randomized trials to help further assess AI's application in the real-world clinical setting.

4. Limitations and ethical considerations of AI in dermatology

AI research in dermatology is still in its infancy and encounters a myriad of challenges. From biases and lack of interpretability to regulatory hurdles and difficulty in integrating with existing clinical workflows, these issues are complex and need to be tackled before AI can become ubiquitous in clinical practice. Robust, transparent, and equitable AI algorithms are needed in order to truly enhance patient care without introducing new problems.

4.1. Datasets

AI algorithms learn by identifying features and patterns found in their training datasets and then use this knowledge to make future predictions. However, the presence of confounders in these datasets can influence the validity of AI algorithms. Confounders are features that may be correlated with the AI algorithm's outcome through spurious associations. An illustrative example involves the presence of surgical pen markers or rulers on clinical dermatology images. As demonstrated in research by Winkler et al., lesions marked with surgical pen markers are more likely to be classified as malignant by the models (81). This finding is due to the fact that these markings are frequently used during biopsy procedures, which are typically performed on lesions suspected of malignancy. Therefore, an algorithm may incorrectly learn an

association between these markers and malignancy, when in fact the markings only indicate which lesions were biopsied, not necessarily those that are malignant. This example highlights the importance of identifying and managing confounders during the training phase of AI models to ensure their accuracy and validity.

Bias in training datasets can also inadvertently perpetuate pre-existing inequities in healthcare. In dermatology, this issue is particularly highlighted by early AI models trained on datasets that predominantly featured lighter Fitzpatrick skin types (I-IV). Daneshjou et al. has shown that some of these existing algorithms tend to underperform when assessed with images of darker Fitzpatrick skin types (V-VI) (15). Fortunately, fine-tuning these original algorithms with a dataset featuring darker Fitzpatrick skin types improved their performance, effectively closing the gap in performance between different skin types. Diverse and equitable data representation in the training dataset is primordial to ensure accurate and fair outputs in AI algorithms.

4.2. Image quality and image capturing modalities

Standardizing images in Dermatology AI research is important to preserve data quality. Images can originate from diverse sources, including various devices (e.g., iPhones, Android smartphones, or professional cameras) and with or without the help of diagnostic tools such as dermatoscopes (82). Additionally, the images may be captured under various settings (e.g., at home or in a clinic) and by different individuals (e.g., patients or healthcare providers). These factors result in a highly heterogeneous dataset comprising images of differing quality. Just as human interpretation can be affected by image quality, AI algorithms are equally sensitive. Blurry images with poor lighting have been shown to negatively impact the performance of AI models (83). Simple image manipulations such as rotation can change the output of an algorithm (84). These considerations underscore the importance of establishing robust image capturing standards and Digital Imaging and Communications in Medicine (DICOM) standards similar to those in other medical fields such as cardiology and radiology (85).

4.3. Black box

The mechanisms behind traditional medical devices are often transparent and logical in nature. In contrast, AI algorithms appear more mysterious and impenetrable, like a “black box.” This phenomenon makes it difficult for humans to understand its reasoning process and to trust its outputs. Various techniques have been developed by researchers to tackle this problem including saliency maps (e.g., highlighting relevant areas on a picture) and content-based image retrieval (e.g., retrieving similar images from a database based on the query image). As AI penetrates high stake fields such as medicine, it becomes increasingly important to bring transparency and interpretability to AI models.

4.4. Implementation

Implementing AI into clinical practice presents a number of challenges that extend beyond technological complexity. The rapid advancement of AI technologies has created a complex landscape of

medical-legal challenges regarding its use in the healthcare sector, spanning from concerns about patient consent and data privacy to liability in the event of AI-induced medical errors (86, 87). Scholars and professionals must work collaboratively to devise sound and comprehensive guidance to navigate the ethical and legal intricacies of integrating AI into our healthcare systems (87). Medical AI devices, by their very nature, will evolve as they learn from newly acquired data, a process that may continue long after receiving approval from regulatory bodies such as the Food and Drug Administration (FDA). This continual learning and adaptation, while a strength in many respects, also presents a challenge in ensuring the devices' sustained reliability and performance over time. Without vigilant monitoring and a robust framework for ongoing validation, there may be unforeseen shifts in the accuracy or effectiveness of these tools, which could potentially negatively impact patient care. Moreover, there is a lack of high quality prospective randomized controlled trials of AI algorithms. While AI holds immense promise in dermatology, the absence of prospective trials hinders the validation of AI models in real-world clinical situations where there will be a diverse photo quality, image capturing modalities and demographically diverse population (33). For these reasons, an AI model validated in a hospital in Asia might not perform similarly in another hospital in North America. Wu et al. have shown that 126 out of 130 FDA approved medical AI devices were trained on retrospective data at the time of their approval (88). Most of the datasets used are not publicly available, thus preventing regulatory bodies and researchers from auditing their algorithms. Future AI models should undergo multi-site validation on a diverse and representative population in order to assess the generalizability of AI models. Furthermore, establishing trust among AI and various stakeholders will be vital in realizing AI's full potential in the field. While model accuracy is very important, research has shown that dermatologists and patients value the potential of augmented intelligence in dermatology and also put a high priority on the human physician-patient relationship (89, 90).

5. Future directions and opportunities

5.1. LLMs and the advent of generalist medical AI

In recent months, advanced language models, in the form of chatbots, have gained popularity in medicine (4, 91–93). For dermatology, an extension of these models—Vision-Language Models (VLMs) and multi-modal models—offer immense potential. VLMs are large-scale models adept at associating visual inputs, such as images and videos, with text data (5). Their capabilities span generative tasks (creating new content), retrieval of information, and navigation. Recent studies underscore their impact on dermatology. For instance, Skin-GPT4, a VLM, can provide descriptions and diagnosis from clinical skin lesion photos (94). Further, research by Moor et al. and Tu et al. show the accuracy of VLMs in medical visual question-answering tasks (5, 24). In a related vein, Kim et al.'s study on FMs underscore the capacity of this new class of models to generate accurate skin images annotations (95).

The rapid advancements in this domain have the potential to usher a future of a generalist medical AI (96). These generalist models could be capable of giving approximate diagnoses from clinical photos, generate treatment options, and offer deeper insights into

patient data by integrating demographics, visual inspection, and genetic data when applicable. Their potential applications can range from patient chatbots to triage tools (96). Additionally, the inclusion of genetic data could improve the diagnosis of orphan skin conditions. As dermatological datasets expand and computing power increases, FMs are on track to become more accurate and prove utility in dermatology. They could augment the practice of dermatology to provide more precise and holistic care.

5.2. Federated learning and the possibility of local models

Medical data, including skin images, are difficult to access largely due to privacy, legal, and the ethical risks associated with sharing health data. Currently, many dermatological images reside in data silos within healthcare institutions all over the world. Also, medical data is hard to collate, and often requires years of planning with significant costs (97). This is even more pronounced in resource-limited settings, where there is less infrastructure to support collection and sharing of data. Since the DL model's performance significantly improves with more diverse data (98, 99), new approaches are needed to expand model access to more distributed high-quality datasets.

Federated learning (FL) is a concept that enables DL models to be trained on different datasets without the need to leave their original locations (100). In FL, multiple collaborators can train a model on separate institutional datasets. It is an approach that can enable the preservation of data privacy, and it has already demonstrated similar performance—compared to centralizing the data—in fields like radiology and oncology (100). Although in some cases, there have been drawbacks in which the model sometimes memorizes the data inputs (101). Appropriately implemented, FL has the potential to enable fairer and more generalizable dermatology models by incorporating diverse demographics, thereby capturing the nuances in skin conditions across different societies. This is crucial in dermatology where the popular models significantly perform worse on underrepresented skin types of Fitzpatrick IV–VI (15).

Beyond FL, the concept of FMs introduces the possibility of local models. FMs have the distinct capability to learn from unlabeled data and can be adapted for a variety of downstream tasks without the necessity of specific training (96). This characteristic allows FMs to be fine-tuned with local data, from which they can glean insights and achieve impressive performance across diverse tasks. Given that the fine-tuning procedure is more cost-efficient than full-scale training (27), it amplifies the appeal of FMs within institutional contexts. Consequently, dermatology institutions can harness bespoke models attuned to their unique demographics and guidelines. While promising, progress towards this will require resolving data quality, aggregation, and infrastructural challenges. However, these new techniques could be instrumental in building invaluable dermatology-AI models.

5.3. Improvements in model architecture and metrics evaluation

Recent years have witnessed notable advancements in the architectures of AI models, leading to enhanced performance across

numerous medical tasks as previously discussed. As the industry attracts more investment and data generation surges, new architectures will likely further improve on existing tasks and expand into new areas. However, with these advancements arises a vital question: how should we holistically evaluate these models? While metrics like accuracy, area under the curve are common, comprehensive model evaluation will need to go beyond mere percentages. Clinical value needs to be demonstrated. As reported by Wornow et al., standard evaluations are lacking for evaluating emerging models (102). In addition, many models fail to be evaluated on fairness and transparency metrics, and in many cases there's no standard for this evaluation frameworks (103). Holistic model evaluation is likely to emerge in the near future as the desire for clinical integration increases. This could include uncertainty, model interpretability, and subpopulation analysis—which is important for dermatology.

Developing these types of model benchmarks will require collaboration among dermatologists, researchers, and patients. We posit that soon, more robust consensus guidelines are likely to emerge.

5.4. Regulation, clinical utility, and usability in resource-poor settings

The current rapid model evolution underscores the pressing need for robust regulation (104). Such regulatory measures serve a two-fold purpose: Firstly, they shield the dermatology community from prematurely adopting under-tested models by establishing stringent benchmarks. Secondly, they foster trust, ensuring that AI tools resonate with the foundational clinical values practitioners hold dear. For AI models to achieve widespread adoption, especially in dermatology, they must be both reproducible and generalizable (105). As these models seek to bridge the dermatology access gap, especially in resource-limited settings, generalizability becomes even more pivotal. Also, standardizing the data collection process is another important factor towards optimizing model training and thus, performance. As highlighted in the position statement by the American Academy of Dermatology Augmented Intelligence working group, research in the dermatology-AI space needs to be directed towards prospective and randomized clinical trials that rigorously vet models before deployment (29, 106). Also, although most DL models are in the form of “black-boxes” (107), emerging FMs could further obfuscate their inner workings, making the issue of explainability vital. Research addressing explainability will be invaluable for model advancement and deployment.

6. Conclusion

Dermatology presents both opportunities and challenges to integrate AI into its daily workings. Whilst in its infancy, with many regulatory standards that are specific to the field yet to be developed, current trajectories of innovation and advance showcase the potential of AI is likely to emerge a critical element of the dermatologists workflow, with the need for the clinician to have a global understanding of its workings. Steering this ship towards a future of a transparent, fair, safe, and responsible dermatology-AI will be an

interdisciplinary effort that involves the leadership of the dermatology community.

Author contributions

JO: Conceptualization, Project administration, Supervision, Writing – original draft, Writing – review & editing. HG: Conceptualization, Visualization, Writing – original draft, Writing – review & editing. RD: Resources, Supervision, Writing – review & editing. ZC: Supervision, Writing – original draft, Writing – review & editing. VM: Conceptualization, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

References

- Dorr DA, Adams L, Embí P. Harnessing the promise of artificial intelligence responsibly. *JAMA*. (2023) 329:1347–8. doi: 10.1001/jama.2023.2771
- Rajpurkar P, Lungren MP. The current and future state of AI interpretation of medical images. *N Engl J Med*. (2023) 388:1981–90. doi: 10.1056/NEJMra2301725
- Goh KH, Wang L, Yeow AYK, Poh H, Li K, Yeow JLL, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nat Commun*. (2021) 12:711. doi: 10.1038/s41467-021-20910-4
- Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. *N Engl J Med*. (2023) 388:1233–9. doi: 10.1056/NEJMs2214184
- Moor M, Huang Q, Wu S, Yasunaga M, Zakka C, Dalmia Y, et al. (2023) 'Med-flamingo: a multimodal medical few-shot learner', arXiv [cs.CV]. Available at: <http://arxiv.org/abs/2307.15189>.
- Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. *Nat Med*. (2019) 25:24–9. doi: 10.1038/s41591-018-0316-z
- Bohr A, Memarzadeh K. Chapter 2 – The rise of artificial intelligence in healthcare applications In: A Bohr and K Memarzadeh, editors. *Artificial intelligence in healthcare*. Cambridge, MA: Academic Press (2020). 25–60.
- Castiglioni I, Rundo L, Codari M, di Leo G, Salvatore C, Interlenghi M, et al. AI applications to medical images: from machine learning to deep learning. *Phys Med*. (2021) 83:9–24. doi: 10.1016/j.ejomp.2021.02.006
- Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. Synthetic data in machine learning for medicine and healthcare. *Nat Biomed Eng*. (2021) 5:493–7. doi: 10.1038/s41551-021-00751-8
- Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst*. (2014) 2:3. doi: 10.1186/2047-2501-2-3
- Feng H, Berk-Krauss J, Feng PW, Stein JA. Comparison of dermatologist density between urban and rural counties in the United States. *JAMA Dermatol*. (2018) 154:1265–71. doi: 10.1001/jamadermatol.2018.3022
- Kimball AB, Resneck JS Jr. The US dermatology workforce: a specialty remains in shortage. *J Am Acad Dermatol*. (2008) 59:741–5. doi: 10.1016/j.jaad.2008.06.037
- Coustasse A, Sarkar R, Abodunde B, Metzger BJ, Slater CM. Use of Telemedicine to improve dermatological access in rural areas. *Telemed J E Health*. (2019) 25:1022–32. doi: 10.1089/tmj.2018.0130
- Tsang MW, Resneck JS Jr. Even patients with changing moles face long dermatology appointment wait-times: a study of simulated patient calls to dermatologists. *J Am Acad Dermatol*. (2006) 55:54–8. doi: 10.1016/j.jaad.2006.04.001
- Daneshjou R, Voderhals K, Novoa RA, Jenkins M, Liang W, Rotemberg V, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci Adv*. (2022) 8:eabq6147. doi: 10.1126/sciadv.abq6147
- 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) 542:115–8. doi: 10.1038/nature21056
- Emam S, Du AX, Surmanowicz P, Thomsen SF, Greiner R, Gniadecki R. Predicting the long-term outcomes of biologics in patients with psoriasis using machine learning. *Br J Dermatol*. (2020) 182:1305–7. doi: 10.1111/bjd.18741
- Pierce EJ, Boytsov NN, Vasey JJ, Sudaria TC, Liu X, Lavelle KW, et al. A qualitative analysis of provider notes of atopic dermatitis-related visits using natural language processing methods. *Dermatol Ther*. (2021) 11:1305–18. doi: 10.1007/s13555-021-00553-5
- LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. (2015) 521:436–44. doi: 10.1038/nature14539
- Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, et al. Recent advances in convolutional neural networks. *Pattern Recogn*. (2018) 77:354–77. doi: 10.1016/j.patcog.2017.10.013
- Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. (2020) 'Transformers: state-of-the-art natural language processing', in Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. Online: Association for Computational Linguistics, pp. 38–45.
- Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc*. (2011) 18:544–51. doi: 10.1136/amiajnl-2011-000464
- Introducing ChatGPT. (no date). Available at: <https://openai.com/blog/chatgpt> (Accessed November 30, 2022).
- Tu T, Azizi S, Driess D, Schaeckermann M, Amin M, Chang P.-C., et al. (2023) 'Towards generalist biomedical AI', arXiv [cs.CL]. Available at: <http://arxiv.org/abs/2307.14334>.
- Li C, Wong C, Zhang S, Usuyama N, Liu H, Yang J, et al. (2023) 'LLaVA-med: training a large language-and-vision assistant for biomedicine in one day', arXiv [cs.CV]. Available at: <http://arxiv.org/abs/2306.00890>.
- Zhu D, Chen J, Shen X, Li X, Elhoseiny M. (2023) 'MiniGPT-4: enhancing vision-language understanding with advanced large language models', arXiv [cs.CV]. Available at: <http://arxiv.org/abs/2304.10592>.
- Bommasani R, Hudson D. A., Adeli E., Altman R., Arora S., von Arx S., et al. (2021) 'On the opportunities and risks of foundation models', arXiv [cs.LG]. Available at: <http://arxiv.org/abs/2108.07258>
- du-Harpur X, Watt FM, Luscombe NM, Lynch MD. What is AI? Applications of artificial intelligence to dermatology. *Br J Dermatol*. (2020) 183:423–30. doi: 10.1111/bjd.18880
- Gomolin A, Netchiporouk E, Gniadecki R, Litvinov IV. Artificial intelligence applications in dermatology: where do we stand? *Front Med*. (2020) 7:100. doi: 10.3389/fmed.2020.00100
- Young AT, Xiong M, Pfau J, Keiser MJ, Wei ML. Artificial intelligence in dermatology: a primer. *J Invest Dermatol*. (2020) 140:1504–12. doi: 10.1016/j.jid.2020.02.026
- Jones OT, Matin RN, van der Schaar M, Prathivadi Bhayankaram K, Ranmuthu CKI, Islam MS, et al. Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review. *Lancet Digit Health*. (2022) 4:e466–76. doi: 10.1016/S2589-7500(22)00023-1
- Marchetti MA, Codella NCF, Dusza SW, Gutman DA, Helba B, Kallou A, et al. Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: comparison of the accuracy of computer algorithms

Conflict of interest

RD has served as an advisor to MDAlgorithms and Revea and received consulting fees from Pfizer, L'Oreal, Frazier Healthcare Partners, and DWA, and research funding from UCB.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

to dermatologists for the diagnosis of melanoma from dermoscopic images. *J Am Acad Dermatol.* (2018) 78:270–277.e1. doi: 10.1016/j.jaad.2017.08.016

33. Marchetti MA, Cowen EA, Kurtansky NR, Weber J, Dauscher M, DeFazio J, et al. Prospective validation of dermoscopy-based open-source artificial intelligence for melanoma diagnosis (PROVE-AI study). *NPJ Digit Med.* (2023) 6:127. doi: 10.1038/s41746-023-00872-1

34. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J Invest Dermatol.* (2018) 138:1529–38. doi: 10.1016/j.jid.2018.01.028

35. Gutman D., Codella N. C. F., Celebi E., Helba B., Marchetti M., Mishra N., et al. (2016) 'Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC)', arXiv [cs.CV]. Available at: <http://arxiv.org/abs/1605.01397>.

36. Jansen P, Bager DO, Duschner N, Arrastia JLC, Schmidt M, Landsberg J, et al. Deep learning detection of melanoma metastases in lymph nodes. *Eur J Cancer.* (2023) 188:161–70. doi: 10.1016/j.ejca.2023.04.023

37. Lu J, Kazmierczak E, Manton JH, Sinclair R. Automatic segmentation of scaling in 2-D psoriasis skin images. *IEEE Trans Med Imaging.* (2013) 32:719–30. doi: 10.1109/TMI.2012.2236349

38. Mohan S, Kasthuri N. Automatic segmentation of psoriasis skin images using adaptive chimp optimization algorithm-based CNN. *J Digit Imaging.* (2023) 36:1123–36. doi: 10.1007/s10278-022-00765-x

39. Shrivastava VK, Londhe ND, Sonawane RS, Suri JS. Computer-aided diagnosis of psoriasis skin images with HOS, texture and color features: a first comparative study of its kind. *Comput Methods Prog Biomed.* (2016) 126:98–109. doi: 10.1016/j.cmpb.2015.11.013

40. Shrivastava VK, Londhe ND, Sonawane RS, Suri JS. A novel and robust Bayesian approach for segmentation of psoriasis lesions and its risk stratification. *Comput Methods Prog Biomed.* (2017) 150:9–22. doi: 10.1016/j.cmpb.2017.07.011

41. Keser G, Bayraktar İŞ, Pekiner FN, Çelik Ö, Orhan K. A deep learning algorithm for classification of oral lichen planus lesions from photographic images: a retrospective study. *J Stomatol Oral Maxillofac Surg.* (2023) 124:101264. doi: 10.1016/j.jormas.2022.08.007

42. Yu Z, Kaizhi S, Jianwen H, Guanyu Y, Yonggang W. A deep learning-based approach toward differentiating scalp psoriasis and seborrheic dermatitis from dermoscopic images. *Front Med.* (2022) 9:965423. doi: 10.3389/fmed.2022.965423

43. Lee LT-J, Yang HC, Nguyen PA, Muhtar MS, Li YCJ. Machine learning approaches for predicting psoriatic arthritis risk using electronic medical records: population-based study. *J Med Internet Res.* (2023) 25:e39972. doi: 10.2196/39972

44. du AX, Ali Z, Ajeiyi KK, Dalager MG, Dam TN, Egeberg A, et al. Machine learning model for predicting outcomes of biologic therapy in psoriasis. *J Am Acad Dermatol.* (2023) 88:1364–7. doi: 10.1016/j.jaad.2022.12.046

45. Zhan Y-P, Chen B-S. Drug target identification and drug repurposing in psoriasis through systems biology approach, DNN-based DTI model and genome-wide microarray data. *Int J Mol Sci.* (2023) 24:10033. doi: 10.3390/ijms241210033

46. Zhou Y, Wang Z, Han L, Yu Y, Guan N, Fang R, et al. Machine learning-based screening for biomarkers of psoriasis and immune cell infiltration. *Eur J Dermatol.* (2023) 33:147–56. doi: 10.1684/ejd.2023.4453

47. Guo P, Luo Y, Mai G, Zhang M, Wang G, Zhao M, et al. Gene expression profile based classification models of psoriasis. *Genomics.* (2014) 103:48–55. doi: 10.1016/j.ygeno.2013.11.001

48. McMullen EP, Syed SA, Espiritu KD, Grewal RS, Elder GA, Morita PP, et al. The therapeutic applications of machine learning in atopic dermatitis: a scoping review. *J Cutan Med Surg.* (2023) 27:286–7. doi: 10.1177/12034754231168846

49. Rasheed A, Umar AI, Shirazi SH, Khan Z, Nawaz S, Shahzad M. Automatic eczema classification in clinical images based on hybrid deep neural network. *Comput Biol Med.* (2022) 147:105807. doi: 10.1016/j.combiomed.2022.105807

50. Gustafson E, Pacheco J, Wehbe F, Silverberg J, Thompson W. A machine learning algorithm for identifying atopic dermatitis in adults from electronic health records. *IEEE Int Conf Healthc Inform.* (2017) 2017:83–90. doi: 10.1109/ICHI.2017.31

51. Hurault G, Domínguez-Hüttlinger E, Langan SM, Williams HC, Tanaka RJ. Personalized prediction of daily eczema severity scores using a mechanistic machine learning model. *Clin Exp Allergy.* (2020) 50:1258–66. doi: 10.1111/cea.13171

52. Wilm A, Kühnl J, Kirchmair J. Computational approaches for skin sensitization prediction. *Crit Rev Toxicol.* (2018) 48:738–60. doi: 10.1080/10408444.2018.1528207

53. Liu S, Fan Y, Duan M, Wang Y, Su G, Ren Y, et al. AcneGrader: an ensemble pruning of the deep learning base models to grade acne. *Skin Res Technol.* (2022) 28:677–88. doi: 10.1111/srt.13166

54. Liu S, Chen R, Gu Y, Yu Q, Su G, Ren Y, et al. AcneTyper: an automatic diagnosis method of dermoscopic acne image via self-ensemble and stacking. *Technol Health Care.* (2023) 31:1171–87. doi: 10.3233/THC-220295

55. Zhang H, Ma T. Acne detection by ensemble neural networks. *Sensors.* (2022) 22:6828. doi: 10.3390/s22186828

56. Seit S, Khammari A, Benzaquen M, Moyal D, Dréno B. Development and accuracy of an artificial intelligence algorithm for acne grading from smartphone photographs. *Exp Dermatol.* (2019) 28:1252–7. doi: 10.1111/exd.14022

57. Hernández Montilla I, Medela A, MacCarthy T, Aguilar A, Gómez Tejerina P, Vilas Sueiro A, et al. Automatic International Hidradenitis Suppurativa Severity System (AIHS4): a novel tool to assess the severity of hidradenitis suppurativa using artificial intelligence. *Skin Res Technol.* (2023) 29:e13357. doi: 10.1111/srt.13357

58. Liu TJ, Wang H, Christian M, Chang CW, Lai F, Tai HC. Automatic segmentation and measurement of pressure injuries using deep learning models and a LiDAR camera. *Sci Rep.* (2023) 13:680. doi: 10.1038/s41598-022-26812-9

59. Manohar Dhane D, Maity M, Mungle T, Bar C, Achar A, Kolekar M, et al. Fuzzy spectral clustering for automated delineation of chronic wound region using digital images. *Comput Biol Med.* (2017) 89:551–60. doi: 10.1016/j.combiomed.2017.04.004

60. Mukherjee R, Manohar DD, Das DK, Achar A, Mitra A, Chakraborty C. Automated tissue classification framework for reproducible chronic wound assessment. *Bio Med Res Int.* (2014) 2014:851582. doi: 10.1155/2014/851582

61. Wang L, Pedersen PC, Agu E, Strong DM, Tulu B. Area determination of diabetic foot ulcer images using a cascaded two-stage SVM-based classification. *IEEE Trans Biomed Eng.* (2017) 64:2098–109. doi: 10.1109/TBME.2016.2632522

62. Toffaha KM, Simsekler MCE, Omar MA. Leveraging artificial intelligence and decision support systems in hospital-acquired pressure injuries prediction: a comprehensive review. *Artif Intell Med.* (2023) 141:102560. doi: 10.1016/j.artmed.2023.102560

63. Dwekat OY, Lam SS, McGrath L. A hybrid system of Braden scale and machine learning to predict hospital-acquired pressure injuries (bedsores): a retrospective observational cohort study. *Diagnostics (Basel, Switzerland).* (2022) 13:31. doi: 10.3390/diagnostics13010031

64. Dwekat OY, Lam SS, McGrath L. An integrated system of multifaceted machine learning models to predict if and when hospital-acquired pressure injuries (bedsores) occur. *Int J Environ Res Public Health.* (2023) 20:828. doi: 10.3390/ijerph20010828

65. Stern L, Roshan Fekr A. In-bed posture classification using deep neural network. *Sensors.* (2023) 23:2430. doi: 10.3390/s23052430

66. Gorman BG, Lifson MA, Vidal NY. Artificial intelligence and frozen section histopathology: a systematic review. *J Cutan Pathol.* (2023) 50:852–9. doi: 10.1111/cup.14481

67. Sauter D, Lodde G, Nensa F, Schadendorf D, Livingstone E, Kukuk M. Deep learning in computational dermatopathology of melanoma: a technical systematic literature review. *Comput Biol Med.* (2023) 163:107083. doi: 10.1016/j.combiomed.2023.107083

68. van Zon MCM, van der Waa JD, Veta M, Krekels GAM. Whole-slide margin control through deep learning in Mohs micrographic surgery for basal cell carcinoma. *Exp Dermatol.* (2021) 30:733–8. doi: 10.1111/exd.14306

69. Hekler A, Utikal JS, Enk AH, Berking C, Klode J, Schadendorf D, et al. Pathologist-level classification of histopathological melanoma images with deep neural networks. *Eur J Cancer.* (2019) 115:79–83. doi: 10.1016/j.ejca.2019.04.021

70. Hocke J, Krauth J, Krause C, Gerlach S, Warnemünde N, Affeldt K, et al. Computer-aided classification of indirect immunofluorescence patterns on esophagus and split skin for the detection of autoimmune dermatoses. *Front Immunol.* (2023) 14:111172. doi: 10.3389/fimmu.2023.111172

71. Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med.* (2020) 26:900–8. doi: 10.1038/s41591-020-0842-3

72. Sitaru S, Oueslati T, Schielein MC, Weis J, Kaczmarczyk R, Rueckert D, et al. Automatic body part identification in real-world clinical dermatological images using machine learning. *J Dtsch Dermatol Ges.* (2023) 21:863–9. doi: 10.1111/ddg.15113

73. Okon E, Rachakonda V, Hong HJ, Callison-Burch C, Lipoff JB. Natural language processing of Reddit data to evaluate dermatology patient experiences and therapeutics. *J Am Acad Dermatol.* (2020) 83:803–8. doi: 10.1016/j.jaad.2019.07.014

74. Young JN, Ross O'Hagan, Poplasky D, Levoska MA, Gulati N, Ungar B, et al. The utility of ChatGPT in generating patient-facing and clinical responses for melanoma. *J Am Acad Dermatol.* (2023) 89:602–4. doi: 10.1016/j.jaad.2023.05.024

75. Mondal H, Mondal S, Podder I. Using ChatGPT for writing articles for patients' education for dermatological diseases: a pilot study. *Indian Dermatol Online J.* (2023) 14:482–6. doi: 10.4103/idoj.idoj_72_23

76. O'Hern K, Yang E, Vidal NY. ChatGPT underperforms in triaging appropriate use of Mohs surgery for cutaneous neoplasms. *JAAD Int.* (2023) 12:168–70. doi: 10.1016/j.jdin.2023.06.002

77. Jain A, Way D, Gupta V, Gao Y, de Oliveira Marinho G, Hartford J, et al. Development and assessment of an artificial intelligence-based tool for skin condition diagnosis by primary care physicians and nurse practitioners in teledermatology practices. *JAMA Netw Open.* (2021) 4:e217249. doi: 10.1001/jamanetworkopen.2021.7249

78. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit Health.* (2019) 1:e271–97. doi: 10.1016/S2589-7500(19)30123-2

79. 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) 29:1836–42. doi: 10.1093/annonc/mdy166
80. Han SS, Kim YJ, Moon IJ, Jung JM, Lee MY, Lee WJ, et al. Evaluation of artificial intelligence-assisted diagnosis of skin neoplasms: a single-center, paralleled, unmasked, randomized controlled trial. *J Invest Dermatol.* (2022) 142:2353–2362.e2. doi: 10.1016/j.jid.2022.02.003
81. Winkler JK, Fink C, Toberer F, Enk A, Deinlein T, Hofmann-Wellenhof R, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol.* (2019) 155:1135–41. doi: 10.1001/jamadermatol.2019.1735
82. Vodrahalli K, Daneshjou R, Novoa RA, Chiou A, Ko JM, Zou J. TrueImage: a machine learning algorithm to improve the quality of telehealth photos. *Pac Symp Biocomput.* (2020) 26:220–31. doi: 10.1142/9789811232701_0021
83. Maier K, Zaniolo L, Marques O. Image quality issues in teledermatology: a comparative analysis of artificial intelligence solutions. *J Am Acad Dermatol.* (2022) 87:240–2. doi: 10.1016/j.jaad.2021.07.073
84. Young AT, Fernandez K, Pfau J, Reddy R, Cao NA, von Franque MY, et al. Stress testing reveals gaps in clinic readiness of image-based diagnostic artificial intelligence models. *NPJ Digit Med.* (2021) 4:10. doi: 10.1038/s41746-020-00380-6
85. Caffery LJ, Rotemberg V, Weber J, Soyer HP, Malvey J, Clunie D. The role of DICOM in artificial intelligence for skin disease. *Front Med.* (2020) 7:619787. doi: 10.3389/fmed.2020.619787
86. Jones C, Thornton J, Wyatt JC. Artificial intelligence and clinical decision support: clinicians' perspectives on trust, trustworthiness, and liability. *Med Law Rev.* (2023):fwad013. doi: 10.1093/medlaw/fwad013
87. Naik N, Hameed BMZ, Shetty DK, Swain D, Shah M, Paul R, et al. Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility? *Front Surg.* (2022) 9:862322. doi: 10.3389/fsurg.2022.862322
88. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med.* (2021) 27:582–4. doi: 10.1038/s41591-021-01312-x
89. Nelson CA, Pérez-Chada LM, Creadore A, Li SJ, Lo K, Manjaly P, et al. Patient perspectives on the use of artificial intelligence for skin cancer screening: a qualitative study. *JAMA Dermatol.* (2020) 156:501–12. doi: 10.1001/jamadermatol.2019.5014
90. Nelson CA, Pachauri S, Balk R, Miller J, Theunis R, Ko JM, et al. Dermatologists' perspectives on artificial intelligence and augmented intelligence – a cross-sectional survey. *JAMA Dermatol.* (2021) 157:871–4. doi: 10.1001/jamadermatol.2021.1685
91. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. *N Engl J Med.* (2023) 388:1201–8. doi: 10.1056/NEJMr2302038
92. Kassab J, el Dahdah J, Chedid el Helou M, Layoun H, Sarraju A, Laffin LJ, et al. Assessing the accuracy of an online chat-based artificial intelligence model in providing recommendations on hypertension management in accordance with the 2017 American College of Cardiology/American Heart Association and 2018 European Society of Cardiology/European Society of Hypertension Guidelines. *Hypertension.* (2023) 80:e125–7. doi: 10.1161/HYPERTENSIONAHA.123.21183
93. Li H, Moon JT, Purkayastha S, Celi LA, Trivedi H, Gichoya JW. Ethics of large language models in medicine and medical research. *Lancet Digit Health.* (2023) 5:e333–5. doi: 10.1016/S2589-7500(23)00083-3
94. Zhou J, He X, Sun L, Xu J, Chen X, Chu Y, et al. (2023) 'SkinGPT-4: an interactive dermatology diagnostic system with visual large language model', arXiv [eessIV]. Available at: <http://arxiv.org/abs/2304.10691>.
95. Kim C., Gadgil S.U., DeGrave A., Cai Z.R., Daneshjou R., Lee S.I. (2023) 'Fostering transparent medical image AI via an image-text foundation model grounded in medical literature', medRxiv [Preprint]. doi: 10.1101/2023.06.07.23291119
96. Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, et al. Foundation models for generalist medical artificial intelligence. *Nature.* (2023) 616:259–65. doi: 10.1038/s41586-023-05881-4
97. van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, et al. A systematic review of barriers to data sharing in public health. *BMC Public Health.* (2014) 14:1144. doi: 10.1186/1471-2458-14-1144
98. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer.* (2018) 18:500–10. doi: 10.1038/s41568-018-0016-5
99. Wang F, Casalino LP, Khullar D. Deep learning in medicine—promise, progress, and challenges. *JAMA Intern Med.* (2019) 179:293–4. doi: 10.1001/jamainternmed.2018.7117
100. Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. *NPJ Digit Med.* (2020) 3:119. doi: 10.1038/s41746-020-00323-1
101. Carlini N., Liu C., Erlingsson Ú., Kos J., Song D. (2019) 'The secret sharer: evaluating and testing unintended memorization in neural networks', in 28th USENIX security symposium (USENIX security 19), pp. 267–284.
102. Wornow M, Xu Y, Thapa R, Patel B, Steinberg E, Fleming S, et al. The shaky foundations of large language models and foundation models for electronic health records. *NPJ Digit Med.* (2023) 6:135. doi: 10.1038/s41746-023-00879-8
103. Goodman RL. *Ophtho notes: the essential guide*. New York: Thieme (2003).
104. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med.* (2023) 6:120. doi: 10.1038/s41746-023-00873-0
105. Daneshjou R, Smith MP, Sun MD, Rotemberg V, Zou J. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA Dermatol.* (2021) 157:1362–9. doi: 10.1001/jamadermatol.2021.3129
106. Kovarik C, Lee I, Ko J, Adamson A, Otley C, Kvedar J, et al. Commentary: position statement on augmented intelligence (AuI). *J Am Acad Dermatol.* (2019) 81:998–1000. doi: 10.1016/j.jaad.2019.06.032
107. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell.* (2019) 1:206–15. doi: 10.1038/s42256-019-0048-x



OPEN ACCESS

EDITED BY

Justin Ko,
Stanford University, United States

REVIEWED BY

Sam Polesie,
Sahlgrenska University Hospital, Sweden
Chengxu Li,
China-Japan Friendship Hospital, China

*CORRESPONDENCE

Babak Saravi
✉ babak.saravi@jupiter.uni-freiburg.de

[†]These authors have contributed equally to this work

RECEIVED 30 May 2023

ACCEPTED 09 October 2023

PUBLISHED 20 October 2023

CITATION

Shavlokhova V, Vollmer A, Zouboulis CC, Vollmer M, Wollborn J, Lang G, Kübler A, Hartmann S, Stoll C, Roeder E and Saravi B (2023) Finetuning of GLIDE stable diffusion model for AI-based text-conditional image synthesis of dermoscopic images. *Front. Med.* 10:1231436. doi: 10.3389/fmed.2023.1231436

COPYRIGHT

© 2023 Shavlokhova, Vollmer, Zouboulis, Vollmer, Wollborn, Lang, Kübler, Hartmann, Stoll, Roeder and Saravi. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Finetuning of GLIDE stable diffusion model for AI-based text-conditional image synthesis of dermoscopic images

Veronika Shavlokhova^{1†}, Andreas Vollmer^{2†}, Christos C. Zouboulis³, Michael Vollmer⁴, Jakob Wollborn⁵, Gernot Lang⁶, Alexander Kübler², Stefan Hartmann², Christian Stoll¹, Elisabeth Roeder⁷ and Babak Saravi^{5,6*}

¹Maxillofacial Surgery University Hospital Ruppiner-Fehrbelliner Straße Neuruppin, Neuruppin, Germany,

²Department of Oral and Maxillofacial Plastic Surgery, University Hospital of Würzburg, Würzburg, Germany, ³Departments of Dermatology, Venereology, Allergology and Immunology, Städtisches Klinikum Dessau, Medical School Theodor Fontane and Faculty of Health Sciences Brandenburg, Dessau, Germany, ⁴Department of Oral and Maxillofacial Surgery, Tuebingen University Hospital, Tuebingen, Germany, ⁵Department of Anesthesiology, Perioperative and Pain Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States, ⁶Department of Orthopedics and Trauma Surgery, Medical Centre-Albert-Ludwigs-University of Freiburg, Faculty of Medicine, Albert-Ludwigs-University of Freiburg, Freiburg, Germany, ⁷Department of Dermatology, University Hospital of Basel, Basel, Switzerland

Background: The development of artificial intelligence (AI)-based algorithms and advances in medical domains rely on large datasets. A recent advancement in text-to-image generative AI is GLIDE (Guided Language to Image Diffusion for Generation and Editing). There are a number of representations available in the GLIDE model, but it has not been refined for medical applications.

Methods: For text-conditional image synthesis with classifier-free guidance, we have fine-tuned GLIDE using 10,015 dermoscopic images of seven diagnostic entities, including melanoma and melanocytic nevi. Photorealistic synthetic samples of each diagnostic entity were created by the algorithm. Following this, an experienced dermatologist reviewed 140 images (20 of each entity), with 10 samples originating from artificial intelligence and 10 from original images from the dataset. The dermatologist classified the provided images according to the seven diagnostic entities. Additionally, the dermatologist was asked to indicate whether or not a particular image was created by AI. Further, we trained a deep learning model to compare the diagnostic results of dermatologist versus machine for entity classification.

Results: The results indicate that the generated images possess varying degrees of quality and realism, with melanocytic nevi and melanoma having higher similarity to real images than other classes. The integration of synthetic images improved the classification performance of the model, resulting in higher accuracy and precision. The AI assessment showed superior classification performance compared to dermatologist.

Conclusion: Overall, the results highlight the potential of synthetic images for training and improving AI models in dermatology to overcome data scarcity.

KEYWORDS

GLIDE, text-to-image, stable diffusion, dermoscopy, cancer, dermatology

1. Introduction

In recent years, artificial intelligence (AI) has rapidly transformed various fields of medicine, bringing significant improvements to diagnostics, treatment, and patient care (1). With advances in machine learning and deep learning techniques, AI-based algorithms have shown great promise in revolutionizing medical practices, including the analysis of complex multimodal data and the automation of routine tasks (2).

Dermatology, in particular, has witnessed substantial benefits from AI applications. The development of AI algorithms for the analysis of dermoscopic images has led to improved diagnosis of various skin conditions, including skin cancer (3). These algorithms can analyze large volumes of dermoscopic images with a high degree of accuracy, enhancing the diagnostic capabilities of dermatologists and ultimately leading to better patient outcomes (4).

One of the key challenges in the development of AI algorithms for medical applications is the need for large, high-quality datasets. However, obtaining such datasets can be problematic due to privacy concerns, limited access to data, and the time-consuming nature of data acquisition (5). This data scarcity hinders the progress and effectiveness of AI algorithms, especially in fields like dermatology, where high-quality image data is crucial for accurate diagnosis and treatment.

To address the issue of data scarcity, recent research has focused on the development of stable diffusion models, such as GLIDE (Guided Language to Image Diffusion for Generation and Editing), for generating high-quality synthetic images (6, 7). Kather et al. recently proposed to apply these algorithms to the medical field (8). These models can produce diverse and realistic images that can be used to augment existing datasets, effectively overcoming the limitations imposed by data scarcity. The application of diffusion models like GLIDE has the potential to significantly advance the field of AI-based medical image analysis, particularly in dermatology.

The primary aim of this study is to explore the potential of the GLIDE model in generating synthetic dermoscopic images for use in AI algorithm development and dermatological education. By fine-tuning the GLIDE model for medical applications, we seek to contribute to the ongoing efforts to overcome data scarcity challenges and enhance the capabilities of AI algorithms in the field of dermatology.

