nature medicine

ARTICLES

https://doi.org/10.1038/s41591-020-0842-3

A deep learning system for differential diagnosis of skin diseases

Yuan Liu[1], Ayush Jain[1], Clara Eng[1], David H. Way[1], Kang Lee[1], Peggy Bui[1,2], Kimberly Kanada[3], Guilherme de Oliveira Marinho[4], Jessica Gallegos[1], Sara Gabriele[1], Vishakha Gupta[1], Nalini Singh[1,5], Vivek Natarajan[1], Rainer Hofmann-Wellenhof[6], Greg S. Corrado[1], Lily H. Peng[1], Dale R. Webster[1], Dennis Ai[1], Susan J. Huang[3], Yun Liu[1]✉, R. Carter Dunn[1,7] and David Coz[1,7]

Skin conditions affect 1.9 billion people. Because of a shortage of dermatologists, most cases are seen instead by general practitioners with lower diagnostic accuracy. We present a deep learning system (DLS) to provide a differential diagnosis of skin conditions using 16,114 de-identified cases (photographs and clinical data) from a teledermatology practice serving 17 sites. The DLS distinguishes between 26 common skin conditions, representing 80% of cases seen in primary care, while also providing a secondary prediction covering 419 skin conditions. On 963 validation cases, where a rotating panel of three board-certified dermatologists defined the reference standard, the DLS was non-inferior to six other dermatologists and superior to six primary care physicians (PCPs) and six nurse practitioners (NPs) (top-1 accuracy: 0.66 DLS, 0.63 dermatologists, 0.44 PCPs and 0.40 NPs). These results highlight the potential of the DLS to assist general practitioners in diagnosing skin conditions.

Skin and subcutaneous disease is the fourth leading cause of nonfatal disease burden globally, affecting 30–70% of individuals and prevalent in all geographies and age groups[1]. Skin disease is also one of the most common chief complaints in primary care, with 8–36% of patients presenting with at least one skin complaint[2,3]. However, dermatologists are consistently in short supply, particularly in rural areas, and consultation costs are rising[4,5]. Thus, the burden of triage and diagnosis commonly falls on non-specialists such as primary care physicians (PCPs), nurse practitioners (NPs) and physician assistants[6–8]. Because of limited knowledge and training in a specialty with hundreds of conditions[9], diagnostic accuracy of non-specialists is only 24–70%[10–13], despite the availability and use of references such as dermatology textbooks, UpToDate[14] and online image search engines[15]. Low diagnostic accuracy can lead to poor patient outcomes such as delayed or improper treatment.

To expand access to specialists, store-and-forward teledermatology has become more popular, with a 48% increase in US non-governmental programs between 2011 and 2016[16]. In store-and-forward teledermatology, digital images of affected skin areas, typically captured with digital cameras or smartphones, are transmitted along with other medical information to a dermatologist. The dermatologist then remotely reviews the case and provides consultation on the diagnosis, work-up, treatment and recommendations for follow-up. This approach has been shown to result in similar clinical outcomes to conventional consultation in dermatology clinics[17], and improved satisfaction from both patients and providers[18].

The use of artificial intelligence tools may be another promising method of broadening the availability of dermatology expertise. Recent advances in deep learning have facilitated the development of artificial intelligence tools to assist in diagnosing skin disorders from images. Many previous works have focused on the visual recognition of skin lesions from dermoscopic images[19–26], which requires a dermatoscope. However, dermatoscopes are usually inaccessible outside of dermatology clinics and are unnecessary for many common skin diseases. By contrast, others have considered clinical photographs for skin cancers[27], onychomycosis[28] and skin lesions from an educational website[29,30]. Several works have also reported comparable performance to experts on binary classification tasks (benign versus malignant) or on skin lesion conditions[22–24,27]. Despite the focus on individual skin lesions, routine practice more commonly sees non-cancerous conditions such as inflammatory dermatoses and pigmentary issues[31]. These skin problems have yet to be addressed despite their high prevalence and similarly low diagnostic accuracy by non-specialists[19–21,27–29,32–34]. Moreover, previous works have focused on predicting a single diagnosis, instead of a full differential diagnosis. A differential diagnosis is a ranked list of diagnoses that is used to plan treatments in the common setting of diagnostic ambiguity in dermatology, and can capture a more comprehensive assessment of a clinical case than a single diagnosis[35].

In this article, we report a deep learning system (DLS) to identify 26 of the most common skin conditions in adult cases that were referred for teledermatology consultation. As a secondary prediction, the DLS also outputs predictions for the full set of 419 skin conditions seen in this work. Our DLS provides several advances relative to previous work. First, instead of a single classification between a small number of conditions, our DLS provides a differential diagnosis across 26 conditions, including various dermatitides, dermatoses, pigmentary conditions, alopecia and lesions, to aid clinical decision making. Second, instead of relying only on images, the DLS leverages 45 types of data that are available to dermatologists in a teledermatology service, such as demographic information and medical history. Third, the DLS supports a variable number of input images, and the benefit of using multiple images was assessed. Finally, to understand the potential value of the DLS, we compared

[1]Google Health, Palo Alto, CA, USA. [2]University of California, San Francisco, San Francisco, CA, USA. [3]Advanced Clinical, Deerfield, IL, USA. [4]Adecco Staffing, Santa Clara, CA, USA. [5]Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Medical University of Graz, Graz, Austria. [7]These authors contributed equally: R. Carter Dunn, David Coz. ✉e-mail: liuyun@google.com

