HHS Public Access

Author manuscript

J Eur Acad Dermatol Venereol. Author manuscript; available in PMC 2022 February 01.

Published in final edited form as:

J Eur Acad Dermatol Venereol. 2021 February ; 35(2): 546–553. doi:10.1111/jdv.16979.

Performance of a deep neural network in teledermatology: a single-centre prospective diagnostic study

C. Muñoz-López[#1], **C. Ramírez-Cornejo**[#1], **M.A. Marchetti**[2], **S. S. Han**[3], **P. Del Barrio-Díaz**[1], **A. Jaque**[1], **P. Uribe**[1,5], **D. Majerson**[1], **M. Curi**[1], **C. Del Puerto**[1], **F. Reyes-Baraona**[1], **R. Meza-Romero**[1], **J. Parra-Cares**[1], **P. Araneda-Ortega**[1], **M. Guzmán**[1], **R. Millán-Apablaza**[1], **M. Nuñez-Mora**[1], **K. Liopyris**[4], **C. Vera-Kellet**[1,‡], **C. Navarrete-Dechent**[1,5,*,‡]

[1]Department of Dermatology, Escuela de Medicina, Pontificia Universidad Católica de Chile, Santiago, Chile

[2]Dermatology Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA

[3]Dermatology Clinic, Seoul, Korea

[4]Department of Dermatology, University of Athens, Andreas Syggros Hospital of Skin and Venereal Diseases, Athens, Greece

[5]Melanoma and Skin Cancer Unit, Escuela de Medicina, Pontificia Universidad Católica de Chile, Santiago, Chile

[#] These authors contributed equally to this work.

Abstract

Background—The use of artificial intelligence (AI) algorithms for the diagnosis of skin diseases has shown promise in experimental settings but has not been yet tested in real-life conditions.

Objective—To assess the diagnostic performance and potential clinical utility of a 174-multiclass AI algorithm in a real-life telemedicine setting.

Methods—Prospective, diagnostic accuracy study including consecutive patients who submitted images for teledermatology evaluation. The treating dermatologist chose a single image to upload to a web application during teleconsultation. A follow-up reader study including nine healthcare providers (3 dermatologists, 3 dermatology residents and 3 general practitioners) was performed.

Results—A total of 340 cases from 281 patients met study inclusion criteria. The mean (SD) age of patients was 33.7 (17.5) years; 63% (n = 177) were female. Exposure to the AI algorithm results was considered useful in 11.8% of visits (n = 40) and the teledermatologist correctly modified the real-time diagnosis in 0.6% (n = 2) of cases. The overall top-1 accuracy of the algorithm (41.2%) was lower than that of the dermatologists (60.1%), residents (57.8%) and

[*]**Correspondence**: C. Navarrete-Dechent. ctnavarr@gmail.com.
[‡]Co-senior authors.

IRB approval status: Reviewed and approved by Pontificia Universidad Católica IRB; approval # IRB 200421005

Conflicts of Interest
None declared.

general practitioners (49.3%) (all comparisons $P < 0.05$, in the reader study). When the analysis was limited to the diagnoses on which the algorithm had been explicitly trained, the balanced top-1 accuracy of the algorithm (47.6%) was comparable to the dermatologists (49.7%) and residents (47.7%) but superior to the general practitioners (39.7%; $P = 0.049$). Algorithm performance was associated with patient skin type and image quality.

Conclusions—A 174-disease class AI algorithm appears to be a promising tool in the triage and evaluation of lesions with patient-taken photographs via telemedicine.

Introduction

The use of artificial intelligence (AI) for the diagnosis of skin diseases has shown significant promise to improve health outcomes.[1] Recently developed AI algorithms using convolution neural networks (CNNs) have been shown to classify clinical images equally or better than dermatologists in artificial study settings.[2–4] Despite the promising results coming from reader studies, these AI algorithms have not yet been applied to dermatology clinical practice and have not yet shown efficacy in improving healthcare outcomes, while only one study has evaluated their performance in real-life conditions.[5]

During the lockdowns that occurred due to the COVID-19 pandemic, our dermatology service implemented live, video-based telemedicine visits to maintain patient care while minimizing the risk of infection and avoiding in-person encounters.[6–8] Given the limitations and diagnostic challenges associated with telemedicine visits, an accurate AI algorithm could help as a screening aid or as a clinical decision support tool for healthcare providers.[9,10] To date, the diagnostic performance of AI algorithms using patient-submitted photographs in telemedicine workflows is poorly characterized.