2. Methods

2.1. GLIDE model fine-tuning

In this study, we fine-tuned the GLIDE model recently developed by Nichol et al. (6). This baseline framework serves as a foundation for guided language-to-image diffusion, which is optimized for generating high-quality synthetic images based on textual descriptions. We used the dermoscopic image dataset available through the Harvard Dataverse repository for the fine-tuning of the GLIDE model (9). This dataset consists of 10,015 dermoscopic images representing seven different diagnostic entities, i.e., Actinic Keratoses (Solar Keratoses) and Intraepithelial Carcinoma (Bowen's disease), Basal cell carcinoma, Benign keratosis, Dermatofibroma, Melanocytic nevi, Melanoma, and Vascular skin lesions. Each image in the dataset is annotated with the

corresponding diagnostic entity. Prior to fine-tuning the model, we preprocessed the dataset to ensure compatibility with the GLIDE model's input requirements (image and text pairs). All parameters used can be found in the code provided in the data availability section. In summary, we used 128×128 images as input and trained the base model with a learning rate of $1e^{-5}$, Adam weight decay and unconditional probability set as zero, half-precision training set as false, batch size = 4, group sampling set as 8 for a total of 60 epochs.

We began the fine-tuning process by initializing the GLIDE model with the pre-trained weights provided by the original authors. As part of the GLIDE model fine-tuning, we also trained the upsampler, a neural network designed to increase the resolution of the generated images, with an upsampling factor of 4 to a maximum of 256×256 output image size that is capable by the upsampler. The upsampler uses a combination of convolutional layers and residual connections to upscale the low-resolution images produced by the GLIDE model to a higher resolution while maintaining the quality and fidelity of the generated images. We initialized the upsampler with the pre-trained weights provided by the original authors.

2.2. Model evaluation

The evaluation of generated images was based on a combination of image quality metrics, including Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), Mean Squared Error (MSE), Frechet Inception Distance (FID), and Inception Score (IS). Ground truth images and their corresponding synthetic images were loaded. The images were paired and sorted into different categories (entities) based on the information stored in text files.

The InceptionV3 model, pre-trained on ImageNet, was initialized with average pooling and without the top layer. The model was used to calculate FID and IS scores. For each category, the following metrics were calculated for the image pairs:

- SSIM: calculated separately for each color channel and averaged. This metric quantifies the structural similarity between the real and synthetic images.
- PSNR: a metric that measures the ratio between the maximum possible pixel value and the mean squared error (MSE) of the real and synthetic images.
- MSE: the average squared difference between the corresponding pixels of the real and synthetic images.
- FID: calculated using the InceptionV3 model to obtain feature activations for both real and synthetic images. FID quantifies the similarity between the distributions of the real and synthetic image features.
- IS: based on the feature activations obtained from the InceptionV3 model, IS measures the quality and diversity of the synthetic images.

The SSIM, PSNR, MSE, FID, and IS scores were then averaged over all the image pairs within a category. The results were obtained for each category. Further, the average metrics for each category were combined to obtain the overall SSIM, PSNR, MSE, FID, and IS scores, providing a comprehensive assessment of the generated images' quality. This evaluation approach ensures a thorough assessment of the generated images' quality and similarity to the ground truth,

considering various aspects such as structural similarity, pixel-level differences, feature distributions, and the diversity of the generated images.

2.3. Dermatologist assessment

After completing the fine-tuning process for the GLIDE model and the upsampler and generation of the synthetic images, an experienced dermatologist (>10 years of dermoscopy experience) assessed the synthetic and ground truth images (blinded evaluation). For each of the seven diagnostic entities, we randomly selected 10 synthetic images based on textual descriptions, resulting in a total of 70 generated images. Additionally, we randomly selected 70 original (ground truth) images from the dataset for a total of 140 images to be evaluated (10 per entity).

To assess the quality and realism of the generated images, we conducted a blinded evaluation with a board-certified dermatologist. Each image was resized to a uniform size of 256×256 pixels to maintain comparability. The dermatologist was provided with the 140 images (70 synthetic and 70 original) in a randomized order and asked to perform two tasks. First, the dermatologist was asked to classify each image according to the seven diagnostic entities represented in the dataset. Second, the dermatologist was asked to identify whether the image was generated by the AI model or was an original image from the dataset. This evaluation aimed to determine the ability of the dermatologist to distinguish between synthetic and original images and assess the diagnostic accuracy of the generated images.

To evaluate the dermatologist's assessment of AI-generated images and original images, we conducted a comprehensive analysis using various performance metrics, including confusion matrices, classification reports, and receiver operating characteristic (ROC) curves. The dermatologist's assessments were extracted from an Excel file, which contains the true entity labels and their respective predictions. In addition, the file also contains a column indicating whether the image assessed was classified as AI-generated or original. We then computed the classification report for AI-generated vs. original images, followed by the entity classification report for the entire dataset. Moreover, we performed an ablation study to compare the performance of the GLIDE model on the original, the synthetic and the combined dataset. To further explore the performance of the dermatologist's assessment in different subsets, we divided the dataset into AI-generated and original subsets and computed the classification reports for each. Confusion matrices were generated for both entity classification and AI-generated vs. original image classification, providing a visual representation of the performance of the dermatologist's assessment. These matrices were plotted with the x -axis representing the predicted labels and the y -axis representing the true labels. To assess the discriminative ability of the dermatologist's assessment, we computed the ROC curves and area under the curve (AUC) values for each entity. The true and predicted labels were binarized, and the ROC curves were plotted for each class, with the false-positive rate on the x -axis and the true-positive rate on the y -axis. Additionally, the ROC curve for AI-generated vs. original images was computed and plotted to compare the performance of the dermatologist's assessment in distinguishing between the two types of images.

2.4. Deep learning assessment

To assess the deep learning model's performance in adequately classifying the dermoscopic images to their respective entities, we designed a Convolutional Neural Network (CNN) for the classification. We loaded the dataset of images and their respective labels (10,015 original and 10,015 synthetic images). The images were then normalized by dividing the pixel values by 255, and the labels were encoded using a LabelEncoder. We divided the dataset into training and testing sets with an 80–20% ratio. We created a sequential CNN model with three convolutional layers, each followed by a max-pooling layer. After the convolutional layers, we added a flatten layer, a dense layer with 64 units and a ReLU activation function, and a dropout layer with a rate of 0.5. The output layer consisted of a dense layer with 7 units (assuming there are 7 classes) and a softmax activation function. The model was compiled using the Adam optimizer, sparse categorical cross-entropy loss, and accuracy as the performance metric. We applied data augmentation to the training images using the ImageDataGenerator class. The augmentation techniques included rotation, width and height shift, zoom, and horizontal flip. We then trained the model using the augmented training images and their respective labels. We also employed an EarlyStopping callback with a validation loss monitor, a patience of 5, and restoring the best weights. The model was trained for a maximum of 100 epochs with a batch size of 32. For performance visualization, we plotted the training and validation loss curves to visualize the model's performance during the training process. The x -axis represents the epochs, while the y -axis represents the loss values. We evaluated the model's performance using the test set. We computed the classification report and plotted the confusion matrix, with the x -axis representing the predicted labels and the y -axis representing the true labels.

2.5. Metrics calculation, programming framework, and web application

All analyses were performed in Python. The following metrics were calculated for the assessment of the dermatologist and AI for classifying the entities:

Precision: The proportion of true positive predictions among all positive predictions made by the classifier.

- **Recall:** the proportion of true positive predictions among all actual positive instances in the dataset.
- **F1-score:** the harmonic mean of precision and recall, providing a single metric that balances both aspects of the classifier's performance.
- **Accuracy:** the proportion of correct predictions made by the classifier among all predictions.
- **Macro avg.:** the average of a particular metric (e.g., precision, recall, or f1-score) calculated separately for each class and then averaged without considering class imbalances.
- **Weighted avg.:** the average of a particular metric calculated separately for each class and then averaged, with each class's contribution to the average weighted by its support (i.e., number of occurrences).

TABLE 1 Evaluation metrics for the fine-tuned GLIDE model.

Dermatological lesion category	SSIM	PSNR	MSE	FID	IS
Melanoma	0.2186	60.1854	0.0698	115.1804	1.3630
Melanocytic nevi	0.2229	61.0407	0.0604	99.2504	1.4739
Actinic keratoses and intraepithelial carcinoma/Bowen's disease	0.0612	62.2664	0.0435	174.9675	1.2991
Benign keratosis-like lesions (solar lentigines/seborrheic keratoses and lichen-planus like keratoses)	0.1174	61.7934	0.0506	203.4957	1.3379
Basal cell carcinoma	0.1347	64.2892	0.0289	189.7611	1.3193
Dermatofibroma	0.0873	63.0862	0.0358	275.1849	1.2379
Vascular lesions (angiomas)	0.1589	60.1402	0.0785	252.0546	1.3672
Overall Metrics	0.1430	61.8288	0.0525	187.1278	1.3426

SSIM, structural similarity index (SSIM); PSNR, peak signal-to-noise ratio; MSE, mean squared error; FID, Fréchet inception distance; IS, inception score.

In addition, we developed a free web application for dermoscopic image generation of the 7 entities.¹ The web application uses a CPU for image generation, which can take up to 20 min per image. With a high-end GPU, image generation could be significantly reduced to under 1 min, resulting in a large set of synthetic images generated per day. Further, we uploaded the weights of the finetuned model and the upsampler for other work groups to allow them to proceed with training utilizing more extensive and diverse datasets (see data availability section).

3. Results

3.1. Evaluation metrics for synthetic images

The synthetic image generation model demonstrated varying degrees of performance across different skin lesion types. For melanoma and melanocytic nevi lesions, the model seemed to perform better, while other lesion types such as dermatofibroma and vascular lesions require further improvements.

Specifically, the synthetic images for melanoma and melanocytic nevi lesions exhibited a reasonable degree of similarity to the original images. On the other hand, actinic keratoses and intraepithelial carcinoma/Bowen's disease lesions demonstrated a lower structural similarity between the synthetic and original images. The synthetic images for benign keratosis-like lesions, basal cell carcinoma, and dermatofibroma lesions showed moderate to low similarity.

The average metrics for all lesion types suggest that the model can generally reproduce the structural and visual features of the original

lesions to a fair extent, albeit with room for further refinement. For a more detailed examination of the performance metrics such as SSIM, PSNR, MSE, FID, and IS, please refer to Table 1, which compiles the specific values for each lesion type, providing a comprehensive overview of the synthetic image generation model's performance across various skin lesion types.

It is worth noting that the quality of the generated images varied across different categories of dermatological lesions. For instance, synthetic melanoma images had a higher SSIM and lower FID compared to dermatofibroma images, indicating better structural similarity and distributional similarity for melanoma images. Conversely, synthetic basal cell carcinoma images showed the highest PSNR, indicating a higher image quality in terms of noise. Table 1 shows the metrics obtained for the evaluation. Figure 1 illustrates a random set of original and synthetic images. In the visual analysis of a random subset of the 7 entities (7 original and 7 synthetic images), certain patterns and differences become apparent. Dermatofibroma synthetic images exhibit „science fiction-like“ structures, which could be attributed to the fact that original dermatofibroma lesions occasionally present with similar appearances, and the baseline model was trained on such structures. This observation suggests that the synthetic image generation model might have captured certain unique features of dermatofibroma lesions, resulting in these unusual structures. Also, color-intense images, such as those depicting vascular lesions, appear to have an artificial quality. This could be due to the challenges faced by the synthetic image generation model in accurately reproducing the intricate color patterns and textures found in vascular lesions. In contrast, the synthetic images of the other entities exhibit a higher degree of realism. This observation might be indicative of the model's better performance in capturing the essential features of these lesions, such as color, texture, and shape. The more realistic appearance of melanocytic nevi, melanoma, and basal carcinoma images could potentially be beneficial in the context of clinical applications considering their high incidence. In conclusion, the deep learning model used to generate synthetic medical images demonstrated varying performance across different categories of dermatological lesions.

3.2. Dermatologist's assessment

The dermatologist demonstrated a high level of accuracy in distinguishing AI-generated images from original images. The overall accuracy in this classification task reached 96%. A balanced performance with a precision of 0.99 and 0.95, and recall of 0.94 and 0.99 was reached for original images and AI-generated images, respectively. The macro-average and weighted average f1-scores were 0.96 for both.

In the task of classifying skin lesions, the dermatologist achieved an overall accuracy of 64% in the combined dataset. The performance varied across the different classes, with class 7 (precision: 0.82, recall: 0.90) achieving the highest f1-score of 0.86, and class 2 (precision: 0.62, recall: 0.50) exhibiting the lowest f1-score of 0.56. The macro-average and weighted average f1-scores were both 0.64.

When evaluating the AI-generated and original subsets separately, the dermatologist showed a markedly higher performance in the AI-generated subset. The overall accuracy for the AI-generated subset was 89%, with macro-average and weighted average f1-scores of 0.88 and 0.89, respectively. In contrast, the overall accuracy for the original subset was 40%, with macro-average and weighted average f1-scores of

¹ https://huggingface.co/spaces/Freiburg-AI-Research/dermoscopic_image_generation

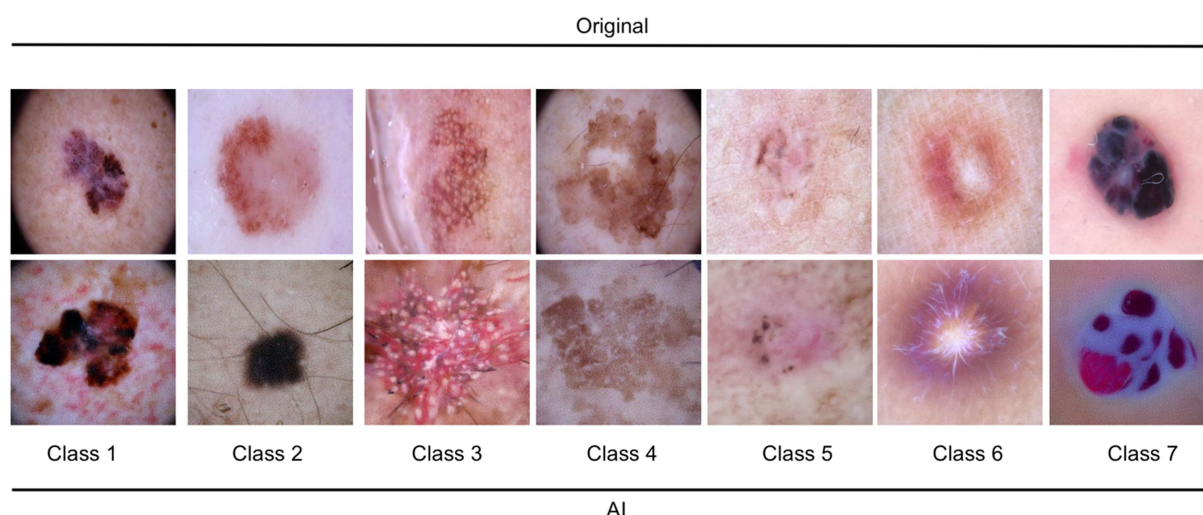


FIGURE 1

Illustration of 7 random original and AI-generated images for the entities. Class 1: melanoma; Class 2: melanocytic nevi; Class 3: Actinic keratoses and intraepithelial carcinoma/Bowen disease; Class 4: benign keratosis-like lesions (solar lentigines/seborrheic keratoses and lichen-planus like keratoses); Class 5: basal cell carcinoma; Class 6: dermatofibroma; Class 7: vascular lesions.

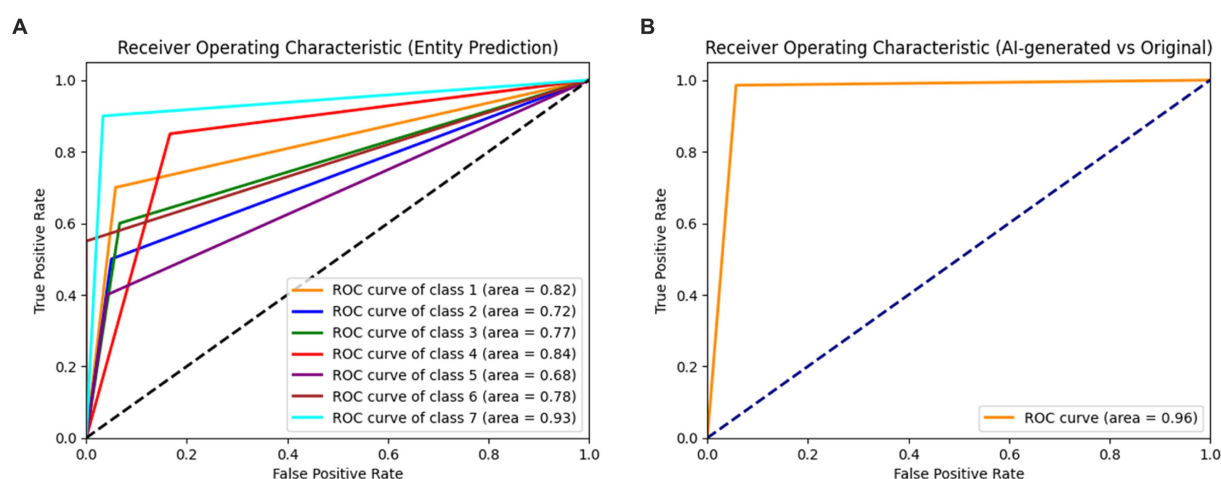


FIGURE 2

Receiver operating characteristic curves (ROC) for the dermatologist assessment of entities (A) and AI versus original (B). Class 1: melanoma; Class 2: melanocytic nevi; Class 3: Actinic keratoses and intraepithelial carcinoma/Bowen disease; Class 4: benign keratosis-like lesions (solar lentigines/seborrheic keratoses and lichen-planus like keratoses); Class 5: basal cell carcinoma; Class 6: dermatofibroma; Class 7: vascular lesions.

0.37 and 0.40, respectively. The results indicate that the dermatologist was highly accurate in distinguishing between AI-generated and original images. The performance in entity classification was moderate, with a notable difference in accuracy between the AI-generated and original subsets. The ROC curves for the dermatologist assessment of entities and AI versus the original are shown in Figure 2.

3.3. Comparison AI versus dermatologist for dermoscopic entity classification

Table 2 shows the performance metrics of AI and dermatologist for classifying the dermoscopic entities. The AI model achieved an overall accuracy of 0.86, with varying performance across different

lesion types. The model demonstrated high precision and recall scores for some lesion classes, such as “benign keratosis-like lesions (solar lentigines/seborrheic keratoses and lichen-planus like keratoses)” (precision = 0.87, recall = 0.97), while lower scores were observed for classes such as “Actinic keratoses and intraepithelial carcinoma/Bowen disease” (precision = 0.91, recall = 0.55).

The dermatologist achieved an overall accuracy of 0.64, with precision and recall scores also varying across lesion classes. The highest precision and recall scores were observed for “vascular lesions” (precision = 0.82, recall = 0.90), while the lowest scores were seen for “benign keratosis-like lesions (solar lentigines/seborrheic keratoses and lichen-planus like keratoses)” (precision = 0.46, recall = 0.85).

Comparing the AI model assessment to the dermatologist assessment, the AI model demonstrated a higher overall accuracy

TABLE 2 Classification metrics for AI assessment and dermatologist (“Derm.”) Assessment.

Class	AI Precision	AI Recall	AI F1-Score	Derm. Precision	Derm. Recall	Derm. F1-Score
1	0.81	0.60	0.69	0.67	0.70	0.68
2	0.83	0.70	0.76	0.62	0.50	0.56
3	0.80	0.63	0.71	0.60	0.60	0.60
4	0.91	0.55	0.69	0.46	0.85	0.60
5	0.87	0.97	0.92	0.62	0.40	0.48
6	0.82	0.61	0.70	1.00	0.55	0.71
7	0.96	0.76	0.85	0.82	0.90	0.86
Accuracy			0.86			0.64
Macro Avg	0.86	0.69	0.76	0.68	0.64	0.64
Weighted Avg	0.86	0.86	0.85	0.68	0.64	0.64

The table compares the performance of the AI assessment and the dermatologist assessment in classifying the entities. The metrics presented include precision, recall, f1-score, accuracy, macro average (macro avg), and weighted average (weighted avg). Class 1: melanoma; Class 2: melanocytic nevi; Class 3: Actinic keratoses and intraepithelial carcinoma/Bowen disease; Class 4: benign keratosis-like lesions (solar lentigines/seborrheic keratoses and lichen-planus like keratoses); Class 5: basal cell carcinoma; Class 6: dermatofibroma; Class 7: vascular lesions.

(0.86) compared to the dermatologist (0.64). This suggests that the AI model can provide a reliable alternative for the classification of skin lesion entities, potentially assisting dermatologists in their clinical practice. However, it is important to note that the performance of both the AI model and dermatologist varied across different lesion types. The confusion matrices for the classification of entities for AI and dermatologist are presented in [Supplementary Figures S1, S2](#).

3.4. Ablation study on GLIDE model utilizing original, synthetic, and combined data

[Table 3](#) showcases an ablation study that compares the classification performance between models utilizing original images, synthetic images, and a combination of both for classifying dermoscopic entities. The specific effects on different lesion types are detailed below:

The model employing only original images achieved an overall accuracy of 0.65. Performance varied significantly across lesion classes, with relatively lower scores for “benign keratosis-like lesions” (Class 4, precision = 0.30, recall = 0.23) and higher scores for “basal cell carcinoma” (Class 5, precision = 0.70, recall = 0.75).

The synthetic-only approach yielded an overall accuracy of 0.80. Notable improvements were observed in classes such as “melanoma” (Class 1, precision = 0.75, recall = 0.65) and “vascular lesions” (Class 7, precision = 0.70, recall = 0.60).

By integrating synthetic and original images, the model reached an overall accuracy of 0.86. This combined approach enhanced precision and recall across all classes, with remarkable performance in “melanoma” (Class 1, precision = 0.81, recall = 0.60), and “vascular lesions” (Class 7, precision = 0.96, recall = 0.76). “Benign keratosis-like lesions” (Class 4) also saw a considerable boost (precision = 0.91, recall = 0.55).

4. Discussion

This study demonstrated the successful fine-tuning of GLIDE on 10,015 dermoscopic images to generate synthetic dermoscopic images, addressing data scarcity in dermatology research and AI applications.

The results indicate that the generated images possess varying degrees of quality and realism, with melanocytic nevi and melanoma having higher similarity to real images than other classes. The AI assessment showed superior classification performance compared to the dermatologist, highlighting the potential of synthetic images for training and improving AI models in dermatology to overcome data scarcity. Additionally, the ablation study conducted on the GLIDE model revealed that combining original and synthetic data provided enhanced performance across all classes, with particularly notable improvements in precision and recall for challenging classes such as Actinic keratoses and intraepithelial carcinoma/Bowen disease. The combined approach yielded an accuracy of 0.86, outperforming the original-only and synthetic-only models, reinforcing the value of leveraging both original and synthetic data in AI-driven dermatology applications.

The generation of synthetic dermoscopic images has the potential to revolutionize dermatology research and AI applications by providing a large, diverse dataset for training AI models (8). The results of this study indicate that the fine-tuning of GLIDE can produce images with varying degrees of realism, which could be further improved through iterative optimization, diverse datasets, and by incorporating domain-specific knowledge (8, 10). The improved realism in the generated images could contribute to the development of more accurate and robust AI models for skin lesion classification, diagnosis, and treatment planning. Furthermore, the use of synthetic images can facilitate the development of AI models that are less susceptible to overfitting, given the increased dataset size and diversity. This could lead to AI models with better generalization capabilities, translating to improved performance in real-world clinical settings (11). Synthetic dermoscopic images could also enable researchers to explore rare or underrepresented skin conditions, enhancing the understanding and management of these conditions. Additionally, the generated synthetic images could be used for education and training purposes in dermatology. Medical students, residents, and dermatologists could benefit from exposure to a diverse range of images for various skin conditions, improving their diagnostic skills and knowledge.

Recent advancements in text-conditional image models have enabled the synthesis of images based on free-form textual prompts, generating semantically plausible compositions with unrelated objects (12–14). However, these models have not yet reached the capability of

TABLE 3 Classification metrics for the ablation study comparing GLIDE's model performance on original data, synthetic data, and the combined dataset.

Class	Combined			Original only			Synthetic only		
	AI Precision	AI Recall	AI F1-Score	AI Precision	AI Recall	AI F1-Score	AI Precision	AI Recall	AI F1-Score
1	0.81	0.60	0.69	0.40	0.35	0.37	0.75	0.65	0.70
2	0.83	0.70	0.76	0.60	0.55	0.57	0.78	0.70	0.74
3	0.80	0.63	0.71	0.58	0.50	0.54	0.80	0.75	0.77
4	0.91	0.55	0.69	0.30	0.23	0.26	0.70	0.60	0.65
5	0.87	0.97	0.92	0.70	0.75	0.73	0.85	0.87	0.86
6	0.82	0.61	0.70	0.50	0.40	0.44	0.76	0.68	0.72
7	0.96	0.76	0.85	0.60	0.50	0.55	0.70	0.60	0.65
Accuracy			0.86			0.65			0.80
Macro Avg	0.86	0.69	0.76			0.49			0.73
Weighted Avg	0.86	0.86	0.85			0.64			0.80

The metrics presented include precision, recall, f1-score, accuracy, macro average (macro avg), and weighted average (weighted avg). Class 1: melanoma; Class 2: melanocytic nevi; Class 3: Actinic keratoses and intraepithelial carcinoma/Bowen disease; Class 4: benign keratosis-like lesions (solar lentigines/seborrheic keratoses and lichen-planus like keratoses); Class 5: basal cell carcinoma; Class 6: dermatofibroma; Class 7: vascular lesions.

generating images with full photorealism that accurately represent all aspects of the corresponding textual descriptions. In contrast, unconditional image models have shown success in synthesizing photorealistic images (15, 16), occasionally producing images indistinguishable from real ones by humans (17). Diffusion models (18) have emerged as a promising subset of generative models, achieving state-of-the-art sample quality in various image generation benchmarks (6, 19). Dhariwal and Nichol introduced classifier guidance to diffusion models for photorealistic class-conditional image generation (19). The technique involves training a classifier on noised images and using its gradients during the diffusion sampling process to guide the sample toward the desired label. Ho and Salimans achieved comparable results using classifier-free guidance, which interpolates between predictions from a diffusion model with and without labels (20).

Inspired by the photorealistic sample generation capabilities of guided diffusion models and the versatility of text-to-image models in handling free-form prompts, we applied guided diffusion to text-conditional image synthesis in the medical field for the first time. Nichols et al. trained a 3.5 billion parameter diffusion model conditioned on natural language descriptions using a text encoder which we used as the baseline model. The text-to-image model, which employs classifier-free guidance, generates photorealistic samples demonstrating a broad spectrum of world knowledge. Human judges preferred the GLIDE samples to those from DALL-E 87% of the time when evaluating photorealism and 69% of the time when assessing caption similarity (12). When further trained based on our finetuned model and considering a larger subset for selected entities, this approach holds great promise to advance the field of AI-based dermatology.

Despite the promising results, this study has some limitations. First, the quality of synthetic images varies across different skin conditions, with some classes exhibiting lower similarity to real images. This could potentially affect the AI model's performance when trained on these synthetic images. Future research should aim to refine the image generation process for some entities and include a

larger subset for these entities to ensure more consistent quality across all classes. Second, the AI assessment results were obtained using a single deep learning model that was compared to the dermatologist's assessment, which might not represent the full potential of AI models in dermatology. Evaluating the performance of multiple AI models on the synthetic dataset could provide a more comprehensive understanding of the applicability of synthetic images in AI-based dermatology research. Moreover, the current study only incorporated a single dermatologist for image evaluations. Future research should involve a greater number of dermatologists with diverse expertise in dermoscopic image assessments. Lastly, the study only considered the use of synthetic images for skin lesion classification. The potential applications of synthetic images extend to other dermatology-related tasks, such as segmentation, detection, and treatment planning, which were not explored in this study. Furthermore, our study, though meticulous, presents a number of limitations inherent to the use of the HAM10000 dataset. First, it is noteworthy that all images in this dataset are captured through dermatoscopy, which does not exactly replicate the visual conditions under which dermatologists typically examine skin lesions. Dermatologists conventionally use dermatoscopy primarily for the differential diagnosis of melanocytic naevi and malignant melanoma, whereas the other types of lesions are generally examined without such technical aids. Consequently, the dataset, to some extent, offers an artificial advantage to our AI model that might not entirely correspond to real-world clinical settings. Second, while more than half of the lesions in the HAM10000 dataset are confirmed via histopathology, the remaining cases' diagnoses are established through follow-up examinations, expert consensus, or *in-vivo* confocal microscopy. Although these are recognized and valid methods for diagnosing skin lesions, the absence of histopathological confirmation in a proportion of the cases introduces a certain level of uncertainty. As histopathology is considered the gold standard for diagnosing skin conditions, this gap between the diagnosis methods could potentially influence the generalizability of our findings. In light of these considerations, while the HAM10000 dataset presents a valuable resource for developing and testing AI models for diagnosing

skin lesions, future studies might benefit from incorporating natural lesion images and increasing the proportion of lesions confirmed through histopathology to further enhance the model's real-world applicability and reliability.

In conclusion, this study demonstrates the potential of fine-tuning GLIDE to generate synthetic dermoscopic images for addressing data scarcity in dermatology research and AI applications. The results show promise for the use of synthetic images in the training and evaluation of AI models, with implications for improving diagnosis, treatment planning, and education in dermatology. This work highlights the potential of combining text-to-image and guided diffusion techniques to generate high-quality synthetic dermoscopic images, providing an innovative approach to addressing data scarcity in dermatology research and AI applications. Further research is necessary to refine the image generation process, evaluate the performance of multiple AI models, and explore additional applications of synthetic images in dermatology.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary material](#), further inquiries can be directed to the corresponding author.

Author contributions

AV, BS, MV, CS, ER, and VS: conceptualization. AV, MV, and BS: data curation. AV, VS, CZ, AK, JW, SH, and GL: formal analysis of results and datasets. AV, VS, CZ, and BS: methodological conception. AK, JW, CS, ER, and GL: resources for studies. AV, MV, VS, CZ, AK, and GL: validation of results. AV, MV, VS, and BS: visualization of results and writing – original draft. SH, CZ, AK, JW, CS, ER, and GL:

writing – review and editing. All authors contributed to the article and approved the submitted version.

Funding

The article processing charge was funded by the Baden-Wuerttemberg Ministry of Science, Research and Art, and the University of Freiburg in the funding program Open Access Publishing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2023.1231436/full#supplementary-material>

References

1. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med.* (2022) 28:31–8. doi: 10.1038/s41591-021-01614-0
2. Saravi B, Hassel F, Ülkümen S, Zink A, Shavlokhova V, Couillard-Despres S, et al. Artificial intelligence-driven prediction modeling and decision making in spine surgery using hybrid machine learning models. *J Pers Med.* (2022) 12:509. doi: 10.3390/jpm12040509
3. Jeong HK, Park C, Henao R, Kheterpal M. Deep learning in dermatology: a systematic review of current approaches, outcomes, and limitations. *JID Innov.* (2023) 3:100150. doi: 10.1016/j.xjidi.2022.100150
4. Baig R, Bibi M, Hamid A, Kausar S, Khalid S. Deep learning approaches towards skin lesion segmentation and classification from dermoscopic images – a review. *Curr Med Imaging.* (2020) 16:513–33. doi: 10.2174/1573405615666190129120449
5. Alzubaidi L, Bai J, Al-Sabaawi A, Santamaria J, Albahri AS, Al-dabbagh BSN, et al. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *J Big Data.* (2023) 10:46. doi: 10.1186/s40537-023-00727-2
6. Nichol A, Dhariwal P, Ramesh A, Shyam P, Mishkin P, McGrew B, et al. *GLIDE: towards photorealistic image generation and editing with text-guided diffusion models.* ARXIV. (2021) 2:10741. doi: 10.48550/ARXIV.2112.10741
7. Zhang C, Zhang C, Zhang M, Kweon IS. *Text-to-image diffusion models in generative AI: a survey* ARXIV. (2023) 3:7909. doi: 10.48550/ARXIV.2303.07909
8. Kather JN, Ghaffari Laleh N, Foersch S, Truhn D. Medical domain knowledge in domain-agnostic generative AI. *npj Digit Med.* (2022) 5:90. doi: 10.1038/s41746-022-00634-5
9. Tschandl P. The HAM10000 dataset, a large collection of multi-source dermoscopic images of common pigmented skin lesions. *Earth Syst Sci Data.* (2018) 5:180161. doi: 10.7910/DVN/DBW86T
10. Xie X, Niu J, Liu X, Chen Z, Tang S, Yu S. A survey on incorporating domain knowledge into deep learning for medical image analysis. *Med Image Anal.* (2021) 69:101985. doi: 10.1016/j.media.2021.101985
11. Man K, Chahl J. A review of synthetic image data and its use in computer vision. *J Imaging.* (2022) 8:310. doi: 10.3390/jimaging8110310
12. Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al. *Zero-shot text-to-image generation* ARXIV. (2021) 2:12092. doi: 10.48550/ARXIV.2102.12092
13. Tao M, Tang H, Wu F, Jing X-Y, Bao B-K, C Xu: *A simple and effective baseline for text-to-image synthesis.* ARXIV. (2020) 8:5865. doi: 10.48550/ARXIV.2008.05865
14. Zhang H, Koh JY, Baldridge J, Lee H, Yang Y. Cross-modal contrastive learning for text-to-image generation. *ARXIV.* (2021) 101:4702. doi: 10.48550/ARXIV.2101.04702
15. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of styleGAN. *ARXIV.* (2019) 12:4958. doi: 10.48550/ARXIV.1912.04958
16. Razavi A, Oord A, Den Van, Vinyals O. *Generating diverse high-fidelity images with VQ-VAE-2.* ARXIV. (2019) 6:446. doi: 10.48550/ARXIV.1906.00446
17. Zhou S, Gordon ML, Krishna R, Narcomey A, Fei-Fei L, Bernstein MS. *HYPE: A Benchmark for Human eYe Perceptual Evaluation of Generative Models* (2019) 4:1121. doi: 10.48550/ARXIV.1904.01121
18. Song Y, Ermon S. Generative modeling by estimating gradients of the data distribution. *ARXIV.* (2019) 7:5600. doi: 10.48550/ARXIV.1907.05600
19. Dhariwal P, Nichol A. Diffusion models beat GANS on image synthesis. *ARXIV.* (2021) 5:5233. doi: 10.48550/ARXIV.2105.05233
20. Ho J, Salimans T. Classifier-free diffusion guidance. *ARXIV.* (2022) 7:12598. doi: 10.48550/ARXIV.2207.12598



OPEN ACCESS

EDITED BY

Antonio Costanzo,
Humanitas Research Hospital, Italy

REVIEWED BY

Clare Primiero,
The University of Queensland, Australia
Eugenio Vocaturo,
University of Calabria, Italy

*CORRESPONDENCE

Dilraj Kalsi
✉ dilraj@skinanalytics.co.uk

RECEIVED 21 July 2023

ACCEPTED 10 October 2023

PUBLISHED 31 October 2023

CITATION

Thomas L, Hyde C, Mullarkey D, Greenhalgh J, Kalsi D and Ko J (2023) Real-world post-deployment performance of a novel machine learning-based digital health technology for skin lesion assessment and suggestions for post-market surveillance. *Front. Med.* 10:1264846. doi: 10.3389/fmed.2023.1264846

COPYRIGHT

© 2023 Thomas, Hyde, Mullarkey, Greenhalgh, Kalsi and Ko. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Real-world post-deployment performance of a novel machine learning-based digital health technology for skin lesion assessment and suggestions for post-market surveillance

Lucy Thomas¹, Chris Hyde², Dan Mullarkey³, Jack Greenhalgh³, Dilraj Kalsi ^{3*} and Justin Ko⁴

¹Chelsea and Westminster Hospital NHS Foundation Trust, London, United Kingdom, ²Exeter Test Group, Department of Health and Community Sciences, University of Exeter Medical School, Exeter, United Kingdom, ³Skin Analytics Ltd., London, United Kingdom, ⁴Department of Dermatology, Stanford Medicine, Stanford, CA, United States

Introduction: Deep Ensemble for Recognition of Malignancy (DERM) is an artificial intelligence as a medical device (AIaMD) tool for skin lesion assessment.

Methods: We report prospective real-world performance from its deployment within skin cancer pathways at two National Health Service hospitals (UK) between July 2021 and October 2022.

Results: A total of 14,500 cases were seen, including patients 18–100 years old with Fitzpatrick skin types I–VI represented. Based on 8,571 lesions assessed by DERM with confirmed outcomes, versions A and B demonstrated very high sensitivity for detecting melanoma (95.0–100.0%) or malignancy (96.0–100.0%). Benign lesion specificity was 40.7–49.4% (DERM-vA) and 70.1–73.4% (DERM-vB). DERM identified 15.0–31.0% of cases as eligible for discharge.

Discussion: We show DERM performance in-line with sensitivity targets and pre-marketing authorisation research, and it reduced the caseload for hospital specialists in two pathways. Based on our experience we offer suggestions on key elements of post-market surveillance for AIaMDs.

KEYWORDS

artificial intelligence, skin cancer, AI for skin cancer, AI as a medical device, DERM, deep ensemble for the recognition of malignancy, Skin Analytics

Introduction

One in every three cancers diagnosed is skin cancer (1). Melanoma is responsible for 90% of skin cancer deaths despite accounting for only ~1% of skin cancers (2). In the United Kingdom (UK), suspected cancer cases are referred to the urgent 2-week-wait (2WW) pathway, in which guidelines suggest that the patient should be seen by a specialist within 2 weeks. Setting this target has been shown to improve the average 5-year melanoma survival by 20%, when compared to historical data (3); however UK cancer registry data

shows that the number of 2WW referrals for skin cancer has increased by more than 200% over the last decade, from 159,430 patients in 2009/2010 to 506,456 patients in 2019/2020 (4), leading to significant access pressures and challenges to achieve standards for timely assessment. Adding to the challenge, approximately 25% of melanoma are found in routine (non-urgent) dermatology referrals or follow-up appointments (5). While in 2009/2010, >94% of patients referred for routine dermatology assessment were seen within the target of 18 weeks, only 80% were seen within this target in 2019/20. Increased patient backlogs since the COVID-19 pandemic mean waiting times have increased with routine clinics often cancelled in order to accommodate additional 2WW activity, leading to downstream delays in the skin cancers, including melanomas, presenting in the routine pathway (6). The increase in skin cancer referrals is expected to continue to rise in the coming decades across Europe and the USA due to ageing populations (7).

Artificial intelligence as a medical device (AIaMD) has the potential to help increase workflow efficiency through triage and supporting clinical decisions in skin cancer pathways (8–13); however, evidence for AIaMDs has largely reflected performance using retrospective data (13–16). There remains the need to understand how appropriately regulated AIaMD platforms perform in real-world clinical settings, including how algorithmic improvements or optimisation for different patient populations affects performance over time. Implementing AI systems in real-world settings reveal often-unforeseen complexities (17). Post-market surveillance (PMS) of medical devices, including AIaMDs, is mandated by regulatory agencies, including the UK Medicines and Healthcare Regulatory Agency (MHRA) and the United States Food and Drug Administration (FDA), but these bodies do not stipulate specific approaches on what data should be collected with what frequency, how it should be analysed, or what auditing and quality control processes should take place (Figure 1) (18–20).

Deep Ensemble for Recognition of Malignancy (DERM; Skin Analytics, London, UK) is an AIaMD that uses deep learning techniques to assess dermoscopic images of skin lesions, identify features associated with malignancies and support referral decisions for patients ≥ 18 years (8–13). DERM is intended to be used for the screening, triage, and assessment of skin lesions, and outputs a suggested diagnosis and referral recommendation. DERM can output a suggested diagnosis of melanoma, squamous cell carcinoma (SCC), basal cell carcinoma (BCC), intraepidermal carcinoma (IEC), actinic keratosis, atypical naevus, or benign, alongside a referral recommendation as agreed for the pathway with local clinical teams. In June 2022 it became the first and only AIaMD for dermatology to be certified as a Class IIa UKCA medical device after an in-depth assessment of Skin Analytics' quality management system and technical documentation by a UK approved body (SGS United Kingdom Ltd, Leicester, UK) designated by the MHRA (previously DERM was a Class I CE device). This manuscript describes the real-world deployment of DERM in clinical practice at two National Health Service (NHS) Trusts in the UK and proposes an approach for the prospective collection and presentation of real-world PMS data from AIaMDs deployed within clinical pathways for ongoing post-deployment monitoring and quality control.

Materials and methods

Study type and location

The analysis is part of the ongoing PMS protocol for DERM to assess its performance in the identification of malignant skin lesions. The data was collected for the service evaluation of DERM commercial deployments in line with its approved intended use. Consistent with medical device regulations, the analysis did not require additional institutional ethics committee approval. All data were collected and analysed according to good clinical practice guidelines and the relevant national laws. All participating patients provided informed consent for their assessment using DERM as part of the service provided by Skin Analytics (data used for case-level analysis), and nearly all (96.7%) provided additional written informed consent for their data to be used for purposes of research and education (data used in the lesion-level analysis).

The data were collected from commercial deployments at University Hospitals Birmingham NHS Foundation Trust (UHB) and West Suffolk NHS Foundation Trust (WSFT). UHB is a large Trust in England treating over 2.8 million patients each year (21). WSFT serves a smaller and predominantly rural geographical area with a population of around 280,000 (22).

DERM software deployment

During the time covered by the analysis, there were two versions of DERM deployed and we refer to them as DERM-version A (DERM-vA) (July 2021 to April 2022), and version B (DERM-vB) (April 2022 to October 2022). Each version used fixed sensitivity thresholds in order to meet sensitivity targets of at least 95% for melanoma and squamous cell carcinoma (SCC) and 90% for basal cell carcinoma (BCC), intraepidermal carcinoma (IEC) and actinic keratosis. The decision to update to DERM-vB was based on confidence in the revised version's ability to maintain target threshold sensitivity for malignancy diagnoses while increasing specificity for benign lesions.