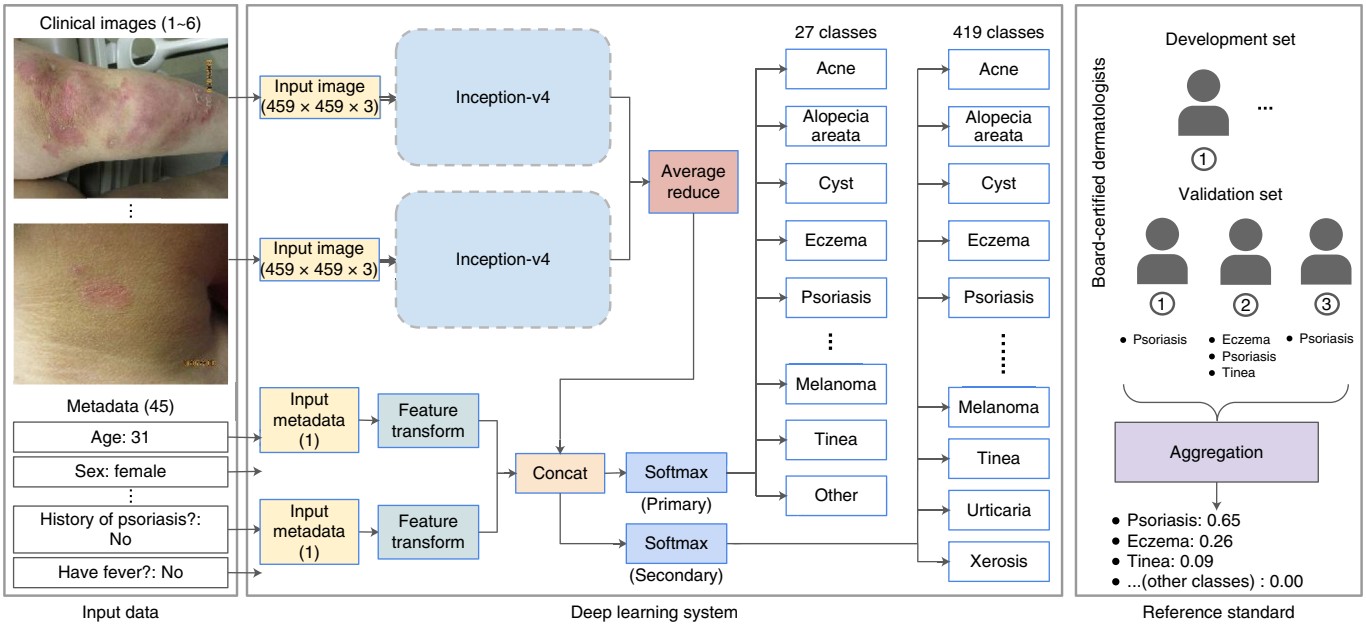


Fig. 1 | Overview of the development and validation of our DLS. For each case, the DLS takes as input one to six de-identified skin photographs and 45 metadata variables such as demographic information and medical history (left). The DLS then processes the images using Inception-v4 modules with shared weights before applying an average pool and concatenating with the metadata features. The primary output of the classification layer of the DLS is the relative likelihood of 27 categories (26 skin conditions plus 'other', Table 1). The secondary output is the relative likelihood of the full set of 419 skin conditions seen in this work. These conditions were chosen based on a granularity that could guide a non-dermatologist clinician to next steps in clinical care. The labels used to develop and validate the DLS were provided by board-certified dermatologists (one or more dermatologists per case for training and three dermatologists per case for the validation set). For each case, each dermatologist provided their top three differential diagnoses. The multiple differential diagnoses are then aggregated into a single ranked list (see 'Reference standard labeling validation set' in the Methods). During training, the aggregated ranked list of dermatologist-provided diagnoses have an associated aggregated 'confidence' score per diagnosis, and these confidences are the target 'soft' labels for the DLS. The DLS therefore learns from both the primary (top-ranked) diagnosis as well as the lower-ranked diagnoses. In this way, the DLS was trained to provide a differential diagnosis instead of a single prediction output.

its diagnostic accuracy with that of board-certified clinicians with three different levels of training: dermatologists, PCPs and NPs.

Results

Overview of approach. Our DLS has two major components: a variable number of deep convolutional neural network modules to process a flexible number of input images, and a shallow module to process metadata such as demographic information and medical history (Fig. 1 and Supplementary Table 1). To develop and validate our DLS, we applied a temporal split[36] to teledermatology cases: the first 80% of the cases (years 2010–2017) for development and the last 20% (years 2017–2018) for validation (Table 1 and Supplementary Fig. 1). The reference standard for each case was determined by the aggregated opinions of multiple dermatologists who reviewed the case independently (see Methods). After excluding cases with multiple skin conditions and those that were non-diagnosable, 16,114 cases (64,837 images) were used for development and 3,756 cases (14,883 images) for validation (validation set 'A'; a smaller subset 'B' was used for comparison with clinicians and is described in the relevant sections). In total, 64,878 dermatologist reviews were collected for development and 11,268 reviews for validation.