In 2020, Han *et al* published a publicly testable CNN algorithm trained using 220,680 images of 174 dermatological disorders and validated on 3,501 images and 134 diseases.[11] The algorithm was able to classify skin diseases and improved the diagnostic performance of healthcare providers in a reader study. Here, we aimed to assess the diagnostic performance and potential clinical utility of this AI algorithm in a prospective study conducted in a real-life telemedicine setting during the visits.

Patients and methods

This was an IRB-approved, single-centre, prospective, diagnostic accuracy study conducted between 27 March and 30 April 2020. All data collection was explicitly planned prior to the performance of the index tests and reference standard and the study was reported following the STARD 2015 guidelines.[12] Consecutive patients who submitted clinical images of one or more new skin conditions for evaluation in our telemedicine dermatology clinic were considered eligible for inclusion. Both new and existing patients were eligible with no age restrictions. Cases were excluded if no images were submitted by patients, images were rated as insufficient quality, patients submitted images but did not present for telemedicine visit, patients had a telemedicine visit but did not subsequently adhere to the recommendation for an in-person visit or diagnostic testing, or upon evaluation, they had no skin condition present for examination.

Telemedicine clinical workflow

Patients submitted ≥ 1 clinical images of the skin condition to the consulting dermatologist prior to, or during the telemedicine visit. All patients received standardized imaging instructions, which provided lighting and flash guidelines in addition to recommended poses. Images were sent to the dermatologist via secure email. If necessary, the dermatologist could request an additional set of images. Quality of images was rated by the dermatologist as 'insufficient' vs. 'sufficient'. Images of sufficient quality were further rated as 'high quality' or 'average quality'. All telemedicine visits were real-time video encounters conducted using a video conferencing platform (Zoom Video Communication, Inc.). After reviewing the images and interviewing the patient, the dermatologist recorded their telemedicine clinical diagnosis and treatment plan into a study-specific data collection form. Patient demographics (age and sex), Fitzpatrick skin type (I-VI) and medical history were recorded. If the dermatologist was unable to reach a diagnosis, the patient was directed for an in-person visit, skin biopsy and/or additional laboratory workup.

Artificial intelligence index test

During the telemedicine visit, a validated, public, web-based AI algorithm (available online at http://modelderm.com) was used as the AI index test. Briefly, the algorithm was trained by one of the authors (S.S.H.) using the ASAN dataset that consisted of 220,680 images. It was validated using part of the ASAN dataset (17,125 images). The test set consisted of images from Korea (2201 images from the Seoul National University Bundang Hospital, Inje University Sanggye Paik Hospital, and Hallym University Dongtan Hospital), and the Edinburgh dataset (1300 images). The network structure includes SENet, SE-ResNet-50, and visual geometry group (VGG-19) parallelly concatenated and trained with 174 disease classes. These neural networks (SENet, SE-ResNet-50, and VGG-19) are convolutional neural networks, and the number of hidden layers is 154, 50 and 19, respectively. ImageNet pretrained models of each network were finetuned end-to-end separately. Details can be found in Han *et al.*[11]

The AI web application allowed uploading images at different magnifications. As per the instructions for the algorithm, images were uploaded with the skin condition occupying at least 80% of the field of view and with the region of interest (ROI) centred. The dermatologist chose a single image to upload to the web application. Dermatologists were instructed to use the most-representative and highest-quality image for each skin condition. The algorithm output consisted of three diagnoses (from a possible list of 174 skin conditions), ranked in order of probability (ranging from 0 to 1). All algorithm outputs were recorded. This procedure was conducted in real-time, during the visit, with the potential of affecting the outcome of the visit. Given the nature and setting of the AI index test, the reference standard as well as the composition of test classes was inherently blinded. Finally, dermatologists recorded if exposure to the AI algorithm: (a) was clinically useful to them during the telemedicine visit by either increasing their diagnostic confidence and/or expanding their differential diagnosis (yes, no) and (b) led to a change in their telemedicine clinical diagnosis (yes, no).

Healthcare provider index test

Subsequent to the telemedicine visit, clinical images were independently evaluated by nine healthcare providers in a reader study. We invited three board-certified dermatologists, three dermatology residents, and three general practitioners (GPs), and all agreed to participate. The dermatologists had a median (min-max) of 3 (2–5) years of postresident clinical experience. One resident was in their second year and two were in their third year of dermatology training. The GPs had no formal training in dermatology but had an interest in skin diseases. GPs had a median (min-max) of 7 (1–11) years of postmedical school clinical experience. In this experiment, readers were only exposed to the images submitted by the patient and were blinded to clinical history/metadata as well as the results of the AI index test and the reference standard. There were no time restrictions.