Urgent skin cancer referral pathway

Patient selection for DERM deployment

Figure 2 shows the deployment workflow at UHB and WSFT where DERM was used as a triage tool within the urgent 2WW referral pathway. The referral pathways incorporating DERM were designed in collaboration with the clinical teams at both hospitals and consistent with regulated intended use. Patients with suspicious skin lesions were referred by their general practitioner (GP) to attend a teledermatology hub where a clinical photographer or healthcare assistant (CP/HCA) captured standardised photographic images of their lesion(s) and recorded their medical history. Fitzpatrick skin type was optionally assessed and recorded by the CPs/HCA in conjunction with the patient (23). The imaging team members were also responsible for recording patient consent and assessing whether the patient's lesions were suitable for assessment by DERM according to its intended use (Table 1).

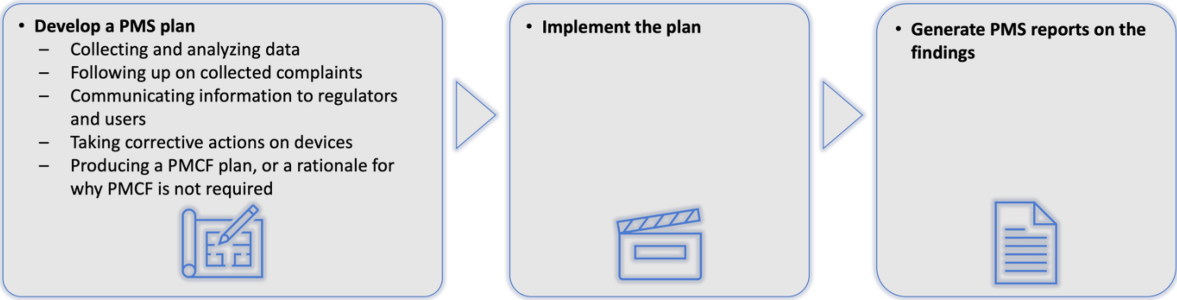


FIGURE 1 Current post-market surveillance requirements for AIaMDs (11, 12). AIaMD, artificial intelligence as a medical device; PMS, post-market surveillance; PMCF, post-market clinical follow-up.

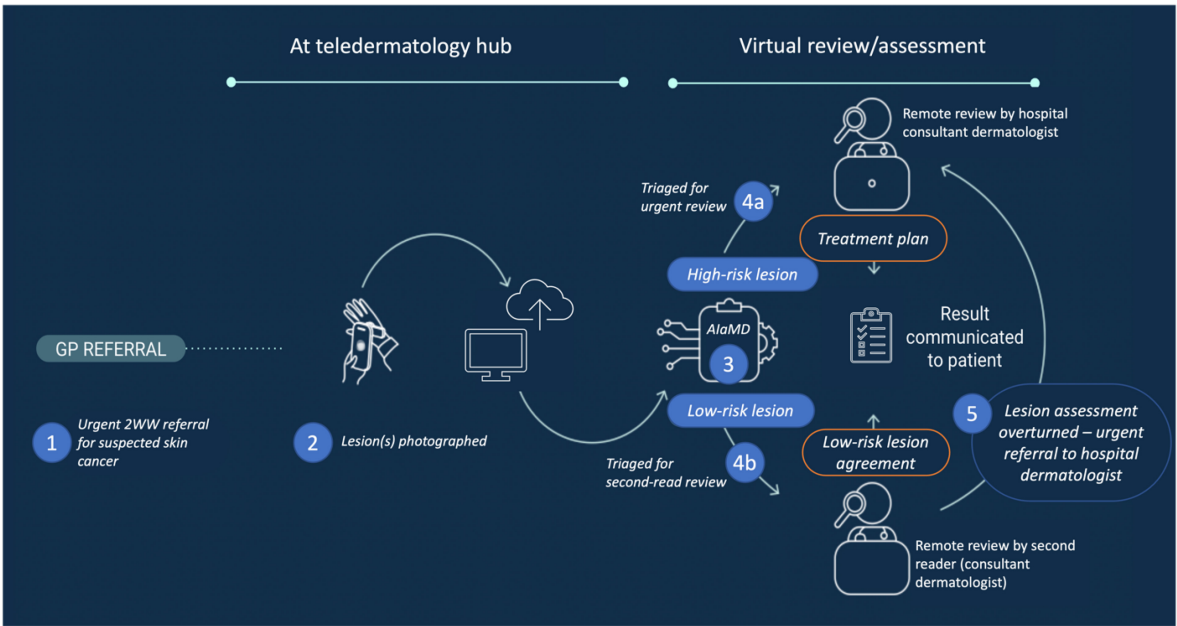


FIGURE 2 Post-referral pathway for DERM. 2WW, 2-week-wait; AIaMD, artificial intelligence as a medical device; DERM, deep ensemble for recognition of malignancy.

TABLE 1 Eligibility criteria for assessment by DERM according to its intended use.

Inclusions	Exclusions
Lesions are eligible to be assessed by DERM if they are: <ul style="list-style-type: none">Located on adults =18 yearsBetween 1 and 3 suspicious lesions which are not larger than the dermatoscopic lens (=15 mm)	<ul style="list-style-type: none">Patients <18 yearsSkin lesions that are not potentially malignant (e.g., rashes, eczema, infectious diseases, lupus)Skin lesions requiring monitoring for treatment responseSkin lesions that require staging of diseaseNon-dermoscopic images of skin lesionsOpen ulcerated skin lesionsSkin lesions too large to be entirely imaged within the dermoscopic device (=15 mm)Lesions obscured by hair, tattoos or scarsLesions which are subungual, or on mucosal, genital or palmoplantar surfacesLesions that have been previously biopsied

DERM, deep ensemble for recognition of malignancy.

Lesion imaging

Patients had locating, macroscopic and dermoscopic digital images of their lesion(s) captured by CPs/HCAs using a smartphone (iPhone 6S or 11; Apple, CA, USA) and polarised

dermoscopic lens attachment (Dermlite DL1 basic, Schuco, UK). For some patients, additional images were captured using a digital single-lens reflex camera (DSLR) with a dermoscopic lens attachment for clinical use (the DSLR images were not assessed

by DERM). Routine post-market auditing identified that a small number of images were captured in error using an unapproved non-polarised dermoscopy tool; however, none of these were excluded from this analysis.

DERM assessment and triage recommendation

The dermoscopic image of eligible patients' lesion(s) was assessed by DERM, which provided a suggested diagnosis and corresponding recommendation, e.g., discharge from pathway or refer to the hospital-based consultant dermatologist for review. DERM classified lesions as melanoma, SCC, BCC, IEC, actinic keratosis, atypical naevus, or benign (six subcategories of benign lesions were grouped together aligned with patient management). DERM's output regarding suggested diagnosis corresponded to the highest risk possibility rather than the most likely classification, e.g., if a lesion was more likely to be a seborrhoeic keratosis but also crossed the defined threshold for melanoma, DERM would output melanoma. Patients for whom all lesions were assessed by DERM and classified as benign were eligible for discharge. Patients with any lesion classified by DERM as not benign or excluded from DERM assessment remained on the urgent 2WW pathway.

Human in the loop: second-read review for benign lesions

Although not required given the Class IIa medical device designation, a second-read review of all cases marked for discharge by DERM was conducted within 48 h by a consultant dermatologist, listed on the UK General Medical Council's Specialist Register (second-read reviewer), working with Skin Analytics and who could agree with or overturn the recommendation to discharge from the 2WW skin cancer pathway. The second-read reviewer had access to the patient's clinical information and smartphone-captured images but not the DSLR images. If the second-read reviewer overturned the recommendation to discharge, the case was referred for hospital dermatologist review.

Cases marked for urgent referral directly by DERM, indirectly via the second-read review, or excluded from DERM assessment were assessed virtually by a hospital consultant dermatologist to provide a clinical diagnosis and final recommendation, e.g., discharge, surgery/biopsy, or clinical follow-up. All hospital dermatologists had access to the patient's clinical information, smartphone images, and additional DSLR images (if available).

Lesion- and case-level analysis

Two different populations were analysed: (1) DERM-assessed lesions that had a final diagnosis (defined by histology for malignant lesions and by dermatologist clinical assessment or histology if available for non-malignant lesions) and the patient had provided additional research consent allowing for assessment of performance of DERM on specific lesions; and (2) case-level data gathered from all patients who were assessed within the pathways described above allowing for assessment of performance of the service integrating DERM overall. The latter includes cases with no DERM assessment (e.g., due to exclusions or technical issues) and where the final diagnosis is still pending. The two populations are expected to be sufficiently similar for interpretation of results to be meaningful with a high patient uptake for additional research consent.

Performance of DERM lesion classification (lesion-level population)

The performance of DERM was evaluated by comparing its lesion classification and management recommendation with the final diagnosis. The performance of DERM compared to the final diagnosis was analysed as to whether it correctly classified lesions as: (1) melanoma or not, whereby a true positive is a histology-confirmed melanoma labelled melanoma by DERM; (2) malignancy or not, whereby a true positive is a histology-confirmed melanoma, SCC, BCC or rare skin cancer labelled as melanoma, SCC or BCC by DERM; and (3) refer or not, whereby a true positive is a histology-confirmed melanoma, SCC, BCC or rare skin cancer or a histology/clinically confirmed Bowen's disease, actinic keratosis, atypical naevus or other premalignant lesion labelled as anything other than benign by DERM ([Supplementary Table 1](#)). Sensitivity, specificity, negative predictive value (NPV), positive predictive value (PPV), and number needed to biopsy/refer/treat (NNB) with their 95% confidence intervals were calculated for all three levels of lesion classification.

Performance of service (case-level population)

The 2WW skin cancer pathway involving DERM was assessed in terms of the proportion of patients with lesions who were safely discharged after DERM, second-read review and hospital dermatologist assessment, respectively. Cancers confirmed from DERM-discharged cases overturned by the second-read and instances where lesions were discharged but histologically confirmed as a cancer on a subsequent presentation ("repeat presentations") were identified and underwent a root cause analysis including a panel review (three dermatologists and an AI expert). Sensitivity for the overall service is reported, whereby repeat presentations of lesions occurring within 6 months of initial discharge are considered false negatives.

Proposal for monitoring post-market surveillance

Based on this experience of deploying an AIaMD in real-world clinical practice, the authors present the current, as well as proposed framework for post-deployment monitoring and quality control of AI in real-world clinical settings.

Results

Patient population

In total, 8,809 cases (patients) at UHB and 2,116 cases at WSFT were assessed by DERM (case-level population; [Figure 3](#)). The number of lesions with a final diagnosis and patient consent for research was 7,220 at UHB and 1,351 at WSFT (lesion-level population). A broad age range of patients were included (18–100 years) and all Fitzpatrick skin types were represented with the majority being skin types I–IV ([Table 2](#)), reflecting skin cancer incidence among these populations ([24](#)).

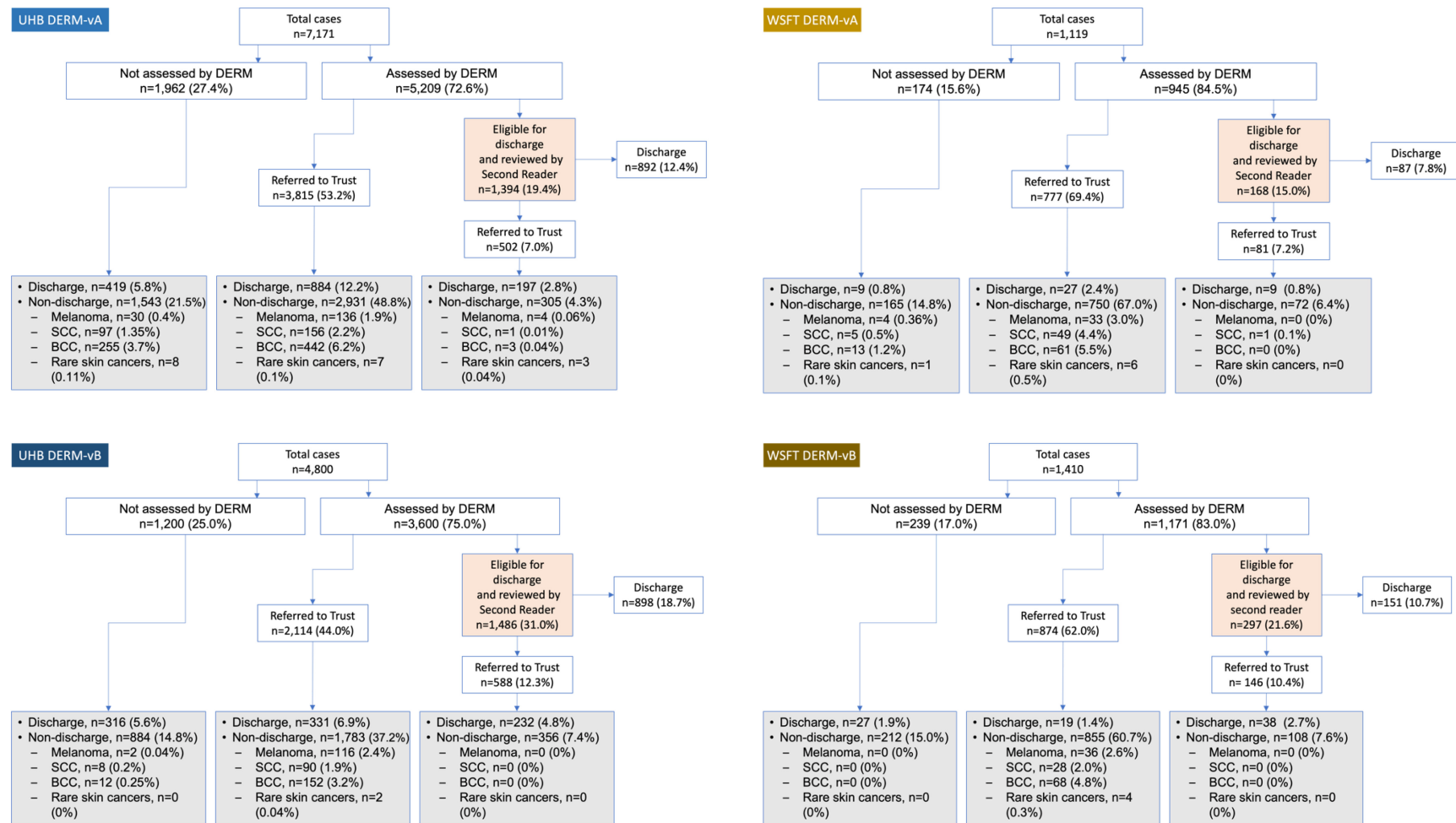


FIGURE 3

Post-deployment DERM referral pathway at two secondary care hospitals in England, United Kingdom (case-level population). Lesions that were not assessed by DERM were referred straight to hospital teledermatology review. Grey- and orange-shaded boxes indicate hospital teledermatology or second-reader review, respectively. For non-discharged lesions, details are provided only for lesions with a skin cancer diagnosis. BCC, basal cell carcinoma; DERM-vA or -vB, deep ensemble for recognition of malignancy version A or B; SCC, squamous cell carcinoma.

TABLE 2 Skin type and age of participants (lesion-level population*).

	UHB DERM-vA (<i>n</i> = 4,635)	WSFT DERM-vA (<i>n</i> = 709)	UHB DERM-vB (<i>n</i> = 2,585)	WSFT DERM-vB (<i>n</i> = 642)
Skin type				
Fitzpatrick type I	248 (5.4%)	125 (17.6%)	215 (8.3%)	74 (11.5%)
Fitzpatrick type II	721 (15.6%)	425 (59.9%)	656 (25.4%)	345 (53.7%)
Fitzpatrick type III	607 (13.1%)	149 (21%)	619 (23.9%)	205 (31.9%)
Fitzpatrick type IV	127 (2.7%)	7 (1%)	132 (5.1%)	14 (2.2%)
Fitzpatrick type V	25 (0.5%)	1 (0.1%)	46 (1.8%)	4 (0.6%)
Fitzpatrick type VI	3 (0.1%)	1 (0.1%)	14 (0.5%)	0 (0%)
Not recorded	2904 (62.7%)	1 (0.1%)	903 (34.9%)	0 (0%)
Age range, years				
18–29	393 (8.5%)	43 (6.1%)	247 (9.6%)	41 (6.4%)
30–39	502 (10.8%)	56 (7.9%)	320 (12.4%)	61 (9.5%)
40–49	505 (10.9%)	60 (8.5%)	302 (11.7%)	71 (11.1%)
50–59	805 (17.4%)	106 (15%)	461 (17.8%)	116 (18.1%)
60–69	874 (18.9%)	123 (17.3%)	503 (19.5%)	129 (20.1%)
70–79	983 (21.2%)	192 (27.1%)	458 (17.7%)	147 (22.9%)
≥80	573 (12.4%)	129 (18.2%)	294 (11.4%)	77 (12%)
Not recorded	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Data are presented as *n* (%). *Lesions were included in the analysis if there was a confirmed final diagnosis (histology for malignant lesions and dermatologist opinion or histology for non-malignant lesions). Lesions were excluded from the analysis if they did not fulfil the inclusion criteria for lesion assessment by DERM, were not analysed by DERM for any technical reason or were pending final diagnosis defined by histology for malignant lesions and by dermatologist clinical assessment or histology if available for non-malignant lesions. DERM, deep ensemble for recognition of malignancy.

Performance of DERM lesion classification (lesion-level population)

Post-deployment performance of DERM-vA and DERM-vB are reported in [Table 3](#). Both versions of DERM performed with very high levels of sensitivity for skin cancer detection (96.0–100.0%). DERM-vB labelled 246 out of 248 lesions as skin cancer; the remaining two lesions were referred with a label of Bowen's disease and later confirmed to be BCC. Specificity was 40.7–49.4% for DERM-vA and 70.1–73.4% for DERM-vB. A total of 159 lesions were assessed in patients with Fitzpatrick skin types V and VI, for which 94 lesions had a final diagnosis, including BCC (*n* = 1) and IEC (*n* = 1), and actinic keratosis (*n* = 1), all correctly referred by DERM, and atypical naevus (*n* = 3) pending face-to-face assessment, and the remainder were benign with a benign specificity of 44.3% (39/88).

Rare skin cancers

Among the lesions assessed, 19 rare skin cancers (defined as not melanoma, SCC or BCC and comprised trichilemmal carcinoma, dermal sarcoma, atypical fibroxanthoma and marginal zone lymphoma) were identified, of which DERM-vA and -vB labelled 13/16 and 3/3 lesions as melanoma or SCC, respectively. Three lesions were labelled “benign” by DERM-vA: two subdermal foci of melanoma with no cutaneous changes [during root cause analysis (RCA) these lesions were assessed as not having been suitable for the service] and one marginal zone lymphoma with cutaneous changes. Complete confusion matrixes for DERM lesion classifications are provided in [Supplementary Table 3](#).

Performance of service (case-level population)

Second-read review

For DERM-vA, 1,393/5,209 cases assessed (26.8%) were labelled as eligible for discharge at UHB. The second-read reviewer overturned 502/1,393 cases (36.0%), of which the hospital dermatologist discharged 197/502 (39.2%). A total of 11 skin cancers (2.0%) were found among these cases. At WSFT, 168/945 (17.8%) cases evaluated by DERM-vA were labelled as eligible for discharge. The second read overturned 81/168 (48.2%) cases, of which the hospital dermatologist discharged 9/81 (11.1%). One skin cancer was found (1.2%) among these cases.

For DERM-vB, 1,486/3,603 cases assessed (41.2%) were labelled as eligible for discharge at UHB. The second read overturned 588/1,486 cases (39.6%), of which the hospital dermatologist discharged 232/588 (39.5%). No skin cancers were found (0% conversion) among lesions marked eligible for discharge by DERM-vB. At WSFT, 297/1,410 cases (25.4%) evaluated by DERM-vB were labelled as eligible for discharge. The second read overturned 146/297 cases (49.2%), of which the hospital dermatologist discharged 38/146 (26.0%). No skin cancers were found (0% conversion).

Repeat presentations

No lesions have been assessed by DERM-vA or -vB and discharged from these pathways with a subsequent re-presentation and diagnosis of cancer (service sensitivity 100% to date); however, there have been four lesions that presented twice to the UHB

TABLE 3 Post-deployment performance of DERM (lesion-level population).

Lesions, % (n/N) [95% confidence interval]	Melanoma or not	Malignant or not	Refer or not
Sensitivity			
DERM-vA (UHB)	95.0% (133/140) [90–97.6%]	96.0% (722/752) [94.4–97.2%]	93.4% (1667/1784) [92.2–94.5%]
DERM-vA (WSFT)	97.0% (32/33) [84.7–99.5%]	99.3% (149/150) [96.3–99.9%]	94.9% (316/333) [92–96.8%]
DERM-vB (UHB)	100.0% (58/58) [93.8–100%]	98.9% (178/180) [96–99.7%]	87.4% (673/770) [84.9–89.6%]
DERM-vB (WSFT)	100.0% (18/18) [82.4–100%]	100.0% (68/68) [94.7–100%]	89.5% (222/248) [85.1–92.7%]
Specificity			
DERM-vA (UHB)	58.8% (2643/4495) [57.4–60.2%]	45.0% (1747/3883) [43.4–46.6%]	49.4% (1408/2851) [47.6–51.2%]
DERM-vA (WSFT)	63.2% (427/676) [59.5–66.7%]	33.1% (185/559) [29.3–37.1%]	40.7% (153/376) [35.8–45.7%]
DERM-vB (UHB)	80.9% (2045/2527) [79.3–82.4%]	64.8% (1559/2405) [62.9–66.7%]	73.4% (1333/1815) [71.4–75.4%]
DERM-vB (WSFT)	80.4% (502/624) [77.2–83.4%]	60.6% (348/574) [56.6–64.5%]	70.1% (276/394) [65.4–74.4%]
Negative predictive value			
DERM-vA (UHB)	99.7% (2643/2650) [99.5–99.9%]	98.3% (1747/1777) [97.6–98.8%]	92.3% (1408/1525) [90.9–93.6%]
DERM-vA (WSFT)	99.8% (427/428) [98.7–100%]	99.5% (185/186) [97–99.9%]	90.0% (153/170) [84.6–93.7%]
DERM-vB (UHB)	100.0% (2045/2045) [99.8–100.0%]	99.9% (1559/1561) [99.5–100.0%]	93.2% (1333/1430) [91.8–94.4%]
DERM-vB (WSFT)	100% (502/502) [99.2–100%]	100% (348/348) [98.9–100.0%]	91.4% (276/302) [87.7–94.1%]
Positive predictive value			
DERM-vA (UHB)	6.7% (133/1985) [5.7–7.9%]	25.3% (722/2858) [23.7–26.9%]	53.6% (1667/3110) [51.8–55.3%]
DERM-vA (WSFT)	11.4% (32/281) [8.2–15.6%]	28.5% (149/523) [24.8–32.5%]	58.6% (316/539) [54.4–62.7%]
DERM-vB (UHB)	10.7% (58/540) [8.4–13.6%]	17.4% (178/1024) [15.2–19.8%]	58.3% (673/1155) [55.4–61.1%]
DERM-vB (WSFT)	12.9% (18/140) [8.3–19.4%]	23.1% (68/294) [18.7–28.3%]	65.3% (222/340) [60.1–70.2%]
Number needed to biopsy, treat or refer			
DERM-vA (UHB)	14.9 (1985/133) [12.7–17.6]	4 (2858/722) [3.7–4.2]	1.9 (3110/1667) [1.8–1.9]
DERM-vA (WSFT)	8.8 (281/32) [6.4–12.2]	3.5 (523/149) [3.1–4]	1.7 (539/316) [1.6–1.8]
DERM-vB (UHB)	9.3 (540/58) [7.3–11.9]	5.8 (1024/178) [5.0–6.6]	1.7 (1155/673) [1.6–1.8]
DERM-vB (WSFT)	7.8 (140/18) [5.2–12.1]	4.3 (294/68) [3.5–5.4]	1.5 (340/222) [1.4–1.7]

DERM, deep ensemble for recognition of malignancy; UHB, University Hospital Birmingham; WSFT, west sussex foundation trust.

pathway before July 2021, with the second presentation resulting in a histologic diagnosis of skin cancer (melanoma, $n = 2$; BCC, $n = 2$; [Supplementary Table 2](#)), though only one (a melanoma) was within 6 months. These were all either triaged by DERM to Trust teledermatology review ($n = 4$) or excluded from assessment by DERM at either the first ($n = 2$) or second presentation ($n = 2$) to the pathway.

Discussion

Herein, we present a real-world deployment performance evaluation for the AIaMD, DERM, which uses deep learning techniques to assess dermoscopic images of skin lesions for patients who were referred to an urgent skin cancer pathway. During the assessed period, DERM performed at or above the expected level for all malignant and pre-malignant lesion types based on 1,150 confirmed malignancies, including 249 melanomas and 19 rare malignancies. DERM-vB correctly referred all skin cancers in these pathways and had a specificity greater than the previous DERM-vA version. During this period, no patients were discharged

from the service and re-presented later with the same lesion being diagnosed as skin cancer. While other published evidence demonstrate a gap in model performance when evaluating real-world prospective clinical use compared with *in silico* data (25–27), our analysis demonstrates that DERM can be deployed safely in live clinical services accessible to patients from a broad range of age groups and skin types, with sensitivity and specificity in-line with target thresholds and performance demonstrated in pre-marketing authorisation studies (8–13).

A critical issue is whether the estimates of performance are valid in this real-world deployment. We examined this by considering the validity and applicability issues identified in the QUADAS-2 tool ([Supplementary Appendix A](#)), the most commonly used quality assessment tool for test accuracy studies (28). This reveals that the general openness to bias is similar to many studies included in systematic reviews, particularly those produced by the Cochrane Collaboration. The area of greatest concern is patient selection, whereby there is not a perfectly consecutive series of patients due to current exclusion criteria; however, given that ~80% of all patients referred for suspected skin cancer to UHB and WSFT were seen by these pathways, there is a high level of consecutiveness.

Other concerns relate to the information provided by DERM being available to those making the reference standard diagnosis, although arguably this is unlikely to introduce bias because of the current general scepticism about the value of AI by the medical community and the positioning of dermatologists in the pathway to either review lesions already identified as high risk or to actively screen for false negatives. Finally, there is differential verification in the reference standard (ground truth), but this is a near universal problem for the evaluation of the accuracy of skin cancers because it is unethical to biopsy all patients in a study, particularly those deemed as having a low likelihood of cancer and this mirrors limitations within any evaluation of current standards of care. Concerning applicability, the study scores highly, and this should be seen as a particular strength for a real-world deployment.

Although the DERM PMS programme was established before the CLEAR consensus guidelines were published for evaluation of AI studies in dermatology (29), our post-deployment data collection methods align with the relevant checklist items, including prospective data collection, and providing details of image acquisition, patient skin colour, deployment referral pathway, hierarchical outputs, and technical assessments of performance. We did not collect ethnicity or patient sex as we operated on the principle of only collecting data necessary to inform or evaluate DERM performance as part of DERM's PMS. We plan to re-evaluate the future role of collecting and reporting on these demographic data elements.

Although DERM used images captured using an iPhone camera, it is not a smartphone app *per se*. In contrast, there are numerous smartphone apps intended to classify skin lesions (30). An analysis of 43 such apps showed that these had a mean sensitivity of 0.28 [95% confidence interval (CI) 0.17–0.39], mean specificity of 0.81 (95% CI 0.71–0.91) and mean accuracy of 0.59 (95% CI 0.55–0.62) for the detection of melanoma (31). Direct-to-consumer products do not meet the standards necessary for utilisation in clinical pathways. Direct-to-consumer products generally are not integrated into healthcare services that enable definitive diagnosis, management recommendation and treatment.

Our real-world evidence suggests that DERM can make autonomous decisions to discharge patients with benign skin lesions from the urgent cancer pathway. The second-read reviewer overturned 40–50% of cases that DERM had marked as eligible for discharge; however, for DERM-vA, only 1.2% of these cases resulted in skin cancer diagnosis and with DERM-vB, none resulted in a skin cancer diagnosis. Cost-benefit and economic analyses for the service are ongoing and supported by a 2021 NHS AI in Health and Care Award (32). Adherence to regulatory standards and continuous monitoring need to ensure that autonomous decisions made by AIaMDs are carried out safely while augmenting the non-specialist clinicians' involvement in care, including in the appropriate counselling of patients.

Suggestions for post-market surveillance for AI medical devices

Medical device regulations which govern AIaMDs are in place to support access to safe and effective devices and limit access to products that are unsafe. This includes the requirement

that manufacturers must submit vigilance reports to the relevant regulatory agency when certain incidents occur involving their device. Although all medical devices require PMS as part of the manufacturer's obligations to ensure that their device continues to meet appropriate standards of safety and performance for as long as it is in use, these requirements are not specific and there is currently limited transparency on how PMS is being conducted by manufacturers. As such, we recommend that manufacturers monitor and publish real-world evaluations of their AIaMDs within a clinically relevant timeframe. There is a need for PMS alignment to reduce variability of surveillance design and analysis and to improve comparability with other AIaMDs or to monitor the same device over time. Guidelines for best-practice evaluation of image-based AI development in dermatology (CLEAR Derm consensus) provide a checklist to ensure consistency but these are aimed at clinical development as opposed to post-deployment data collection. Nevertheless, many of the items listed in the checklist are pertinent to PMS (29). Manufacturers of AIaMDs may also benefit from specific, tangible advice to support their PMS development plans and regulators and adopters (users) should have a good understanding of what to expect from real-world evidence collected as part of PMS plans (Table 4). PMS processes need to have automatic safeguards or systems in place to ensure rigorous monitoring for robust performance of the AIaMD. Moreover, collecting, analysing, and publishing PMS data requires significant collaboration between the manufacturer, healthcare provider partners, healthcare professionals and patients. Automatic systems, such as electronic patient records that auto-populate a registry database may improve the collection of long-term patient outcomes that go beyond monitoring the specificity and sensitivity of the AIaMD.

Data management

Post-market surveillance data collection methods need to be planned before AIaMD deployment, including what is needed to ensure ongoing performance and any baseline values that would be useful. The manufacturer needs to put in place plans for auditing and data quality assurance.

A period of continuous monitoring is required to ensure that the AIaMD is performing as expected, especially when there are software updates or changes to the deep learning algorithms that may affect performance. As such, processes need to be able to quickly identify and analyse performance errors so that these can be corrected, and future occurrences prevented (33). For example, during initial deployment, a second-read review would provide a safety net until performance is at or above the expected targets. A statistically significant amount of continuous data with performance at or above expected targets is achieved in alignment with regulatory standards and intended use; for DERM, the demonstration across two distinct locations may support its deployment without a human second-read.

Manufacturers need to start conversations with healthcare providers as early as possible, to consider contractual obligations or incentives to ensure the manufacturer has access to data required for PMS in a timely manner. There is considerable variation in terms of which stakeholder owns or can access the data required in any given organisation. Data requirements need to be agreed with all stakeholders, with ongoing discussions and iterations to ensure the data being collected and analysed remain relevant for

TABLE 4 Post-market surveillance recommendations for monitoring AIaMDs deployed in real-world settings.

When	Responsibility	Recommendation	Process consideration	Dermatology example
Prior to deployment	Manufacturer	Document and share PMS Plan with healthcare provider partners. This should include final outcome definitions, data sources and cadence of performance reports	Time and resource implications for healthcare providers to acknowledge/review the PMS plan	Agree all skin cancer outcomes will be based on histopathologically confirmed cases to mitigate for inter-clinician variation and mirror clinical practice
	Manufacturer and deploying organisation	Agree how data will be shared with the manufacturer to support	Time and resource implications for healthcare providers. Data privacy and data sharing compliance with patient consent and local laws	Access to histology reports for cases assessed through the service
	Manufacturer	Agree RCA process for false negatives	Process may also be applicable for further investigation of other areas of interest, e.g., low incidence populations, rare diseases, common false positives	Consideration of patient history vs. macro imaging vs. dermoscopic imaging as key factors in cancer diagnosis of a false negative
During deployment and as set out in PMS plan	Manufacturer	Agreement on how many cases should be reviewed initially with second-read review (human-in-the-loop) as a safety net, with a performance review before removal	Time and resource savings should only be considered once the AIaMD has proven to operate within acceptable safety limits	Performance at or above stated target sensitivity for skin cancer over a 6-month+ period at =2 deployment sites
	Manufacturer and/or deploying organisation	Active search for repeat presentations	Will patients always present through the same pathway? If not, does the deploying organisation have better data to search for patients presenting with the same complaint more than once?	Has the same patient presented to the service twice regarding the same lesion?
	Manufacturer	Follow RCA Process for all false negatives and share findings with deploying organisation	Time and financial cost associated with conducting process	Multi-step process including detailed review of histology, review of case by panel of dermatologists, adversarial testing
	Manufacturer	Publish performance report including reference to any available benchmark data (i.e., to allow comparison with other health providers and performance over time; ideally, data would be published in a peer-reviewed journal)	Peer-review publication may introduce delays and so as a minimum the performance should be made available to existing partners or upon request by health organisations considering using the AIaMD	Quarterly Performance Report shared with partners including comparison of new pathway performance vs. nationally available conversion rates
	Manufacturer and/or deploying organisation	Risk-registry database to identify common themes and to investigate if agreed thresholds are breached*	Quickly identify any performance issues and their cause	Ensure correct hardware is in use to collect skin lesion images

AIaMD, artificial intelligence as medical device; PMS, post-market surveillance; RCA, root cause analysis. *This should build on existing quality management system and clinical risk management requirements already mandated for medical device manufacturers.

performance assessment. Consideration also needs to be given to liability and data privacy issues, including General Data Protection Regulation (GDPR) or equivalent local legislature and the patient's right to withdraw consent.

Root cause analysis for quality control issues, false negative classifications, and "near-misses"

Deep Ensemble for Recognition of Malignancy has now classified more than 60,000 skin lesions in real-world settings across eleven NHS pathways in the UK that have identified 5,385 histology-confirmed malignant lesions (34–36). Specific guidance on AI quality control and improvement in hospitals has been recently published, which describes detection of errors in AI

algorithms, monitoring software updates, cause-and-effect analysis for a drop in performance, monitoring changes to input or target, the challenges in monitoring AI system variables, and adapting the FDA's existing Sentinel Initiative for monitoring AIaMDs after deployment (37).

In terms of reviewing a false negative, case review should be undertaken by a relevant specialist. When a false negative was identified for DERM post-deployment, a root cause analysis was conducted. Histology reports were reviewed for factors such as uncertainty of diagnosis, staging of disease, subtype of disease and perineural and perivascular invasion. A panel of three dermatologists plus an AI expert reviewed all case details including clinical and dermoscopic images and histology reports,

and assessed which factor(s) contributed to the false negative result. Current labels include whether the lesion should have been excluded from DERM assessment or other technical factors, had an unusual presentation or was due to AI performance issues. Any lesion(s) that were a repeat presentation and were confirmed to have cancer were also identified for false negative review and for these cases the panel was asked to comment on whether the malignancy was likely to be present at the point of the first assessment or whether the transformation took place in the interval between appointments. These false negatives should be collated in a “risk” registry and assessed to identify common themes with thresholds for escalation for more in-depth review.

Considerations arising from assessment of openness to bias

Our reflection on the validity of our data also suggests ways in which the process of PMS data collection could be optimised to maximise validity. Careful attention to documenting and describing legitimate losses to follow-up, patients who are ineligible for assessment and technical failures is particularly important for the credibility of the information. Moreover, documenting repeat presentations provides reassurance that cancers are not being missed. As such, PMS protocols should clearly describe the time intervals that are being used to confirm that a repeat presentation has not occurred. Clear information about how the AIaMD is being used in the final diagnosis would also be helpful to alert to the possibility of bias if there appears to be heavy reliance on its assessment.

Future directions—Looking beyond AIaMD performance at patient outcomes

We are looking to make improvements to the quality of care provided to patients with suspicious skin lesions. Currently, PMS of AIaMDs is focussed on performance, but ultimately data collected as part of PMS should include clinically meaningful metrics, such as reporting the timeliness of diagnosis of malignant lesions after the initial GP referral, time to excision/treatment, provide more information about lesion characteristics (e.g., staging) and importantly longer-term outcomes such as progression-free or overall survival.

Limitations

Deep Ensemble for Recognition of Malignancy is not intended to provide a definitive diagnosis for skin cancer, as the final diagnosis is confirmed by histopathology or a dermatologist for the case of high-risk lesions. Future opportunities exist to realise further potential of DERM to allow patients with benign lesions to be discharged as quickly as possible, including reducing the exclusion rate (e.g., by using larger dermatoscopic lenses) and using additional data to develop and validate its use on mucosal, palmoplantar and subungual lesions. Human factors and user interaction including explainability could also be assessed in future but was outside the scope of this analysis (38). More explainable outputs could include techniques such as saliency maps, differential diagnosis using conformal predictions, or argumentation approaches (39, 40). However, any

additional outputs would need to be validated by human factors and reader studies.

Skin cancers are less common in people with skin of colour (Fitzpatrick skin types V and VI) (24, 41). The current exclusion of palmoplantar and subungual lesions means that DERM cannot be used on the areas where patients with darker skin colour are most likely to develop melanoma (42). Continued surveillance is needed to ensure that patients with darker skin tones have equitable access to the DERM service particularly because patients with darkly pigmented skin often have a more advanced initial melanoma and higher mortality rate than fair-skinned patients (43). This is, however, not a concern that is exclusive to AIaMD-powered skin cancer pathways but rather that appropriate vigilance is required for any skin cancer service.

There is currently a lack of robust baseline operational data from prior to developing and implementing the DERM pathway for UHB and WSFT for number of biopsies, non-melanoma skin cancers diagnosed and pre-malignant diagnoses, or discharge rates for patients with non-malignant lesions. As such, we cannot currently determine how these metrics have changed since the deployment of the DERM pathway.

Overall conclusion

The real-world implementation of DERM, an AIaMD, in two NHS skin cancer pathways, demonstrates high levels of performance. DERM is accessible to adults of all ages (18–100 years) and has been used to assess potential malignant skin lesions in all Fitzpatrick skin types I–VI. The performance of DERM will continue to be assessed as part of its PMS, including continued consideration of accessibility across the whole population. The performance demonstrated to date provides sufficient evidence to support the removal of the second-read for low-risk lesions in order to maximise health system benefits safely. Based on our experience we offer some suggestions on key elements of post-market surveillance for AIaMDs.

Data availability statement

The datasets presented in this article are not readily available because the data included in this manuscript have been collected as part of the routine post-market surveillance programme for DERM, conducted by Skin Analytics, London. Requests to access the datasets should be directed to DK, dilraj@skinanalytics.co.uk and DM, dan@skinanalytics.co.uk.

Ethics statement

Ethical approval was not required for the studies involving humans because this study was conducted as part of ongoing service evaluation of skin cancer pathways and post-market surveillance of a medical device. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

LT: Conceptualization, Formal Analysis, Investigation, Methodology, Supervision, Writing – review and editing. CH: Conceptualization, Formal Analysis, Investigation, Methodology, Supervision, Writing – review and editing. DM: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Visualization, Writing – original draft, Writing – review and editing. JG: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Supervision, Validation, Writing – review and editing. DK: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing. JK: Conceptualization, Formal Analysis, Investigation, Methodology, Supervision, Writing – review and editing.

Acknowledgments

Mr. Christopher Felix Brewer, Dr. Sahar Abdulrahman, Dr. Rachel Jenkins, Dr. Rabia Rashid, Dr. Irshad Zaki, Dr. Simona Ungureanu, and Dr. Zahra Haider, who supported DERM deployment and/or data acquisition at UHB and/or WSFT. Medical writing support was provided by Celia J. Parkyn, Ph.D. The role of the medical writer was to attend meetings between the authors when discussing the manuscript content, write the meeting minutes of author discussion, provide writing support as directed by the authors, coordinate author review and amendments during manuscript development, data check the manuscript, format references, and collect and check the documents required for submission.