DLS performance. The DLS's top differential diagnosis in validation set A had a 'top-1 accuracy' (accuracy across all cases for 26 skin conditions + 'other') of 0.71 and a 'top-1 sensitivity' (sensitivity computed for the 26 conditions separately and averaged) of 0.58 (Fig. 2a and Supplementary Table 2). When the DLS was allowed three diagnoses (for example, to mimic a clinical decision support tool that suggests a few possibilities for the clinician's consideration), the DLS's top-3 accuracy rose to 0.93 and its top-3 sensitivity across 26 conditions rose to 0.83. To ensure that the DLS was not biased against different skin tones, we evaluated DLS accuracy stratified by estimated Fitzpatrick skin type (Table 1) and self-reported race/ethnicity (Supplementary Table 3). Among the Fitzpatrick skin types that comprised at least 2% of the data (types II–V), the top-1 accuracy ranged from 0.70 to 0.74 and the top-3 accuracy ranged from 0.91 to 0.95. Additional subanalyses based on self-reported demographic information (that is, age, sex, race, ethnicity) are presented in Supplementary Table 3. Evaluation of the DLS's overall differential diagnosis using the average overlap (AO) metric[37–39] yielded 0.68 overall (Fig. 2b) and 0.66–0.69 when stratified by Fitzpatrick skin types (types II–V, Supplementary Table 3). The DLS performance for each of the 26 conditions is presented in Extended Data Fig. 1a and the results across 419 skin conditions are presented in Extended Data Fig. 2.

DLS performance compared with clinicians. To compare the DLS performance with that of clinicians, validation set A was randomly subsampled (stratified by condition). This produced 963 cases (3,707 images; validation set B) enriched for rarer conditions (for example, 2–5% prevalance in B compared to below 1% in A). Eighteen clinicians of three different levels of training (dermatologists, PCPs and NPs, all of whom were board-certified) graded validation set B. On this smaller dataset, the DLS achieved a top-1 accuracy of 0.66, compared to 0.63 for dermatologists, 0.44 for PCPs and 0.40 for NPs (Fig. 2a). The DLS was non-inferior to the dermatologists at a 5% margin ($P < 0.001$). The top-3 accuracy was substantially higher at 0.90 for the DLS, compared to 0.75 for dermatologists, 0.60 for PCPs and 0.55 for NPs. Consistent with the top-1 and top-3 accuracies, evaluation of the full differential diagnosis using

Table 1 | Dataset characteristics

Characteristics	Development set	Validation set A	Validation set B (enriched subset of 'A')
Years	2010 to 2017	2017 to 2018	2017 to 2018
Total no. of cases	16,530	4,146	N/A
No. of cases with multiple skin conditions (excluded from study)	142	224	N/A
No. of cases indicated as not-diagnosable by dermatologists (excluded from study)	271	165	N/A
No. of cases included in study	16,114	3,756	963
No. of images included in study	64,837	14,883	3,707
No. of patients included in study	12,399	3,241	933
Age[a], years: median (25th, 75th percentiles)	40 (28, 54)	40 (28, 54)	43 (30, 56)
Female (%)	10,014 (62.1%)	2,371 (63.1%)	615 (63.9%)
Race and ethnicity			
American Indian or Alaska Native (%)	142 (0.1%)	42 (0.1%)	9 (0.9%)
Asian (%)	1,775 (11.0%)	473 (12.6%)	97 (10.1%)
Black or African American (%)	1,087 (6.8%)	229 (6.1%)	61 (6.3%)
Hispanic or Latino (%)	7,044 (43.7%)	1,631 (43.4%)	409 (42.5%)
Native Hawaiian or Pacific Islander (%)	224 (1.4%)	61 (1.6%)	19 (2.0%)
White (%)	5,475 (34.0%)	1,175 (31.3%)	329 (34.2%)
Not specified (%)	367 (2.2%)	145 (3.9%)	39 (4.0%)
Fitzpatrick skin types (6 types)[b]			
Type I (%)	46 (0.3%)	9 (0.2%)	0 (0.0%)
Type II (%)	2,807 (17.4%)	383 (10.2%)	104 (10.8%)
Type III (%)	6,641 (41.2%)	2,412 (64.2%)	607 (63.0%)
Type IV (%)	5,040 (31.3%)	724 (19.3%)	195 (20.2%)
Type V (%)	510 (3.2%)	101 (2.7%)	24 (2.5%)
Type VI (%)	46 (0.3%)	1 (0.0%)	0 (0.0%)
Unknown (%)	1,024 (10.2%)	126 (3.4%)	33 (3.4%)
Skin conditions based on primary diagnosis (26 conditions, plus 'other')[c]			
Acne (%)	1,645 (10.2%)	407 (10.8%)	40 (4.2%)
Actinic keratosis (%)	223 (1.4%)	53 (1.4%)	37 (3.9%)
Allergic contact dermatitis (%)	173 (1.1%)	36 (0.9%)	25 (2.6%)
Alopecia areata (%)	323 (2.0%)	96 (2.5%)	37 (3.8%)
Androgenetic alopecia (%)	185 (1.1%)	50 (1.3%)	33 (3.4%)
Basal cell carcinoma (%)	285 (1.8%)	45 (1.2%)	28 (2.9%)
Cyst (%)	279 (1.7%)	86 (2.3%)	31 (3.2%)
Eczema (%)	2,368 (14.7%)	659 (17.5%)	50 (5.2%)
Folliculitis (%)	361 (2.2%)	87 (2.3%)	32 (3.3%)
Hidradenitis (%)	161 (1.0%)	45 (1.2%)	35 (3.6%)
Lentigo (%)	113 (0.7%)	33 (0.9%)	32 (3.3%)
Melanocytic nevus (%)	792 (4.9%)	183 (4.9%)	34 (3.6%)
Melanoma (%)	93 (0.6%)	22 (0.6%)	19 (1.9%)
Post inflammatory hyperpigmentation (%)	162 (1.0%)	51 (1.4%)	29 (3.0%)
Psoriasis (%)	1,983 (12.3%)	335 (8.9%)	39 (4.1%)
Squamous cell carcinoma/squamous cell carcinoma in situ (SCC/SCCIS) (%)	148 (0.9%)	37 (1.0%)	33 (3.5%)
Seborrheic keratosis/irritated seborrheic keratosis (SK/ISK) (%)	745 (4.6%)	211 (5.6%)	38 (4.0%)
Scar condition (%)	297 (1.8%)	60 (1.6%)	33 (3.4%)
Seborrheic dermatitis (%)	365 (2.3%)	98 (2.6%)	37 (3.8%)
Skin tag (%)	239 (1.5%)	70 (1.9%)	33 (3.4%)
Stasis dermatitis (%)	124 (0.8%)	26 (0.7%)	25 (2.6%)
Tinea (%)	229 (1.4%)	34 (0.9%)	31 (3.2%)