Reference standard

The reference standard was defined in one of two ways. First, if the patient was recommended to return for an in-person clinic visit, the diagnosis from this visit (and any associated laboratory testing or skin biopsies) was used as the reference standard. Second, if no in-person clinic visit was performed, a panel of 6 dermatologists evaluated the case and established the reference standard based on consensus agreement. The panel dermatologists had a median (min-max) of 8 (5–13) years of postresident clinical experience; there was no overlap between dermatologists who participated in the reader study and the reference standard. During the consensus evaluation, the clinical images and medical history were shown to the panel of dermatologists. All index test results were unavailable during consensus evaluations. Only one diagnosis was permitted for the reference standard. There were no cases of disagreement.

Statistical analysis

Descriptive statistics were used to characterize the study data. Diagnoses were grouped into 5 categories and 13 subcategories: (1) ‘inflammatory’ (subcategories: dermatitis, acne/rosacea, autoimmune, papulosquamous and other); (2) ‘infectious’ (subcategories: bacterial, viral, fungal and parasitic); (3) ‘neoplastic’ (subcategories: malignant and benign); (4) ‘alopecia’ (subcategories: scarring and non-scarring); and (5) ‘other’ (e.g. burn, scar, striae and among others), using a system adapted from Liu *et al.*[13]

The primary outcomes were the ‘top-1 accuracy’ and the ‘balanced top-1 accuracy’. The ‘top-1 accuracy’ measures how frequently the top-1 prediction of the index test matched the reference standard diagnosis. Similar but not identical diagnoses to the reference standard (i.e. herpes simplex vs. herpes zoster) were considered incorrect. The ‘balanced top-1 accuracy’ is the top-1 accuracy computed for each unique diagnosis separately and then averaged. This was performed to account for the variant prevalence of certain diagnoses (e.g. acne vs hidradenitis suppurativa). Because algorithm fundamentally always predicted incorrect answers for the untrained cases (‘out of distribution’), we performed sub-analyses for both measures restricted to the 174 conditions on which the index AI algorithm was explicitly trained (i.e. ‘in distribution’ vs. ‘overall’ conditions). A secondary outcome was the ‘top-3 accuracy’, which measures if the reference standard diagnosis was included in any of 3 diagnoses of the AI algorithm output.

Shapiro–Wilk test was used for tests of normality, Z-tests were used to compare proportions, and Chi-square and Fisher exact tests were used to evaluate the association of categorical variables. Statistical analysis was performed using STATA 14.0®. An alpha value below 0.05 was considered statistically significant.

Results

A total of 380 skin conditions were evaluated in the telemedicine clinic during the study period. After exclusion of 40 skin conditions (18 no clinical images, 8 insufficient quality images, 8 lost to follow-up after recommendation of in-person visit, and 6 no active disease), 340 skin conditions (87 unique diagnoses) from 281 patients met the study inclusion criteria. The mean (SD) age of the patients was 33.7 (17.5) years; 37% (n = 104) were male and 63% (n = 177) were female. Of the 340 skin conditions, 190 (55.9%) occurred in Fitzpatrick skin phototype III patients, 84 (24.7%) in type IV patients, 59 (17.3%) in type II patients and 7 (2.1%) in type I patients.

In 7.1% (n = 24) of visits, additional clinical images were requested and provided by the patients. Overall, 87.4% (n = 297) and 12.6% (n = 43) of images were rated as having ‘high quality’ and ‘average quality’, respectively. There were no differences in image quality by diagnostic category (P = 0.971) (Table S1). The reference standard was determined from in-person clinic visits (18.5%, n = 63), skin biopsy (13.8%, n = 47) and consensus agreement of panel dermatologists (81.5%, n = 277). Overall, per the reference standard, 225 of the skin conditions were inflammatory (66.2%), 52 infectious (15.3%), 32 neoplastic (9.4%), 15 alopecia (4.4%) and 16 other (4.7%). There were 87 unique diagnoses; the most frequent were acne (18.5%), contact dermatitis (8.2%), psoriasis (3.8%) and warts (3.6%) (Table 1 and Table S2).