Conflict of interest

LT is a clinical advisor to Skin Analytics Ltd., has received Skin Analytics shares or share options; has received research

funding support from Skin Analytics (salaries and equipment) and AIaMD deployment programme; has received reimbursement of conference fees, travel and accommodation costs from Skin Analytics to present research results; LT has received financial remuneration for separate programme of work as a consultant by Skin Analytics; has received grant funding from NHSX and CW+; has received paid honoraria to lecture for Almirall; was supported to attend a conference by Abbvie and Janssen; and holds multiple unpaid leadership roles. CH is a clinical advisor to Skin Analytics Ltd., and has received research funding to undertake a health economic model of the impact of the use of DERM in the NHS. DM is an employee of Skin Analytics Ltd., and has received Skin Analytics shares or share options. DK is an employee of Skin Analytics Ltd., and has received Skin Analytics shares or share options. JG is an employee of Skin Analytics Ltd; has received Skin Analytics shares or share options; and is named as an inventor on patents (pending) relating to DERM. JK is a clinical advisor to Skin Analytics Ltd and has received Skin Analytics shares or share options. The authors declare that this study received funding from Skin Analytics, London, UK. The funder had the following involvement in the study: study design, data collection and analysis, decision to publish, and preparation of the manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

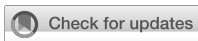
Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2023.1264846/full#supplementary-material>

References

1. World Health Organization [WHO]. *Radiation: Ultraviolet (UV) Radiation and Skin Cancer* [Internet]. (2017). Available online at: [https://www.who.int/news-room/questions-and-answers/item/radiation-ultraviolet-\(uv\)-radiation-and-skin-cancer](https://www.who.int/news-room/questions-and-answers/item/radiation-ultraviolet-(uv)-radiation-and-skin-cancer) (accessed September 20, 2023).
2. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA Cancer J Clin.* (2020) 70:7–30. doi: 10.3322/caac.21590
3. Pacifico M, Pearl R, Grover R. The UK Government two-week rule and its impact on melanoma prognosis: an evidence-based study. *Ann R Coll Surg Engl.* (2007) 89:609–15. doi: 10.1308/003588407X205459
4. Smith L, Sansom N, Hemphill S, Bradley S, Shinkins B, Wheatstone P, et al. Trends and variation in urgent referrals for suspected cancer 2009/2010–2019/2020. *Br J Gen Pract.* (2022) 72:34–7. doi: 10.3399/bjgp.22X718217
5. National Cancer Registration and Analysis Service. *Routes to Diagnosis.* (2023). England: National Cancer Registration and Analysis Service.
6. Levell N. *Dermatology GIRFT Programme National Specialty Report.* (2021). London: National Health Service.
7. Garbe C, Amaral T, Peris K, Hauschild A, Arenberger P, Basset-Seguin N, et al. European consensus-based interdisciplinary guideline for melanoma. Part 1: diagnostics: update 2022. *Eur J Cancer.* (2022) 170:236–55. doi: 10.1016/j.ejca.2022.03.008
8. Phillips M, Greenhalgh J, Marsden H, Palamaras I. Detection of malignant melanoma using artificial intelligence: an observational study of diagnostic accuracy. *Dermatol Pract Concept.* (2020) 10:e2020011. doi: 10.5826/dpc.1001a11
9. Marsden H, Morgan C, Austin S, DeGiovanni C, Venzi M, Kemos P, et al. Effectiveness of an image analyzing AI-based digital health technology to identify Non-Melanoma Skin Cancer (NMSC) and other skin lesions: results of the DERM-003 study. *Front Med Sec Dermatol.* (2023) 10:1288521. doi: 10.3389/fmed.2023.1288521
10. Marsden H, Palamaras I, Kemos P, Greenhalgh J. P63 Effectiveness of an image-analysing artificial intelligence-based digital health technology to diagnose nonmelanoma skin cancer and benign skin lesions. *Br J Dermatol.* (2023) 188 (Supplement_4):ljad113.091. doi: 10.1093/bjd/ljad113.091
11. Marsden H, Kemos P, Venzi M, Noy M, Maheswaran S, Francis N. Accuracy of an Artificial Intelligence as a medical device as part of a UK-based skin cancer teledermatology service. *Front Med.* (2023) 10:1288521.

12. Kawsar A, Hussain K, Kalsi D, Kemos P, Marsden P, Thomas L, et al. Patient Perspectives of Artificial Intelligence as a Medical Device in a Skin Cancer Pathway. Pending publication.
13. Phillips M, Marsden H, Jaffe W, Matin R, Wali G, Greenhalgh J. Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. *JAMA Netw Open*. (2019) 2:e1913436. doi: 10.1001/jamanetworkopen.2019.13436
14. Esteva A, Kuprel B, Novoa R, Ko J, Swetter S, Blau H, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. (2017) 542:115–8. doi: 10.1038/nature21056
15. Fink C, Blum A, Buhl T, Mitteldorf C, Hofmann-Wellenhof R, Deinlein T. Diagnostic performance of a deep learning convolutional neural network in the differentiation of combined naevi and melanomas. *J Eur Acad Dermatol Venereol* (2020) 34:1355–61. doi: 10.1111/jdv.16165
16. Wells A, Patel S, Lee JB, Motaparthy K. Artificial intelligence in dermatopathology: diagnosis, education, and research. *J Cutan Pathol* (2021) 48:1061–8. doi: 10.1111/cup.13954
17. Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P. A human-centred evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu, HI: (2020). doi: 10.1145/3313831.3376718
18. Legislation.gov.uk. *The Medical Devices Regulations 2002*. (2023). Available online at: <https://www.legislation.gov.uk/uksi/2002/618/contents/made> (accessed March 3, 2023).
19. European Union. *REGULATION (EU) 2017/745 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 5 April 2017 on Medical devices, Amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and Repealing Council Directives 90/385/EEC and 93/42/EEC*. (2023). Brussels: European Union.
20. FDA. *Postmarket Surveillance Under Section 522 of the Federal Food, Drug, and Cosmetic Act Guidance for Industry and Food and Drug Administration Staff*. (2023). Silver Spring, MD: FDA.
21. University Hospitals Birmingham NHS Foundation Trust. *Annual Report and Accounts 2018/19*. (2023). Birmingham: University Hospitals Birmingham NHS Foundation Trust.
22. West Suffolk NHS Foundation Trust. *Annual Report and Accounts 2018/19*. (2023). Bury St Edmunds: West Suffolk NHS Foundation Trust
23. Fitzpatrick T. The validity and practicality of sun-reactive skin types I through VI. *Arch Dermatol*. (1988) 124:869–71. doi: 10.1001/archderm.124.6.869
24. Delon C, Brown K, Payne N, Kotrotsios Y, Vernon S, Shelton J, et al. Differences in cancer incidence by broad ethnic group in England, 2013–2017. *Br J Cancer*. (2022) 126:1765–73. doi: 10.1038/s41416-022-01718-5
25. Li C, Fei W, Shen C, Wang Z, Jing Y, Meng R. Diagnostic capacity of skin tumor artificial intelligence-assisted decision-making software in real-world clinical settings. *Chin Med J*. (2020) 133:2020–6. doi: 10.1097/CM9.0000000000001002
26. Zech JR, Badgeley M, Liu M, Costa A, Titano J, Oermann E, et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. (2018) 15:e1002683. doi: 10.1371/journal.pmed.1002683
27. Lin D, Xiong J, Liu C, Zhao L, Li Z, Yu S. Application of Comprehensive Artificial intelligence Retinal Expert (CARE) system: a national real-world evidence study. *Lancet Digit Health*. (2021) 3:e486–95. doi: 10.1016/S2589-7500(21)00086-8
28. Whiting PF, Rutjes A, Westwood M, Mallett S, Deeks J, Reitsma J, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. (2011) 155:529–36. doi: 10.7326/0003-4819-155-8-201110180-00009
29. Daneshjou R, Barata C, Betz-Stablein B, Celebi M, Codella N, Combalia M. 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) 158:90–6. doi: 10.1001/jamadermatol.2021.4915
30. Vocaturo E, Zumpano E. Smart apps for risk assessment of skin cancer. *Proceedings of the 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. Melbourne: (2020). doi: 10.1109/WIIAT50758.2020.00106
31. Sun M, Kentley J, Mehta P, Duszka S, Halpern A, Rotemberg V. Accuracy of commercially available smartphone applications for the detection of melanoma. *Br J Dermatol*. (2022) 186:744–6. doi: 10.1111/bjd.20903
32. NHS England. *Transformation Directorate*. (2023). Leeds: NHS England.
33. United States Food and Drug Administration. *Corrective and Preventive Actions (CAPA)*. (2023). Silver Spring, MD: FDA.
34. Jenkins R, Brewer C, Kalsi D, Mullarkey D. BT09 Clinical performance of an artificial intelligence-based medical device deployed within an urgent suspected skin cancer pathway. *Br J Dermatol*. (2023) 188 (Suppl_4):ljad113.375. doi: 10.1093/bjd/ljad113.375
35. NHS England. *Case Study: Artificial Intelligence Helping to Speed Up Skin Cancer Diagnosis in Leicester, Leicestershire, and Rutland Integrated Care System*. (2023). Available online at: <https://www.england.nhs.uk/long-read/artificial-intelligence-helping-to-speed-up-skin-cancer-diagnosis-in-leicester-leicestershire-and-rutland-integrated-care-system/> (accessed August 25, 2023).
36. Abu Baker K, Roberts E, Harman K, Mullarkey D, Kalsi D. BT06 Using artificial intelligence to triage skin cancer referrals: outcomes from a pilot study. *Br J Dermatol*. (2023) 188 (Suppl_4):ljad113.372. doi: 10.1093/bjd/ljad113.372
37. Feng J, Phillips R, Malenica I, Bishara A, Hubbard A, Celi L. Clinical artificial intelligence quality improvement: towards continual monitoring and updating of AI algorithms in healthcare. *NPJ Digit Med*. (2022) 5:66. doi: 10.1038/s41746-022-00611-y
38. Caroprese L, Vocaturo E, Zumpano E. Argumentation approaches for explainable AI in medical informatics. *Intellig Syst Applic*. (2022) 16:200109. doi: 10.1016/j.iswa.2022.200109
39. Singh A, Sengupta S, Lakshminarayanan V. Explainable deep learning models in medical image analysis. *J Imaging*. (2020) 6:52. doi: 10.3390/jimaging6060052
40. Lu C, Lemay A, Chang K, Höbel K, Kalpathy-Cramer J. Fair conformal predictors for applications in medical imaging. *Proc AAAI Conf Artif Intellig*. (2022) 36:12008–16. doi: 10.1609/aaai.v36i1.21459
41. Hogue L, Harvey V. Basal cell carcinoma, squamous cell carcinoma, and cutaneous melanoma in skin of color patients. *Dermatol Clin*. (2019) 37:519–26. doi: 10.1016/j.det.2019.05.009
42. Basurto-Lozada P, Molina-Aguilar C, Castaneda-Garcia C, Vázquez-Cruz M, Garcia-Salinas O, Álvarez-Cano A, et al. Acral lentiginous melanoma: basic facts, biological characteristics and research perspectives of an understudied disease. *Pigment Cell Melanoma Res*. (2021) 34:59–71. doi: 10.1111/pcmr.12885
43. Stubblefield J, Kelly B. Melanoma in non-caucasian populations. *Surg Clin North Am*. (2014) 94:1115–1126.ix. doi: 10.1016/j.suc.2014.07.008



OPEN ACCESS

EDITED BY
Giusto Trevisan,
University of Trieste, Italy

REVIEWED BY
Diana Crisan,
University of Ulm, Germany
Mara Giavina-Bianchi,
Albert Einstein Israelite Hospital, Brazil

*CORRESPONDENCE
Anusuya Kawsar
✉ Anusuya.kawsar1@nhs.net

†These authors share first authorship

RECEIVED 16 July 2023
ACCEPTED 27 October 2023
PUBLISHED 16 November 2023

CITATION
Kawsar A, Hussain K, Kalsi D, Kemos P,
Marsden H and Thomas L (2023) Patient
perspectives of artificial intelligence as a
medical device in a skin cancer pathway.
Front. Med. 10:1259595.
doi: 10.3389/fmed.2023.1259595

COPYRIGHT
© 2023 Kawsar, Hussain, Kalsi, Kemos, Marsden
and Thomas. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in this
journal is cited, in accordance with accepted
academic practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Patient perspectives of artificial intelligence as a medical device in a skin cancer pathway

Anusuya Kawsar^{1*†}, Khawar Hussain^{1†}, Dilraj Kalsi²,
Polychronis Kemos^{2,3}, Helen Marsden² and Lucy Thomas¹

¹Chelsea and Westminster Hospital NHS Foundation Trust, London, United Kingdom, ²Skin Analytics Ltd., London, United Kingdom, ³Blizard Institute, Queen Mary University of London, London, United Kingdom

The use of artificial intelligence as a medical device (AlaMD) in healthcare systems is increasing rapidly. In dermatology, this has been accelerated in response to increasing skin cancer referral rates, workforce shortages and backlog generated by the COVID-19 pandemic. Evidence regarding patient perspectives of AlaMD is currently lacking in the literature. Patient acceptability is fundamental if this novel technology is to be effectively integrated into care pathways and patients must be confident that it is implemented safely, legally, and ethically. A prospective, single-center, single-arm, masked, non-inferiority, adaptive, group sequential design trial, recruited patients referred to a teledermatology cancer pathway. AlaMD assessment of dermoscopic images were compared with clinical or histological diagnosis, to assess performance (NCT04123678). Participants completed an online questionnaire to evaluate their views regarding use of AlaMD in the skin cancer pathway. Two hundred and sixty eight responses were received between February 2020 and August 2021. The majority of respondents were female (57.5%), ranged in age between 18 and 93 years old, Fitzpatrick type I-II skin (81.3%) and all 6 skin types were represented. Overall, there was a positive sentiment regarding potential use of AlaMD in skin cancer pathways. The majority of respondents felt confident in computers being used to help doctors diagnose and formulate management plans (median = 70; interquartile range (IQR) = 50–95) and as a support tool for general practitioners when assessing skin lesions (median = 85; IQR = 65–100). Respondents were comfortable having their photographs taken with a mobile phone device (median = 95; IQR = 70–100), which is similar to other studies assessing patient acceptability of teledermatology services. To the best of our knowledge, this is the first comprehensive study evaluating patient perspectives of AlaMD in skin cancer pathways in the UK. Patient involvement is essential for the development and implementation of new technologies. Continued end-user feedback will allow refinement of services to ensure patient acceptability. This study demonstrates patient acceptability of the use of AlaMD in both primary and secondary care settings.

KEYWORDS

artificial intelligence, skin cancer, dermatology, patient perspectives, medical device, AI as a medical device, deep ensemble for the recognition of malignancy (DERM), Skin Analytics

Introduction

The use of artificial intelligence (AI) is currently being explored across the field of medicine and the deployment of AI as a medical device (AIaMD) in healthcare systems is rapidly expanding. In dermatology, this has been accelerated in response to increasing skin cancer referral rates, workforce shortages and backlog generated by the COVID-19 pandemic (1). AI is making significant contributions in dermatology especially with automated skin lesion analysis, triage of cutaneous lesions, skin cancer detection, and dermatological image recognition, offering benefits to enhance various aspects of patient care (2). However while AI holds great potential, it is not without its challenges with potential concerns regarding patient data privacy and confidentiality, while ensuring these technologies are validated, reliable and accurate (3).

The deep ensemble for recognition of malignancy (DERM) device, designed by Skin Analytics, is an AIaMD that analyzes images of skin lesions to support the identification and appropriate management of skin cancers, premalignant lesions, and benign conditions. DERM was the first UKCA Class IIa certified AIaMD dermatology device on the UK market (4).

Evidence regarding patient perspectives of AIaMD being used by doctors to help make decisions about their care, is currently lacking in the literature. Patient acceptability is fundamental if this novel technology is to be effectively integrated into care pathways and patients must be confident that it is implemented safely, legally, and ethically.

The aim of this study was to explore patients' perspectives on the use of AI as part of their skin cancer management pathway.

Methods

A prospective, single-center, single-arm, masked, non-inferiority, adaptive, group sequential design trial, designed to demonstrate the potential of DERM to reduce unnecessary referrals, was conducted at Chelsea and Westminster Hospital in London, UK (ClinicalTrials.gov: NCT04123678). Patients over the age of 18, who were referred to a teledermatology skin cancer clinic with at least one skin lesion that could be photographed, were eligible for the study. Patients provided written informed consent for the study, and there was no financial compensation. Ethical approval for the study was granted by the West Midlands, Edgbaston Research Ethics Committee.

Patients attended an appointment with a clinical photographer based within the hospital. In addition to images of the lesions captured for standard of care assessment, macroscopic and dermoscopic images of each skin lesion were taken by a healthcare assistant using an iPhone X smartphone with DermLite DL1 basic dermoscopic lens attachment. Captured images were uploaded for analysis by DERM, which was certified as a Class I AIaMD at the time. The DERM analysis result was not shared with the patient or the dermatologist, and the patient's care continued in accordance with routine standard of care. Information on lesion history, risk factors for skin cancer, number of appointments needed to diagnose, and the final diagnosis were collected (5).

After their assessment participants were sent a link by email to an online questionnaire (Supplementary material) which was designed to evaluate their views regarding potential use of AIaMD in the skin cancer pathway. The questionnaire was hosted on an electronic Case Report Form that could be linked with the study record. Reminders were sent to patients who had not completed the survey after at least 1 week. The questionnaire included 4 questions on healthcare appointments prior to their teledermatology

appointment, and 14 questions that evaluated patient acceptance of: (i) clinic and photography appointments (ii) AI as a service tool, which were worded both positively and negatively to minimize bias. A visual analog scale (VAS) was used to assess respondents' satisfaction. The VAS ranged from 0 to 100 with a score of >50 taken to indicate an agreement with a given statement. The impact of patient factors (age, sex, and Fitzpatrick skin type) and management outcome on the patient's response were evaluated using a Kruskal-Wallis (KW) test, with statistical significance set at $p < 0.05$. Statistical analysis was conducted using the R language version 4.1.3 and environment for statistical computing.

Results

Seven hundred patients were recruited between February 2020 and August 2021, including 12 patients who consented twice. Two hundred and sixty eight questionnaire responses were received (38.2% response rate), including two patients who completed the questionnaire twice. Respondents ranged in age between 18 and 93 years old. Most respondents were female ($n = 154$, 57.5%) and had Fitzpatrick type I-II skin ($n = 218$, 81.3%); however all 6 skin types were represented (Table 1).

Most patients ($n = 207$, 77.5%) attended the teledermatology clinic within 14 days of their GP appointment and reported that this time was "about right" for them (Table 2). Most patients ($n = 191$, 71.3%) reported never, or only once or twice, visiting a doctor about the same skin lesions in the past 5 years, with the median number of prior

TABLE 1 Baseline characteristics of study population.

	Respondents (N, %) Total <i>n</i> = 268
Gender	
Male	114 (43%)
Female	154 (57%)
Age groups	
18–29 years	35 (13%)
30–39 years	33 (12%)
40–49 years	28 (10%)
50–59 years	36 (13%)
60–69 years	50 (19%)
70–79 years	62 (23%)
80+	24 (9%)
Fitzpatrick skin type	
I	72 (27%)
II	146 (54%)
III	44 (16%)
IV	3 (1%)
V	1 (0.3%)
VI	2 (0.7%)
Management outcome	
Discharge	83 (31%)
Routine appointment	66 (25%)
Biopsy/urgent follow up	118 (44%)

healthcare appointments, to assess the skin lesion/s included in the study, being 1 (IQR = 1–1, max 6).

Overall, there was a positive sentiment regarding potential use of AIaMD in skin cancer pathways (Table 3). The majority of respondents felt confident in ‘computers’ being used to help doctors diagnose and formulate management plans (median = 70; interquartile range (IQR) = 50–95) and as a support tool for general practitioners when assessing skin lesions (median = 85; IQR = 65–100). The majority

would rather have had their skin assessed by a computer than wait weeks to see an in-person dermatologist (median = 70; IQR = 50–97.5).

Responses for most questions (9 out of 14) were comparable across the sub-groups assessed, with no significant variation in the median scores. Differences in responses were most frequently associated with the outcome of the teledermatology assessment, reaching statistical significance for four questions. Women were found to be less comfortable having photographs of their lesions taken, compared to men, while no statistically significant differences in responses were associated with respondent’s age (Table 4).

TABLE 2 Appointments made by respondents prior to attending teledermatology clinic.

Question	Option	Respondents (N, %)
Number of visits to GP about lesions on my skin, over the last 5 years	Never	73 (27%)
	Once or twice	118 (44%)
	A couple of times	38 (14%)
	Several times	31 (12%)
	Quite a lot	6 (2%)
Number of days since my GP appointment	2 days	5 (2%)
	5 days	24 (9%)
	7 days	42 (16%)
	14 days	136 (51%)
	28 days	27 (10%)
	More than 28 days	24 (9%)
The time between seeing the GP and attending the teledermatology clinic was...	Too short	3 (1%)
	About right	224 (84%)
	Too long	20 (7%)
	Far too long	10 (4%)
Number of previous healthcare appointments to assess these lesions	Mean	1.36
	Median	1
	IQR	1–1
	Max	6

Discussion

AI has demonstrated potential to enhance skin cancer detection and improve efficiency in urgent cancer pathways (5), through the development of several machine learning algorithms to distinguish malignant from benign skin lesions (6). While ongoing technologies are being developed, it is paramount that patient perspectives of AI are explored in parallel, to ensure acceptability of this new technology, and to help inform successful large scale deployment into clinical pathways. Structured feedback from patients who are involved in clinical research and early deployments of AIaMD is one way in which this data can be collected.

To the best of our knowledge, this is the first comprehensive study evaluating patient perspectives of AIaMD in skin cancer pathways in the UK. Our cohort involved a large group of patients that reflect the local population who are referred on a cancer pathway, with all six Fitzpatrick skin types being represented.

Overall our study revealed a positive sentiment regarding potential use of AIaMD in skin cancer pathways. This complements a qualitative study conducted in Germany reporting 75% would recommend AI tools for skin cancer screening to family and friends, with 94% of patients expressing acceptance of the symbiosis between clinicians and AI systems (7).

The majority of our respondents felt confident in computers being used to help doctors diagnose and formulate management plans and

TABLE 3 Summary table of results from patient satisfaction of AIaMD in skin cancer pathways.

	Median	IQR
I feel confident in ‘computers’ being used to help doctors diagnose and formulate management plans	70	50–95
I think having computers assess my photographs to help guide my GP is a good way of dealing with my problem	85	65–100
I would rather have my skin assessed by a computer than wait weeks to see an in-person dermatologist	70	50–97.5
I felt comfortable having my photographs taken with a mobile phone device	95	70–100
The prospect of having my lesions assessed by a computer made me feel uncomfortable	10	0–46
I have confidence that a computer can help me and my doctor by analyzing photographs of lesions	85	60–100
I feel more confident in my diagnosis when it is made by a dermatologist compared to a computer	50	50–75
The photography service is an efficient use of my time	85	62.5–100
I found it embarrassing having my photographs taken	0	0–5
I would have preferred to see a dermatologist face to face rather than have a computer assess my lesion	50	25–80
Having a computer assess photographs of my lesion saves time in comparison to a face-to-face consultation	75	50–95
I would have preferred to have my photographs taken in my GP practice rather than in hospital	50	10–71.25
I felt the time needed to take photographs was too long	0	0–10
I would recommend the teledermatology service to friends and family.	80	50–100

TABLE 4 Survey questions with statistically significant variation in median scores across sub-groups of respondents.

Question	Subgroup	Median	KW <i>p</i> -value
I feel confident in computers being used to help my doctor determine my diagnosis and management plan	Discharge	61.96	0.005
	Routine appointment	66.50	
	Biopsy/urgent follow up	74.61	
I think having computers assess my photographs to help guide my GP is a good way of dealing with my problem	Fitzpatrick I	79.89	0.021
	Fitzpatrick II	81.39	
	Fitzpatrick III	74.77	
	Fitzpatrick IV–VI	44.60	
I felt comfortable having my photographs taken with a mobile phone device	Discharge	75.47	0.017
	Routine appointment	86.73	
	Biopsy/urgent follow up	84.59	
	Male	85.00	0.018
	Female	80.62	
	Fitzpatrick I	84.23	0.04
	Fitzpatrick II	83.01	
	Fitzpatrick III	83.04	
	Fitzpatrick IV–VI	53.00	
I have confidence that a computer can help me and my doctor by analyzing photographs of my lesions	Discharge	71.20	0.023
	Routine appointment	79.46	
	Biopsy/urgent follow up	81.47	
I found it embarrassing having my photographs taken	Discharge	12.04	0.001
	Routine appointment	6.33	
	Biopsy/urgent follow up	7.4	
	Male	4.58	0.009
	Female	11.35	

as a support tool for general practitioners when assessing skin lesions. Importantly, our survey highlighted acceptability of AIaMD alongside clinicians as a decision-making support tool, however further assessment for stand-alone autonomous applications is required. Respondents were comfortable having their photographs taken, which is similar to other studies assessing patient acceptability of teledermatology services (8), though the differences in responses between sexes may be relevant for the wider deployment of teledermatology. Differences in responses across the Fitzpatrick skin types may be influenced by the comparatively small number of responders with Fitzpatrick skin types IV–VI, and the significance of these results should be interpreted with caution.

The differences in responses associated with the outcome of the teledermatology review is interesting as those patients who were referred for a biopsy or urgent referral were consistently more willing to accept the AIaMD as part of their assessment than those who were discharged or referred for a routine appointment. This suggests patients are more amenable to new technologies being used to inform their care when they feel their condition is being more actively managed.

A common limitation of patient surveys is low participation rates, and the resultant self-selection bias with feedback missing from those patients who are unwilling or unable to participate, or those who simply forget to complete the questionnaire. The response rate for this survey was almost 40%, which is similar to response rates to online

surveys elsewhere (9, 10). However, it remains possible that the results presented here are not wholly representative of the views of all patients recruited into the clinical study, and indeed the wider population of patients attending teledermatology clinics.

Further work is required to evaluate the psychological status of patients whose care involves an assessment by an AIaMD, compared to those who just attend face to face consultations. Patient feedback will continue to be important as products like DERM develop, and the clinical patient pathways in which they are deployed evolve. Further, larger studies are needed to capture patient feedback from more diverse populations, including different socio-economic groups and a wider variety of ethnicities and skin colors also focusing on acceptability of autonomous AIaMD in clinical pathways.

Conclusion

To the best of our knowledge, this is the first comprehensive study evaluating patient perspectives of AIaMD in skin cancer pathways in the UK. Patient involvement is essential for the successful development and implementation of new technologies. Continued end-user feedback will allow refinement of services to ensure patient acceptability. This study demonstrates patient acceptability of AIaMD in both primary and secondary care settings.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by the West Midlands, Edgbaston Research Ethics Committee. The studies were conducted in accordance with the local legislation and institutional requirements. The patients provided their written informed consent to participate in this study.

Author contributions

AK: Writing – original draft, Writing – review & editing. KH: Writing – review & editing. DK: Data curation, Investigation, Methodology, Resources, Validation, Writing – review & editing. PK: Formal analysis, Writing – review & editing. HM: Formal analysis, Writing – review & editing, Conceptualization, Data curation, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization. LT: Conceptualization, Methodology, Project administration, Supervision, Visualization, Writing – review & editing.

Funding

The study was funded by Skin Analytics Ltd., as part of a grant from Innovate UK.

Acknowledgments

The authors would like to thank all patients who consented to the study and completed the patient survey; Chelsea and Westminster Hospital staff involved in patient recruitment, data collection, and diagnosis confirmation; and Skin Analytics staff involved in the development of the AIaMD and the operations of the study.

References

1. Mahmood F, Bendayan S, Ghazawi FM, Litvinov IV. Editorial: the emerging role of artificial intelligence in dermatology. *Front Med (Lausanne)*. (2021) 8:751649. doi: 10.3389/fmed.2021.751649
2. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med*. (2019) 25:30–6. doi: 10.1038/s41591-018-0307-0
3. Liopyris K, Gregoriou S, Dias J, Stratigos AJ. Artificial intelligence in dermatology: challenges and perspectives. *Dermatol Ther (Heidelb)*. (2022) 12:2637–51. doi: 10.1007/s13555-022-00833-8
4. 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) 542:115–8. doi: 10.1038/nature21056
5. Thomas L, Hyde C, Mullarkey D, Greenhalgh J, Kalsi D, Ko JM, et al. Real-world post-deployment performance of a novel machine learning-based digital health technology for skin lesion assessment and suggestions for post-market surveillance. *Front Med*. (2023) 10:1264846.
6. Marsden H, Kemos P, Venzi M, Noy M, Maheswaran S, Francis N, et al. Accuracy of an artificial intelligence as a medical device as part of a UK-based skin cancer teledermatology service. *Front Med*. (2023).
7. Phillips M, Marsden H, Jaffe W, Matin RN, Wali GN, Greenhalgh J, et al. Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. *JAMA Netw Open*. (2019) 2:e1913436. doi: 10.1001/jamanetworkopen.2019.13436
8. Nelson CA, Pérez-Chada LM, Creadore A, Li SJ, Lo K, Manjaly P, et al. Patient perspectives on the use of artificial intelligence for skin Cancer screening: a qualitative study. *JAMA Dermatol*. (2020) 156:501–12. doi: 10.1001/jamadermatol.2019.5014
9. Nicholson P, Macedo C, Fuller C, Thomas L. Patient satisfaction with a new skin cancer teledermatology service. *Clin Exp Dermatol*. (2020) 45:691–8. doi: 10.1111/ced.14191
10. Meyer VM, Benjamins S, Moumni ME, Lange JFM, Pol RA. Global overview of response rates in patient and health care professional surveys in surgery: a systematic review. *Ann Surg*. (2022) 275:e75–81. doi: 10.1097/SLA.00000000000004078

Conflict of interest

LT is a clinical advisor to Skin Analytics Ltd., has received Skin Analytics shares or share options; has received research funding support from Skin Analytics (salaries and equipment) and AIaMD deployment programme; has received reimbursement of conference fees, travel and accommodation costs from Skin Analytics to present research results; LT has received financial remuneration for separate programme of work as a consultant by Skin Analytics; has received grant funding from NHSX and CW+; has received paid honoraria to lecture for Almirall; was supported to attend a conference by Abbvie and Janssen; and holds multiple unpaid leadership roles. HM is an employee of Skin Analytics Ltd., and has received Skin Analytics shares or share options. DK is an employee of Skin Analytics Ltd., and has received Skin Analytics shares or share options. PK was previously a contractor with Skin Analytics Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Skin Analytics, London, UK sponsored and funded this study, as part of an Innovate UK BioMedical Catalyst project, and was involved with the study design, data collection, statistical analysis and interpretation of the data.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2023.1259595/full#supplementary-material>



OPEN ACCESS

EDITED BY
Zubair Shah,
Hamad Bin Khalifa University, Qatar

REVIEWED BY
Ionela Manole,
Colentina Clinical Hospital, Romania
Giusto Trevisan,
University of Trieste, Italy

*CORRESPONDENCE
Brunna C. R. S. Furriel
✉ brunna.silva@ifg.edu.br

RECEIVED 02 October 2023
ACCEPTED 12 December 2023
PUBLISHED 08 January 2024

CITATION

Furriel BCRS, Oliveira BD, Prôa R, Paiva JQ,
Loureiro RM, Calixto WP, Reis MRC and
Giavina-Bianchi M (2024) Artificial intelligence
for skin cancer detection and classification for
clinical environment: a systematic review.
Front. Med. 10:1305954.
doi: 10.3389/fmed.2023.1305954

COPYRIGHT

© 2024 Furriel, Oliveira, Prôa, Paiva, Loureiro,
Calixto, Reis and Giavina-Bianchi. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted which
does not comply with these terms.

Artificial intelligence for skin cancer detection and classification for clinical environment: a systematic review

Brunna C. R. S. Furriel ^{1,2,3*}, Bruno D. Oliveira¹, Renata Prôa¹,
Joselisa Q. Paiva¹, Rafael M. Loureiro¹, Wesley P. Calixto^{2,3},
Márcio R. C. Reis^{1,3} and Mara Giavina-Bianchi¹

¹Imaging Research Center, Hospital Israelita Albert Einstein, São Paulo, Brazil, ²Electrical, Mechanical and Computer Engineering School, Federal University of Goiás, Goiânia, Brazil, ³Studies and Researches in Science and Technology Group (GCITE), Federal Institute of Goiás, Goiânia, Brazil

Background: Skin cancer is one of the most common forms worldwide, with a significant increase in incidence over the last few decades. Early and accurate detection of this type of cancer can result in better prognoses and less invasive treatments for patients. With advances in Artificial Intelligence (AI), tools have emerged that can facilitate diagnosis and classify dermatological images, complementing traditional clinical assessments and being applicable where there is a shortage of specialists. Its adoption requires analysis of efficacy, safety, and ethical considerations, as well as considering the genetic and ethnic diversity of patients.

Objective: The systematic review aims to examine research on the detection, classification, and assessment of skin cancer images in clinical settings.

Methods: We conducted a systematic literature search on PubMed, Scopus, Embase, and Web of Science, encompassing studies published until April 4th, 2023. Study selection, data extraction, and critical appraisal were carried out by two independent reviewers. Results were subsequently presented through a narrative synthesis.

Results: Through the search, 760 studies were identified in four databases, from which only 18 studies were selected, focusing on developing, implementing, and validating systems to detect, diagnose, and classify skin cancer in clinical settings. This review covers descriptive analysis, data scenarios, data processing and techniques, study results and perspectives, and physician diversity, accessibility, and participation.

Conclusion: The application of artificial intelligence in dermatology has the potential to revolutionize early detection of skin cancer. However, it is imperative to validate and collaborate with healthcare professionals to ensure its clinical effectiveness and safety.

KEYWORDS

Skin cancer, artificial intelligence, melanoma, detection, classification, feature extraction

1 Introduction

The role of technology and artificial intelligence has gained increasing prominence in the field of dermatology. Techniques such as convolutional neural networks and image processing have been extensively examined for their capacity to identify specific features in skin lesion images, with the potential to aid in the recognition of suspicious lesions and the diagnosis of conditions like melanoma.

Skin cancer is the most common form of cancer worldwide (1). Over the past decade, there has been a concerning 27% increase in the annual diagnosis of invasive melanoma cases (2). Alarming, more than 5,400 people die from non-melanoma skin cancer every month (3). In the United States alone, the annual financial burden of treating skin cancer is estimated at a staggering US\$8.1 billion, with approximately US\$4.8 billion allocated to non-melanoma skin cancer and US\$3.3 billion to melanoma (4). Among skin cancer types, basal cell carcinoma ranks as the most common, followed by squamous cell carcinoma and melanoma, which stands out as the most aggressive and lethal type of skin cancer (5, 6). Merkel cell carcinoma also stands out among aggressive tumors (7). These tumors can arise anywhere on the body but are frequently observed in regions more exposed to the sun, including the face, neck, arms, and hands. Thus, there is an imperative need for sustained efforts to promote awareness and prevention of skin cancer (8–10).

Conventional techniques for detecting these diseases include patient data analysis, as well as visual and histopathological analysis of the lesions (11). Visual assessment relies on the clinical inspection of the lesion, taking into consideration factors such as its appearance, size, shape, location, and evolution. On the other hand, histopathological analysis entails the collection of a sample of the lesion for laboratory examination, typically through techniques such as biopsy. Additionally, devices like the dermatoscope are used to facilitate the examination of the lesion and the identification of features such as pigmentation, vascularity, and regression (12). Another example is the use of confocal microscopy, a technique that allows the analysis of skin layers without the need for sample collection (13, 14).

These techniques have proven effective in the detection and diagnosis of skin diseases. However, they may present limitations, including subjectivity in visual analysis and the need for invasive sample collection procedures. Confocal microscopy incurs high financial costs and is relatively inaccessible to medical professionals, even among specialists.

It is also important to highlight that diagnosing these diseases poses a significant challenge to the healthcare system, especially in regions lacking specialized professionals or adequate equipment for skin lesion identification (15, 16). An alternative approach involves initial screening by general practitioners, who may not always possess the necessary training for early skin cancer detection (17).

The implementation of Computer-Aided Diagnosis (CAD) solutions powered by Artificial Intelligence (AI) holds the potential to address some of these limitations and offer a promising alternative for accurate and non-invasive skin disease diagnosis. Existing literature suggests that AI systems can classify skin cancers competently on par with dermatologists. Notably, the diagnostic capabilities of the dermatologist vary based on experience, i.e., it is not a uniform basis of reference. Moreover, studies highlight the feasibility of leveraging mobile devices equipped with neural networks to broaden the access of dermatological expertise, offering low-cost access to vital diagnostic care (18, 19).

While numerous solutions are being developed for skin cancer detection and classification, those are usually not evaluated and validated in real clinical settings, which limits their practical applicability. The review study conducted by Goyal et al. (20) provides an updated assessment of the performance of artificial

intelligence algorithms in skin cancer classification and diagnosis. It also delves into the challenges faced by these systems and future opportunities to enhance of dermatologists' diagnostic abilities through AI support.

However, for these technologies to become effective and applicable in clinical settings, several challenges must be addressed. These challenges include the need for standardization in image acquisition and processing techniques, the requirement for extensive training datasets, and the creation of robust and representative databases (20–24). Prior studies in skin cancer classification have demonstrated restricted generalizability due to insufficient data and an emphasis on standardized tasks (19). Furthermore, it is essential to evaluate the effectiveness and safety of these tools in diverse contexts, taking into account variables such as the ethnic and genetic diversity of the population and the specific type of skin cancer under consideration, among other factors. In this regard, it is imperative for research in this field to adhere rigorously to scientific and ethical standards. Finally, it is crucial to emphasize that automated skin disease detection should not replace clinical evaluation by medical professionals but rather complement it.

The aim of this systematic review is to investigate studies focused on the detection, classification, and evaluation of skin cancer images in a clinical setting. The main approaches and challenges encountered while implementing these techniques must be identified to do this. The importance of this systematic review lies in its ability to aggregate and thoroughly examine all pertinent research in this field, thus offering a comprehensive view of the subject. In turn, researchers can assess the quality and credibility of existing studies, identify knowledge gaps, and propose innovative research directions. Furthermore, this systematic review can provide valuable information for doctors and healthcare professionals looking to harness the potential of AI in aiding the diagnosis and treatment of skin diseases.

2 Methods

This section outlines the methodology employed for the systematic literature review, encompassing the following stages: (i) research identification, (ii) selection, (iii) eligibility, (iv) data extraction, and (v) synthesis.

2.1 Step 1: study identification

First, we established the objectives and questions that frame this literature review. The primary goal of this systematic review is to highlight research involving the implementation of AI in clinical settings. Our aim is to gain insights into the methodologies employed in previous research and the outcomes achieved when using AI in this context.

For this review, we registered a protocol with the International Prospective Register of Systematic Reviews (PROSPERO) under ID CRD42023411211 on April 4, 2023, and PRISMA guidelines were followed. PROSPERO is a global registry for systematic review protocols, where researchers publish their research methods in

advance. This process promotes transparency, prevents publication bias, and improves the reproducibility of studies.

The search databases used for the literature review include PubMed, Scopus, Embase, and Web of Science, and topics are analyzed using the following search terms: (“skin cancer” OR “skin lesion” OR “dermatology” OR “dermatoscopy” OR “melanoma”) AND (“artificial intelligence” OR “neural network*” OR “deep learning” OR “convolutional neural network*” OR “transfer learning” OR “machine learning” OR “Computer aided diagnostic*” OR “CAD” OR “image classification” OR “image processing” OR “Internet of things” OR “Data mining” OR “Iot”) AND (“real-time” or “real time” OR “real-world” OR “real world” OR “smartphone”) AND NOT (“Meta-Analysis” OR “Meta Analysis” OR “Systematic Review”).

2.2 Step 2: study selection

Secondly, we defined the search terms and established inclusion/exclusion criteria. In this literature review, we used the terms highlighted in the previous section, with the sole restriction being the inclusion of journal articles and conference proceedings only.

Our initial search yielded 760 results, of which 457 were identified as duplicates and therefore removed. This resulted in a pool of 303 distinct studies, which were subsequently evaluated for eligibility.

2.3 Step 3: study relevance and quality assessment

In the third step, we assessed the relevance and quality of the selected studies. Two authors (BCRSF and MRCR) were responsible for reading each title and abstract in order to assess the relevance and quality of each previously selected study. The criteria used to determine eligibility is as follows:

- The document’s abstract presents clear objectives, methodology, and results.
- The study addresses computer-aided diagnostic solutions for skin cancer with a focus on real clinical applications.
- The study reports the accuracy, sensitivity, specificity, and/or overall accuracy of artificial intelligence systems for skin cancer.
- The study describes the development and/or validation process of the systems.
- The study provides a critical analysis of the results obtained by artificial intelligence systems and discusses their limitations and potential biases.

Based on the inclusion criteria stated above, a total of 282 studies were eliminated from consideration. Following a comprehensive review of the entire texts, three more studies were removed from consideration due to their limited content, which included only abstracts or incomplete texts. Ultimately, 18 studies

have been retained. [Figure 1](#) presents the study identification flowchart.

For the study, Mendeley and Rayyan tools were used.

2.4 Step 4: data extraction

To facilitate data extraction in our literature review, we utilized a spreadsheet to document the metadata of each selected study. The following metadata was analyzed:

1. Publication year and study objective.
2. Regarding the data used: Types of data, source, and quantity.
3. Resources used to assist in the detection and/or classification of skin lesions.
4. Technique for the detection and/or classification of skin lesions.
5. Study function.
6. Key findings and study perspectives.
7. Information regarding ethnic and genetic diversity of the population.
8. Information regarding system accessibility and availability.
9. Relationship and/or involvement of dermatologists and other medical professionals.

2.5 Step 5: data synthesis

The concluding phase of our study encompasses data synthesis, which was subdivided into two key steps. Initially, we conducted a systematic analysis of the raw data obtained through the literature review process. Subsequently, we compiled metadata pertaining to the articles chosen in our literature review.