Continued

Table 1 | Dataset characteristics (Continued)

Characteristics	Development set	Validation set A	Validation set B (enriched subset of 'A')
Tinea versicolor (%)	200 (1.2%)	36 (0.9%)	35 (3.6%)
Urticaria (%)	127 (0.8%)	34 (0.9%)	33 (3.4%)
Verruca vulgaris (%)	374 (2.3%)	83 (2.2%)	34 (3.5%)
Vitiligo (%)	217 (1.3%)	74 (2.0%)	36 (3.7%)
Other (%)	3,902 (24.2%)	809 (21.5%)	96 (10.0%)

The dataset contained clinical cases from a teledermatology practice serving 17 primary care and specialist sites from two states in the United States. The dataset was split temporally into a development set (cases seen between 2010 and 2017) and validation set A (cases seen between 2017 and 2018). Validation set B was a subset of set A that was enriched for rarer skin conditions in this study and was reviewed by three groups of clinicians for comparison. See Supplementary Fig. 1 for more details. [a]Ages were truncated at 90 years as part of the de-identification process. For each dataset, the minimum age was 18 years and the maximum age was 90 years. [b]Fitzpatrick skin type was obtained via the majority opinion of three raters trained by dermatologists to distinguish skin types and involved an assessment of both exposed skin and skin that was visible under clothing. Importantly, this was an estimate that does not fully recapitulate the melanin index as objectively measured from reflectance spectrophotometry or patient-reported response to sun exposure. Some cases' skin types were labeled as 'unknown' because of reasons such as lack of majority agreement among raters, inconsistent skin types observed in different images and insufficient visible skin regions. [c]When multiple primary diagnoses exist, the contribution of each condition in the list towards its total count is fractionalized, such that the total number of cases over all conditions sums up to the size of each dataset. This causes a slight difference when compared to the numbers as part of the *x*-axes labels in Extended Data Fig. 1, where each condition is treated independently.

the AO metric yielded 0.63 for the DLS, compared with 0.58 for dermatologists, 0.46 for PCPs and 0.42 for NPs. The average top-1 and top-3 sensitivities across the 26 conditions followed the same trend (Extended Data Fig. 1b and Supplementary Table 2). When the comparator clinicians' confidence in their primary diagnosis was lower, the clinicians' accuracies dropped, whereas the DLS's top-1 accuracy remained high (Fig. 2c,d). Representative examples of cases that were missed by PCPs or NPs are shown in Fig. 3a–e and Supplementary Fig. 2. Corresponding results for the clinicians compared with the DLS across all 419 conditions are presented in Extended Data Fig. 2 and Supplementary Table 4.

Subgroup analysis. Next, we assessed the ability of DLS to distinguish between conditions that present similarly and can be misidentified in clinical settings (see the 'Conditions in the subcategory' column of Table 2 for definitions of the subgroups). The first analysis distinguished between malignant versus benign growths. Note that, in this and subsequent subanalyses, the DLS and clinicians could have determined the case as belonging to neither category (for example, neither a malignant nor a benign growth; that is, not a growth at all). Because the decision to biopsy depends on whether malignant conditions are in the differential, in this 'growths' subgroup analysis, we focused on the top-3 sensitivity for malignant growths. The DLS's top-3 sensitivity of 0.90 was comparable with that of dermatologists (0.89) and higher than that of both PCPs and NPs (0.69 and 0.72, respectively; Table 2a). Similar trends were observed when restricting the analysis to cases with biopsy confirmation (Table 2b and Supplementary Table 5).

The second subgroup analysis distinguished between infectious versus non-infectious cases of erythematosquamous and papulosquamous skin diseases. The DLS had a higher top-1 sensitivity than the PCPs and NPs at identifying the infectious subcategory (0.71 for the DLS versus 0.54 for PCPs and 0.49 for NPs) and comparable to dermatologists. For the non-infectious subcategory, the DLS had a higher top-1 sensitivity than all clinicians (0.63 for the DLS versus 0.42–0.48 for clinicians).