AI performance

The AI algorithm had overall top-1 and balanced top-1 accuracies of 41.2% and 35.1%, respectively (Table 2). Accuracy was associated with the Fitzpatrick skin type with better performance in darker skin types and was not associated with sex (Table S3). Of the 87 unique skin diagnoses included, 63 (72.4%) were part of the AI algorithm training dataset. Therefore, 305 out of the 340 (89.7%) skin conditions were ‘in distribution’ and 35 were ‘out of distribution’ images. When limiting the analysis to the ‘in distribution’ skin conditions, the top-1 and balanced top-1 accuracies increased to 45.3% and 47.6%, respectively (Table 3).

AI performance by disease category/subcategory

Top-1 accuracy was highest in the ‘other’ category (68.8%), followed by the ‘infectious’ (50%), ‘inflammatory’ (38.2%), ‘neoplastic’ (37.5%) and ‘alopecia’ (33.3%) categories (P = 0.089). For the subcategories, ‘non-scarring alopecia’ had the highest top-1 accuracy (71.4%), followed by ‘papulosquamous’ (52.9%), ‘fungal’ (52.4%), ‘viral’ (50%), ‘bacterial’ (50%), ‘acne/rosacea’ (47.4%), benign neoplasms (42.9%), ‘dermatitis’ (34.1%), ‘autoimmune’ (26.5%), ‘parasitic’ (0%), ‘malignant neoplasms’ (0%) and ‘scarring alopecia’ (0%) (P = 0.035) (Table 2).

AI vs. Physician performance

The overall top-1 accuracy of the AI algorithm (41.2%) was lower than that of the dermatologists (60.1%; $P < 0.001$), residents (57.8%; $P < 0.001$) and GPs (49.3%; $P = 0.034$) (Table 2, Figure 1). The balanced top-1 accuracy of the AI algorithm was 35.1%, lower than the dermatologists (45.2%; $P = 0.007$) and the residents (42.6%: $P = 0.045$), but not different than the GPs (33.1%; $P = 0.582$).

The overall top-1 'in distribution' accuracy of the AI algorithm (45.3%) was lower than that of the dermatologists (63.1%) and residents (60.9%) ($P < 0.001$) but similar to GPs (52.9%, $P = 0.061$) (Table 3, Figure 1). The balanced top-1 'in distribution' accuracy of the AI algorithm (47.6%) was similar to that of the dermatologists (49.7%, $P = 0.604$) and residents (47.7%, $P = 0.980$) and superior than the GPs (39.7%, $P = 0.049$).

Accuracy by image quality

The algorithm achieved a higher top-1 accuracy in images of 'high quality' than in images of 'average quality' (44.4% vs. 18.6%, $P = 0.001$). In contrast, there were no differences in the performance of the readers by image quality (Table S4).

AI utility

Overall, use of the AI algorithm was considered useful in 11.8% ($n = 40$) of cases, by increasing diagnostic confidence and/or expanding the differential diagnosis (Figure 2). The top-1 accuracy of these 40 photographs was 50%, similar to the overall top-1 accuracy of the 340 images (41.2%, $P = 0.286$). The top-3 accuracy was 75.9%. The 'balanced' top-1 accuracy for these 40 images was 54.9%, higher than the balanced top-1 accuracy of the 340 images (35.1%, $P = 0.014$). 'In distribution' pathologies included in these 40 cases were inflammatory ($n = 17$), infectious ($n = 12$) and neoplastic ($n = 8$) (Table S5). Exposure to the AI algorithm modified the real-time diagnosis of the teledermatologist in 0.9% ($n = 3$) of cases; in 2 cases the diagnosis was correctly modified (tinea corporis and perioral dermatitis) and in 1 case neither the clinician nor the algorithm matched the reference standard (sarcoidosis).

Discussion

In this prospective, real-life study, we evaluated an AI algorithm's performance in 340 skin conditions from 281 patients assessed via teledermatology and compared the results to physicians with different levels of experience in a reader study. The top-1 diagnosis accuracy, which indicates the accuracy of the algorithm's diagnosis with the highest probability, was inferior to the accuracy of dermatologists, residents and GPs. However, when narrowing the analysis to the diseases included in the algorithm training dataset (i.e. 'in distribution' cases) and adjusting for the variant prevalence of certain conditions, the 'balanced top-1 accuracy' of the AI program was similar to the dermatologists and residents, and better than the GPs, suggesting growing potential for the use of AI in dermatology care.

Our results are similar to a recent study by Han *et al*[11] that tested the same algorithm. They observed a top-1 accuracy of 44.8%, similar to the dermatologists. Liu *et al*[13] used a CNN

algorithm to classify clinical images obtained retrospectively from cases of a teledermatology service within 26 pathologies. The AI program achieved a top-1 accuracy of 66%, exceeding that of primary care physicians and nurse practitioners and similar to dermatologists.