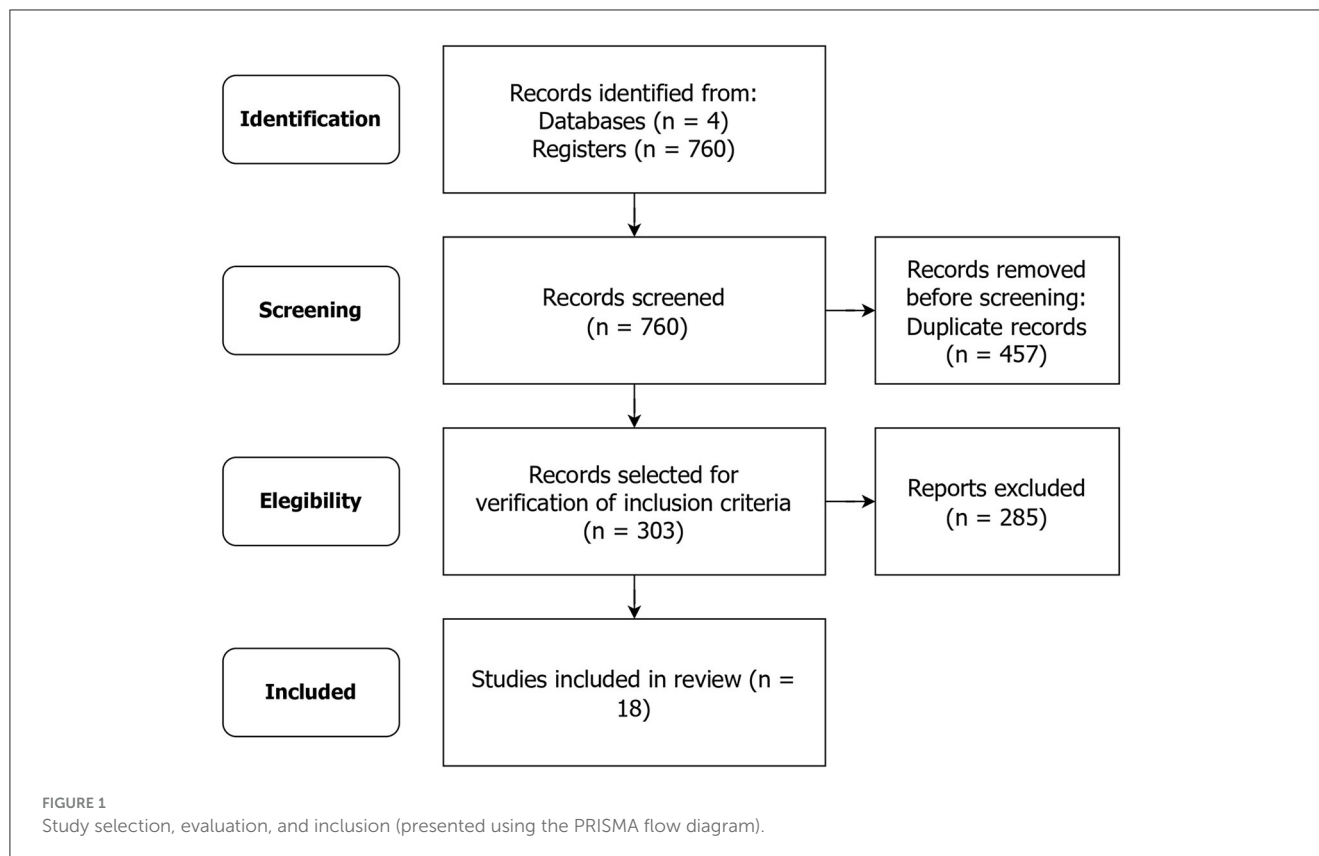
3 Results and discussion

This section outlines the results obtained through the search strategies describes in the methodology.

3.1 Descriptive analysis

The first section of our analysis pertains to descriptive information. As part of this analysis, we examined the objectives of the selected studies.

The primary objective of all the mentioned studies is to develop, implement, and/or validate systems for the detection, diagnosis, and classification of skin cancers, particularly melanoma, using mobile devices or computers. These systems aim to improve the early detection of skin lesions and enhance diagnostic accuracy, assisting healthcare professionals and providing more accessible and efficient screening for patients. Furthermore, they explore the use of advanced techniques such as image processing, pattern recognition, and deep learning to automate the analysis process and deliver real-time results.



3.2 Data scenario

For our study, it is of utmost importance to analyze the quality and representativeness of the data, given that these factors play a critical role in developing reliable algorithms and models for skin lesion diagnosis. Diversity in data sources is key to ensure broader model generalization since different sources can provide specific and varied information about the lesions. Furthermore, the availability of large datasets containing hundreds of thousands of images can be extremely advantageous in creating more robust and accurate machine learning models. Table 1 presents important information about the data from each article. The study information is listed in chronological order based on the publication date.

The analysis of the studies reveals a remarkable diversity of approaches in the diagnosis of skin tumors, with a significant emphasis on the detection of melanoma and other dermatological conditions. Among these research studies, there is a notable convergence in the preference for the use of clinical (macroscopic) images and/or dermoscopic images for analysis. The choice of these images demonstrates a consensus in the scientific community regarding the importance of this data in developing more effective and accessible diagnostic methods.

In the context of the types of data employed, Roy et al. (28) and Alizadeh and Mahloojifar (29) used dermoscopic images from established databases like PH2 and ISIC, adding to the reliability of the results. Meanwhile, Dulmage et al. (35) relies on clinical images collected by healthcare professionals, reflecting real-world conditions.

The discrepancy in the size of datasets is evident, with some studies using relatively small datasets, such as Ramlakhan and Shang (25) and Afifi et al. (26), which have 83 and 356 images, respectively. This limitation in size may restrict the models' capacity for generalization and accuracy. On the other hand, Udrea et al. (31) and Pangti et al. (36) present massive datasets containing 131,873 and 17,408 images, respectively. This provides a more solid foundation for model generalization and learning. Furthermore, Thissen et al. (27) works with a dataset of 341 images, which is still considerably limited compared to the larger datasets. This difference in dataset size directly impacts the models' ability to generalize, emphasizing the importance of carefully assessing effectiveness at different scales.

An additional disparity is observed when considering the specific focus versus the breadth of conditions addressed in studies of skin lesion diagnosis. While some studies have a narrow focus on melanoma and non-melanoma lesions (29), others adopt a broader approach, covering various categories of skin diseases (36). This distinction highlights the decision between targeting a specific condition or taking a more comprehensive approach, which directly influences the clinical applicability of the developed models.

However, there are less ideal scenarios to consider. Afifi et al. (26) and Ramlakhan and Shang (25) use clinical images without specifying their origin, which can negatively impact data quality and representativeness. Additionally, Thissen et al. (27) relies on images obtained from a commercial application, potentially resulting in limitations regarding image quality and diversity. The absence of specification of image origin in Francese

TABLE 1 Overview of studies on skin cancer image analysis: data type, origin, and quantity by year and author.

References	Data type	Data origin	Total amount of initial data
Ramlakhan and Shang (25)	Clinical (macroscopic) images acquired by mobile device.	Randomly collected images from the internet.	Dataset of 37 images of benign skin lesions and 46 images of malignant lesions.
Afifi et al. (26)	Clinical (macroscopic) images.	Not specified how they are acquired.	Dataset of 356 images, including 168 melanoma images.
Thissen et al. (27)	Clinical (macroscopic) images acquired by mobile device.	SkinVision application.	Dataset of 341 images of melanocytic and non-melanocytic lesions, with 239 undergoing histopathological examination, while the other 102 lesions were clinically diagnosed as benign and not removed.
Roy et al. (28)	Dermoscopic images.	PH2 database.	Dataset of 200 dermoscopic images from the PH2 database, including 80 common nevi, 80 atypical nevi, and 40 melanomas.
Alizadeh and Mahloojifar (29)	Dermoscopic images.	ISIC Database.	Dataset of 150 dermatoscopy images from the ISIC website, consisting of 75 images for non-melanoma lesions and 75 images for melanoma lesions.
Fujisawa et al. (30)	Clinical (macroscopic) images from digital cameras.	Patient data from the University of Tsukuba Hospital from 2003 to 2016.	Dataset of 6,009 images from 2,296 patients, including 14 diagnoses, both malignant and benign conditions.
Udrea et al. (31)	Clinical (macroscopic) images acquired by mobile device.	Data obtained from the University Hospital of Munich and a hospital in Eindhoven, funded by SkinVision BV.	Dataset of 131,873 images acquired from 31,449 users of the app. It included 285 histopathologically validated skin cancer cases, including 138 malignant melanomas.
Bakheet and Al-Hamadi (32)	Dermoscopic images.	PH2 public dataset.	Dataset of 200 images, including 40 malignant and 160 benign lesions.
Abbas (33)	Dermoscopic images.	Various public and private sources, including EDRA-CDROM, ISIC, DermNet, and PH2.	Total of 2,200 dermatoscopy images, including 1,100 malignant melanomas (MM) and 1,100 benign tumors.
Bakheet and El-Nagar (34)	Dermoscopic images.	PH2 public dataset.	Dataset of 200 images, including 80 common nevi, 80 atypical nevi, and 40 melanomas.
Dulmage et al. (35)	Clinical (macroscopic) images.	Images collected by primary care professionals.	Dataset of 76,926 images annotated by dermatologists from the VisualDx privately curated image database, focusing on lesion morphology analysis.
Pangti et al. (36)	Clinical (macroscopic) images.	Raw images from public databases (http://www.hellenicdermatlas.com/en and http://www.danderm.dk/atlas), as well as images from dermatologists in India.	Initial total dataset of 17,718 images. Of these, 310 images were discarded during preprocessing due to poor resolution or multiple lesions. Of the remaining 17,408 images, 1,990 images belonged to the non-specific category, and 15,418 images fell within the 40 selected disease categories.
Giavina-Bianchi et al. (17)	Clinical (macroscopic) and dermoscopic images.	Clinical Model: Teledermatology Project. Dermatoscopic Model: ISIC2019 and PH2 datasets.	Clinical Model: Dataset of 14,000 images belonging to seven classes. Dermatoscopic Model: Dataset of 26,342 images.
Francese et al. (37)	Clinical (macroscopic) images acquired by mobile device.	Despite lack of specification, there is an assumption that the images originate from the authors.	Dataset of 8,000 melanoma or non-melanoma images.
Felmingham et al. (24)	Dermoscopic images.	Two Australian tertiary centers: Skin Health Institute and Alfred Hospital in Melbourne, Australia.	The study aims to recruit 220 participants and provide a minimum of three lesions per participant for final analysis.
Sangers et al. (38)	Clinical (macroscopic) images acquired by mobile device.	University hospital in the Netherlands.	Dataset of 785 skin lesion images, including 418 suspected lesions and 367 benign lesions used as controls.
Jahn et al. (39)	Clinical (macroscopic) images acquired by mobile device.	Dermatology Department at University Hospital Basel, Switzerland.	Dataset of 1,204 pigmented skin lesions.
Kränke et al. (40)	Clinical (macroscopic) images acquired by mobile device.	Tertiary reference center in Graz, Austria.	Dataset of 1,171 images.

et al. (37) is also a factor that can influence data quality and validity.

Finally, the study phase also presents divergences, with some studies still ongoing (24), while others already have final results. The preliminary nature of ongoing studies may limit the availability of conclusive results and the validity of analyses.

It is essential to recognize that both the quantity and type of data play crucial roles in the development of accurate and reliable cutaneous diagnostic models. Larger and more diverse datasets, coupled with high-quality images and reliable sources, tend to produce more robust and generalizable results. Therefore, the careful selection of these elements is fundamental to the effectiveness and clinical applicability of the developed models.

3.3 Techniques and processing

Next, we describe the resources employed in image processing, the classification algorithms used, and the devices on which these approaches were implemented. The resources employed in image processing are used to perform manipulation and feature extraction operations, aiming to prepare the images for analysis. Classification algorithms play the role on categorizing skin lesions based on the extracted features, enabling the precise identification of different classes. Furthermore, these algorithms can assist in clinical decision making, guiding healthcare professionals in choosing the best treatment approaches. These approaches are implemented on devices such as computers, servers, or mobile devices, providing efficient execution of algorithms and practical application of diagnostic techniques on skin lesion images.

Extracting this information from the studies presented here is crucial to guide the development of effective applications, allowing the appropriate selection of preprocessing methods, reliable classifiers, and suitable devices for achieving accurate detection and clinical assessment of skin lesions. [Table 2](#) describes the resources used in image processing, the classification algorithms used, and the main purpose of the study.

It is notable that several studies aim to utilize image segmentation, feature extraction, and classification techniques, as observed in Ramlakhan and Shang (25), Afifi et al. (26), Roy et al. (28), Alizadeh and Mahloojifar (29), Bakheet and Al-Hamadi (32) and Abbas (33). These steps are often fundamental for proper processing of skin lesion images and subsequent diagnostic decision-making.

On the other hand, there are differences regarding the choice of classifiers and processing devices. While some studies, such as Afifi et al. (26), employ Support Vector Machines (SVM) as classifiers, others, like Roy et al. (28), opt for more recent approaches like YOLOv2. The research by Roy et al. (28), Bakheet and Al-Hamadi (32), and Giavina-Bianchi et al. (17) presents a variety of approaches, ranging from the use of traditional machine learning algorithms to deep neural networks, such as Convolutional Neural Networks (CNNs). This diversity of techniques allows for a rich comparative analysis, enabling the identification of the most promising approaches for skin tumor detection. Additionally, the detailed description of the resources used and processing devices provides valuable insights for the development of effective applications.

Regarding processing devices, there is a distinction between approaches that perform detection and classification directly on mobile devices, such as Alizadeh and Mahloojifar (29), and approaches that send extracted features to a server for further analysis, as in the case of Giavina-Bianchi et al. (17). This difference highlights the variety of options available for implementing skin lesion detection solutions.

Finally, some studies do not provide complete information about the resources used, such as Dulmage et al. (35), which limits the understanding of the methodologies employed.

3.4 Main results and perspectives

In this section, we present the main outcomes and prospective insights derived from the various studies analyzed. The primary classification results demonstrate the accuracy, sensitivity, and specificity achieved by different approaches, allowing an assessment of how reliable these methods are in detecting malignant and benign lesions. Furthermore, the perspectives highlight the unique contributions of each study, such as the use of deep learning algorithms, real-time detection effectiveness, and the potential for screening in populations with limited access to dermatologists.

In the context of medicine and healthcare, this information assists medical professionals in choosing the most suitable approaches for early detection of malignant skin lesions, contributing to more precise and rapid diagnosis. Additionally, these results and perspectives also have significant implications for the future development of healthcare applications, guiding research and innovations in the field of artificial intelligence applied to dermatology.

[Table 3](#) provides details related to the main results and perspectives.

The analysis of [Table 3](#) highlights the positive aspects of recent advances in the detection, classification, and evaluation of skin cancer applications using machine learning and image processing, achieving high sensitivity and specificity in identifying malignant lesions. Furthermore, mobile applications offer an accessible approach to screening in populations with limited access to dermatologists.

However, more robust clinical validation is needed, considering the testing stage and comparison with traditional diagnosis. Performance variation between devices and the possibility of unnecessary excisions are also issues to be addressed. These advancements represent significant potential, but it is essential to balance opportunities with challenges, prioritizing ongoing research and validations for effective implementations in medical practice.

Among the studies presented, the YOLOv2 model, proposed by Roy et al. (28), stands out by demonstrating high precision and sensitivity in the detection of melanoma in dermoscopic images, processing in real-time efficiently. Additionally, Udrea et al. (31) present a machine learning-based method that achieves significant results in sensitivity and specificity for the detection of melanomas and basal cell carcinomas and squamous cell carcinomas. In turn, Giavina-Bianchi et al. (17) develop dermatoscopy models to assist dermatologists, offering positive perspectives for improving the detection and management of skin lesions. Furthermore, an innovative approach by Francese et al. (37) uses augmented reality and deep learning in a lesion analysis system, with the potential to facilitate dermatological diagnosis.

It is important to note that, although all the approaches highlighted in [Table 3](#) show promising results, many of them are still undergoing testing and clinical validation phases. Therefore, it is crucial to continue rigorous research and in-depth evaluations, as emphasized by various researchers, before considering the widespread and effective implementation of these approaches in medical practice. These innovations have the potential to revolutionize early detection and diagnosis of skin cancer, but

TABLE 2 Summary of techniques and classifiers used in skin cancer image analysis studies for clinical settings by year and author.

References	Resources used	Classifier	Purpose
Ramlakhan and Shang (25)	Image segmentation, feature calculation, and classification.	K-Nearest Neighbors (K-NN).	Classify malignant and benign lesions.
Afifi et al. (26)	Pre-processing, segmentation, feature extraction, and classification.	Support Vector Machine (SVM)	Melanoma detection.
Thissen et al. (27)	Lesion area, mean grayscale value, standard deviation over the lesion, and lesion circularity extracted from fractal map.	The evaluation algorithm is based on fractal and classical image.	Classification of low, medium, or high-risk lesions (where proven benign skin lesions should fall into the low or medium-risk class, and melanoma and non-melanoma skin cancer, along with melanoma <i>in situ</i> , actinic keratosis, and Bowen's disease, should fall into the high-risk class).
Roy et al. (28)	Image segmentation, feature calculation, and classification.	YOLOv2.	Melanoma detection.
Alizadeh and Mahloojifar (29)	Pre-processing, segmentation, lesion detection, and classification algorithms.	Normal Bayes and Support Vector Machine (SVM).	Melanoma detection.
Fujisawa et al. (30)	Pre-processing and feature extraction.	GoogLeNet DCNN deep convolutional neural network (DCNN).	Classify malignant and benign lesions.
Udrea et al. (31)	Pre-processing, segmentation, and feature extraction.	Conditional generative adversarial network to segment skin lesions in images. For classification, Support Vector Machine Classifier with radial basis kernel function was used.	Detection of (pre)malignant and malignant conditions.
Bakheet and Al-Hamadi (32)	Image pre-processing, skin lesion segmentation, feature extraction, and classification.	Multilevel Neural Network (MNN)	Melanoma detection.
Abbas (33)	Image pre-processing, skin lesion segmentation, feature extraction, and classification.	The Smart-Dermo system is proposed in this article using image processing and applies clinical rules using the ABC clinical technique. It also uses Fuzzy technique for classification.	Melanoma detection.
Bakheet and El-Nagar (34)	Image pre-processing, adaptive lesion segmentation, and feature extraction.	Deep Neural Network (DNN).	Classification of malignant vs. benign lesions.
Dulmage et al. (35)	Not specified	Deep convolutional neural network (CNN) architecture, including DenseNet and NASNetMobile, as well as proprietary models developed by VisualDx.	Detection of skin lesion morphology.
Pangti et al. (36)	Pre-processing and image optimization resources, normalization algorithms, and custom loss function for training the neural network.	Convolutional Neural Networks (CNN).	Detection of 40 common skin diseases.
Giavina-Bianchi et al. (17)	Similarity networks and Data Augmentation.	In the clinical model, image features are extracted through a convolutional network (VGG16), and then the K-Nearest Neighbor (KNN) algorithm is used to classify the images based on these features. In the dermatoscopic model, images are processed using generative adversarial networks (GANs), and classification is performed through an ensemble model that combines the results of five EfficientNetB6 models.	Melanoma detection.
Francese et al. (37)	Real-time analysis process of skin lesions involves acquiring camera frames, tracking device position relative to the patient's skin, cropping the nevus, image pre-processing, feature extraction, nevus classification using a CNN, pose estimation, rendering, and displaying augmented images.	Convolutional Neural Network (CNN)	Melanoma detection.
Felmingham et al. (24)	Not specified.	Convolutional Neural Network (CNN) developed by MoleMap Ltd and Monash eResearch.	Classification into benign, uncertain, or malignant lesions.
Sangers et al. (38)	Not specified.	The study used a mobile health app called SkinVision, which utilizes Convolutional Neural Network (CNN).	Classification into suspicious and benign lesions.

(Continued)

TABLE 2 (Continued)

References	Resources used	Classifier	Purpose
Jahn et al. (39)	Not specified.	The study used a mobile health app called SkinVision, which utilizes Convolutional Neural Network (CNN).	Melanoma detection.
Kränke et al. (40)	Not specified.	Two CNNs: one classical CNN and the other region proposal network (RPN)-based CNN for stratification.	Classification of various skin lesions.

ensuring their reliability and clinical utility through robust studies is fundamental.

3.5 Diversity, accessibility, and medical collaboration

Ethnic diversity, the involvement of medical professionals, and ethical considerations play a pivotal and indispensable role in the development of applications designed for the detection and classification of skin lesions. These factors significantly contribute to the efficacy, validity, and accessibility of these technological solutions, thereby ensuring their widespread acceptance and adoption within the medical community, characterized by both confidence and equity. The continuous advancement within this scientific domain necessitates a multidisciplinary approach that seamlessly amalgamates the expertise of dermatologists, data scientists, and healthcare practitioners with the overarching objective of further enhancing the precision and impact of these pioneering applications.

Within this context, the systematic incorporation of a comprehensive array of ethnicities and genotypes into the training and evaluation datasets assumes fundamental importance. This strategic inclusion is essential to ensure the capability of such applications to meticulously identify and classify lesions across diverse skin types. This strategic approach contributes profoundly to the reduction of potential biases and affirms the technology's reliability for a broad and variegated spectrum of end-users.

Additionally, the active involvement of seasoned healthcare professionals plays a critical role in the formulation of the training parameters for AI models and the meticulous review of the decisions emanating from these applications. This collaborative synergy serves as an anchor to guarantee diagnostic precision while also facilitating the identification of intricate cases that may pose challenges to the technology. Furthermore, the validation of these applications by dermatologists is of paramount importance in the comprehensive evaluation of their effectiveness in comparison to conventional diagnostic methodologies.

In this manner, Table 4 presents a repository of pertinent information pertaining to the ethnic and genetic diversity of the study population, in conjunction with a meticulous assessment of the participation levels of dermatologists and other healthcare professionals in each research study.

The studies present diverse approaches in their research endeavors. For instance, Udrea et al. (31) emphasizes the

inclusion of data origin information, indicating that the data predominantly comes from countries such as the United Kingdom, the Netherlands, Australia, and New Zealand. On the other hand, Pangti et al. (36) mentions the scarcity of clinical images from different ethnicities as a challenge but addresses this issue by using locally generated data to mitigate class imbalance and racial bias in public datasets.

Another notable difference lies in the validation approach. While Fujisawa et al. (30) and Pangti et al. (36) highlight comparative validation with diagnoses performed by healthcare professionals, Francese et al. (37) focuses on evaluation by dermatologists through post-test questionnaires. Each of these studies adopts a unique strategy to verify the effectiveness and accuracy of the applications.

Moreover, Dulmage et al. (35) draws attention to image classification based on the Fitzpatrick skin type, emphasizing specific considerations for variations in skin tone in their assessments. Conversely, Bianchi et al. (17) utilizes data collected through tele dermatology for their project, highlighting a different data acquisition approach.

In summary, the studies exhibit differences in terms of data origin, validation strategies, considerations regarding ethnic diversity, and specific data collection approaches, showcasing the diversity and innovation in the approaches taken to create skin lesion detection applications. However, a central characteristic is the close collaboration with dermatologists and medical professionals, as evidenced in multiple studies. This direct interaction ensures the clinical validity of the applications by aligning the AI decisions with specialized medical knowledge.

Furthermore, comparing results with assessments by dermatologists reinforces the diagnostic accuracy of these technologies. Notably, the explicit consideration of ethnic and genetic diversity within the population, as discussed in Fujisawa et al. (30) and Pangti et al. (36), also stands out as a significant strength. By encompassing various skin types and demographic characteristics, such applications become more comprehensive and reliable in real-world scenarios. Taken together, these aspects underscore the relevance of these applications in medical practice and their potential to significantly contribute to early and accurate skin lesion detection.

When analyzing the studies, a consensus becomes evident regarding the importance of accessibility and availability of systems and applications for skin lesion detection and classification. However, many systems still fail fully meet these requirements due to resource limitations, technical complexity, or the absence

TABLE 3 Overview of classification results and potential implications of skin cancer studies for clinical settings by year and author.

References	Key classification results	Prospects
Ramlakhan and Shang (25)	Sensitivity of 80.5% for benign lesions and 60.7% for malignant lesions.	Demonstrates the ability to perform image segmentation, calculate features, and classify lesions on a smartphone with good recognition accuracy.
Afifi et al. (26)	No results presented regarding the classifier.	The system can detect melanoma in real-time with high accuracy and low power consumption, proposed for use in primary care settings, using a high-level hardware design methodology to implement the SVM classifier quickly and efficiently on an FPGA.
Thissen et al. (27)	Achieved 80% sensitivity and 78% specificity in detecting (pre)malignant conditions.	The evaluated app can support less experienced professionals in differentiating between benign and malignant lesions. It analyzes data related to texture, color, geometric features extracted from images, as well as lesion characteristics (lesion age, pain, itching, bleeding, among others).
Roy et al. (28)	The proposed model, YOLOv2, achieved an average precision of 0.89, average recall of 0.91, overall accuracy of 86.00%, recall of 86.35%, specificity of 85.90%, and a frame rate of 21 FPS, indicating high precision and recall in detecting melanoma in dermoscopic images, as well as efficiency in terms of time.	YOLOv2 is presented as a more efficient and accurate approach than other works in automatic melanoma detection in dermoscopic images. The model can process images in real-time with high precision and recall in melanoma detection, and it is invariant to the presence of hair in the images.
Alizadeh and Mahloojifar (29)	Average accuracy, sensitivity, and specificity were 95%, 98%, and 92.19%, respectively.	Development of a mobile application for skin lesion detection using image processing and machine learning techniques.
Fujisawa et al. (30)	The overall accuracy of the trained DCNN was 76.5%. The DCNN achieved a sensitivity of 96.3% (correctly classified as malignant), and a specificity of 89.5% (correctly classified as benign). Although the accuracy of malignancy classification by certified dermatologists was statistically higher than that of dermatology trainees ($85.3\% \pm 3.7\%$ and $74.4\% \pm 6.8\%$, $P < 0.01$), the DCNN achieved higher accuracy.	Classifying skin tumor images into 14 different diagnoses with higher accuracy than certified dermatologists. However, the authors state that it should be validated in a prospective clinical study before considering its use for screening in general medical practice.
Udrea et al. (31)	The machine learning-based skin lesion risk classification algorithm showed sensitivity of 95.1% for melanoma detection and 90.2% for basal cell carcinomas and squamous cell carcinomas. The algorithm's specificity was 78.3% for melanomas and 92.0% for basal cell carcinomas and squamous cell carcinomas. The overall accuracy of the algorithm was 86.1% for melanomas and 79.0% for basal cell carcinomas and squamous cell carcinomas. The study also showed that the algorithm's performance was consistent across different mobile devices and user groups. Additionally, the study demonstrated that the smartphone app could be a useful tool for skin lesion screening in populations with limited access to dermatologists.	Evaluates the accuracy of the latest version of a smartphone app for skin lesion risk assessment and provides an accessible and user-friendly screening tool for individuals with limited access to dermatologists.
Bakheet and Al-Hamadi (32)	The method achieved an area under the ROC curve (AUC) of 0.94, indicating good performance in distinguishing between benign and malignant lesions. Additionally, the method showed sensitivity of 100%, specificity of 95-99%, positive predictive value (PPV) of 86-90%, and negative predictive value (NPV) of 100%.	Developing an effective and fast method with promising performance and 100% sensitivity. The premise is that the detection of malignant melanoma in skin lesion images can be improved through image processing and machine learning techniques. The proposed method uses specific lesion features, such as color and asymmetry, to classify lesions as benign or malignant.
Abbas (33)	The proposed Smart-Dermo achieved 92% accuracy in classifying malignant melanomas and benign tumors.	The Smart-Dermo app aims to assist dermatologists and healthcare professionals in diagnosing skin lesions, enabling early detection and patient monitoring for skin cancer risk. The work is based on using smartphones as processing devices and training the machine learning algorithm with a database of pre-classified dermoscopy images.
Bakheet and El-Nagar (34)	The method achieved an average accuracy rate of 97.5%, sensitivity of 96.67%, and specificity of 100.0% on a dermoscopy image dataset.	The study promises an efficient and real-time approach for melanoma detection in dermoscopy images, with results comparable to or superior to state-of-the-art methods. The work's premises include using a well-established dermoscopy image dataset and validating the proposed method on a test set.
Dulmage et al. (35)	The main results of the study show that the AI system can categorize skin lesion morphology with 68% accuracy. When considering the top three classifications predicted by the AI system, accuracy increases to 80%. Additionally, the study reveals that the AI system performed similarly to primary care physicians who used visual guidance to assist in lesion morphology categorization.	The work aims to develop an AI system capable of categorizing skin lesion morphology with high accuracy, which can be useful for primary care and emergency physicians in diagnosing skin diseases.
Pangti et al. (36)	The machine learning model achieved an overall accuracy of 76.93% ($\pm 0.88\%$) in top-1 and an average area under the curve (AUC) of 0.95 (± 0.02) on clinical images in an <i>in silico</i> validation study. In a clinical study with patients of color, the app achieved an overall accuracy of 75.07% (95% CI = 73.75-76.36) in top-1, 89.62% (95% CI = 88.67-90.52) in top-3, and an average AUC of 0.90 (± 0.07).	The model was trained on a large dataset of skin lesion images and evaluated in three different clinical settings, including an internal validation dataset, an external validation dataset, and a multicenter prospective clinical study, providing a diagnostic tool for 40 types of skin lesions.

(Continued)

TABLE 3 (Continued)

References	Key classification results	Prospects
Giavina-Bianchi et al. (17)	Dermoscopy models achieved an accuracy of 89.3% for melanoma, while the clinical model achieved an accuracy of 84.7%. Sensitivity for these models was 0.91 and 0.89, and specificity reached 0.89 and 0.83, respectively. Both models demonstrated a remarkable area under the curve (AUC) exceeding 0.9.	Developed a mobile application with a data collection protocol (photos, demographic information, and brief medical history) and AI to classify clinical and dermoscopic images. The app generates reports for each lesion with images, indicative heatmaps, estimated probability of melanoma or malignancy, likely diagnosis, and management suggestions.
Francesca et al. (37)	The results are related to the usability of the application: clarity of tasks (100% of dermatologists found tasks clear), ease of use of the app (5 dermatologists found it easy to use), the need for technical support (100% of dermatologists felt they would not need support), and integration of system functions (100% of dermatologists found functions well-integrated).	It is possible to identify that the work proposes a system for skin lesion analysis that uses augmented reality and deep learning techniques to assist dermatologists in diagnosing skin lesions. The system was evaluated through a post-test questionnaire answered by dermatologists, and the results indicated that the system is easy to use and does not require additional technical support.
Felmingham et al. (24)	The study is still ongoing.	The promises and premises of the work are to assess the effectiveness of the CNN in assisting physicians in diagnosing and managing skin lesions in a real-world clinical environment. The study also aims to evaluate the safety of the AI algorithm before its use in post-intervention settings and assess the acceptance of the AI algorithm by physicians and patients.
Sangers et al. (38)	The app showed an overall sensitivity of 86.9% and specificity of 70.4%. Sensitivity was significantly higher on iOS devices compared to Android devices (91.0% vs. 83.0%). Furthermore, specificity was considerably higher for control benign lesions compared to suspicious skin lesions (80.1% vs. 45.5%). It was also observed that sensitivity was higher in skin fold areas compared to smooth skin areas (92.9% vs. 84.2%), while specificity was higher for lesions in smooth skin areas (72.0% vs. 56.6%).	The study evaluated the effectiveness of the app in detecting skin lesions at risk of skin cancer and concluded that the app has the potential to help patients assess their skin lesions before consulting a healthcare professional.
Jahn et al. (39)	The study assessed the diagnostic accuracy of the SkinVision® smartphone app in melanoma detection and found that the app classified a significantly higher number of lesions as high-risk compared to dermatologists, potentially leading to unnecessary excisions. Additionally, the diagnostic performance of the app was below the advertised rates, with low sensitivity and specificity.	The text highlights the importance of evaluating apps for certification with real-world prospective evidence.
Kränke et al. (40)	The detection algorithm showed a sensitivity of 96.4% and specificity of 94.85%, while the analysis algorithm achieved a sensitivity of 95.35% and specificity of 90.32%.	To evaluate the accuracy of two new neural networks for diagnosing skin cancer on currently available smartphones. The study also aimed to provide a low-cost and easily accessible screening tool for early skin cancer detection.

of clear guidelines. To address this issue, broader collaboration among companies, accessibility experts, programmers, and users is crucial in translating intentions into practical actions. Such collaborative effort will result in significant benefits for all parties involved.

Finally, it is essential to ensure that AI applications are developed and tested ethically and responsibly. This includes safeguarding patient data privacy and security, as well as ensuring transparency in the development and training processes of algorithms.

4 Conclusion

The application of artificial intelligence in dermatology has the potential to revolutionize the detection and diagnosis of skin lesions, especially in the case of melanoma, a severe and potentially fatal disease.

This review highlights that several studies are making significant advancements in improving image processing

capabilities, pattern recognition, and deep learning. These advancements enable rapid and accurate analyses that can lead to real-time diagnoses. This evolution contributes to early detection of skin cancer, expanding the prospects for cure and minimizing the reliance on invasive procedures.

However, it is important to note that the vast majority of the solutions presented have not yet been validated in clinical settings or developed in collaboration with dermatologists and other healthcare professionals to ensure they meet patients' needs and are effective in clinical practice.

In summary, the solutions presented can help enhance the efficiency of healthcare services, reducing the time required for examinations and diagnoses. This can be especially important in areas with a shortage of healthcare professionals or in emergency situations where time is critical. However, they should be used with caution and responsibility, in collaboration with dermatologists and other healthcare professionals, to ensure they meet patients' needs and are effective in clinical practice.

TABLE 4 Diversity considerations and medical professional involvement in skin cancer studies for clinical settings by year and author.

References	Ethnic and genetic diversity of the population	Participation of dermatologists and other medical professionals
Ramlakhan and Shang (25)	Does not present data on this aspect.	Does not present data on this aspect.
Afifi et al. (26)	Does not present data on this aspect.	Does not present data on this aspect.
Thissen et al. (27)	Does not present data on this aspect.	The text mentions that consecutive patients were seen by both a dermatologist and a dermatology resident.
Roy et al. (28)	Does not present data on this aspect.	Does not present data on this aspect.
Alizadeh and Mahloojifar (29)	Does not present data on this aspect.	The text mentions that results are displayed to dermatologists on smartphones, suggesting that the system may be used by healthcare professionals.
Fujisawa et al. (30)	The study mentions that it was conducted in the Division of Dermatology at the University of Tsukuba Hospital but does not provide additional information about the studied population.	The authors compare results with interns and dermatologists, implying that the system may be developed to assist medical professionals in their diagnoses.
Udrea et al. (31)	The data primarily come from countries such as the United Kingdom, the Netherlands, Australia, and New Zealand. However, it does not provide additional information about the studied population's diversity.	Yes, the study mentions that each pair of image and corresponding risk classification undergoes a quality control check performed by a dermatologist. Moreover, for lesions classified as high-risk or for cases that have been upgraded or downgraded by a dermatologist, the user will receive a message from the Customer Care team within 48 hours, indicating the level of urgency. This indicates that there is dermatologist support and involvement in the skin lesion assessment process.
Bakheet and Al-Hamadi (32)	Does not present data on this aspect.	The study mentions that the methodology was developed in collaboration with dermatologists and other medical professionals.
Abbas (33)	Does not present data on this aspect.	The Smart-Dermo application was developed to assist dermatologists and healthcare professionals in diagnosing skin lesions. However, it does not provide detailed information about the specific support of dermatologists and other medical professionals during the application's development. It can be inferred that the application aims to provide an additional tool to assist healthcare professionals in diagnosing and monitoring patients at risk of developing skin cancer.
Bakheet and El-Nagar (34)	Does not present data on this aspect.	Does not present data on this aspect.
Dulmage et al. (35)	The study mentions concerns about the potential for artificial intelligence technology to exacerbate health inequalities among patients of different ethnicities but does not provide specific data on the ethnic and genetic diversity of the studied population. Additionally, the images were classified by Fitzpatrick skin type and separated into darker skin types (Fitzpatrick skin type IV - VI) and lighter skin types (Fitzpatrick skin type I - III).	The study mentions that the artificial intelligence system was developed in collaboration with dermatologists and other medical professionals. The study also mentions that skin lesion images used to train the system were manually labeled by dermatologists.
Pangti et al. (36)	The work mentions the scarcity of clinical images (macroscopic) from different ethnicities as one of the major challenges in developing deep learning-based skin disease classifiers. Additionally, the text mentions that using locally generated data helped address the issue of class imbalance and racial bias in public datasets. However, the text does not provide specific information about the ethnic and genetic diversity of the population used.	Dermatologists were involved in the study to assess the accuracy of the skin disease diagnostic application compared to human dermatologists.
Giavina-Bianchi et al. (17)	Does not present data on this aspect.	Data was obtained for a teledermatology project, meaning it utilized data collected by dermatologists.
Francese et al. (37)	Does not present data on this aspect.	The system was evaluated by dermatologists through a post-test questionnaire.
Felmingham et al. (24)	Does not present data on this aspect.	The study is led by dermatologists and involves other medical professionals, including pathologists and nurses.
Sangers et al. (38)	Does not present data on this aspect.	A set of 239 cases were confirmed through dermatological evaluation and/or histopathology.
Jahn et al. (39)	Does not present data on this aspect.	Seven dermatologists participated in the study as evaluators of the lesions.
Kränke et al. (40)	Does not present data on this aspect.	The study was conducted by the Department of Dermatology at the Medical University of Graz, Austria, suggesting the involvement of dermatologists and other medical professionals in the study's execution.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

BF: Writing—original draft, Writing—review & editing. BO: Writing—review & editing. RP: Writing—review & editing. JP: Writing—review & editing. RL: Writing—original draft, Writing—review & editing. WC: Writing—review & editing. MR: Writing—review & editing. MG-B: Writing—original draft, Writing—review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article.