The last subgroup deals with two types of hair loss: alopecia areata and androgenetic alopecia. The top-1 sensitivity of the DLS for alopecia areata (0.84) was higher than for PCPs (0.56) and NPs (0.48) and similar to dermatologists (0.84). For androgenetic alopecia, the DLS's top-1 sensitivity of 0.82 was higher than that of the clinicians (range 0.27–0.71).

Importance of input data. We examined the importance of different input data to the DLS. Among the 45 types of non-image metadata (demographic information and medical history, see Supplementary Table 1), the type of self-reported skin problem (for example, acne, hair loss or rash) and history of psoriasis had the greatest impact on accuracy (Fig. 4a and Supplementary Fig. 3).

For image inputs, the performance of DLS improved dramatically when more than one image was provided and plateaued when there were at least five images (Fig. 4b, blue line). This trend was preserved when the non-image metadata were also withheld from the DLS (Fig. 4b, red line). Compared to withholding metadata from the DLS that was developed in the presence of metadata, training another DLS that uses only images (so it does not 'rely' on metadata) yielded a small improvement (Fig. 4b, green line). Finally, saliency analysis via integrated gradients[40] highlighted those regions of the image where a skin condition was visible, suggesting the DLS had generally learned to focus on the correct region of interest when making the prediction (Fig. 3a–e).

We also examined the effect of training dataset size on the performance of the DLS and observed that more training data led to a better top-1 accuracy, though with diminishing return after 10,000 cases (Supplementary Fig. 4).

Discussion

In this study, we have developed and validated a DLS to identify 26 common skin conditions that were referred by primary care for a teledermatology consultation, representing ~80% of cases seen in primary care[1,31,41–43]. The DLS's top-1 diagnostic accuracy was non-inferior to dermatologists and higher than PCPs and NPs. Moreover, the DLS's high top-3 accuracy and AO metric suggest that the DLS's full differential diagnosis is relatively complete, and may help alert clinicians to differential diagnoses that they may not have considered. As a secondary prediction, the DLS also tackles a broader set of 419 skin conditions, with encouragingly similar results.

Providing assistance with a differential diagnosis instead of a single diagnosis is particularly important in dermatology. Because most skin conditions are not verified with pathology, the differential diagnosis is used for decision making around work-up and treatment. If all conditions in the differential diagnosis share the same treatment, a single diagnosis may not be clinically necessary. If the diagnoses on the differential have opposing treatments (for example, treatment for one condition on the differential may aggravate another diagnosis on the differential), a clinician can still consider this group of diagnoses together to determine a work-up or initiate treatment. In all these situations, the DLS may be an effective aid to non-specialists by helping them to arrive at both a