Previous studies have used AI algorithms to distinguish melanoma from benign and malignant neoplasms, using datasets with high-quality, standardized clinical or dermoscopic images, mostly in experimental-reader environments. Those studies have shown similar or better AI diagnostic performances when compared with physicians.[14–17] This study is the first to apply an AI algorithm to patient-taken clinical photographs in a prospective manner and in a real-life telemedicine scenario. In 2020, Tschandl *et al*[18] evaluated an algorithm on dermoscopic images taken by patients at high risk for skin cancer and showed a lower performance compared with physician-taken dermoscopic images. Training an algorithm on dermatological clinical images has inherently more difficulties due to the great variability on image acquisition of those images compared with dermoscopic images alone. Only one study has evaluated the performance of an AI algorithm in classifying lesions as benign or malignant in a prospective real-life setting. In that study, the performance of the AI algorithm was significantly lower than obtained on a test set during algorithm training.[5]

Our study is novel in several aspects. We examined the performance on 87 different diagnoses; images were taken by patients mainly with cellphones and not by a health professional or professional cameras; patients were consecutive and not selected; and the study was conducted in a real-life teledermatology setting, with untrained dermatologists on the use of an AI algorithm. Interestingly, the performance of the AI algorithm was poorer in images of lower quality. This suggests that for optimal use of AI in clinical practice, image quality standards are needed.[19,20] In contrast, image quality did not appear to affect the performance of physicians. We also observed better performance of the AI algorithm in darker skin types. This may be related to the population used to train the algorithm, which was mostly Asian. Our patient population was mainly comprised of Hispanic/Latino patients, which more commonly have darker skin types. This highlights the importance of the inclusion of patients with all skin colours in the development of AI algorithms, as skin type might affect diagnostic accuracy.[21]

Our results suggest that the use of AI algorithms could be a potentially useful tool for GPs to broaden their differential diagnosis of skin diseases. For the dermatologist, in contrast, it could be a useful tool to in atypical, challenging or doubtful cases. In fact, the use of the AI algorithm was deemed as useful in more than 10% of the cases, even though it did not modify the dermatologists' primary diagnosis in most cases. The benefit of exposure to the algorithm may be greater in groups with less experience, considering that Tschandl *et al*[18] showed an inverse relationship between the gain from AI-based support and rater experience.

Limitations

Due to the sample size and consecutive case recruitment, we did not obtain representative results of less common diseases. The AI algorithm output was based on a single photograph, which differs from other AI algorithms that consider more than one photograph or additional

clinical metadata.[22,23] Finally, we did not use formally validated tools to measure the physicians' confidence in their diagnosis as well as the perceived utility of the algorithm.

Conclusion

In this prospective real-life study, we demonstrated that the accuracy of an AI algorithm was inferior to dermatologists for images of lesions submitted by patients during teledermatology encounters. Nonetheless, the use of an algorithm could enhance the confidence and accuracy of physicians as a human-algorithm collaboration tool. This is especially relevant due to the recent increase in teledermatology clinics worldwide.[24,25] To further improve the performance of AI algorithms, they may need to be customized regionally. More studies are needed before AI can be incorporated into daily clinical practice.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

'The patients in this manuscript have given written informed consent to the publication of their case details'.

Funding sources

Dr. Marchetti's research is funded in part through the MSKCC institutional NIH/NCI Cancer Center Support Grant P30 CA008748. The funders had no role in the design and conduct of the study; collection, management, analysis and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

References

1. Gomolin A, Netchiporouk E, Gniadecki R, Litvinov IV. Artificial intelligence applications in dermatology: where do we stand? Front Med 2020; 31: 7.
2. Aractingi S, Pellacani G. Computational neural network in melanocytic lesions diagnosis: artificial intelligence to improve diagnosis in dermatology? Eur J Dermatol 2019; 29: 4–7. [PubMed: 31017580]
3. Dick V, Sinz C, Mittlböck M, Kittler H, Tschandl P. Accuracy of computer-aided diagnosis of melanoma: a meta-analysis. JAMA Dermatology. 2019; 155: 1291–1299. [PubMed: 31215969]
4. Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. Eur J Cancer 2019; 111: 148–154. [PubMed: 30852421]
5. Dreiseitl S, Binder M, Hable K, Kittler H. Computer versus human diagnosis of melanoma: evaluation of the feasibility of an automated diagnostic system in a prospective clinical trial. Melanoma Res 2009; 19: 180–184. [PubMed: 19369900]
6. Smith AC, Thomas E, Snoswell CL, Haydon H, Mehrotra A, Clemensen J. Telehealth for global emergencies: Implications for coronavirus disease 2019 (COVID-19). J Telemed Telecare 2019; 2020: 2019.
7. Villani A, Scalvenzi M, Fabbrocini G. Teledermatology: a useful tool to fight COVID-19. J Dermatolog Treat 2020;31(4): 325. [PubMed: 32238000]
8. Ohannessian R, Duong TA, Odone A. Global telemedicine implementation and integration within health systems to fight the COVID-19 pandemic: a call to action. JMIR Public Heal Surveill 2020; 6: e18810.
9. Lee JJ, English JC. Teledermatology: a review and update. Am J Clin Dermatol 2018; 19: 253–260. [PubMed: 28871562]