References

- Balaha HM, Hassan AE-S. Skin cancer diagnosis based on deep transfer learning and sparrow search algorithm. *Neural Comp Appl.* (2023) 35:815–53.
- In T. *Facts & Figures 2019: US Cancer Death rate has Dropped 27% in 25 Years.* Washington, DC: American Cancer Society. (2019).
- Fitzmaurice C, Akinyemiju TF, Al Lami FH, Alam T, Alizadeh-Navaei R, Allen C, et al. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2016: a systematic analysis for the global burden of disease study. *JAMA Oncol.* (2018) 4:1553–68. doi: 10.1200/JCO.2018.36.15_suppl.1568
- Guy Jr GP, Machlin SR, Ekwueme DU, Yabroff KR. Prevalence and costs of skin cancer treatment in the US, 2002–2006 and 2007–2011. *Am J Prev Med.* (2015) 48:183–7. doi: 10.1016/j.amepre.2014.08.036
- Gordon R. Skin cancer: an overview of epidemiology and risk factors. *Semi Oncol Nurs.* (2013) 29:160–9. doi: 10.1016/j.soncn.2013.06.002
- Leiter U, Keim U, Garbe C. Epidemiology of skin cancer: update 2019. *Sunlight Vitamin D Skin Cancer.* (2020) 123–39.
- di Meo N, Vernoni S, Longone M, Trevisan G. Image gallery: Merkel cell carcinoma under the rainbow. *Br J Dermatol.* (2017) 177:e166–e166. doi: 10.1111/bjd.15815
- Molina DAC, Rios-Duarte JA, Mu noz-Ordo nez S, Fierro-Lozada JD, Celorio-Murillo WJ, Hernandez-Amaris ME, et al. 33734 Patient education as the main target in skin cancer prevention: knowledge, attitudes, and practices toward sun exposure and use of sun protection. *J Am Acad Dermatol.* (2022) 87:AB189. doi: 10.1016/j.jaad.2022.06.789
- Hammad SSAEH, Gaber MA. Knowledge, attitudes and practices of the general public toward the harmful effects of sun exposure and protection. *J Health Sci.* 6:14007–20. doi: 10.53730/ijhs.v6n2.8675
- Besch-Stokes J, Brumfiel CM, Patel MH, Harvey J, Montoya J, Severson KJ, et al. Skin cancer knowledge, attitudes and sun protection practices in the hispanic population: a cross-sectional survey. *J Racial Ethn Health Disparitie.* (2022) 10:1293–303. doi: 10.1007/s40615-022-01314-6
- Jiang S, Li H, Jin Z. A visually interpretable deep learning framework for histopathological image-based skin cancer diagnosis. *IEEE J Biomed Health Inform.* (2021) 25:1483–94. doi: 10.1109/JBHI.2021.3052044
- Parsons SK, Chan JA, Winifred WY, Obadan N, Ratichek SJ, Lee J, et al. *Noninvasive Diagnostic Techniques for the Detection of Skin Cancers.* (2011).
- Malvey J, Pellacani G. Dermoscopy, confocal microscopy and other non-invasive tools for the diagnosis of non-melanoma skin cancers and other skin conditions. *Acta Derm Venereol.* (2017) 97:22–30. doi: 10.2340/00015555-2720
- Narayanamurthy V, Padmapriya P, Noorasafrin A, Pooja B, Hema K, Nithyakalyani K, et al. Skin cancer detection using non-invasive techniques. *RSC Adv.* (2018) 8:28095–130. doi: 10.1039/C8RA04164D
- Garbe C, Peris K, Soura E, Forsea AM, Hauschild A, Arenbergerova M, et al. The evolving field of Dermato-oncology and the role of dermatologists: position paper of the EADO, EADV and task forces, EDE, IDS, EBDV-UEMS and EORTC cutaneous lymphoma task force. *J Eur Acad Dermatol Venereol.* (2020) 34:2183–97. doi: 10.1111/jdv.16849
- Frisinger A, Papachristou P. The voice of healthcare: introducing digital decision support systems into clinical practice—a qualitative study. *BMC Primary Care.* (2023) 24:67. doi: 10.1186/s12875-023-02024-6
- Giavina-Bianchi M, de Sousa RM, Paciello VZdA, Vitor WG, Okita AL, Prôa R, et al. Implementation of artificial intelligence algorithms for melanoma screening in a primary care setting. *PLoS ONE.* (2021) 16:e0257006. doi: 10.1371/journal.pone.0257006
- Esteve 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) 542:115–8. doi: 10.1038/nature21056
- Kwiatkowska D, Kluska P, Reich A. Convolutional neural networks for the detection of malignant melanoma in dermoscopy images. *Adv Dermatol Allergol.* (2021) 38:412. doi: 10.5114/ada.2021.107927
- Goyal M, Knackstedt T, Yan S, Hassanpour S. Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. *Comput Biol Med.* (2020) 127:104065. doi: 10.1016/j.combiomed.2020.104065
- Hosny KM, Kassem MA, Foad MM. Classification of skin lesions using transfer learning and augmentation with Alex-net. *PLoS ONE.* (2019) 14:e0217293. doi: 10.1371/journal.pone.0217293
- Barros WK, Morais DS, Lopes FF, Torquato MF, Barbosa RdM, Fernandes MA. Proposal of the CAD system for melanoma detection using reconfigurable computing. *Sensors.* (2020) 20:3168. doi: 10.3390/s20113168
- Pyun SH, Min W, Goo B, Seit S, Azzi A, Wong DYS, et al. Real-time, in vivo skin cancer triage by laser-induced plasma spectroscopy combined with a deep learning-based diagnostic algorithm. *J Am Acad Dermatol.* (2022). doi: 10.1016/j.jaad.2022.06.1166
- Felmingham C, MacNamara S, Cranwell W, Williams N, Wada M, Adler NR, et al. Improving Skin cancer Management with Artificial Intelligence (SMARTI): protocol for a preintervention/postintervention trial of an artificial intelligence system used as a diagnostic aid for skin cancer management in a specialist dermatology setting. *BMJ Open.* (2022) 12:e050203. doi: 10.1136/bmjopen-2021-050203

This work was supported by the Program of Support for the Institutional Development of the Unified Health System (PROADI-SUS,01/2020; NUP: 25000.161106/2020-61) and Hospital Israelita Albert Einstein.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

25. Ramlakhan K, Shang Y. A mobile automated skin lesion classification system. In: *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*. Boca Raton: IEEE. (2011) p. 138–41.
26. Afifi S, GholamHosseini H, Sinha R. A low-cost FPGA-based SVM classifier for melanoma detection. In: *2016 IEEE EMBS Conference on Biomedical Engineering and Sciences (IECBES)*. Kuala Lumpur: IEEE. (2016) p. 631–6.
27. Thissen M, Udrea A, Hacking M, von Braunmühl T, Ruzicka T. mHealth app for risk assessment of pigmented and nonpigmented skin lesions—a study on sensitivity and specificity in detecting malignancy. *Telemed e-Health*. (2017) 23:948–54. doi: 10.1089/tmj.2016.0259
28. Roy SS, Haque AU, Neubert J. Automatic diagnosis of melanoma from dermoscopic image using real-time object detection. In: *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*. Princeton, NJ: IEEE. (2018) p. 1–5.
29. Alizadeh SM, Mahloojifar A. A mobile application for early detection of melanoma by image processing algorithms. In: *2018 25th National and 3rd International Iranian Conference on Biomedical Engineering (ICBME)*. Qom: IEEE. (2018) p. 1–5.
30. Fujisawa Y, Otomo Y, Ogata Y, Nakamura Y, Fujita R, Ishitsuka Y, et al. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. *Br J Dermatol*. (2019) 180:373–81. doi: 10.1111/bjd.16924
31. Udrea A, Mitra G, Costea D, Noels E, Wakkee M, Siegel D, et al. Accuracy of a smartphone application for triage of skin lesions based on machine learning algorithms. *J Eur Acad Dermatol Venereol*. (2020) 34:648–55. doi: 10.1111/jdv.15935
32. Bakheet S, Al-Hamadi A. Computer-aided diagnosis of malignant melanoma using gabor-based entropic features and multilevel neural networks. *Diagnostics*. (2020) 10:822. doi: 10.3390/diagnostics10100822
33. Abbas Q. Smart-Dermo: a computerize tool for classification of skin cancer using smartphone through image processing and fuzzy logic. *Int J Comp Sci Netw Secur*. (2020).
34. Bakheet S, El-Nagar A. A deep neural approach for real-time malignant melanoma detection. *Appl Math*. (2021) 15:89–96. doi: 10.18576/amis/150111
35. Dulmage B, Tegtmeier K, Zhang MZ, Colavincenzo M, Xu S, A. point-of-care, real-time artificial intelligence system to support clinician diagnosis of a wide range of skin diseases. *J Investigat Dermatol*. (2021) 141:1230–5. doi: 10.1016/j.jid.2020.08.027
36. Pangti R, Mathur J, Chouhan V, Kumar S, Rajput L, Shah S, et al. A machine learning-based, decision support, mobile phone application for diagnosis of common dermatological diseases. *J Eur Acad Dermatol Venereol*. (2021) 35:536–45. doi: 10.1111/jdv.16967
37. Francese R, Frasca M, Risi M, Tortora G, A. mobile augmented reality application for supporting real-time skin lesion analysis based on deep learning. *J Real-Time Image Proc*. (2021) 18:1247–59. doi: 10.1007/s11554-021-01109-8
38. Sangers T, Reeder S, van der Vet S, Jhingoer S, Mooyaart A, Siegel DM, et al. Validation of a market-approved artificial intelligence mobile health app for skin cancer screening: a prospective multicenter diagnostic accuracy study. *Dermatology*. (2022) 238:649–56. doi: 10.1159/000520474
39. Jahn AS, Navarini AA, Cerminara SE, Kostner L, Huber SM, Kunz M, et al. Over-detection of melanoma-suspect lesions by a CE-certified smartphone app: performance in comparison to dermatologists, 2D and 3D convolutional neural networks in a prospective data set of 1204 pigmented skin lesions involving patients' perception. *Cancers*. (2022) 14:3829. doi: 10.3390/cancers14153829
40. Kränke T, Tripolt-Droschl K, Röd L, Hofmann-Wellenhof R, Koppitz M, Tripolt M. New AI-algorithms on smartphones to detect skin cancer in a clinical setting—a validation study. *PLoS ONE*. (2023) 18:e0280670. doi: 10.1371/journal.pone.0280670



OPEN ACCESS

EDITED BY

Justin Ko,
Stanford University, United States

REVIEWED BY

Gerardo Cazzato,
University of Bari Aldo Moro, Italy
Jana Lipkova,
University of California, Irvine, United States

*CORRESPONDENCE

Maria L. Wei
✉ maria.wei@ucsf.edu

RECEIVED 01 November 2023

ACCEPTED 26 February 2024

PUBLISHED 19 March 2024

CITATION

Wei ML, Tada M, So A and Torres R (2024)
Artificial intelligence and skin cancer.
Front. Med. 11:1331895.
doi: 10.3389/fmed.2024.1331895

COPYRIGHT

© 2024 Wei, Tada, So and Torres. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Artificial intelligence and skin cancer

Maria L. Wei^{1,2*}, Mikio Tada³, Alexandra So⁴ and Rodrigo Torres²

¹Department of Dermatology, University of California, San Francisco, San Francisco, CA, United States,

²Dermatology Service, San Francisco VA Health Care System, San Francisco, CA, United States,

³Institute for Neurodegenerative Diseases, University of California, San Francisco, San Francisco, CA, United States, ⁴School of Medicine, University of California, San Francisco, San Francisco, CA, United States

Artificial intelligence is poised to rapidly reshape many fields, including that of skin cancer screening and diagnosis, both as a disruptive and assistive technology. Together with the collection and availability of large medical data sets, artificial intelligence will become a powerful tool that can be leveraged by physicians in their diagnoses and treatment plans for patients. This comprehensive review focuses on current progress toward AI applications for patients, primary care providers, dermatologists, and dermatopathologists, explores the diverse applications of image and molecular processing for skin cancer, and highlights AI's potential for patient self-screening and improving diagnostic accuracy for non-dermatologists. We additionally delve into the challenges and barriers to clinical implementation, paths forward for implementation and areas of active research.

KEYWORDS

artificial intelligence, skin cancer, melanoma, dermatology, dermatopathology

Introduction

Artificial intelligence (AI) stands at the forefront of technological innovation and has permeated into almost every industry and field. In dermatology, significant progress has been made toward the application of AI in skin cancer screening and diagnosis. Notably, a milestone that marked the era of modern artificial intelligence in dermatology was the demonstration of skin cancer classification abilities by deep learning convolutional neural networks (CNNs), which was on par with the performance of board-certified dermatologists (1). This CNN was trained on a dataset that was two orders of magnitude greater than those previously utilized. The dermatologist-level classification ability has since been experimentally validated by other papers (2, 3). Recent progress in the field of AI enables models to not only analyze image data but also integrate clinical information, including patient demographics and past medical history (4–6). Advancements allow for the simultaneous evaluation and identification of multiple lesions from wide-field images (7, 8). Moreover, models can now gain information from whole slide images without having to use costly pixel-wise human-made annotations (9). Despite these advancements, research has found that AI models lack robustness to simple data variations, have proven inadequate in real-world dermatologic practice performance, and that barriers remain before achieving clinical readiness (2, 10–14).

Clinical applications

Artificial intelligence has been employed to predict the most common types of skin cancers, melanoma (1) and non-melanoma skin cancer (1), through image analysis. In addition, machine learning has been used on RNA datasets to develop classifiers that also predict skin cancer, as well as the prognosis of skin lesions. Several of these methods can be, or have the potential to be, readily deployed by patients, primary care practitioners, dermatologists, and dermatopathologists.

Patients

With the rising prevalence of smartphone usage, patients can directly screen for and monitor lesions with AI applications. These applications can run AI models on patients' own local devices, which ensures the protection of patient data (15). The feasibility of an AI model to assist patients' with self-assessed risk using smartphones has been validated with a model that was trained on pictures captured from patients' smartphones, and which exhibited comparable performance to general practitioners' ability to distinguish lower-risk vs. higher-risk pigmented lesions (16). Moreover, AI significantly increased the abilities of 23 non-medical professionals to correctly determine a diagnosis of malignancy from 47.6 to 87.5% without compromising specificity (12). In the future, AI models may assist with overseeing and assessing changes to lesions as they progress (17) and collaborate with apps that allow patients to examine themselves and document moles (18, 19).

Despite progress with these AI models, there is no smartphone application that is endorsed on the market in the United States for non-professionals to evaluate their lesions as they do not have satisfactory performance or generalizability (20). Limitations include biases introduced due to the narrow range of lesion types, skin pigmentation types, and low number of high-quality curated images used in training. Further, inadequate follow-up has been a limitation with regards to identifying false negative diagnoses (21). Notably, users may not be adequately protected from the risks of using smartphone diagnostic apps by Conformit Européenne (CE) certification, which endorsed two apps with flaws (SkinVision and TeleSkin's skinScan app). A prospective trial of SkinVision found low sensitivity and specificity for melanoma classification (22). In contrast to CE, the US Food and Drug Administration's (FDA) requirements for endorsement are more stringent (21).

Primary care

Artificial intelligence applications can enhance skin cancer screening in the primary care setting and streamline referrals to dermatologists. Referral data from primary care practitioners to teledermatology consultations were used to train a model capable of a top-3 accuracy and specificity of 93 and 83%, respectively, given 26 skin conditions that makeup 80% of encountered primary care cases (4). This performance was on par with dermatologists and surpassed primary care physicians (PCPs) and nurse practitioners. This type of model could assist PCPs in diagnosing patients more accurately and broadening their differential diagnoses. In cases in which the top 3 diagnoses from the model have the same management strategy, patients may start treatment while awaiting further workup or follow-up with dermatology. Nevertheless, further testing on

populations with a low prevalence of skin cancer is essential to demonstrate efficacy in the broader population (23).

Dermatology

Models have been trained to use electronic health record (EHR) data and/or gene sequencing data to predict an individual's likelihood of developing melanoma (24–27) or nonmelanoma skin cancer (27–31). While AI models could potentially flag patients at high risk of skin cancer to be screened, studies are limited by the variability of included predictive factors, inconsistent methods of evaluating models, and inadequate validation (32). Moreover, EHRs often do not include some of the most important risk determinants for skin cancer, such as exposure to UV light and the patient's familial history; the omission of such data may result in decreased performance (28).

Artificial intelligence has the potential to supplement dermatologists' diagnostic and treatment capabilities in what is known as augmented intelligence (AuI). For diagnosis, AuI might assist dermatologists in more effectively managing teledermatology referrals (4) and increase the efficacy of in-person visits (33). However, in a prospective trial comparing AI to dermatologists in a teledermatology setting, dermatologists outperformed the AI (13). Despite AI currently underperforming dermatologists, AI could provide a new perspective that could still be beneficial as AI and humans exhibit distinct types of errors. For instance, models may provide insights into certain images' classification ambiguity, whereas humans are better able to distinguish variability in image quality such as blurriness or shadowing (12).

Augmented intelligence can also assist with suggesting clinical decisions given inputted images, such as recommending whether a lesion warrants excision (34). The integration of AuI into dermatologic patient management resulted in a 19.2% reduction in unnecessary excisions of benign lesions (35). Although current CNNs' performance has been shown to fall short when compared with using sequential dermatoscopic photography in predicting melanoma, AuI may be used in the future by dermatologists to evaluate and monitor lesion change (36). Of interest, in this study, neither dermatologists nor the CNN had satisfactory diagnostic performance levels on baseline images, but both dermatologists and CNN had improved performances when follow-up images were provided, and the best performance was combining CNN and dermatologist assessment together.

Integration of AI into advanced imaging techniques may reduce the extent of training necessary to use them (37). One area of application is in the detection of the dermal-epidermal junction, which is crucial in a non-invasive method of skin cancer diagnosis called reflectance confocal microscopy (RCM) imaging (38). Furthermore, there are ongoing efforts to analyze RCM images with AI (39).

The FDA has not approved any medical devices or algorithms based on artificial intelligence in the field of dermatology (40, 41). On the other hand, the FotoFinder Moleanalyzer Pro, an AI application for dermatology, was approved in the European market. It demonstrated performance on par with dermatologists in store-and-forward dermatology (42) and a prospective diagnostic study (43), however, the latter had extensive exclusion criteria, e.g., excluding patients of skin type IV and greater. The first randomized controlled trial comparing AI skin lesion prediction to dermatologists' assessment reported that AI did not exceed attending dermatologists in skin cancer detection (44).

Dermatopathology

With the growing application of whole slide imaging (WSI) in the field of dermatopathology (45), AI can potentially support dermatopathologists in several ways, particularly skin cancer recognition. Among the AI models trained to detect melanoma from digitized slides (5, 46–50), two models were able to match the performance of pathologists in an experimental setting. These models were limited in that they were only given either a part of (46) or a single (49) hematoxylin and eosin (H&E)-stained slide. In contrast, pathologists can utilize supplementary data such as immunohistochemistry or relevant patient data. However, integrating patient information, such as age, sex, and lesion location, into CNN models did not enhance performance (5). One limitation to implementing AI in dermatopathology is the unreliable prediction that may be made when a model is given an input that differs from the training dataset. One potential solution is the use of conformal prediction, which has been shown to increase accuracy of prostate biopsy diagnosis by flagging unreliable predictions (51).

Studies have been done to evaluate AI's ability for diagnosing basal cell carcinoma (BCC) using WSI (9, 52, 53). Campanella et al. showed the ability of a convolutional neural network to achieve 100% sensitivity for detecting BCC, on the test set; importantly, a multiple instance learning approach was introduced that obviated the necessity of time-consuming pixel-level slide annotations to distinguish between areas with and without disease (9). Kimeswenger et al. subsequently incorporated an "attention" function to draw attention to areas of digital slides that include indications of BCC. Interestingly, CNN pattern recognition varied from that of pathologists for BCC diagnosis as tissues were flagged based on different image regions (53). These CNNs could also be applied to identify and filter slides for Mohs micrographic surgery (52). In the setting of rising caseloads, AI can help to decrease pathologists' workload generated by these commonly diagnosed, low risk entities. Duschner et al. applied AI to automated diagnosis of BCCs, and demonstrated both sensitivity and specificity of over 98%. Notably, the model demonstrated successful generalization to samples from other centers with similar sensitivity and specificity (54).

Artificial intelligence has also had some success in predicting sentinel lymph node status (55), visceral recurrence, and death (56) based on histology of primary melanoma tumors. In the future, AI could be utilized to identify mitotic figures, delineate tumor margins, and determine the results of immunohistochemistry stains; further, AI could recommend more immunostaining or genetic panels that could be of use diagnostically (57). While AI predictions have not been consistently successful for melanoma (58), AI has been demonstrated to identify the mutation given a lung adenocarcinoma slide that has been stained with H&E (59–61).

Machine learning applied to RNA profiles

While AI in dermatology is most often associated with using deep learning techniques on clinical and histological images, machine learning methods have been utilized in developing gene expression profile (GEP) classifiers for predicting skin cancer diagnosis and prognosis. Generally, simpler machine learning models that require tuning of fewer parameters compared to more complex neural nets have been employed to analyze GEP. They still, however, share the benefits of the ability to use iterative learning optimized to find patterns in complex non-linear relationships not possible in traditional statistical and linear

models, assuming sufficient data is available. Some common models include many Kernel methods such as support vector machines (SVM) or tree-based models, e.g., Random Forest and XGBoost that have often been found to produce the best performance for tabular gene expression data. These models also often use some method to feature select (62) to both maximize performance and find the most relevant features for the classification task. This also allows for a better sense of interpretability as with fewer features there is the ability to assess their relevance individually. Reproducibility is of great concern and has often been the critique of many biomarker and classifier studies, since there is often little to no overlap in targets, which understandably can lead to general skepticism of the results, especially considering the generally small sample sizes employed in many studies. Despite this, there has been a push to make use of molecular profiling to assist in different aspects of melanoma management.

Currently, the GEPs developed for use in melanoma management fall into two categories. First, some GEPs are used as a diagnostic tool to help determine the malignancy of a pigmented lesion either pre- or post-biopsy. Pre-biopsy there is an epidermal tape sampling test that can predict melanoma with 94% sensitivity and 69% specificity (63) with an improved sensitivity of 97%, when TERT mutation assessment is included (64). There are, however, reported limitations to this test as it cannot be used on mucous membranes or acral skin and there is the possibility of non-actionable results due to insufficient sample collected for testing (65). Post-biopsy GEPs can be used to help with diagnostically difficult cases such as Spitz nevi, but have poorer performance on Spitz melanomas and pediatric patients (66). Machine learning has also been applied with success to miRNA profiles to differentiate melanomas from nevi (67).

Second, there are GEPs, derived from biopsy material, that are used as prognostic tools to stratify the risk of melanoma recurrence or metastasis (68), however subsequent management protocols for high risk early-stage disease are not in place (68). Despite optimism for prognostic use of prognostic GEP classifiers, the expert consensus is that there is currently insufficient evidence to support routine use (69). The climate, however, is evolving, with new reports incorporating additional clinicopathological data together with patient outcomes (70). Overall, there remains a lack of consensus on the use of the GEP biopsy and tape sampling tests (71, 72). Further studies are needed, such as non-interventional retrospective studies, followed by prospective interventional trials, but there remains promise that they can become additional tools in providers' arsenal of available tests.

Barriers to clinical implementation

Image quality

Image quality significantly impacts the prediction performance of AI computer vision (73). Several factors can result in subpar images, including inadequate focus or lighting, color misrepresentations, unfavorable angles or framing, obstructing objects, and poor resolution. Moreover, while humans can readily ignore items such as blurred focus, scale bars, and surgical markings, these artifacts affect AI prediction performance (11, 74, 75).

Obtaining consistently high-quality images in the fast-paced environment of a clinic presents many challenges. Barriers such as limited time, insufficient training, inadequate imaging equipment, and other constraints may hinder the process. Guidelines for skin lesion

imaging have been suggested to facilitate the capture of high-quality images (76, 77). These guidelines include suggestions for adequate lighting, background, field of view, image orientation, and color calibration. Additional recommendations are suggested for photographing skin of color (78).

A comprehensive, multifaceted solution is necessary to enhance image quality. Educating dermatology residents in photography might contribute to improving image quality in a clinical setting (79). Moreover, a study done in United Kingdom primary care facilities showed enhanced photo quality when patients were educated with the “4 Key Instructions” (Framing—requesting at least one near and one distant image; Flash—educating about the use of flash to enhance image sharpness, emphasizing not to use it too closely; Focus—educating patients to give the camera time to auto-focus; Scale—asking for a comparison like a ruler or a coin) (73). Among 191 digital applications for skin imaging, 57% included one or more strategies to enhance quality, but it was rare for applications to have more than one (80). An immediate feedback feature for image quality shows promise, although it is still in the early stages of development (81).

Algorithmic bias and health equity

There is a risk for indiscriminately implemented AI to potentially exacerbate health inequities by incorporating pre-existing and newly emerging biases (82) (Table 1). Pre-existing biases include pre-coding biases in datasets used to train the model or personal biases inadvertently introduced by developers. Emergent biases can be introduced by relying on models in new or unexpected contexts and not adjusting models for new knowledge and shifting cultural norms.

Artificial intelligence models for early melanoma detection have relied on large datasets from individuals with mostly lighter pigmented skin. While melanoma is more prevalent among individuals with lighter skin, those with darker skin frequently come in with a more severe stage of disease and experience lower survival rates. An AI model trained on lighter skin tones for melanoma prediction had lower performance for lesions on darker skin tones (83). The International Skin Imaging Collaboration (ISIC) archive, one of the most extensive and widely used databases for individuals in the United States, Europe, and Australia, and a prospective diagnostic accuracy paper comparing an AI model with other noninvasive imaging techniques did not include individuals with Fitzpatrick phototype III or higher (43, 84). Efforts to collect lesions from individuals of all skin tones should be a priority, and transparency in the characteristics of training datasets as well as the quality and range of disease labels should be disclosed (85).

AI model validation

It is crucial to carefully validate AI models before applying them in real-world settings (Table 1). Computational stress testing is necessary to guarantee efficacy in actual clinical scenarios (2). Validation should be performed using large amounts of external data as determining performance solely on internal data has been shown to often lead to overestimation (2, 86). The reason for the lower model performance on external validation datasets can arise from training data that is not representative of the general population or from leakage of additional

TABLE 1 Challenges in AI in dermatology.

Challenges	Summary
Model validation	Many models fail to have a true external validation set so can fail to be representative of the general population. In addition, standardized benchmarks that can be used across models are not readily available due to limitations with few public datasets that serve as good benchmarks.
Quality of data	Model performance can be limited by quality of data, which can be affected at initial collection through user error creating data artifacts or with intrinsic deficiencies of the source limiting diversity and creating class imbalances that are not accounted for by the model.
Algorithmic bias and health equity	Models can contain biases based on the selection of data used to train that can affect generalizability to different demographics both racial and socioeconomic.
Implementation and user confidence	Acceptance of AI can be limited not only by governmental agencies such as FDA approving use, but also at the clinician and patient level where mistrust or uncertainty can dissuade use.

data, either between the training and testing data or from the future drift of data (86). Unfortunately, most models are not open code, limiting research into the external validation of these models. On the other hand, Han et al. share the use of their models publicly, setting a standard that should be followed (7, 12, 87). Along with publicly shared models, having publicly shared benchmarks such as the melanoma classification benchmark (88) and accessible databases (such as DataDerm) is crucial for comprehensive validation (89). Few public datasets have representation of all skin types. A rigorous testing of outcome metrics with and without the support of an AI model in randomized controlled trials would be optimal.

Though CNNs routinely and autonomously identify image features pertinent for classification, this ability can lead to the incorporation of unintended biases. An example of possible bias is the use of ink markings (75) or scale bars (74) in melanoma identification. It is important to assess whether and how changes to inputted images can affect the prediction output. Changes to test include image quality, rotation, brightness/contrast adjustments, adversarial noise, and the presence of artifacts, such as those aforementioned (2, 10, 74, 75, 90, 91). Testing for robustness given such uncertainties can assist users in understanding the model’s scope and reasons for error (92).

The path to clinical implementation

Given the rapid pace of advancements in AI in the medical field, the American Academy of Dermatology (AAD) issued a position statement regarding how to integrate augmented intelligence into dermatologic clinical settings (93). The AAD underscored the importance of high-quality validated models, open transparency to patients and providers, and efforts to actively engage stakeholders.

For AI to be broadly accepted in dermatology, studies need to demonstrate a significant improvement in health outcomes. The first randomized controlled trial of an AI’s ability to augment clinicians’ diagnostic accuracy on skin lesions highlighted the potential for AI to augment non-dermatologists diagnostic performance in a real-world setting, but not that of dermatology residents in training, and found

superior performance by experienced dermatologists—who use patient metadata as well as images—compared to the AI model (44). It also noted that if the model's top 3 diagnoses were incorrect, trainees' diagnostic accuracy fell after consulting the AI model, highlighting a pitfall of using current AI models.

Increasing access to dermatological care

AI offers hope for increasing health equity through increasing access, and democratizing skin screenings. Access to dermatologists is a problem, especially in rural areas, where it may take longer for a patient to obtain a biopsy of suspected melanoma (94). As of 2018, 69% of counties in the United States do not have access to dermatologists (95). Further exacerbating the issue, many dermatology clinics closed during the COVID-19 pandemic (96). AI-augmented teledermatology may be able to enhance accessibility by streamlining referrals and reducing waiting times, and it could help increase the accessibility in areas with a scarcity of dermatopathologists. AI may also help dermatologists more accurately diagnose skin disease in patients whose skin is not well-represented in the local population (97).

Human-computer collaboration

Clinicians are indispensable to synthesizing relevant context and offering patient counseling and subsequent care. Furthermore, given the enhanced accuracy of diagnosis when integrating AI into decision-making, the future of dermatology will likely entail human-computer collaboration (98). Embedding Collective Human Intelligence (CoHI) or even swarm intelligence (CoHI with interaction between participating humans) as checkpoints within an AI model may help overcome the limited ability of AI to contextualize and generalize (99).

When interacting with AI, potential cognitive errors and biases may be exacerbated, especially when there is discordance in diagnosis between clinicians and AI (100). The use of AI introduces a new kind of bias called automation bias, in which humans tend to unquestioningly trust automated decisions from AI (100). When physicians used AI decision support for reading chest X-rays, experienced physicians rated diagnostic advice as lower quality when they thought the advice was generated by AI, but not physicians with less experience (101). Though rated as less trustworthy, inaccurate advice by AI still led to decreased diagnostic accuracy (101). It will be important for AI developers and medical educators, the latter when teaching AI applications, to take such human factors into account.

Areas of active research

There are several areas of active computational research that are anticipated to aid in bringing validated image analysis models to clinical use (Table 2).

Federated learning

A problem with training models for clinical use to detect skin cancer or other disorders is the limitation in sharing clinical images

due to privacy concerns and the inherent limitations in collecting sufficient images of rare skin cancer types and disorders and of different skin pigmentation. The current approach for multi-institution model training necessitates the forwarding of patient data to a centralized location, termed collective data sharing (102). Alternatively, federated learning uses a decentralized training system in which a shared global model learns collaboratively while keeping data locally. Each device's data comes with its own inherent bias and different properties due to demographic variations. Instead of sending data to a central server, the model itself travels to each device, learns from the locally-stored data, and then updates the global model with this newly acquired training. By not sharing the training data across devices, federated learning enables the preservation of privacy of sensitive data (103). In a study across 10 institutions, the performance from federated learning was shown to be better than that of a single institution model and shown to be comparable to that of collective data sharing (102). Moreover, the federated learning approach would be a method to virtually aggregate data on rare skin cancers or disorders from different centers, such as Merkel cell cancer, or data from patients with rarer subtypes of skin cancers, such as mucosal or acral melanoma. An analogy of federated learning is a team of dermatologists who visit multiple clinics to learn and share knowledge, rather than asking patients to visit a single central hospital to see the team. A model trained with federated learning can offer more accurate diagnoses on rare skin cancer types and disorders, including lesions found on differing skin pigmentations, and still maintain patient privacy.

Deploying federated learning faces several challenges. Ensuring fairness across different demographic groups and data security while optimizing the overall performance of the global model is computationally complicated. Establishing computational infrastructure capable of seamless communication, such as transmission of a model, may require additional IT assistance. These obstacles pose a barrier to the practical implementation of federated learning (104).

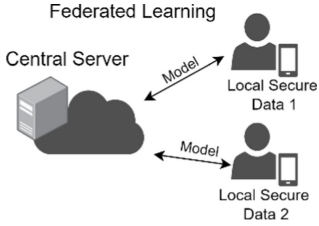
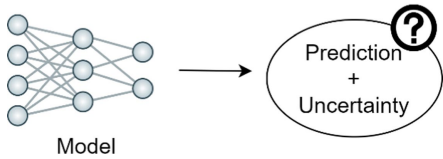
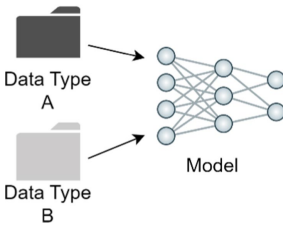
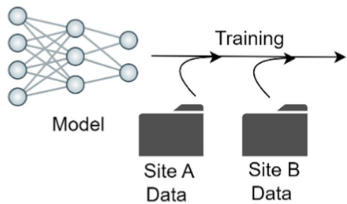
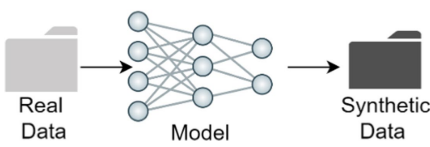
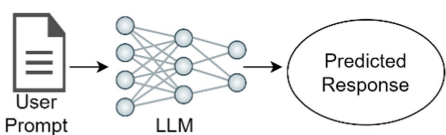
Uncertainty estimation

Whereas many studies on the applications of skin cancer classification models have reported high accuracy, these models rarely concurrently report uncertainty estimates for the predictions and when assessed, models have been found to be overconfident (2). As a result, medical practitioners may hesitate to incorporate these models into their diagnostic workflow. Uncertainty estimation provides a meaningful confidence level, with regards to when to trust a model prediction. To safely deploy a computer-aided diagnostic system in a clinical setting, it is crucial to incorporate not only a model's prediction but also a confidence score. Clinicians are then equipped to decide whether to trust the prediction or alternatively disregard the AI prediction and rely on provider assessment (94).

Multimodal learning

Most skin disease diagnosis models are trained only on one data modality: clinical or histological images or RNA sequence data.

TABLE 2 Future advances in AI.

Method	Description
<p>Federated Learning</p> 	Uses decentralized training where a global model is trained on locally-stored data and then updated while preserving the privacy of local data.
<p>Uncertainty Estimations</p> 	To calculate an uncertainty estimate for model predictions so model confidence can be interpreted by end user.
<p>Multimodal Learning</p> 	Allows the use of multiple types of data to train a combined model to take advantage of unique differences in data.
<p>Incremental Learning</p> 	Enables a model to continue learning on a new stream of data.
<p>Generative Model</p> 	Model is used to train compressed representations of data so new instances can be recreated.
<p>Large Language Model</p> 	Model uses text based prompt to generate a response based on language based learning.

However, medical data is inherently multimodal by nature, and dermatologists use patient information in addition to clinical images to make a diagnosis. Metadata from patients, such as age, ethnicity, and anatomic location of lesion, can also be useful to enhance skin cancer classification models. Multimodal learning is a technique where a single model learns from multiple types of data simultaneously (105). One skin disease classifier that integrated up to six clinical images and 45 demographic items and medical history to classify 26 skin conditions as the primary prediction outperformed six primary care physicians and six nurse practitioners (4). Another study showed that a model integrating dermatoscopic and macroscopic images with three patient metadata variables

outperformed models with just one image modality for binary and multiclass classification setting (106, 107).

Incremental learning

Current skin disease diagnostic models are static, wherein data distribution is already known and the target skin diseases are pre-set. However, in the clinical setting, as the database size grows over time, with the accumulation of new images, a shift in data distribution can occur, for example after the inclusion of new skin disease classes, or with improved or new devices. Changes or differences in image

acquisition tools, such as mobile phone cameras, also can shift dataset distribution by changing the quality of images captured. This results in the need to adapt models to new images while not degrading model performance on the pre-existing data. Incremental learning enables a model to continue learning the attributions of new data while preserving learned features from the data acquired before; successful incremental learning strategies on dermatology images have been recently reported (97, 108, 109).

Generative adversarial networks modeling

The ability to synthesize new data that closely resembles real skin lesion images can augment training on rare skin diseases and create a diverse and balanced dataset (110). While the potential to fill the data gaps is promising, models' performance does not show significant improvement when trained on synthesized data (111). The stylized images should be used cautiously, so as to not degrade the quality or reliability of the dataset and model by adding unintentional bias, and also ensure alignment with real-world conditions for clinical application (111, 112).

Emerging new model architectures—vision transformers

Vision transformer has emerged as an advanced model architecture, challenging the traditional dominance of convolutional neural networks (CNNs). CNNs have been the default choice for in both medical imaging and natural image tasks (113, 114). However, inspired by the success of Transformer in natural language processing (NLP), researchers have increasingly utilized ViTs or hybrid models of CNN and ViT and demonstrated promising results across various medical imaging tasks (115, 116). Concurrently, a resurgence of CNN is occurring with advanced CNN architectures such as ConveNeXt, showcasing competitive performance alongside Transformers in natural image task (117). These ongoing explorations and adaptations of ViTs address the challenges and uncertainties in deciding on model architecture.

Applications of large language model

Large language model is a type of natural language processing model that is trained to “understand” and generate human-like text, and has potential applications in enhancing clinical decision-making and overall patient care. For example, ChatGPT-style LLMs designed only for clinical diagnosis can accelerate clinical diagnoses by helping patients better understand their medical conditions and communicate with doctors remotely (118). Another application of LLM in clinic is AI-enabled digital scribes that can record and summarize patients visit information for treatment plans and billing purposes, eliminating the workload due to medical charting (119, 120). While there are positive aspects of LLM utilization for clinical care, there are also concerns such as the need for continued oversight of such models. It is essential to recognize that LLMs and doctors can complement each other, with LLM providing efficiency in processing large amounts of information

while doctors offer interpretation of the data, emotional intelligence and compassion to patients, thus improving patient care (121). However, caution should be used when utilizing LLM for medical advice. A recent study demonstrated that 4 LLM provided erroneous race-based responses to queries designed to detect race-based medical misapprehensions (122). To address this, testing of LLMs is critical before clinical implementation, and human feedback can help to correct errors.

Self-supervised learning

Self-supervised learning offers a promising approach to enhance the robustness and generalizability of models by enabling them to learn meaningful representations from unlabeled data. Traditionally, the efficacy of training deep learning models has relied on access to large-scale labeled datasets (123). However, in the medical field, acquiring such data is costly and requires specialized expertise. As a result, the scarcity of annotated data poses a significant obstacle to the development of robust models for various clinical settings. SSL addresses this challenge by developing a versatile model capable of efficiently adapting to new data distributions with a reduced number of labeled data during fine-tuning, while ensuring strong performance (124). Thus, SSL is a promising method to bridge the gap between AI research in the medical field and its clinical implementation.

Conclusion

Artificial intelligence currently is able to augment non-dermatologists' performance in a synergistic fashion and performs at the level of experienced dermatologists in a randomized controlled trial assessing skin malignancies. This achievement opens the door to aiding primary care physicians' discriminative triaging of patients to dermatologists and likely will decrease referrals for benign lesions, thereby freeing up dermatology practices to address true malignancies in a timely manner. Similarly, the potential for patients to self-refer for lesions concerning for malignancy may be possible in the near future, with models that can assess regional anatomic sites for lesions with concerning features. Through the implementation of AI, access to dermatologic care may become more democratic and accessible to the general population, including underserved subpopulations.

Limitations in performance include misdiagnosis by the model when assessing out of distribution diagnoses, leading clinicians astray; a solution might be for models to provide confidence estimates together with diagnostic predictions. A formidable problem in training models is the large number of diagnoses in dermatology, including numerous low incidence but aggressive malignancies (such as Merkel cell carcinoma, microcytic adnexal carcinoma, dermatofibrosarcoma tuberans, and angiosarcoma), or low incidence chronic malignancies such as cutaneous T cell lymphoma with potential for aggressive progression; one solution is federated training through the collaboration of multiple academic centers, some of which have specialty clinics focused on these diagnoses; or the formation of a central shared databank. In the future, models likely will be utilized to aid experienced dermatologists and dermatopathologists, as well as primary care providers and patients, particularly after training on multimodal datasets.

Author contributions

MW: Writing – review & editing, Writing – original draft, Supervision, Resources, Conceptualization. MT: Writing – review & editing, Writing – original draft. RT: Writing – review & editing, Writing – original draft, Visualization. AS: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Funding provided by Department of Defense grant W81XWH2110982 (MW, RT) and Department of Veterans Affairs grant 1I01HX003473 (MW, MT).