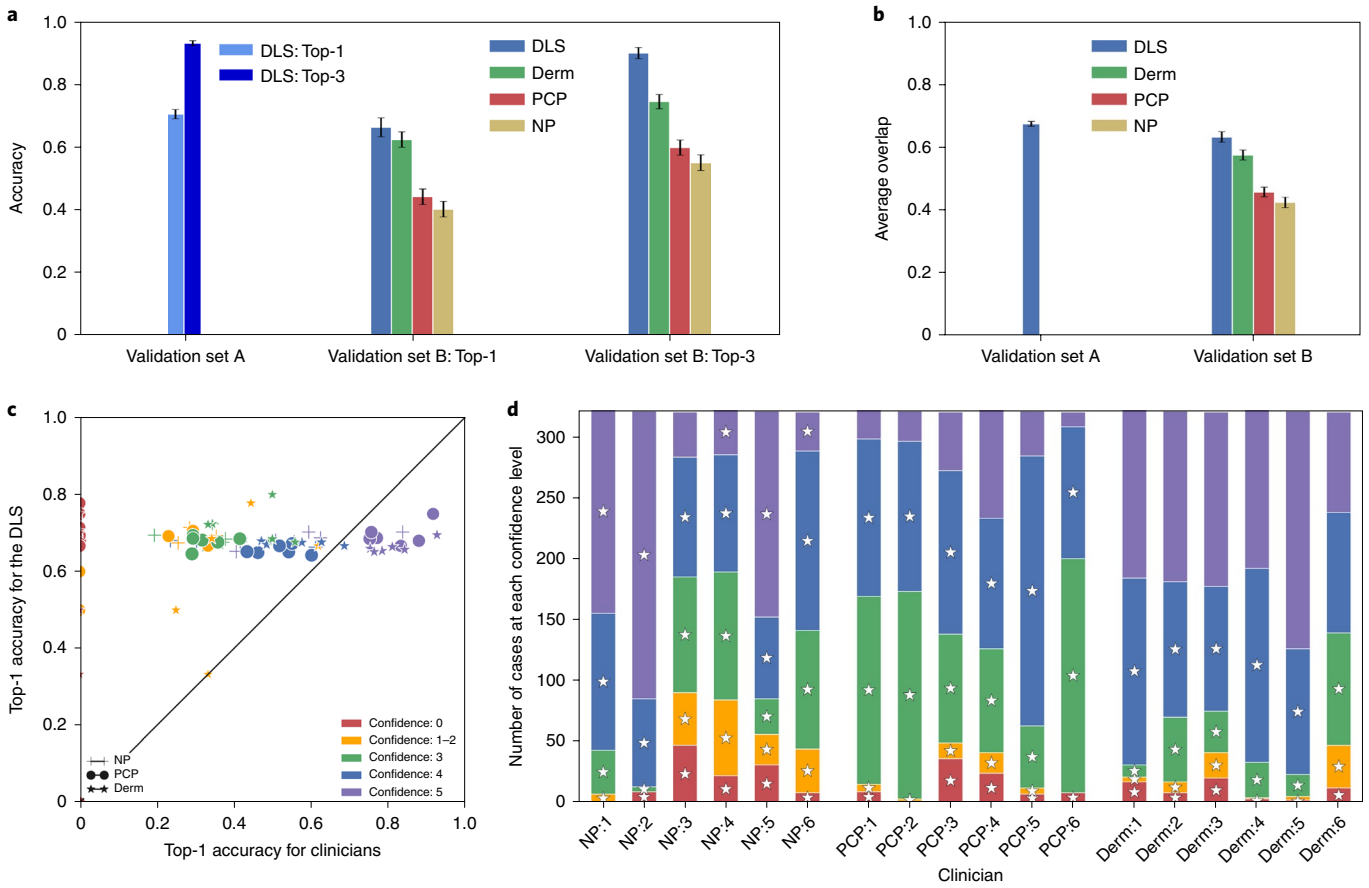


Fig. 2 | Performance of the DLS and the dermatologists (Derm), primary care physicians (PCPs) and nurse practitioners (NPs). a, Top-1 and top-3 accuracy for the DLS and clinicians across all cases and 27 categories of skin conditions on validation set A ($n=3{,}756$) and validation set B ($n=963$). The sensitivity of the DLS for each of the 26 conditions is presented in Extended Data Fig. 1. Results for the full spectrum of 419 conditions are presented in Extended Data Fig. 2. **b**, Average overlap (AO) (to assess the full differential diagnosis) of the DLS and clinicians on both validation sets. The AO ranges from 0 to 1, with higher values indicating better agreement. Error bars in **a** and **b** indicate 95% confidence intervals (see 'Statistical analysis' section). **c**, Top-1 accuracy of the DLS and comparator clinicians on validation set B, stratified by the confidence of the comparator clinician. A confidence of 0 indicates cases that the clinician graded as 'unknown', while confidences 1 and 2 were rarely used and thus merged. Points indicate clinicians (with the category represented by the shape), colors indicate confidence levels, and sizes of symbols are proportional to the number of cases. The diagonal line has slope 1; points to its left indicate DLS performance being better than clinicians and vice versa. The distributions of the number of cases with each confidence level are reported in **d**. **d**, Stacked bar plots of number of cases at each confidence level for each clinician in validation set B. Colors indicate the confidence level of the cases as in **c**. An asterisk indicates that for the cases contained in the sub-bar, the DLS was more accurate than the clinician.

more accurate primary diagnosis and a more complete differential diagnosis. Dermatologists in 'store-and-forward' teledermatology (where dermatologists review cases asynchronously) could also potentially use such a DLS to help triage cases rapidly.

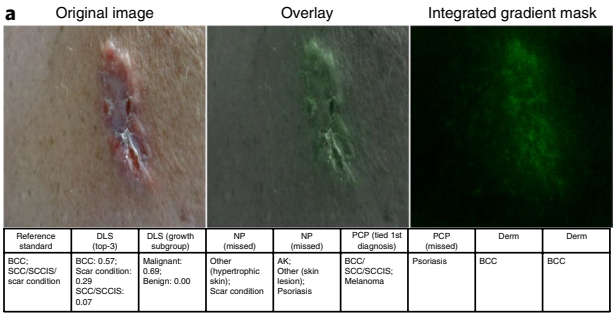
To better understand the impact of the DLS in specific challenging diagnostic situations, our subanalyses examined subgroups containing conditions with similar visual presentations and where distinguishing between conditions in different groups can affect subsequent clinical decisions. The three subgroup analyses were as follows: individual growths, erythematosquamous and papulosquamous skin disease, and hair loss. For growths, malignant lesions should have subsequent biopsy or excision, whereas patients with benign lesions can be reassured. The top-3 sensitivity for malignant lesions is important because the inclusion of a diagnosis of malignancy on the differential diagnosis may prompt a clinician to obtain a specimen for pathology, even if it is not the primary suspected diagnosis. In erythematosquamous and papulosquamous skin disease, these eruptions can be clinically similar to erythema and scaling, although they can have very different etiologies and treatment plans. An accurate diagnosis is particularly important, because first-line treatment of non-infectious entities is often with a topical steroid, which conversely would make an infectious process like tinea more resistant to treatment and can even hinder diagnosis (for example, tinea incognito[44]) at future appointments. The inclusion of tinea as a potential diagnosis can also prompt a clinician to do a KOH exam for confirmation. For hair loss, the two conditions have different etiologies, possible work-up and treatment options. Distinguishing one from the other could allow a clinician to start first-line therapies and possible work-up for these conditions. In the first subgroup, the DLS was very 'specific'; that is, it was able to correctly identify the 'negative' subcategory of benign growths. Despite a lower top-1 sensitivity for malignant lesions, the DLS had a high top-3 sensitivity, which is on par with dermatologists. For the second and third subgroups, the DLS had substantially higher sensitivities than non-specialists, suggesting that the DLS may be particularly valuable in helping determine the work-up or initiate treatment based on a working diagnosis.