10. Coates SJ, Kvedar J, Granstein RD. Teledermatology: From historical perspective to emerging techniques of the modern era: Part I: History, rationale, and current practice. J Am Acad Dermatol 2015; 72: 563–574. [PubMed: 25773407]
11. 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(9): 1753–1761. [PubMed: 32243882]
12. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L et al. STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015; 351: 1–9.
13. 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(6): 900–908. http://www.nature.com/articles/s41591-020-0842-3 [PubMed: 32424212]
14. Tschandl P, Kittler H, Argenziano G. A pretrained neural network shows similar diagnostic accuracy to medical students in categorizing dermatoscopic images after comparable training conditions. Br J Dermatol 2017; 177: 867–869. [PubMed: 28569993]
15. 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: 1836–1842. [PubMed: 29846502]
16. 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: 115–118. [PubMed: 28117445]
17. Marchetti MA, Codella NCF, Dusza SW, Gutman DA, Helba B, Kalloo A et al. Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. J Am Acad Dermatol 2018; 78: 270–277.e1. [PubMed: 28969863]
18. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N. Human – computer collaboration for skin cancer recognition. Nat Med 2020; 26(8): 1229–1234. 10.1038/s41591-020-0942-0 [PubMed: 32572267]
19. Finnane A, Curiel-Lewandrowski C, Wimberley G, Caffery L, Katragadda C, Halpern A et al. Proposed technical guidelines for the acquisition of clinical images of skin-related conditions. JAMA Dermatology 2017; 153: 453–457. [PubMed: 28241182]
20. Hulstaert E, Hulstaert L. Artificial intelligence in dermato-oncology: a joint clinical and data science perspective. Int J Dermatol 2019; 58: 989–990. [PubMed: 31149729]
21. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol 2018; 154: 1247–1248. [PubMed: 30073260]
22. Navarrete-Dechent C, Dusza SW, Liopyris K, Marghoob AA, Halpern AC, Marchetti MA. Automated dermatological diagnosis: Hype or reality? J Invest Dermatol. 2018; 138: 2277–2279. [PubMed: 29864435]
23. Navarrete-Dechent C, Liopyris K, Molenda MA et al. Human surface anatomy terminology for dermatology: a Delphi consensus from the International Skin Imaging Collaboration. J Eur Acad Dermatol Venereol 2020; 34: 2659–2663. [PubMed: 32770737]
24. Hollander JE, Carr BG. Virtually Perfect? Telemedicine for Covid-19. N Engl J Med 2020; 382 18:1679–1681. [PubMed: 32160451]
25. Webster P. Virtual health care in the era of COVID-19. The Lancet 2020;395: 1180–1181.

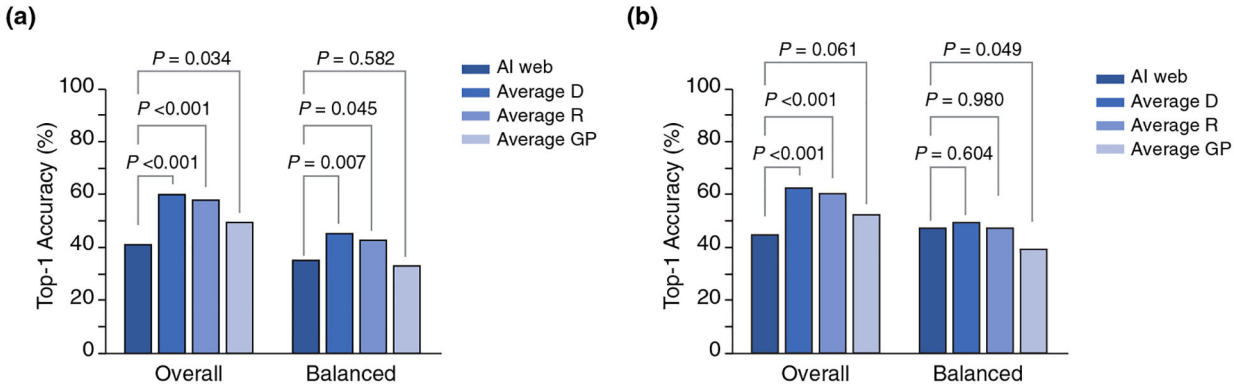

Figure 1.