References

- 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) 542:115–8. doi: 10.1038/nature21056
- Young AT, Fernandez K, Pfau J, Reddy R, Cao NA, von Franque MY, et al. Stress testing reveals gaps in clinic readiness of image-based diagnostic artificial intelligence models. *NPJ Digit Med*. (2021) 4:10. doi: 10.1038/s41746-020-00380-6
- Young AT, Xiong M, Pfau J, Keiser MJ, Wei ML. Artificial intelligence in dermatology: a primer. *J Invest Dermatol*. (2020) 140:1504–12. doi: 10.1016/j.jid.2020.02.026
- Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. *Nat Med*. (2020) 26:900–8. doi: 10.1038/s41591-020-0842-3
- Höhn J, Krieghoff-Henning E, Jutz T, von Kalle C, Utikal JS, Meier F, et al. Combining CNN-based histologic whole slide image analysis and patient data to improve skin cancer classification. *Eur J Cancer*. (2021) 149:94–101. doi: 10.1016/j.ejca.2021.02.032
- Rotemberg V, Kurtansky N, Betz-Stablein B, Caffery L, Chousakos E, Codella N, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Sci Data*. (2021) 8:1–8. doi: 10.1038/s41597-021-00815-z
- Han SS, Moon JJ, Lim W, Suh IS, Lee SY, Na JI, et al. Keratinocytic skin Cancer detection on the face using region-based convolutional neural network. *JAMA Dermatol*. (2020) 156:29–37. doi: 10.1001/jamadermatol.2019.3807
- Soenksen LR, Kassiss T, Conover ST, Marti-Fuster B, Birkenfeld JS, Tucker-Schwartz J, et al. Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Sci Transl Med*. (2021) 13:eabb3652. doi: 10.1126/SCITRANSLMED.ABB3652
- Campanella G, Hanna MG, Geneslaw L, Mirafior A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med*. (2019) 25:1301–9. doi: 10.1038/s41591-019-0508-1
- Maron RC, Haggrenmüller S, von Kalle C, Utikal JS, Meier F, Gellrich FF, et al. Robustness of convolutional neural networks in recognition of pigmented skin lesions. *Eur J Cancer*. (2021) 145:81–91. doi: 10.1016/j.ejca.2020.11.020
- Maier K, Zaniolo L, Marques O. Image quality issues in teledermatology: a comparative analysis of artificial intelligence solutions. *J Am Acad Dermatol*. (2022) 87:240–2. doi: 10.1016/j.jaad.2021.07.073
- Han SS, Park I, Eun Chang S, Lim W, Kim MS, Park GH, et al. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *J Invest Dermatol*. (2020) 140:1753–61. doi: 10.1016/j.jid.2020.01.019
- Muñoz-López C, Ramírez-Cornejo C, Marchetti MA, Han SS, Del Barrio-Díaz P, Jaque A, et al. Performance of a deep neural network in teledermatology: a single-Centre prospective diagnostic study. *J Eur Acad Dermatol Venerol*. (2021) 35:546–53. doi: 10.1111/JDV.16979
- Agarwala S, Mata DA, Hafeez F. Accuracy of a convolutional neural network for dermatological diagnosis of tumours and skin lesions in a clinical setting. *Clin Exp Dermatol*. (2021) 46:1310–1. doi: 10.1111/CED.14688
- Xiong M, Pfau J, Young AT, Wei ML. Artificial intelligence in Teledermatology. *Curr Dermatol Rep*. (2019) 8:85–90. doi: 10.1007/s13671-019-0259-8
- Chin YPH, Hou ZY, Lee MY, Chu HM, Wang HH, Lin YT, et al. A patient-oriented, general-practitioner-level, deep-learning-based cutaneous pigmented lesion risk classifier on a smartphone. *Br J Dermatol*. (2020) 182:1498–500. doi: 10.1111/bjd.18859

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Navarro F, Escudero-Vinolo M, Bescos J. Accurate segmentation and registration of skin lesion images to evaluate lesion change. *IEEE J Biomed Health Inform*. (2019) 23:501–8. doi: 10.1109/JBHI.2018.2825251
- Webster DE, Suver C, Doerr M, Mounts E, Domenico L, Petrie T, et al. The mole mapper study, mobile phone skin imaging and melanoma risk data collected using ResearchKit. *Sci Data*. (2017) 4:1–8. doi: 10.1038/sdata.2017.5
- Kong FW, Horsham C, Ngoo A, Soyer HP, Janda M. Review of smartphone mobile applications for skin cancer detection: what are the changes in availability, functionality, and costs to users over time? *Int J Dermatol*. (2021) 60:289–308. doi: 10.1111/IJD.15132
- 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
- Matin RN, Dinnes J. AI-based smartphone apps for risk assessment of skin cancer need more evaluation and better regulation. *Br J Cancer*. (2021) 124:1749–50. doi: 10.1038/s41416-021-01302-3
- Jahn AS, Navarini AA, Cerminara SE, Kostner L, Huber SM, Kunz M, et al. Over-detection of melanoma-suspect lesions by a CE-certified smartphone app: performance in comparison to dermatologists, 2D and 3D convolutional neural networks in a prospective data set of 1204 pigmented skin lesions involving patients' perception. *Cancers (Basel)*. (2022) 14:3829. doi: 10.3390/cancers14153829
- Jones OT, Matin RN, van der Schaar M, Prathivadi Bhayankaram K, Ranmuthu CKI, Islam MS, et al. Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review. *Lancet Digit Health*. (2022) 4:e466–76. doi: 10.1016/S2589-7500(22)00023-1
- Vuong K, Armstrong BK, Weiderpass E, Lund E, Adami H-O, Veierod MB, et al. Development and external validation of a melanoma risk prediction model based on self-assessed risk factors. *JAMA Dermatol*. (2016) 152:889–96. doi: 10.1001/JAMADERMATOL.2016.0939
- Vuong K, Armstrong BK, Drummond M, Hopper JL, Barrett JH, Davies JR, et al. Development and external validation study of a melanoma risk prediction model incorporating clinically assessed naevi and solar lentigines. *Br J Dermatol*. (2020) 182:1262–8. doi: 10.1111/BJD.18411
- Olsen CM, Pandeya N, Thompson BS, Dusingize JC, Webb PM, Green AC, et al. Risk stratification for melanoma: models derived and validated in a purpose-designed prospective cohort. *J Natl Cancer Inst*. (2018) 110:1075–83. doi: 10.1093/jnci/djy023
- Fontanillas P, Alipanahi B, Furlotte NA, Johnson M, Wilson CH, Pitts SJ, et al. Disease risk scores for skin cancers. *Nature. Communications*. (2021) 12:1–13. doi: 10.1038/s41467-020-20246-5
- Roffman D, Hart G, Girardi M, Ko CJ, Deng J. Predicting non-melanoma skin cancer via a multi-parameterized artificial neural network. *Sci Rep*. (2018) 8:1–7. doi: 10.1038/s41598-018-19907-9
- Wang H-H, Wang Y-H, Liang C-W, Li Y-C. Assessment of deep learning using nonimaging information and sequential medical records to develop a prediction model for nonmelanoma skin Cancer. *JAMA Dermatol*. (2019) 155:1277–83. doi: 10.1001/JAMADERMATOL.2019.2335
- Huang C-W, Nguyen APA, Wu C-C, Yang H-C, Li Y-C. Develop a prediction model for nonmelanoma skin Cancer using deep learning in EHR data. *Stud Comput Intellig*. (2021) 899:11–8. doi: 10.1007/978-3-030-49536-7_2
- Bakshi A, Yan M, Riaz M, Polekhina G, Orchard SG, Tiller J, et al. Genomic risk score for melanoma in a prospective study of older individuals. *JNCI J Natl Cancer Inst*. (2021) 113:1379–85. doi: 10.1093/JNCI/DJAB076
- Kaiser I, Pfahlberg AB, Uter W, Hepp M, Veierod MB, Gefeller O. Risk prediction models for melanoma: a systematic review on the heterogeneity in model development

- and validation. *Int J Environ Res Public Health*. (2020) 17:7919. doi: 10.3390/IJERPH17217919
33. Sies K, Winkler JK, Fink C, Bardehle F, Toberer F, Buhl T, et al. Past and present of computer-assisted dermoscopic diagnosis: performance of a conventional image analyser versus a convolutional neural network in a prospective data set of 1,981 skin lesions. *Eur J Cancer*. (2020) 135:39–46. doi: 10.1016/j.ejca.2020.04.043
34. Abhishek K, Kawahara J, Hamarneh G. Predicting the clinical management of skin lesions using deep learning. *Sci Rep*. (2021) 11:7769–14. doi: 10.1038/s41598-021-87064-7
35. Winkler JK, Blum A, Kommos K, Enk A, Toberer F, Rosenberger A, et al. Assessment of diagnostic performance of dermatologists cooperating with a convolutional neural network in a prospective clinical study: human with machine. *JAMA Dermatol*. (2023) 159:621–7. doi: 10.1001/jamadermatol.2023.0905
36. Winkler JK, Tschandl P, Toberer F, Sies K, Fink C, Enk A, et al. Monitoring patients at risk for melanoma: May convolutional neural networks replace the strategy of sequential digital dermoscopy? *Eur J Cancer*. (2022) 160:180–8. doi: 10.1016/j.ejca.2021.10.030
37. Young AT, Vora NB, Cortez J, Tam A, Yeniy A, Afifi L, et al. The role of technology in melanoma screening and diagnosis. *Pigm Cell Melanoma Res*. (2020) 34:288–300. doi: 10.1111/pcmr.12907
38. Bozkurt A, Kose K, Coll-Font J, Alessi-Fox C, Brooks DH, Dy JG, et al. Skin strata delineation in reflectance confocal microscopy images using recurrent convolutional networks with attention. *Sci Rep*. (2021) 11:12576–11. doi: 10.1038/s41598-021-90328-x
39. Mehrabi JN, Baugh EG, Fast A, Lentsch G, Balu M, Lee BA, et al. A clinical perspective on the automated analysis of reflectance confocal microscopy in dermatology. *Lasers Surg Med*. (2021) 53:1011–9. doi: 10.1002/LSM.23376
40. Benjamins S, Dhunoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ Digit Med*. (2020) 3:118. doi: 10.1038/s41746-020-00324-0
41. The Medical Futurist FDA-approved A.I.-based algorithms. (2022). Available at: <https://medicalfuturist.com/fda-approved-ai-based-algorithms/> (Accessed November 7, 2022)
42. Haenssle HA, Fink C, Toberer F, Winkler J, Stolz W, Deinlein T, et al. Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. *Ann Oncol*. (2020) 31:137–43. doi: 10.1016/j.annonc.2019.10.013
43. MacLellan AN, Price EL, Publicover-Brouwer P, Matheson K, Ly TY, Pasternak S, et al. The use of noninvasive imaging techniques in the diagnosis of melanoma: a prospective diagnostic accuracy study. *J Am Acad Dermatol*. (2021) 85:353–9. doi: 10.1016/j.jaad.2020.04.019
44. Han SS, Kim YJ, Moon IJ, Jung JM, Lee MY, Lee WJ, et al. Evaluation of artificial intelligence-assisted diagnosis of skin neoplasms: a single-center, paralleled, unmasked, randomized controlled trial. *J Invest Dermatol*. (2022) 142:2353–2362.E2. doi: 10.1016/j.jid.2022.02.003
45. Onega T, Barnhill RL, Piepkorn MW, Longton GM, Elder DE, Weinstock MA, et al. Accuracy of digital pathologic analysis vs traditional microscopy in the interpretation of melanocytic lesions. *JAMA Dermatol*. (2018) 154:1159. doi: 10.1001/jamadermatol.2018.2388
46. Hekler A, Utikal JS, Enk AH, Berking C, Klode J, Schadendorf D, et al. Pathologist-level classification of histopathological melanoma images with deep neural networks. *Eur J Cancer*. (2019) 115:79–83. doi: 10.1016/j.ejca.2019.04.021
47. F de LoguUgolini F, Maio V, Simi S, Cossu A, Massi D, et al. Of cutaneous melanoma on digitized histopathological slides via artificial intelligence algorithm. *Front Oncol*. (2020) 10:1559. doi: 10.3389/FONC.2020.01559
48. Wang L, Ding L, Liu Z, Sun L, Chen L, Jia R, et al. Automated identification of malignancy in whole-slide pathological images: identification of eyelid malignant melanoma in gigapixel pathological slides using deep learning. *Br J Ophthalmol*. (2020) 104:318–23. doi: 10.1136/bjophthalmol-2018-313706
49. Ba W, Wang R, Yin G, Song Z, Zou J, Zhong C, et al. Diagnostic assessment of deep learning for melanocytic lesions using whole-slide pathological images. *Transl Oncol*. (2021) 14:101161. doi: 10.1016/j.TRANON.2021.101161
50. del Amor R, Launet L, Colomer A, Moscardó A, Mosquera-Zamudio A, Monteagudo C, et al. An attention-based weakly supervised framework for Spitzoid melanocytic lesion diagnosis in WSI. *Artif Intell Med*. (2021) 121:102197. doi: 10.1016/j.artmed.2021.102197
51. Olsson H, Kartasalo K, Mulliqi N, Capuccini M, Ruusuvaari P, Samarutunga H, et al. Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. *Nat Commun*. (2022) 13:7761. doi: 10.1038/s41467-022-34945-8
52. van Zon MCM, van der Waa JD, Veta M, Krekels GAM. Whole-slide margin control through deep learning in Mohs micrographic surgery for basal cell carcinoma. *Exp Dermatol*. (2021) 30:733–8. doi: 10.1111/EXD.14306
53. Kimeswenger S, Tschandl P, Noack P, Hofmarcher M, Rumetshofer E, Kindermann H, et al. Artificial neural networks and pathologists recognize basal cell carcinomas based on different histological patterns. *Mod Pathol*. (2020) 34:895–903. doi: 10.1038/s41379-020-00712-7
54. Duschner N, Bague DO, Schmidt M, Griewank KG, Hadaschik E, Hetzer S, et al. Applying an artificial intelligence deep learning approach to routine dermatopathological diagnosis of basal cell carcinoma. *J Dtsch Dermatol Ges*. (2023) 21:1329–37. doi: 10.1111/DDG.15180
55. Brinker TJ, Kiehl L, Schmitt M, Jutzi TB, Kriehoff-Henning EI, Krahel D, et al. Deep learning approach to predict sentinel lymph node status directly from routine histology of primary melanoma tumours. *Eur J Cancer*. (2021) 154:227–34. doi: 10.1016/j.ejca.2021.05.026
56. Kulkarni PM, Robinson EJ, Pradhan JS, Gartrell-Corrado RD, Rohr BR, Trager MH, et al. Deep learning based on standard H&E images of primary melanoma tumors identifies patients at risk for visceral recurrence and death. *Clin Cancer Res*. (2020) 26:1126–34. doi: 10.1158/1078-0432.CCR-19-1495
57. Polesie S, McKee PH, Gardner JM, Gillstedt M, Siarov J, Neittaanmäki N, et al. Attitudes toward artificial intelligence within dermatopathology: an international online survey. *Front Med*. (2020) 7:591952. doi: 10.3389/FMED.2020.591952
58. Johansson E, Månérjod F (2021). Segmentation and prediction of mutation status of malignant melanoma whole-slide images using deep learning.
59. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med*. (2018) 24:1559–67. doi: 10.1038/s41591-018-0177-5
60. Fu Y, Jung AW, Torne RV, Gonzalez S, Vöhringer H, Shmatko A, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat Can*. (2020) 1:800–10. doi: 10.1038/s43018-020-0085-8
61. Kather JN, Heij LR, Grabsch HI, Loeffler C, Echle A, Muti HS, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat Can*. (2020) 1:789–99. doi: 10.1038/s43018-020-0087-6
62. Torres R, Judson-Torres RL. Research techniques made simple: feature selection for biomarker discovery. *J Invest Dermatol*. (2019) 139:2068–2074.e1. doi: 10.1016/j.jid.2019.07.682
63. Gerami P, Yao Z, Polsky D, Jansen B, Busam K, Ho J, et al. Development and validation of a noninvasive 2-gene molecular assay for cutaneous melanoma. *J Am Acad Dermatol*. (2017) 76:114–120.e2. doi: 10.1016/j.jaad.2016.07.038
64. Jackson SR, Jansen B, Yao Z, Ferris LK. Risk stratification of severely dysplastic nevi by non-invasively obtained gene expression and mutation analyses. *SKIN J Cutan Med*. (2020) 4:124–9. doi: 10.25251/skin.4.2.5
65. Ludzik J, Lee C, Witkowski A. Potential limitations in the clinical adoption of 3-GEP pigmented lesion assay for melanoma triage by dermatologists and advanced practice practitioners. *Cureus*. (2022) 14:e31914. doi: 10.7759/cureus.31914
66. Estrada S, Shackleton J, Cleaver N, Depcik-Smith N, Cockerell C, Lencioni S, et al. Development and validation of a diagnostic 35-gene expression profile test for ambiguous or difficult-to-diagnose suspicious pigmented skin lesions. *SKIN J Cutan Med*. (2020) 4:506–22. doi: 10.25251/skin.4.6.3
67. Torres R, Lang UE, Hejna M, Shelton SJ, Joseph NM, Shain AH, et al. MicroRNA ratios distinguish melanomas from nevi. *J Invest Dermatol*. (2019) 140:164–173.E7. doi: 10.1016/j.jid.2019.06.126
68. Grossman D, Okwundu N, Bartlett EK, Marchetti MA, Othus M, Coit DG, et al. Prognostic gene expression profiling in cutaneous melanoma: identifying the knowledge gaps and assessing the clinical benefit. *JAMA Dermatol*. (2020) 156:1004–11. doi: 10.1001/JAMADERMATOL.2020.1729
69. Swetter SM, Thompson JA, Albertini MR, Barker CA, Baumgartner J, Boland G, et al. NCCN guidelines[®] insights: melanoma: cutaneous, version 2.2021: featured updates to the NCCN guidelines. *J Natl Compr Cancer Netw*. (2021) 19:364–76. doi: 10.6004/JNCCN.2021.0018
70. Jarell A, Gastman BR, Dillon LD, Hsueh EC, Podlipnik S, Covington KR, et al. Optimizing treatment approaches for patients with cutaneous melanoma by integrating clinical and pathologic features with the 31-gene expression profile test. *J Am Acad Dermatol*. (2022) 87:1312–20. doi: 10.1016/j.jaad.2022.06.1202
71. Varedi A, Gardner LJ, Kim CC, Chu EY, Ming ME, Leachman SA, et al. Use of new molecular tests for melanoma by pigmented-lesion experts. *J Am Acad Dermatol*. (2020) 82:245–7. doi: 10.1016/j.jaad.2019.08.022
72. Kashani-Sabet M, Leachman SA, Stein JA, Arbiser JL, Berry EG, Celebi JT, et al. Early detection and prognostic assessment of cutaneous melanoma. *JAMA Dermatol*. (2023) 159:545–53. doi: 10.1001/jamadermatol.2023.0127
73. Jones K, Lennon E, McCathie K, Millar A, Isles C, McFadyen A, et al. Teledermatology to reduce face-to-face appointments in general practice during the COVID-19 pandemic: a quality improvement project. *BMJ Open Qual*. (2022) 11:e001789. doi: 10.1136/BMJQO-2021-001789
74. Winkler JK, Sies K, Fink C, Toberer F, Enk A, Abassi MS, et al. Association between different scale bars in dermoscopic images and diagnostic performance of a market-approved deep learning convolutional neural network for melanoma recognition. *Eur J Cancer*. (2021) 145:146–54. doi: 10.1016/j.ejca.2020.12.010
75. Winkler JK, Fink C, Toberer F, Enk A, Deinlein T, Hofmann-Wellenhof R, et al. Association between surgical skin markings in Dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol*. (2019) 155:1135–41. doi: 10.1001/JAMADERMATOL.2019.1735
76. Katragadda C, Finnane A, Soyer HP, Marghoob AA, Halpern A, Malvey J, et al. Technique standards for skin lesion imaging: a Delphi consensus statement. *JAMA Dermatol*. (2017) 153:207–13. doi: 10.1001/JAMADERMATOL.2016.3949

77. Daneshjou R, Barata C, Betz-Stablein B, Celebi ME, Codella N, Combalia M, 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) 158:90–6. doi: 10.1001/JAMADERMATOL.2021.4915
78. Lester JC, Clark L, Linos E, Daneshjou R. Clinical photography in skin of colour: tips and best practices. *Br J Dermatol.* (2021) 184:1177–9. doi: 10.1111/BJD.19811
79. Jae HK, Soo HS, Young CK, Hyo HA. The influence of photography education on quality of medical photographs taken by dermatology resident. *Kor J Dermatol.* (2008) 46:1042–7.
80. Sun MD, Kentley J, Wilson BW, Soyer HP, Curiel-Lewandrowski CN, Rotemberg V, et al. Digital skin imaging applications, part I: assessment of image acquisition technique features. *Skin Res Technol.* (2022) 28:623–32. doi: 10.1111/SRT.13163
81. Vodrahalli K, Daneshjou R, Novoa RA, Chiou A, Ko JM, Zou J. TrueImage: a machine learning algorithm to improve the quality of telehealth photos. *Pac Symp Biocomput.* (2021) 26:220–31. doi: 10.1142/9789811232701_0021
82. Chen RJ, Wang JJ, Williamson DFK, Chen TY, Lipkova J, Lu MY, et al. Algorithm fairness in artificial intelligence for medicine and healthcare. *Nat Biomed Eng.* (2023) 7:719. doi: 10.1038/S41551-023-01056-8
83. Daneshjou R, Vodrahalli K, Novoa RA, Jenkins M, Liang W, Rotemberg V, et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci Adv.* (2022) 8:6147. doi: 10.1126/SCIADV.ABQ6147/SUPPL_FILE/SCIADV.ABQ6147_SM.PDF
84. ISIC (2018). ISIC-Archive. Available at: <https://www.isic-archive.com/#/topWithHeader/wideContentTop/main> (Accessed November 1, 2018).
85. Daneshjou R, Smith MP, Sun MD, Rotemberg V, Zou J. Lack of transparency and potential Bias in artificial intelligence data sets and algorithms: a scoping review. *JAMA Dermatol.* (2021) 157:1362–9. doi: 10.1001/JAMADERMATOL.2021.3129
86. Han SS, Moon IJ, Kim SH, Na J-I, 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:e1003381. doi: 10.1371/JOURNAL.PMED.1003381
87. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE, et al. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J Invest Dermatol.* (2018) 138:1529–38. doi: 10.1016/j.jid.2018.01.028
88. Brinker TJ, Hekler A, Hauschild A, Berking C, Schilling B, Enk AH, et al. Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark. *Eur J Cancer.* (2019) 111:30–7. doi: 10.1016/J.EJCA.2018.12.016
89. Van Beek MJ, Swerlick RA, Mathes B, Hruza GJ, Resneck J, Pak HS, et al. The 2020 annual report of DataDerm: the database of the American Academy of Dermatology. *J Am Acad Dermatol.* (2021) 84:1037–41. doi: 10.1016/j.jaad.2020.11.068
90. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science.* (2019) 363:1287–9. doi: 10.1126/science.aaw4399
91. Navarrete-Dechent C, Liopyris K, Marchetti MA. Multiclass artificial intelligence in dermatology: Progress but still room for improvement. *J Invest Dermatol.* (2021) 141:1325–8. doi: 10.1016/J.JID.2020.06.040
92. Lee G-H, Ko H-B, Lee S-W (2021). Joint dermatological lesion classification and confidence modeling with uncertainty estimation. arXiv [Preprint]. doi: 10.48550/arXiv.2107.08770
93. Kovarik C, Lee I, Ko J, Adamson A, Otley C, Kvedar J, et al. Commentary: position statement on augmented intelligence (AuI). *J Am Acad Dermatol.* (2019) 81:998–1000. doi: 10.1016/j.jaad.2019.06.032
94. Cortez JL, Vasquez J, Wei ML. The impact of demographics, socioeconomic, and health care access on melanoma outcomes. *J Am Acad Dermatol.* (2021) 84:1677–83. doi: 10.1016/J.JAAD.2020.07.125
95. Feng H, Berk-Krauss J, Feng PW, Stein JA. Comparison of dermatologist density between urban and rural counties in the United States. *JAMA Dermatol.* (2018) 154:1265–71. doi: 10.1001/jamadermatol.2018.3022
96. Ashrafzadeh S, Nambudiri VE. The COVID-19 Crisis: A Unique Opportunity to Expand Dermatology to Underserved Populations Mosby Inc. *J Am Acad Dermatol.* (2020) 83:e83–e84. doi: 10.1016/j.jaad.2020.04.154
97. Minagawa A, Koga H, Sano T, Matsunaga K, Teshima Y, Hamada A, et al. Dermoscopic diagnostic performance of Japanese dermatologists for skin tumors differs by patient origin: a deep learning convolutional neural network closes the gap. *J Dermatol.* (2021) 48:232–6. doi: 10.1111/1346-8138.15640
98. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human–computer collaboration for skin cancer recognition. *Nat Med.* (2020) 26:1229–34. doi: 10.1038/s41591-020-0942-0
99. Winkler JK, Sies K, Fink C, Toberer F, Enk A, Abassi MS, et al. Collective human intelligence outperforms artificial intelligence in a skin lesion classification task. *J Dtsch Dermatol Ges.* (2021) 19:1178–84. doi: 10.1111/DDG.14510
100. Felmingham CM, Adler NR, Ge Z, Morton RL, Janda M, Mar VJ. The importance of incorporating human factors in the design and implementation of artificial intelligence for skin Cancer diagnosis in the real world. *Am J Clin Dermatol.* (2020) 22:233–42. doi: 10.1007/S40257-020-00574-4
101. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lerner E, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Med.* (2021) 4:1–8. doi: 10.1038/s41746-021-00385-9
102. Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep.* (2020) 10:12598. doi: 10.1038/S41598-020-69250-1
103. McMahan B, Moore E, Ramage D, Hampson S, Arcas BAY (2017). Communication-efficient learning of deep networks from decentralized data. arXiv [Preprint]. doi: 10.48550/arXiv.1602.05629
104. Zhang DY, Kou Z, Wang D (2020). “FairFL: a fair federated learning approach to reducing demographic bias in privacy-sensitive classification models” in *Proceedings—2020 IEEE International Conference on Big Data, Big Data 2020*.
105. Lipkova J, Chen RJ, Chen B, Lu MY, Barbieri M, Shao D, et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell.* (2022) 40:1095. doi: 10.1016/J.CCELL.2022.09.012
106. Yap J, Yolland W, Tschandl P. Multimodal skin lesion classification using deep learning. *Exp Dermatol.* (2018) 27:1261–7. doi: 10.1111/EXD.13777
107. Berkowitz SJ, Kwan D, Cornish TC, Silver EL, Thullner KS, Aisen A, et al. Interactive multimedia reporting technical considerations: HIMSS-SIIM collaborative white paper. *J Digit Imaging.* (2022) 35:817–33. doi: 10.1007/S10278-022-00658-Z
108. Morgado AC, Andrade C, Teixeira LF, Vasconcelos MJM. Incremental learning for dermatological imaging modality classification. *J Imaging.* (2021) 7:180. doi: 10.3390/JIMAGING7090180
109. Gottumukkala VSSPR, Kumaran N, Sekhar VC. BLSNet: skin lesion detection and classification using broad learning system with incremental learning algorithm. *Expert Syst.* (2022) 39:e12938. doi: 10.1111/exsy.12938
110. Bissoto A, Perez F, Valle E, Avila S. Skin lesion synthesis with generative adversarial networks. *Lect Notes Comput Sci.* (2018) 11041. doi: 10.1007/978-3-030-01201-4_32
111. Carrasco Limeros S, Majchrowska S, Zoubi MK, Rosén A, Suvilehto J, Sjöblom L, et al. (2022). “Assessing GAN-Based Generative Modeling on Skin Lesions Images” in *MIDI 2022: Digital Interaction and Machine Intelligence*. pp. 93–102. doi: 10.1007/978-3-031-37649-8_10
112. Salvi M, Branciforti F, Veronese F, Zavattaro E, Tarantino V, Savoia P, et al. DermoCC-GAN: a new approach for standardizing dermatological images using generative adversarial networks. *Comput Methods Prog Biomed.* (2022) 225:107040. doi: 10.1016/j.cmpb.2022.107040
113. Esteve A, Chou K, Yeung S, Naik N, Madani A, Mottaghi A, et al. Deep learning-enabled medical computer vision. *NPJ Digit Med.* (2021) 4:5. doi: 10.1038/s41746-020-00376-2
114. Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, et al. Recent advances in convolutional neural networks. *Pattern Recogn.* (2018) 77:354–77. doi: 10.1016/J.PATCOG.2017.10.013
115. Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, et al. Transformers in medical imaging: a survey. *Med Image Anal.* (2023) 88:102802. doi: 10.1016/j.media.2023.102802
116. Khan S, Ali H, Shah Z. Identifying the role of vision transformer for skin cancer—a scoping review. *Front Artif Intell.* (2023) 6:1202990. doi: 10.3389/FRAI.2023.1202990/BIBTEX
117. Liu Z, Mao H, Wu C-Y, Feichtenhofer C, Darrell T, Xie S, et al. A ConvNet for the 2020s. Available at: <https://github.com/facebookresearch/ConvNeXt> (Accessed February 22, 2024).
118. Zhou J, He X, Sun L, Xu J, Chen X, Chu Y, et al. (2023). Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. medRxiv [Preprint]. doi: 10.1101/2023.06.10.23291127
119. Krishna K, Khosla S, Bigham J, Lipton ZC (2021). “Generating SOAP notes from doctor-patient conversations using modular summarization techniques” in *ACL-IJCNLP 2021—59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*. 4958–4972.
120. May A.I. (2023). Someday Work Medical Miracles. For Now, It Helps Do Paperwork. The New York Times. Available at: <https://www.nytimes.com/2023/06/26/technology/ai-health-care-documentation.html> (Accessed February 22, 2024).
121. Matin RN, Linos E, Rajan N. Leveraging large language models in dermatology. *Br J Dermatol.* (2023) 189:253–4. doi: 10.1093/BJD/LJAD230
122. Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. *NPJ Digit Med.* (2023) 6:195. doi: 10.1038/s41746-023-00939-z
123. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2010). “ImageNet: A large-scale hierarchical image database” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.
124. Azizi S, Culp L, Freyberg J, Mustafa B, Baur S, Kornblith S, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nat Biomed Eng.* (2023) 7:756–79. doi: 10.1038/s41551-023-01049-7



OPEN ACCESS

EDITED BY

Robert Gniadecki,
University of Alberta, Canada

REVIEWED BY

Amanda Oakley,
Waikato District Health Board, New Zealand
Mara Giavina-Bianchi,
Albert Einstein Israelite Hospital, Brazil

*CORRESPONDENCE

Helen Marsden
✉ helen@skinanalytics.co.uk

RECEIVED 26 September 2023

ACCEPTED 27 February 2024

PUBLISHED 22 March 2024

CITATION

Marsden H, Kemos P, Venzi M, Noy M,
Maheswaran S, Francis N, Hyde C,
Mullarkey D, Kalsi D and Thomas L (2024)
Accuracy of an artificial intelligence as a
medical device as part of a UK-based skin
cancer teledermatology service.
Front. Med. 11:1302363.
doi: 10.3389/fmed.2024.1302363

COPYRIGHT

© 2024 Marsden, Kemos, Venzi, Noy,
Maheswaran, Francis, Hyde, Mullarkey, Kalsi
and Thomas. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Accuracy of an artificial intelligence as a medical device as part of a UK-based skin cancer teledermatology service

Helen Marsden^{1*}, Polychronis Kemos², Marcello Venzi¹,
Mariana Noy³, Shameera Maheswaran³, Nicholas Francis⁴,
Christopher Hyde⁵, Daniel Mullarkey¹, Dilraj Kalsi¹ and
Lucy Thomas³

¹Skin Analytics Ltd., London, United Kingdom, ²Blizard Institute, The Faculty of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom, ³Chelsea and Westminster Hospital NHS Foundation Trust, London, United Kingdom, ⁴Imperial College Healthcare NHS Trust, St Mary's Hospital, London, United Kingdom, ⁵Exeter Test Group, Department of Health and Community Sciences, University of Exeter Medical School, Exeter, United Kingdom

Introduction: An artificial intelligence as a medical device (AlaMD), built on convolutional neural networks, has demonstrated high sensitivity for melanoma. To be of clinical value, it needs to safely reduce referral rates. The primary objective of this study was to demonstrate that the AlaMD had a higher rate of correctly classifying lesions that did not need to be referred for biopsy or urgent face-to-face dermatologist review, compared to teledermatology standard of care (SoC), while achieving the same sensitivity to detect malignancy. Secondary endpoints included the sensitivity, specificity, positive and negative predictive values, and number needed to biopsy to identify one case of melanoma or squamous cell carcinoma (SCC) by both the AlaMD and SoC.

Methods: This prospective, single-centre, single-arm, masked, non-inferiority, adaptive, group sequential design trial recruited patients referred to a teledermatology cancer pathway ([clinicaltrials.gov](#) NCT04123678). Additional dermoscopic images of each suspicious lesion were taken using a smartphone with a dermoscopic lens attachment. The images were assessed independently by a consultant dermatologist and the AlaMD. The outputs were compared with the final histological or clinical diagnosis.

Results: A total of 700 patients with 867 lesions were recruited, of which 622 participants with 789 lesions were included in the per-protocol (PP) population. In total, 63.3% of PP participants were female; 89.0% identified as white, and the median age was 51 (range 18–95); and all Fitzpatrick skin types were represented including 25/622 (4.0%) type IV–VI skin. A total of 67 malignant lesions were identified, including 8 diagnosed as melanoma. The AlaMD sensitivity was set at 91 and 92.5%, to match the literature-defined clinician sensitivity (91.46%) as closely as possible. In both settings, the AlaMD identified had a significantly higher rate of identifying lesions that did not need a biopsy or urgent referral compared to SoC (p -value = 0.001) with comparable sensitivity for skin cancer.

Discussion: The AlaMD identified significantly more lesions that did not need to be referred for biopsy or urgent face-to-face dermatologist review, compared to teledermatologists. This has the potential to reduce the burden of unnecessary referrals when used as part of a teledermatology service.

KEYWORDS

artificial intelligence, skin cancer, deep ensemble for the recognition of malignancy (DERM), teledermatology, AI as a medical device, skin analytics

Introduction

The global burden of skin cancer is growing, but healthcare systems lack the necessary capacity, especially in the aftermath of the COVID-19 pandemic. Skin cancers, primarily melanoma, squamous cell carcinoma (SCC), and basal cell carcinoma (BCC), are the most common cancers worldwide. In the United States, 9,500 people are diagnosed daily with annual treatment costs of \$8.1bn (1). Skin cancer accounts for half of all cancers diagnosed in England and Wales and is increasing by 8% annually (2). However, of over 500,000 urgent referrals made to UK Secondary Care in 2019/20, only 6.5% resulted in a skin cancer diagnosis. Moreover, 25% of melanoma are found in non-urgent dermatology referrals (3) and diagnostic delays of 2 weeks or more can lead to a 20% decrease in 5-year survival rates (4). With approximately one in four UK Consultant Dermatologist posts unfilled (2), the situation is unsustainable.

A novel AI as a medical device (AIaMD), built on convolutional neural networks, has previously demonstrated high sensitivity for melanoma, similar to the level of skin cancer specialists (5). Trained using machine learning to recognise the most common malignant, premalignant, and benign skin lesions, the AIaMD analyses a dermoscopic image of a skin lesion and returns a suspected diagnosis of melanoma, SCC, BCC, Bowen's disease/intraepidermal carcinoma (IEC), actinic keratosis (AK), atypical nevus (AN), or benign (labels of individual benign conditions are possible, but as the patient management is often the same, they are grouped into one output), along with a corresponding referral recommendation. The AIaMD applies a risk-based hierarchy so that the most serious potential diagnosis is returned. For example, if the AIaMD identifies a lesion as potentially either a BCC or melanoma, it will return a classification of melanoma.

The AIaMD is the key component of the Skin Analytics' medical device deep ensemble for the recognition of malignancy (DERM), which is intended for use in the screening, triage, and assessment of skin lesions suspicious for skin cancer. DERM is deployed in the United Kingdom National Health Service (NHS) to support skin cancer diagnosis pathways that have assessed over 81,000 patients since 2020. After a period of use as a Class I device for clinical decision support, during which time this study was conducted, DERM received UKCA Class IIa approval in April 2022, allowing it to be used for autonomous decision-making, to further optimise the urgent referral pathways. To be of clinical value, the AIaMD needs to achieve a high specificity for premalignant and benign lesions as well as a high sensitivity for skin cancer. This study compared the rate and accuracy of the AIaMD and teledermatology in identifying premalignant and benign lesions that do not require biopsy or urgent referral while maintaining a high sensitivity for malignancy.

Materials and methods

Study design

This prospective, single-centre, single-arm, masked, non-inferiority design trial (the "Impact study"), with an adaptive group sequential design, was conducted at Chelsea and Westminster NHS Foundation Trust between February 2020 and August 2021. Chelsea and Westminster, which serves a population of 620,000 that has a demographic profile comparable with the London average (6), established an urgent skin cancer teledermatology service in 2017 (7) where patients with suspicious skin lesions can be referred from primary care.

The primary objective of this study was to demonstrate that the AIaMD had a higher rate of correctly classifying premalignant and benign lesions as not needing to be referred for biopsy or urgent face-to-face review compared to teledermatology standard of care (SoC) while achieving the same sensitivity to detect malignancy. Secondary endpoints included the sensitivity, specificity, positive and negative predictive values, number needed to biopsy for malignancy, and number needed to refer for premalignancy (IEC and AK) of the AIaMD and SoC. These performance data were used to conduct a simple cost impact assessment, based on the assumptions that teledermatology reviews cost £115.44 and require 10 min of specialist time per case on average; face-to-face assessments cost £163.41 and require 15 min per case; and biopsies cost £257.43 and require on average 32.5 min per lesion (based on a 50:50 split of excision biopsies which are booked for 45 min and incisional/punch biopsies which are booked for 20 min) (8, 9). This will be used to inform future health economic assessments. Surveys were conducted on patients' perspectives on AIaMD use in their care and are reported in another publication (10). The study was registered on clinicaltrials.gov (NCT04123678) and was approved by the West Midlands-Edgbaston Research Ethics Committee and UK Health Research Authority on 23 December 2019.

Participants

Patients aged 18 or over with at least one suspicious lesion being photographed as part of SoC were invited to consent to the study in a consecutive series. Patients who returned to the teledermatology service with the same or different lesions were able to re-consent to the study. To be eligible for inclusion in the study, lesions needed to be less than 15 mm in diameter (so as to fit within the dermatoscope lens); in an anatomical location suitable for photography (avoiding genital, hair-bearing, mucosal sites, and subungual sites), have no previous trauma including biopsy or excision; and have no visible scarring or tattooing.

Procedures

In the teledermatology service, digital single-lens reflex (DSLR) and dermatoscope images of each suspicious lesion are taken by medical photographers. These images are reviewed remotely, alongside the primary care referral letter and patient-reported medical history, by consultant dermatologists who record a suspected diagnosis, and triage the lesion(s) for surgery, further assessment, or discharge.

Patients who consented to the study had an additional macroscopic and dermoscopic image of each lesion taken, using an iPhone XR smartphone and DermLite DL1 basic dermoscopic lens attachment, by a healthcare assistant (HCA). The suspected diagnosis and management decision recorded by the teledermatologist, and any subsequent patient review by skin cancer specialists, were collected, along with relevant medical history, patients' levels of concern, healthcare resource utilisation data (number of appointments, time required to take images), and histopathology results where biopsies were undertaken. The iPhone XR dermoscopic images were used for AIaMD assessment, while the teledermatology review was conducted utilising DSLR images in accordance with the established SoC at Chelsea and Westminster. Patients completed all study-related activities in one visit, but the AIaMD image analysis occurred outside of the study, so clinicians were blinded to its output, and patient care was unaffected. Dermoscopic images were first quality-checked using an AI tool that assesses whether an image is dermoscopic, blurry, or dark, and rejected images were excluded from the AIaMD assessment.

Statistics and analysis

Based on the reported prevalence of and dermatologist sensitivity for melanoma, SCC, and BCC (5, 11, 12), it was estimated that dermatologists would correctly identify 91.46% of skin cancers. The AIaMD settings were optimised to match this as closely as possible with the aim of achieving a difference of <0.2%. The closest AIaMD settings that could be achieved were 91% (AIaMD-A) and 92.5% (AIaMD-B). As both options were >0.2% of the estimated dermatologist sensitivity, both settings were used for the primary endpoint. For the secondary endpoints, AIaMD-A was used as it was closer to the estimated clinical sensitivity.

The expected specificity of the AIaMD to identify malignancies was 54%, and the expected prevalence rates for MM, SCC, and BCC were 4.12, 5.16, and 21.39%, respectively. To demonstrate that the specificity of the AIaMD was not inferior to the specificity of SoC, using a 1% non-inferiority margin and with 99% power, a sample size of 634 lesions was needed. Assuming 1.2 lesions per patient and allowing for a 10% dropout rate, the sample size required was estimated to be 581 patients.

An interim analysis was conducted when the first third of data had been collected, to allow data-driven sample size reassessments. The primary endpoint was analysed using a one-sided, 2-proportion Z-test, with an overall alpha of 0.05. The final analysis was performed by combining the *p*-values from both phases of the study, using the procedure described by Lehman and Wassmer. The *p*-values of the test statistic from both phases of the study were therefore combined using specific predefined weights set as 0.577 and 0.82 for phases 1 and 2, respectively (13). The one-sided significance level was adjusted to 0.0246 for the final analysis based on the O'Brien-Fleming approach

(14). Statistical estimates of accuracy are reported with 95% confidence intervals (CIs). Statistical analysis was conducted using R language version 4.1.3 (The R Project for Statistical Computing).

The suspected diagnosis and management outcome from the AIaMD and teledermatology were compared with a histologically confirmed diagnosis, where obtained, and failing that, consultant dermatologist diagnosis and management with a second opinion where available. Only histopathological diagnosis was accepted for melanoma, SCC, or BCC diagnosis, with a second review for melanoma. Final opinions on clinical diagnosis were provided by authors MN and LT, both consultant dermatologists, who also checked histopathology reports of all biopsied lesions and confirmed that no cases of rare skin cancer were identified. Patients and lesions that did not meet the inclusion criteria were excluded from the intention-to-treat population (ITT), as were those lesions without a final diagnosis available. Lesions with no AIaMD result available (missing dermoscopic images, and/or where these failed the image quality assessment) were excluded from the per-protocol (PP) population. The specificity of AIaMD was defined as the percentage of lesions diagnosed as IEC, AK, AN, or benign that were labelled as IEC, AK, AN, or benign by the AIaMD. The specificity of dermatologists was defined as the percentage of lesions diagnosed as IEC, AK, AN, or benign that the teledermatologist referred for a routine dermatologist appointment or discharged.

The COVID-19 pandemic began after the study had commenced recruitment and led to the reassessment, often downgrading, of patient management decisions. This was captured; however, the primary analysis is based on the original patient management decisions by the dermatologists. For the secondary analysis, diagnostic accuracy indices (sensitivity, specificity, predictive values) were calculated by evaluating the performance on lesions grouped as malignant vs. premalignant/benign. For instance, an SCC labelled as a melanoma would count as a true positive in the sensitivity calculation for both the AIaMD and clinical (SoC) diagnosis.

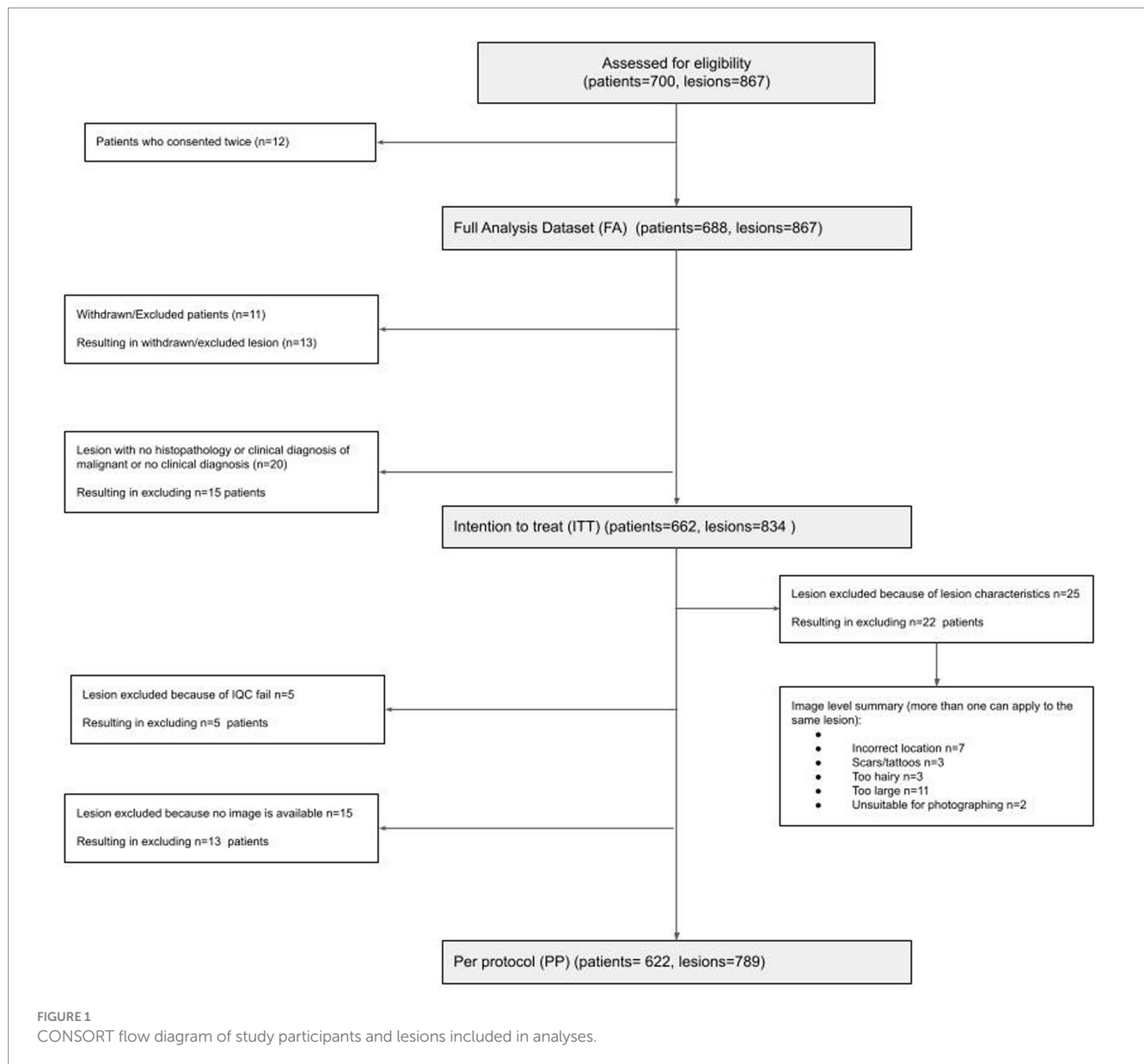
Real-world settings

For comparative purposes, a *post-hoc* analysis was conducted using the same version of the AIaMD with threshold settings that were used in live deployments at the time of the study analysis. These targeted a higher sensitivity of >95% for melanoma and SCC and >90% for BCC (AIaMD-RWS). Accuracy metrics including sensitivity and specificity were calculated using these settings, and the results are presented to provide more insight as to the impact of AIaMD if it had been a real-world deployment.