Diagnostic uncertainty exists in our non-biopsied reference standard and dermatology more generally. Two conditions seemed particularly challenging based on low clinician accuracies (Extended

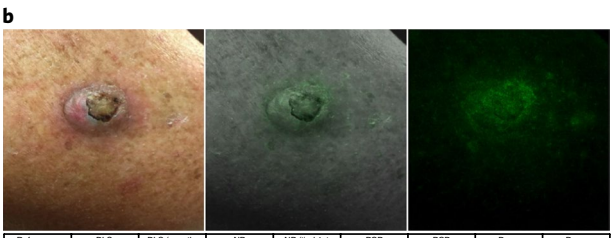
Data Fig. 1b) and inter-dermatologist agreement (Supplementary Table 6): allergic contact dermatitis (ACD) and post-inflammatory hyperpigmentation (PIH). Manual review of those cases by two other dermatologists showed that eight ACD cases were deemed clinically difficult because they lacked a 'classic' visual presentation. Additionally, despite guidance to be as specific as possible, some clinicians used the general term 'contact dermatitis', which also encompasses irritant contact dermatitis, which has different etiology, work-up and treatment interventions. This lack of specificity had prompted us to (a priori) categorize contact dermatitis under 'Other' and consider it 'incorrect'. On the other hand, PIH was commonly 'misdiagnosed' due to attempts to label what the primary process leading to the PIH could have been. To ensure that these complexities did not cause our DLS evaluations to be overly optimistic, we recomputed the sensitivities excluding these two conditions for the DLS and all clinicians (Supplementary Table 7), finding small improvements for both the DLS and clinicians, with no change in conclusions. We also repeated the analysis on cases with less diagnostic ambiguity (that is, where two or three of the reference standard dermatologists agreed on the primary diagnosis) and observed similar results (Supplementary Figs. 5 and 6).

Our study showed that dermatologists were substantially more accurate than PCPs and NPs. The finding was not surprising, as the majority of the cases were sent by primary care providers to a teledermatology service, and presumably the clinician had found them difficult to diagnose. Although not strictly comparable because of differences in study design, the low accuracies we observed (35–49%, Supplementary Table 8) are in line with those previously reported (24–70%[10–13]). These numbers serve to highlight the challenging nature of this classification task that incorporates both visual and non-visual information, and underscores the need for decision support tools for non-specialists.

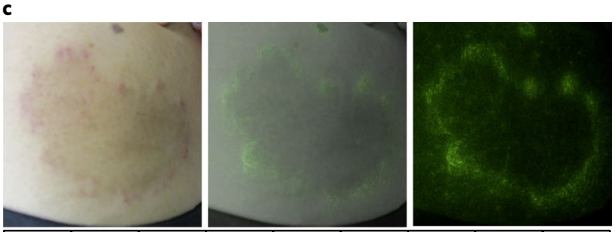
Previous studies in this area generally have not focused on providing diagnostic assistance in a more generalized workflow, but instead have focused on early screening of skin cancer and thus were limited to a narrower scope of conditions (for example, melanoma or not) or on more standardized images that require specialized equipment (that is, dermoscopic images). Of the studies that attempted to tackle a broader range of conditions[29,32,34,45], the datasets were often either educational in nature, leading to potential bias towards cases with more typical presentations or unusually severe cases that prompted pathologic confirmation[29,32,45], or a simplification of labels towards a mix of morphological descriptions



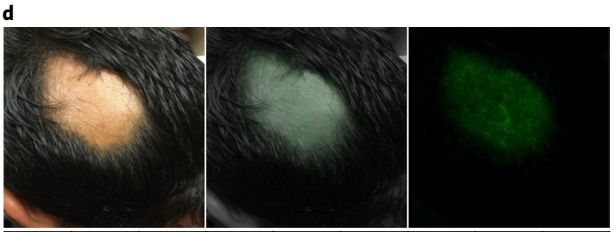
Reference standard	DLS (top-3)	DLS (growth subgroup)	NP (missed)	NP (missed)	PCP (tied 1st diagnosis)	PCP (missed)	Derm	Derm
BCC; SCC/SCCIS/ scar condition	BCC: 0.57; Scar condition: 0.29 SCC/SCCIS: 0.07	Malignant: 0.69; Benign: 0.00	Other (hypertrophic skin); Scar condition	AK; Other (skin lesion); Psoriasis	BCC/ SCC/SCCIS; Melanoma	Psoriasis	BCC	BCC



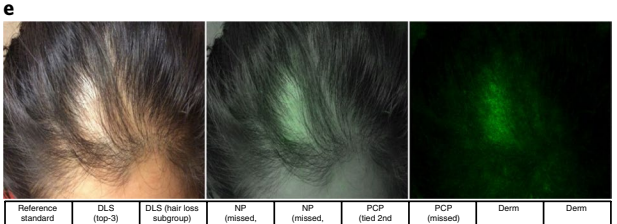
Reference standard	DLS (top-3)	DLS (growth subgroup)	NP (2nd diagnosis)	NP (tied 1st diagnosis)	PCP (missed)	PCP (missed)	Derm	Derm
SCC/SCCIS; BCC	SCC/SCCIS: 0.69; BCC: 0.19; AK: 0.07	Melanoma: 0.93; Benign: 0.07	BCC; SCC/SCCIS; Melanoma	Other (skin lesion)/ SCC/SCCIS; BCC	Cannot diagnose	Other (pyoderma)	SCC/SCCIS; BCC	SCC/SCCIS