Overall and Balanced Top-1 accuracy of the AI algorithm vs. the Top-1 for clinicians. (a) All cases (in distribution + out of distribution) ($n = 340$). (b) In distribution cases ($n = 305$).

Figure 2.

Representative cases seen in the study. (a) 25-year-old female with diagnosis of acne. The tele-dermatologist's diagnosis was 'acne' and the top-1 diagnosis of the algorithm was 'varicella'. The algorithm was considered 'not useful'. (b) 10-year-old male with atopic dermatitis and diagnosis of 'eczema herpeticum'. The tele-dermatologist had diagnostic uncertainty, suspecting 'bacterial superinfection' or 'herpetic eczema'. The second and third diagnoses of the algorithm were 'herpes simplex' and 'herpes zoster' (top-3 correct). The algorithm increased the diagnostic confidence of the tele-dermatologist, who prescribed antivirals. (c) 20-year-old female. A diagnosis of 'pseudolymphoma' or 'cutaneous lupus' was suspected by the tele-dermatologist. The top-1 diagnosis of the algorithm was 'tinea corporis'. A potassium hydroxide examination confirmed the diagnosis of a dermatophyte infection. In this case, the AI algorithm changed the diagnosis correctly.

Table 1

Skin conditions based on the reference standard

Category	Subcategory	Diagnosis
Inflammatory (*n* = 225, 66.2%)	Dermatitis (*n* = 82, 36.4%)	Contact dermatitis *, Unspecific eczema *, Hand eczema *, Seborrheic dermatitis *, Atopic dermatitis *, Nummular eczema, Dyshidrotic eczema *, Cheilitis *, Chronic paronychia *, Erythema annulare centrifugum *, Lichen simplex chronicus *, Lichenoid dermatitis, Pseutotinea amiantacea. *
	Acne/rosacea (*n* = 78, 34,7%)	Acne *, Rosacea *, Hidradenitis suppurativa, Perioral dermatitis
	Autoimmune (*n* = 34, 15.1%)	Morphea *, Cutaneous lupus *, Bullous disease *, Vitiligo *, Poikiloderma * (dermatomyositis), Lichen sclerosus, Amicrobial pustulosis of the folds, Complex aphthosis, Calcinosis cutis.
	Papulosquamous (*n* = 17, 7.6%)	Psoriasis *, Pityriasis rosea *, Lichen nitidus *, Pityriasis lichenoides chronica *.
	Others inflammatory (*n* = 14, 6.2%)	Urticaria *, Insect bite *, Exantema, Postinflammatory hypopigmentation, Postinflammatory hyperpigmentation *, Grover's disease, Pernio Like Eruption, Sarcoidosis, Ulcers *.
Infectious (*n* = 52, 15.3%)	Viral (*n* = 22, 42.3%)	Warts *, Herpes zoster *, Viral rash *, Eczema herpeticum *, Molluscum contagiosum *, Condylomas *.
	Fungal (*n* = 21, 40.4%)	Tinea *, Onychomycosis *, Pityriasis Versicolor *, Balanitis, Angular cheilitis *, Candidal intertrigo.
	Bacterial (*n* = 8, 15,4%)	Ingrown nail *, Folliculitis *, Paronychia *, Furuncle *, Impetigo *, Pyoderma.
	Parasitic (*n* = 1, 1.9%)	Demodicosis.
Neoplastic (*n* = 32, 9.4%)	Benign (*n* = 28, 87.5%)	Melanocytic nevus *, Seborrheic keratosis *, Actinic Keratosis *, Dermatofibroma *, Epidermal cyst *, Keloid *, Organoid nevus *, Inflamed nevus, Telangiectatic granuloma *, Targetoid Hemosiderotic Hemangioma, Milia *, Inverted follicular keratosis, Cutaneous horn *.
	Malignant (*n* = 4, 12,5%)	Basal cell carcinoma *, Bowen's disease *.
Alopecias (*n* = 15, 4.4%)	Scarring (*n* = 8, 53.3%)	Lichen planus pilaris, Folliculitis decalvans, Non-specified.
	Non-scarring (*n* = 7, 46.7%)	Androgenetic alopecia *, Effluvium telogen, Alopecia Areata *.
Other (*n* = 16, 4.7%)		Striae distansae *, Scar *, Acne scar *, Burn *, Hematoma *, Erythema ab igne *, Galli-Galli disease, Portwine stain *.