Results

Patient and lesion populations

A total of 688 participants (12 re-consented so 700 attendances) presenting with 867 lesions (average 1.3 lesions per patient) were recruited; 662 participants with 834 lesions were included in the intention-to-treat (ITT) population; and 622 participants with 789 lesions were included in the per-protocol (PP) population (Figure 1). In the PP population, 63.3% of participants were female; 89.0%



identified as white, and the median age was 51 (range 18–95); and all Fitzpatrick skin types were represented including 25/622 (4.0%) type IV–VI skin (Table 1).

Most lesions were located on the face and scalp (25%), back (18.6%), arms (13.2%), and legs (19%) and had a history of change in the previous 3 months (86.3%). Lesions averaged 6.3 mm (range 0.5–15 mm) in diameter, and patients were most often (68.1%) a little concerned about their lesions (Table 2).

Sixty-seven malignant lesions were identified in the PP population: 8 melanoma, 13 SCCs, and 46 BCCs. Most melanomas were superficial spreading ($N=4$) and <1 mm thick ($N=7$). Most SCCs were well or moderately differentiated ($N=9$), while most BCCs were nodular ($N=22$) (Table 3). Three additional lesions diagnosed as melanoma and four lesions diagnosed as SCCs had been included in the study but were ineligible because no images were available (1x melanoma, 1x SCC), the lesion was located on a scar (1x melanoma), or the lesions were larger than the dermoscopic lens (1x melanoma, 3x SCC).

Primary outcome

The interim analysis of phase 1 included 199 lesions (21 malignant and 178 premalignant or benign). AIaMD-A correctly identified 77.5% of the premalignant and benign lesions (138/178, 95% CI 70.6–83.3%) as lesion types that did not need a biopsy or urgent face-to-face assessment and AIaMD-B identified 74.7% (133/178, 95% CI 67.6–80.8%) compared to 73.6% (131/178, 95% CI 66.4–79.8%) by SoC. The interim analysis of the primary endpoint confirmed the non-futility of the study; however, the required sample size increased to 700 patients, to achieve a statistical power of 95%. In phase 2, there were 590 lesions (46 malignant and 544 premalignant or benign). AIaMD-A correctly identified 85.1% of the premalignant and benign lesions (463 out of 544, 95% CI 81.8–87.9%) as lesions types that did not need a biopsy or urgent face-to-face assessment, and AIaMD-B identified 81.6% (444 out of 544, 95% CI 78.8–84.7%), compared to 71.3% by SoC (388 out of 544, 95% CI 67.3–75.1%). After weighing the two phases across the whole study as described in the Statistics and

TABLE 1 Breakdown of the per-protocol patient population by age group, sex, ethnic group, Fitzpatrick skin type, and past personal history of skin cancer.

		<i>N</i>	%
Total number of patients		622	100
Age group	Mean	51.5	
	Standard deviation	19.6	
	Minimum	18	
	Maximum	95	
Sex	Female	394	63.3
	Male	228	36.7
Ethnic group	White	555	89.2
	Asian	14	2.3
	Black	8	1.3
	Other	10	1.6
	Mixed	22	3.5
	Unknown	13	2.1
Fitzpatrick skin type	I	203	32.6
	II	301	48.4
	III	93	15
	IV	12	1.9
	V-VI	13	2.1
Past personal history of skin cancer	None	490	78.8
	Melanoma	24	3.9
	SCC	10	1.6
	BCC	44	7.1
	Other	43	6.9
	Unknown	11	1.8

Analysis methods, AIaMD-A and AIaMD-B had a significantly higher rate of correctly identifying premalignant and benign lesions as lesions that did not need a biopsy or urgent face-to-face assessment compared to SoC (p -value<0.0246).

Secondary outcomes

The sensitivity, positive and negative predictive values, and false negative and positive rates, of the teledermatologists and AIaMD to identify malignant lesions, were calculated (Table 4).

Of the 8 histology-diagnosed melanomas, seven were sent for urgent biopsy and one was referred to BCC/Mohs clinic by the teledermatologist. Seven were labelled as melanoma by both SoC and AIaMD, while the other melanoma was thought to be a traumatised angioma by SoC and was classified as benign by AIaMD. Of the 13 histology-confirmed SCCs, all 13 were sent to urgent biopsy or urgent face-to-face dermatologist appointment by SoC, 9 with a suspected diagnosis of SCC; and 12 were labelled SCC and 1 was labelled BCC by AIaMD. Of the 46 histology-confirmed BCCs, 43 were sent for biopsy or were referred to BCC/Mohs clinic by the teledermatologist, while 2 lesions were referred to routine face-to-face dermatology; 38

had suspected diagnoses from teledermatology of melanoma or BCC, while the remaining 8 lesions were referred with a suspected premalignant or benign diagnosis; and 31 were labelled as BCC, 11 as melanoma or SCC, and 4 as premalignant or benign by the AIaMD (Figure 2).

In total, 216 lesions were referred directly from teledermatology SoC to urgent or non-urgent biopsy. The number needed to biopsy (NNB) for SoC to diagnose one malignancy was 4.2 (216/51, 95% CI 3.3–5.5). If all lesions classified as malignant by the AIaMD were biopsied, the NNB was 3 (182/61, 95% CI 2.4–3.7) (Table 4).

A total of 268 lesions were referred from teledermatology SoC to biopsy or urgent face-to-face assessment. The number needed to refer (NNR) for SoC to diagnose one case of IEC or AK was 8.6 (268/31, 95% CI 6.2–12.3). If all lesions classified as malignant or premalignant by the AIaMD were referred, the NNR for IEC and AK was 4.5 (249/55, 95% CI 3.6–5.8) (Table 4).

SoC required 688 teledermatology patient reviews, 221 face-to-face assessments, and up to 299 lesion biopsies (240 biopsies were conducted with the missing biopsies mainly due to delays from ongoing pressures following the COVID-19 pandemic meaning histopathology reports were not available within the study data collection window). If lesions had been triaged in accordance with the AIaMD output, 454 patient reviews would not have been required on the skin cancer pathway, 141 face-to-face assessments would have been avoided, and 124 fewer lesions would have been biopsied. This equates to cost savings of £52,409.76 in teledermatology reviews, £23,040.81 in face-to-face assessments, and £31,921.32 in biopsies. In terms of specialist time, this would save 76 h in teledermatology reviews, 35 h of face-to-face appointments, and 67 h of biopsies. In total, this amounts to a cost impact of £107,371.89 and 178 specialist h saved. Extrapolated to per 1,000 patients entering the pathway, this would scale to £156,063.79 and 259 specialist hours saved.

Out of 867 lesions included in the study, 843 (97.2%) had dermoscopic images successfully captured, and 24 lesions could not be imaged dermoscopically using the iPhone X. In total, 837 dermoscopic images (99.3% of those captured) passed the image quality check, and it took the HCA an average of 1 min to capture the study images. No adverse events were reported in the study.

Post-hoc analysis of real-world settings

The sensitivity, positive and negative predictive values, and false negative and positive rates, of the AIaMD-RWS, were also calculated (Table 5).

Of the eight histology-diagnosed melanoma, the RWS-AIaMD correctly identified all eight as melanoma. Of the 13 histology-confirmed SCCs, 11 were correctly labelled as SCC by the RWS-AIaMD, with the remaining 2 classified as melanoma. Of the 46 histology-confirmed BCCs, 21 were labelled as BCC by the RWS-AIaMD, 21 as melanoma or SCC, and 4 as benign (Figure 3).

If all lesions classified as malignant by the AIaMD-RWS were biopsied, the NNB was 4.1 (256/63, 95% CI 3.3–5.1). If all lesions classified as malignant or premalignant by the AIaMD were referred, the NNR for IEC and AK was 5.2 (300/58, 95% CI 4.1–6.6).

TABLE 2 Breakdown of per-protocol lesions by size in millimetres, body location, patient concern, and history of change.

		N	%
Lesion size (mm)	Mean	6.3	
	Standard deviation	3	
	Minimum	0.5	
	Maximum	15	
Lesion location	Face and scalp	197	25
	Neck	34	4.3
	Right arm	48	6.1
	Left arm	56	7.1
	Right palm	3	0.4
	Left palm	1	0.1
	Anterior chest	94	11.9
	Abdomen	54	6.8
	Posterior chest	3	0.4
	Back	147	18.6
	External genitals	1	0.1
	Right leg	72	9.1
	Right sole	0	0
	Left leg	78	9.9
	Left sole	1	0.1
Patient concern	Not concerned	94	11.9
	A little concerned	537	68.1
	Very concerned	144	18.3
	Unknown	14	1.8
Lesion change	None	108	13.7
	Changed colour	53	6.7
	More symptomatic	408	51.7
	New lesion	8	1
	Grown a bit	212	26.9
	Grown a lot	0	0

Discussion

This study demonstrates a high specificity for skin cancer of the AIaMD with a significantly lower rate of premalignant and benign lesion referral for biopsy or urgent face-to-face dermatologist review compared to SoC. AIaMD, therefore, shows potential to improve healthcare resource utilisation (HRU), which will be the subject of further health economic analyses utilising the data from this study. Assuming premalignant or benign AIaMD outputs meant that no further patient management on the urgent suspected cancer pathway was required, there could have been savings of >£100,000 and >150 h of specialist time. There are, however, many other costs and benefits, as well as the potential need to expedite treatment for non-cancerous dermatological conditions, that should be considered when conducting health economic modelling.

While the high specificity of the AIaMD has the potential to improve HRU, this does raise the key question of the possible risk of

a trade-off in sensitivity. The study-specific settings used by the AIaMD were determined to match an expected sensitivity by clinicians of 91.46%, which had been determined by a review of the literature on clinician sensitivity for skin cancer detection (5, 11, 12). The sensitivity achieved by teledermatologists in this study was higher than expected and higher than the sensitivity of the AIaMD to identify malignancies, either when used with the study-specific settings or the settings optimised for live deployment. This may be because the study was carried out at a centre with a well-established teledermatology service and experienced teledermoscopy clinicians, which is unlikely to be representative of UK dermatology more widely, as many centres have yet to implement urgent cancer teledermatology pathways. It is also important to note that the malignant lesions that the AIaMD missed were mostly BCC lesions. One lesion diagnosed as melanoma was classified as benign by the AIaMD with the study-specific settings but correctly identified when the live-deployment settings (AIaMD-RWS) were used, which also had a benign suspected clinical diagnosis by teledermatology, indicating the lesion was difficult to diagnose without a biopsy and that the AIaMD-RWS would have expedited treatment for the melanoma over and above SoC.

Furthermore, patient management being determined by teledermatology means that there was a risk of validation bias towards the outcome that validates the teledermatologist management plan (15). There were 25 lesions discharged by teledermatology but classified as malignant by the AIaMD (35 with RWS-AIaMD) that were not followed up, due to the length of time between patient recruitment into the study and image analysis by the AIaMD. This means there may have been malignant lesions that the AIaMD identified but the teledermatologist discharged. Of the patients who presented at the teledermatology service, and consented to the study, twice, three had a different clinical diagnosis at the second assessment, and two patients (four lesions) were subsequently biopsied. In all but one of the lesions reviewed twice, the AIaMD output was benign for the first assessment and only changed for the four lesions subsequently biopsied, indicating that the AIaMD was picking up similar features in the second assessment that prompted the clinicians to refer for a biopsy. Though beyond the scope of this study, the potential of malignant lesions missed by SoC, but identified by the AIaMD, should be considered for future research, as should the impact of changes in a lesion on the AIaMD classification.

This study builds on previous studies, which found that the AIaMD component of DERM can detect melanoma and non-melanoma skin cancer with accuracy comparable to specialists (5, 16–18), by looking at its accuracy to detect premalignant or benign lesions. The Melanoma Image Analysis Algorithm (MIAA) study evaluated the AIaMD on lesions that dermatologists referred for biopsy or were obviously benign (5), missing out those that GPs were concerned about but a dermatologist could diagnose and manage without a biopsy. While the non-melanoma skin cancer (NMSC) study included suspicious skin lesions that were not referred for a biopsy, the lesions were all assessed by a dermatologist prior to inclusion in the study, again missing lesions that look suspicious to a non-skin cancer specialist, but which a dermatologist is less concerned about (17, 18). The lesion population in this study was primarily based on lesions that had not first been evaluated by a dermatologist and is therefore more representative of the population that the AIaMD may be used on in primary care.

TABLE 3 Breakdown of lesion diagnosis from histology or clinical diagnosis in the per-protocol population, including subtypes of malignant lesions, Breslow thickness of melanoma, and staging of squamous cell carcinoma and basal cell carcinoma.

			Clinical diagnosis	Histopathology
Total number of lesions			789	240
Malignant	(Suspect) melanoma		26	8
	Subtype	<i>In situ</i>		2
		Lentigo maligna		2
		Superficial spreading		4
		Nodular		0
	Breslow thickness	<i>In situ</i>		4
		<1.0 mm		3
		1.01–2.0 mm		1
		>2.0 mm		0
	(Suspect) squamous cell carcinoma		41	13
	Subtype	Well differentiated		4
		Moderately differentiated		5
		Poorly differentiated		0
		Unknown		4
	Stage	T1		6
		T2		1
		T3		0
		T4		1
		Unknown		5
	(Suspect) basal cell carcinoma		51	46
	Stage	Tis		1
		T1		20
		Unknown		25
	Subtype	Superficial		6
		Nodular		21
		Infiltrative		6
		Micronodular		1
		Unknown		12
Premalignant or benign	IEC/SCC <i>in situ</i>		21	7
	Actinic keratosis		35	25
	Atypical/dysplastic nevus		29	10
	Seborrheic keratosis		168	26
	Dermatofibroma		16	3
	Vascular lesion		13	4
	Lentigo		14	3
	Benign melanocytic Nevus		259	39
	Other (Benign)/unknown		116	56

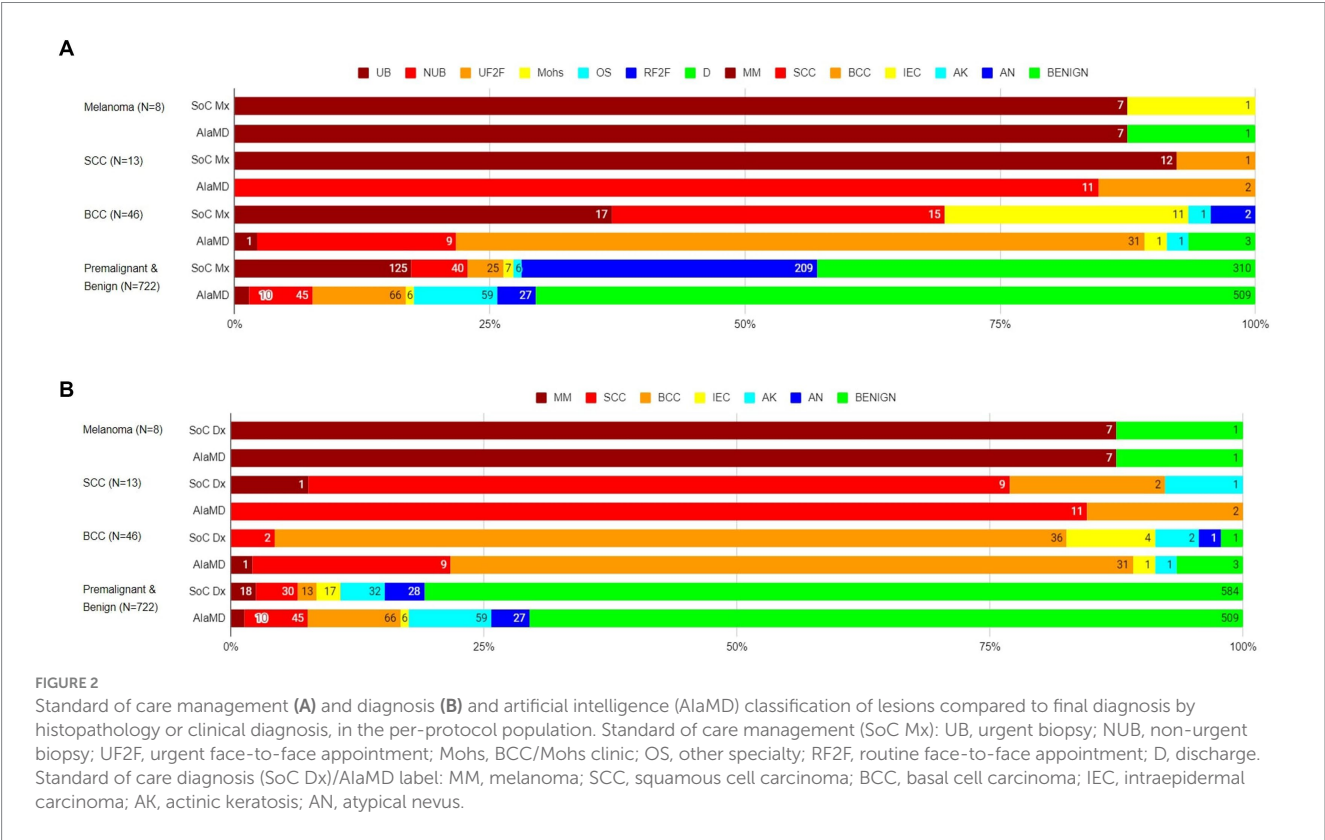
This study was conducted in a single site in North West London with a younger and more ethnically diverse population than the UK overall (6). The incidence of melanoma in the region is half of the national rate of (14 vs. 28 per 100,000) (19). There is a growing body of evidence that shows a drop in AI performance between research and real-world environments (20, 21). This means caution is needed

in extrapolating these results, particularly the NPV, PPV, and NNB, into other settings in which the patient population, incidence and risk of skin cancer, and physician experience are different. The AIaMD has, however, been safely deployed in real-world pathways by incorporating clinical reviews of its outputs. Indeed, real-world evidence of AIaMD performance continues to be collated showing strong performance

TABLE 4 Comparison of the accuracy of the standard of care (SoC) and artificial intelligence (AIaMD) for skin cancer detection in the per-protocol population.

	Sensitivity (% 95 CI)	Specificity (% 95 CI)	PPV (% 95 CI)	NPV (% 95 CI)	FNR (% 95 CI)	FPR (% 95 CI)	NNB (N, 95% CI)	NNR (N, 95% CI)
SoC	97.0, 88.7–99.5	71.9, 68.4–75.1	24.2, 19.3–29.9	99.6, 98.5–99.9	3.0, 0.5–11.3	28.1, 24.9–31.6	4.2 (3.3–5.5)	8.6 (6.2–12.3)
AIaMD	91.0, 80.9–96.3	83.2, 80.3–85.9	33.5, 26.8–40.9	99.0, 97.7–99.6	9.0, 3.7–19.1	16.8, 14.1–19.7	3 (2.4–3.7)	4.5 (3.6–5.8)

CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value; FNR, false negative rate; FPR, false positive rate; NNB, number needed to biopsy to confirm a diagnosis of skin cancer; NNR, number needed to refer to confirm a diagnosis of IEC or AK; IEC, intraepidermal carcinoma (Bowen's disease); AK, actinic keratosis.



and impact (22–26). There is an ongoing study to optimise how it is integrated into clinical pathways and workflows, as well as to evaluate the real-world impact with respect to health economics (27, 28).

The Get-It-Right-First-Time (GIRFT) dermatology workforce recommendations include the uptake of digital technologies to achieve more efficient NHS HRU (29). Implementing AIaMD services could allow trusts and dermatologists to dedicate more time to meeting skin cancer targets; individual cancer patients; addressing the post-pandemic backlog; patients with other severe skin diseases requiring systemic/biologic medication; and teaching/research. Moreover, this could reduce clinician burnout and increase the recruitment/retention of dermatologists with a greater capacity for trainees to see cases in addition to skin cancer referrals. Most importantly, this offers the opportunity to reduce reliance on insourcing and waitlist initiatives, which are short-term solutions to deep-seated and long-term dermatology capacity issues.

The UK Faster Diagnosis Standard (FDS) is currently being implemented, with a target of communicating to patients referred on cancer pathways their diagnosis within 28 days (30). As of June 2023, approximately one in every five NHS trusts was not able to meet this

target for skin cancer referrals (31). The immediacy of AIaMD outputs allows for quicker communication of premalignant and benign lesion classifications as well as the potential for greater surgical capacity to ensure more timely biopsies.

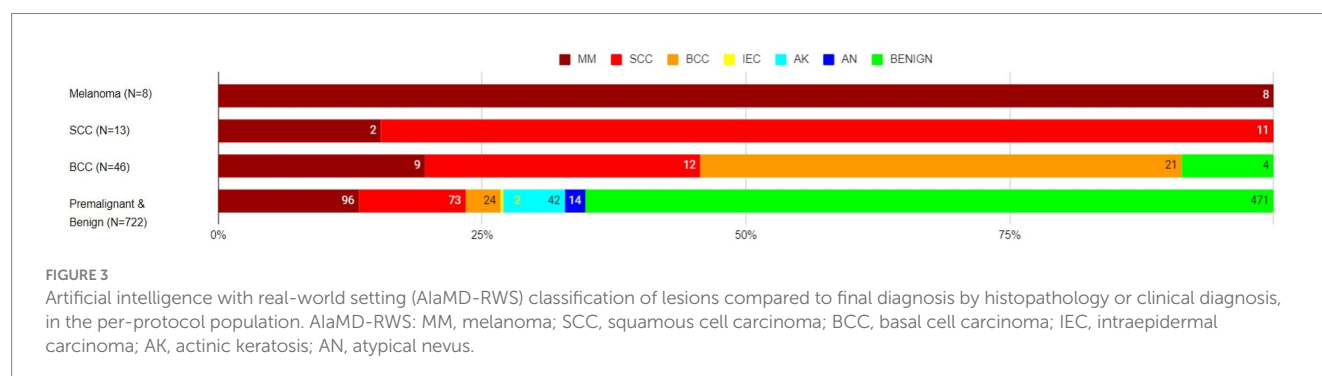
In the US, the American Academy of Dermatology supports skin cancer screenings at community events (32). An action in the FDS is to ‘consider linking the development of Community Locality Image Centres to Community Diagnostic Centres, to provide high-quality images for teledermatology and teledermoscopy activity’ (30). Given almost all lesions in this study were photographed by an HCA within 1 min, both settings could use the AIaMD to provide faster access to care in more remote locations, furthering the potential HRU benefits from fewer patients being unnecessarily referred for specialist assessment.

A recent UK government report highlighting several projects evaluating AI within healthcare stated that there are currently no standardised methods for the real-world evaluation of AI. Independent evaluations of the DERM service are ongoing, but a description of a real-world deployment of DERM at University Hospitals Birmingham is noted to have ‘helped 40% of patients avoid the need for a hospital

TABLE 5 Comparison of the accuracy of the standard of care (SoC) and artificial intelligence with real-world setting (AIaMD-RWS) for skin cancer detection in the per-protocol population.

	Sensitivity (% 95 CI)	Specificity (% 95 CI)	PPV (% 95 CI)	NPV (% 95 CI)	FNR (% 95 CI)	FPR (% 95 CI)	NNB (N, 95% CI)	NNR (N, 95% CI)
SoC	97.0, 88.7–99.5	71.9, 68.4–75.1	24.2, 19.3–29.9	99.6, 98.5–99.9	3.0, 0.5–11.3	28.1, 24.9–31.6	4.2 (3.3–5.5)	8.6 (6.2–12.3)
AIaMD-RWS	94, 84.7–98.1	73.3, 69.9–76.4	24.6, 19.6–30.4	99.2, 98–99.8	6, 1.9–15.3	26.7, 23.6–30.1	4.1 (3.3–5.1)	5.2 (4.1–6.6)

CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value; FNR, false negative rate; FPR, false positive rate; NNB, number needed to biopsy to confirm a diagnosis of skin cancer; NNR, number needed to refer to confirm a diagnosis of IEC or AK; IEC, intraepidermal carcinoma (Bowen's disease); AK, actinic keratosis.



appointment' (33). There are also examples in other therapy areas where AI might increase the speed of diagnosis [e.g., lung cancer (34) and heart failure (35)], but data in this regard are limited and an assessment of their impact on HRU is not yet available.

The COVID-19 pandemic impacted the study: Recruitment was suspended during national lockdowns; 48 patients had their management changed, usually downgraded; and follow-up appointments and non-urgent biopsies were delayed, including some biopsies that occurred more than 3 months after AIaMD assessment (no malignancies were identified in these). Indeed, the final diagnoses of 40 lesions could not be confirmed by biopsy or face-to-face assessment within the timeframe of the study. For these, LT conducted a second teledermatology assessment to provide a final diagnosis. Furthermore, elderly and immunosuppressed patients, who are at high risk of both COVID-19 and skin cancer, were encouraged to isolate during the pandemic and may have delayed seeking medical care during this time, which might account for the lower-than-expected incidence of malignancy in the study.

Connectivity issues led to some initial image-capture difficulties, but very few lesions had no images captured or failed image quality assessment, indicating an improvement in the image-capture process used in the MIAA study (5). This is likely to be due to the accessibility of capturing images using smartphones rather than a DSLR camera, emphasised by the images being captured in a minute on average. Technological deployment issues do remain a challenge that must be addressed for successful real-world deployment of the AIaMD.

There were no rare skin cancers identified in this study. This is not unexpected given the low incidence of rare skin cancers (3.1 per 100,000, 95% CI 3.0–3.2, in the UK in 2018–2020) and even lower incidence of specific rare skin cancers (e.g., 0.62, 95% CI 0.58–0.66 for Merkel cell carcinoma in the UK in 2018–2020) as opposed to skin cancer as a whole (387 per 100,000, 95% CI 386–388) (36). A few cases of rare skin cancers have been included in other studies of AIaMD (18,

26); however, additional data are needed to demonstrate the performance of the AIaMD.

The study was also reflective of the low incidence of skin cancers in higher Fitzpatrick skin types across a large population; however, it was not large enough to identify any malignant lesions in patients with darker skin, nor, therefore, to demonstrate the performance of the AIaMD in these patients. This is again to be expected given that less than 0.5% of skin cancers diagnosed in the UK are in Black and Asian patients (37). These cases often present late or are missed in the conventional care setting, making it difficult to demonstrate the performance of a novel product in patient groups with a low incidence of skin cancer through classical clinical studies. Efforts are ongoing to improve datasets in these under-represented patient groups, including surveillance of deployments and international collaborations.

AI systems can suffer from overfitting, hindering generalisability (38). The AIaMD algorithm has been trained on dermoscopic images of skin lesions from multiple sources. Biases may exist in these datasets, reducing AIaMD performance in different populations; however, the accuracy of the AIaMD observed is similar to previous reports (5, 16–18), demonstrating limited overfitting and good generalisability across novel datasets. Only one smartphone and dermatoscope combination was used, which is different from previous studies (5, 16–18), so no direct comparison of AIaMD performance on images captured by different devices can be made. This is controlled in real-world deployments of AIaMD too, however, whereby specific combinations of smartphones and dermatoscopes are qualified for usage with AIaMD, which is a mechanism of standardising the input to support consistent performance.

Finally, while the MIAA, NMSC, and impact studies show the performance of this particular AIaMD, these results cannot be generalised to the potential impact of other AI-based skin cancer detection tools. Indeed, a study of 25 freely downloadable AI apps found an average sensitivity of <30% for melanoma (39); a multicentre trial across Australia and Austria of a mobile phone-based AI found

its performance was significantly inferior to specialists in a real-world scenario (40); and an AI system studied in Canada identified 6 out of 10 melanoma included in the study (41). Importantly, the AIaMD evaluated here is a component of the first and, at the time of writing, only AI-based skin cancer detection product that is a Class IIa UKCA Medical Device. This is crucial not only as a verification of safety from the Medicines and Healthcare products Regulatory Agency (MHRA) but also for all of the systems in place to monitor and improve the technology. This certification opens up further opportunities for AIaMD to triage patients with skin lesions to the most appropriate next step with the aim of improved access and early diagnosis for all patients with suspected skin cancer.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by West Midlands – Edgbaston Research Ethics Committee and UK Health Research Authority. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

HM: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. PK: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – review & editing. MV: Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – review & editing. MN: Data curation, Investigation, Validation, Writing – review & editing. SM: Data curation, Investigation, Project administration, Validation, Writing – review & editing. NF: Data curation, Validation, Writing – review & editing. CH: Data curation, Methodology, Visualization, Writing – review & editing. DM: Data curation, Investigation, Project administration, Resources, Supervision, Writing – review & editing. DK: Data curation, Formal analysis, Investigation, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. LT: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study received funding from Skin Analytics Ltd. and Innovate UK.

Acknowledgments

The authors would like to thank all the patients who consented to the study and other clinical staff at Chelsea and Westminster NHS Foundation Trust without whom the study would not have been feasible. The authors would also like to acknowledge Dr. Jack Geenhalgh who developed the AIaMD and conducted the AIaMD image analysis for the study.

Conflict of interest

HM is an employee of Skin Analytics Ltd. and has received Skin Analytics shares or share options. PK provided contractual services to Skin Analytics Ltd. and has received Skin Analytics shares or share options. MV is an employee of Skin Analytics Ltd. and has received Skin Analytics shares or share options. SM's role as lead research assistant for the study was funded by payment from Skin Analytics Ltd. to Chelsea & Westminster Hospital NHS Foundation Trust. CH is a clinical advisor to Skin Analytics Ltd. and has received research funding to undertake a health economic model of the impact of the use of DERM in the NHS. DM is an employee of Skin Analytics Ltd. and has received Skin Analytics shares or share options. DK is an employee of Skin Analytics Ltd. and has received Skin Analytics shares or share options. LT is a clinical advisor to Skin Analytics Ltd.; has received Skin Analytics shares or share options; has received research funding support from Skin Analytics (salaries and equipment) and AIaMD deployment programme; has received reimbursement of conference fees, travel and accommodation costs from Skin Analytics to present research results; has received financial remuneration for separate programme of work as a consultant by Skin Analytics; has received grant funding from NHSX and CW+; has received paid honoraria to lecture for Almirall; was supported to attend a conference by Abbvie and Janssen; and holds multiple unpaid leadership roles.

The authors declare that this study received funding from Skin Analytics Ltd. and Innovate UK. The funders had the following involvement in the study: research, authorship, and publication.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2024.1302363/full#supplementary-material>

References

1. The Skin Cancer Foundation. Skin cancer facts & statistics. (2022). Available at: <https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/> (Accessed August 15, 2022).
2. Gettingitrightfirsttime.co.uk. Dermatology overview. (2022). Available at: <https://www.gettingitrightfirsttime.co.uk/wp-content/uploads/2021/11/Dermatology-overview.pdf> (Accessed August 15, 2022)
3. Ncinorguk. Routes to diagnosis. (2022). Available at: http://www.ncin.org.uk/publications/routes_to_diagnosis (Accessed August 15, 2022)
4. Pacifico M, Pearl R, Grover R. The UK government two-week rule and its impact on melanoma prognosis: an evidence-based study. *Ann R Coll Surg England*. (2007) 89:609–15. doi: 10.1308/003588407X205459
5. Phillips M, Marsden H, Jaffe W, Matin R, Wali G, Greenhalgh J, et al. Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. *JAMA Netw Open*. (2019) 2:e1913436. doi: 10.1001/jamanetworkopen.2019.13436
6. Chelwestnhsuk. A picture of health: profile of our trust's local population. (2020). Available at: <https://www.chelwest.nhs.uk/about-us/links/Full-Report-A-Picture-of-Health-Sep-2020.pdf> (Accessed August 15, 2022)
7. Limb M. The BMJ Awards 2020: showcase of this year's winning teams. *BMJ*. (2020):m4341. doi: 10.1136/bmj.m4341
8. NHS England. Approved costing guidance. (2023). Available at: www.england.nhs.uk; <https://www.england.nhs.uk/costing-in-the-nhs/approved-costing-guidance/> (Accessed February 14, 2024)
9. Personal Social Services Research Unit. (2022). Unit costs of health and social care programme (2022 – 2027). Available at: <https://www.pssru.ac.uk/unitcostsreport/>
10. Kawsar A, Hussain K, Kalsi D, Kemos P, Marsden H, Thomas L. Patient perspectives of artificial intelligence as a medical device in a skin cancer pathway. *Front Med*. (2023) 10:1259595. doi: 10.3389/fmed.2023.1259595
11. Rosendahl C, Tschandl P, Cameron A, Kittler H. Diagnostic accuracy of dermatoscopy for melanocytic and nonmelanocytic pigmented lesions. *J Am Acad Dermatol*. (2011) 64:1068–73. doi: 10.1016/j.jaad.2010.03.039
12. Reiter O, Mimouni I, Gdalevich M, Marghoob A, Levi A, Hodak E, et al. The diagnostic accuracy of dermoscopy for basal cell carcinoma: a systematic review and meta-analysis. *J Am Acad Dermatol*. (2019) 80:1380–8. doi: 10.1016/j.jaad.2018.12.026
13. Lehman W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics*. (1999) 55:1286–90. doi: 10.1111/j.0006-341X.1999.01286.x
14. O'Brien P, Fleming T. A multiple testing procedure for clinical trials. *Biometrics*. (1979) 35:549. doi: 10.2307/2530245
15. Pepe M. The statistical evaluation of medical tests for classification and prediction. *Technometrics*. (2005) 47:245–5. doi: 10.1198/tech.2005.s278
16. Phillips M, Greenhalgh J, Marsden H, Palamaras I. Detection of malignant melanoma using artificial intelligence: an observational study of diagnostic accuracy. *Dermatol Pract Conceptual*. (2019):e2020011. doi: 10.5826/dpc.1001a11
17. Marsden H, Palamaras I, Kemos P, Greenhalgh J. P63 Effectiveness of an image-analysing artificial intelligence-based digital health technology to diagnose non-melanoma skin cancer and benign skin lesions. *Br J Dermatol*. (2023) 188:ljad113.091. doi: 10.1093/bjd/ljad113.091
18. Marsden H, Morgan C, Austin S, DeGiovanni C, Venzi M, Kemos P, et al. Effectiveness of an image analyzing AI-based digital health technology to identify non-melanoma skin cancer (NMSC) and other skin lesions: results of the DERM-003 study. *Front Med Sec Dermatol*. (2023) 10:1288521. doi: 10.3389/fmed.2023.1288521
19. CancerData. Cancerdata.nhs.uk. (2022). Available from: https://www.cancerdata.nhs.uk/incidence_and_mortality
20. Zech J, Badgeley M, Liu M, Costa A, Titano J, Oermann E. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med*. (2018) 15:e1002683. doi: 10.1371/journal.pmed.1002683
21. Li C, Fei W, Shen C, Wang Z, Jing Y, Meng R, et al. Diagnostic capacity of skin tumor artificial intelligence-assisted decision-making software in real-world clinical settings. *Chin Med J*. (2020) 133:2020–6. doi: 10.1097/CM9.0000000000001002
22. Andrew K, Price T, Barlow N, Morris E, Ungureanu S, Zaki I, et al. Continued improvement of artificial intelligence in identifying skin Cancer. *EADV*. (2023) POSTER ID(P1005)
23. NHS England. (2023). Case study: artificial intelligence helping to speed up skin cancer diagnosis in Leicester, Leicestershire, and Rutland integrated care system. Available at: <https://www.england.nhs.uk/long-read/artificial-intelligence-helping-to-speed-up-skin-cancer-diagnosis-in-leicester-leicestershire-and-rutland-integrated-care-system/>
24. Jenkins R, Brewer CF, Kalsi D, Mullarkey D. BT09 clinical performance of an artificial intelligence-based medical device deployed within an urgent suspected skin cancer pathway. *Br J Dermatol*. (2023) 188:ljad113–375.
25. Abu Baker K, Roberts E, Harman K, Mullarkey D, Kalsi D. BT06 using artificial intelligence to triage skin cancer referrals: outcomes from a pilot study. *Br J Dermatol*. (2023) 188:ljad113.372. doi: 10.1093/bjd/ljad113.372
26. Thomas L, Hyde C, Mullarkey D, Greenhalgh J, Kalsi D, Ko J. Real-world post-deployment performance of a novel machine learning-based digital health technology for skin lesion assessment and suggestions for post-market surveillance. *Front Med*. (2023) 10:1264846. doi: 10.3389/fmed.2023.1264846
27. NHS England. (2021) AI in health and care award winners. Available at: <https://transform.england.nhs.uk/ai-lab/ai-lab-programmes/ai-health-and-care-award/ai-health-and-care-award-winners/>
28. Edge Health. (2021) Evaluating AI implementation in the NHS: skin analytics AI-powered teledermatology. (2024). Available at: <https://www.edgehealth.co.uk/news-insights/evaluation-nhs-ai-skin-cancer/>
29. Levell N. (2021). Dermatology GIRFT Programme National Specialty Report. NHS England and NHS Improvement. Available at: <https://www.gettingitrightfirsttime.co.uk/wp-content/uploads/2021/09/DermatologyReport-Sept21o.pdf>
30. Eoecitizenssenat.org. Best practice timed pathways: implementing a timed skin cancer diagnostic pathway. (2022). Available at: <https://www.eoecitizenssenat.org/sites/default/files/reports/Skin%20Faster%20Diagnostic%20Pathways%20v.2.2%20FINAL.pdf>
31. NHS England. Data extracts (monthly provider based only). NHS England. Available at: <https://www.england.nhs.uk/statistics/statistical-work-areas/cancer-waiting-times/>
32. Aad.org. Free skin cancer screenings. (2022). Available at: <https://www.aad.org/public/public-health/skin-cancer-screenings>
33. GOV.UK. Data saves lives: reshaping health and social care with data. (2022). Available at: <https://www.gov.uk/government/publications/data-saves-lives-reshaping-health-and-social-care-with-data/data-saves-lives-reshaping-health-and-social-care-with-data>
34. Digital Health. Somerset NHS FT trials AI algorithm for lung cancer diagnosis. (2022). Available at: <https://www.digitalhealth.net/2022/06/somerset-nhs-trials-ai-algorithm-lung-cancer/>
35. Bachtiger P, Petri C, Scott F, Ri Park S, Kelshiker M, Sahemey H, et al. Point-of-care screening for heart failure with reduced ejection fraction using artificial intelligence during ECG-enabled stethoscope examination in London, UK: a prospective, observational, multicentre study. *The Lancet Digital Health*. (2022) 4:e117–25. doi: 10.1016/S2589-7500(21)00256-9
36. NHS. Skin | get data out | CancerData. Available at: www.cancerdata.nhs.uk; <https://www.cancerdata.nhs.uk/getdataout/skin>
37. Delon C, Brown KF, Payne NW, Kotrotsios Y, Vernon S, Shelton J. Differences in cancer incidence by broad ethnic group in England, 2013–2017. *Br J Cancer*. (2022) 126:1765–73. doi: 10.1038/s41416-022-01718-5
38. Obermeyer Z, Emanuel E. Predicting the future — big data, machine learning, and clinical medicine. *N Engl J Med*. (2016) 375:1216–9. doi: 10.1056/NEJMp1606181
39. Sun MD, Kentley J, Mehta P, Duszka S, Halpern AC, Rotemberg V. Accuracy of commercially available smartphone applications for the detection of melanoma. *Br J Dermatol*. (2022) 186:744–6. doi: 10.1111/bjd.20903
40. Menzies SW, Sinz C, Menzies M, Lo SN, Yolland W, Lingohr J, et al. Comparison of humans versus mobile phone-powered artificial intelligence for the diagnosis and management of pigmented skin cancer in secondary care: a multicentre, prospective, diagnostic, clinical trial. *Lancet Digital Health*. (2023) 5:e679–91. doi: 10.1016/S2589-7500(23)00130-9
41. Crawford ME, Kamali K, Dorey RA, MacIntyre OC, Cleminson K, MacGillivray ML, et al. Using artificial intelligence as a melanoma screening tool in self-referred patients. *J Cutan Med Surg*. (2023) 28:37–43. doi: 10.1177/12034754231216967

Frontiers in Medicine

Translating medical research and innovation into
improved patient care

A multidisciplinary journal which advances our
medical knowledge. It supports the translation
of scientific advances into new therapies and
diagnostic tools that will improve patient care.

Discover the latest Research Topics

[See more →](#)

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact



Frontiers in Medicine