* In-distribution diagnoses (explicitly included in algorithm training).

J Eur Acad Dermatol Venereol. Author manuscript; available in PMC 2022 February 01.

Table 2

Accuracy of the AI algorithm, dermatologists, dermatology residents, and general practitioners in overall conditions

	n	Top-1 Accuracy				Top-3 Accuracy
		AI algorithm	Average dermatologist	Average resident	Average general practitioner	AI algorithm
General	340	41.2%	60.1%	57.8%	49.3%	63.5%
Balanced	340	35.1%	45.2%	42.6%	33.1%	55.1%
Inflammatory	225	38.2%	63.3%	59.1%	50.8%	63.6%
Dermatitis	82	34.1%	61.4%	52.0%	48.8%	59.8%
Acne/rosacea	78	47.4%	81.2%	80.8%	73.9%	76.9%
Autoimmune	34	26.5%	37.3%	36.3%	16.7%	41.2%
Papulosquamous	17	52.9%	52.9%	47.1%	33.3%	82.4%
Infectious	52	50.0%	59.6%	58.3%	53.2%	69.2%
Viral	22	50.0%	65.2%	62.1%	51.5%	72.7%
Fungal	21	52.4%	68.3%	68.3%	71.4%	66.7%
Bacterial	8	50.0%	20.8%	25.0%	16.7%	62.5%
Parasitic	1	0%	66.7%	33.3%	0%	0%
Neoplastic	32	37.5%	50.0%	53.1%	41.7%	59.4%
Benign	28	42.9%	56.0%	59.5%	45.2%	64.3%
Malignant	4	0.0%	8.3%	8.3%	16.7%	25.0%
Alopecia	15	33.3%	44.4%	44.4%	31.1%	33.3%
Scarring	8	0%	25.0%	33.3%	16.7%	0%
Non-scarring	7	71.4%	66.7%	57.1%	47.6%	71.4%
Other	16	68.8%	52.1%	60.4%	47.9%	81.3%

AI, Artificial Intelligence; D, dermatologist; R, residents; GP, general practitioners.

J Eur Acad Dermatol Venereol. Author manuscript; available in PMC 2022 February 01.

Table 3

Accuracy of the AI algorithm, dermatologists, dermatology residents, and general practitioners of diseases included in the algorithm ('in-distribution') training dataset

Category	*n*	Top-1 Accuracy				Top-3 Accuracy
		AI algorithm	Average dermatologist	Average resident	Average general practitioner	AI algorithm
General	305	45.3%	63.1%	60.9%	52.9%	69.2%
Balanced	305	47.6%	49.7%	47.7%	39.7%	72.2%
Inflammatory	210	40.5%	65.7%	61.3%	53.0%	67.1%
Dermatitis	80	35.0%	62.5%	52.9%	50.0%	60%
Acne/rosacea	75	49.3%	81.3%	81.8%	74.6%	80%
Autoimmune	30	30.0%	42.2%	41.1%	18.9%	46.7%
Papulosquamous	17	52.9%	52.9%	47.1%	33.3%	82.4%
Infectious	47	55.3%	58.2%	59.6%	56.7%	72.3%
Viral	22	50.0%	65.1%	62.1%	51.5%	68.2%
Fungal	18	61.1%	64.8%	70.4%	79.6%	77.8%
Bacterial	7	57.1%	19.1%	23.8%	14.3%	71.4%
Neoplastic	28	39.3%	52.4%	54.7%	44.1%	64.3%
Benign	24	45.8%	59.7%	62.5%	48.6%	70.8%
Malignant	4	0%	8.3%	8.3%	16.7%	25%
Alopecia	5	100%	80%	80%	66.7%	100%
Scarring	0	N/A	N/A	N/A	N/A	N/A
Non-scarring	5	100%	80%	80%	66.7%	100%
Other	15	73.3%	55.5%	64.4%	51.1%	86.7%

AI, Artificial Intelligence; D, dermatologist; R, residents; GP, general practitioners.