Reference standard	DLS (top-3)	DLS (erythemato-squamous and papulo-squamous subgroup)	NP (missed)	NP (missed)	PCP (tied 1st diagnosis)	PCP (missed)	Derm	Derm
Tinea	Tinea: 0.68; Eczema: 0.14; Other: 0.03	Infectious: 0.80; Non-infectious: 0.16	Eczema/ Other (chronic contact dermatitis); Psoriasis	Other (generalized granuloma annulare)	Other (generalized annulare)/ Tinea	Eczema	Tinea; Other (granuloma annulare)	Tinea



Reference standard	DLS (top-3)	DLS (hair loss subgroup)	NP (3rd diagnosis)	NP (missed)	PCP (2nd diagnosis)	PCP	Derm	Derm
AA	AA: 0.98; Other: 0.01; AGA: 0.01	AA: 0.98; AGA: 0.01	AGA; Other (alopecia localis); AA	AGA	AGA; AA	AA	AA	AA; Other (trichotillomania)



Reference standard	DLS (top-3)	DLS (hair loss subgroup)	NP (missed, non-specific)	NP (missed, non-specific)	PCP (tied 2nd diagnosis)	PCP (missed)	Derm	Derm
AGA; AA; Other	AGA: 0.73; Other: 0.13; Seborrheic dermatitis: 0.08	AGA: 0.78; AA: 0.00	Other (diffuse alopecia)	Other (alopecia)	Other (telogen effluvium); AA/ AGA	AA	AGA; Other (drug-related alopecia)	AGA

Fig. 3 | Representative examples of challenging cases missed by non-dermatologists. For each case, an original image is provided on the left and an image with a saliency mask on the right. The middle image shows the original image, with the saliency overlaid in green. All clinicians were instructed to be as specific as possible when providing the diagnostic labels. Diagnoses for the reference standard and comparator clinicians who reviewed each case are included here and ranked by confidence from top to bottom. **a**, The DLS's primary diagnosis of basal cell carcinoma (BCC) concurs with the reference standard, both comparator dermatologists and one PCP. Both NPs and one PCP missed this diagnosis. **b**, The DLS's primary diagnosis of squamous cell carcinoma (SCC/SCCIS) concurs with the reference standard and both comparator dermatologists. Both NPs considered another diagnosis as more or equally likely, while the PCPs missed this diagnosis. **c**, The DLS's primary diagnosis of tinea concurs with the reference standard and primary diagnoses of the comparator dermatologists. One PCP considered another diagnosis as equally likely, while the other PCP and both NPs missed this diagnosis. **d**, The DLS, comparator dermatologists and one PCP all agreed with the reference standard of alopecia areata (AA). This was missed as a primary diagnosis by both NPs and one of the PCPs. **e**, The DLS's primary diagnosis of androgenetic alopecia (AGA) concurs with the reference standard and both comparator dermatologists. This was missed as a primary diagnosis by both NPs and both PCPs. In the last two cases (**d**,**e**), diagnosing the specific type of alopecia is important because AGA and AA have different treatments. More details about these cases are presented in Supplementary Fig. 2.

Table 2 | Sensitivity of the deep learning system (DLS) and three types of clinician (dermatologist, primary care physician and nurse practitioner) for clinically relevant and challenging subgroups based on appearance on clinical presentation

(a)

Clinically relevant groups	Subcategory	Conditions in the subcategory	No. of cases	Top-1 sensitivity				Top-3 sensitivity			
				DLS	Derm	PCP	NP	DLS	Derm	PCP	NP
Growth	Malignant	BCC, melanoma, SCC/SCCIS	83	0.57	**0.74**	0.52	0.51	**0.90**	0.89	0.69	0.72
	Benign	Actinic keratosis, cyst, lentigo, melanocytic nevus, SK/ISK, skin tag, verruca vulgaris	178	**0.74**	0.66	0.56	0.44	**0.92**	0.76	0.70	0.60
Erythematosquamous and papulosquamous skin disease	Infectious	Tinea, tinea versicolor	66	0.71	**0.72**	0.54	0.49	**0.88**	0.85	0.70	0.60
	Non-infectious	Eczema, psoriasis, stasis dermatitis, allergic contact dermatitis, seborrheic dermatitis	202	**0.63**	0.42	0.48	0.45	**0.92**	0.55	0.62	0.57
Hair loss	Alopecia areata	Alopecia areata	37	**0.84**	**0.84**	0.56	0.48	0.86	**0.91**	0.77	0.64
	Androgenetic alopecia	Androgenetic alopecia	34	**0.82**	0.71	0.30	0.27	**0.97**	0.84	0.43	0.37

(b)

	Validation set A (52 malignancies out of $n=100$ cases)		Validation set B (enriched subset of set A; 37 malignancies out of $n=53$ cases)							
	Sensitivity	PPV[a]	Sensitivity				PPV[a]			
	DLS	DLS	DLS	Derm	PCP	NP	DLS	Derm	PCP	NP
Top-1	0.48	0.96	0.57	**0.74**	0.58	0.55	**0.95**	0.88	0.86	0.88
Top-3	0.88	0.69	**0.92**	0.89	0.81	0.73	0.76	0.81	**0.86**	0.84

a, Performance relative to the dermatologists-determined reference standard. **b**, Performance relative to a biopsy-determined reference standard (where available). Distributions of biopsy cases are reported in Supplementary Table 5. Bold indicates the highest value for each subcategory and each evaluation metric. [a]PPV, positive predictive value; of cases predicted or diagnosed to have a malignant condition in the top-3, what fraction were found to be malignant on biopsy.

(for example, erythema/'redness') and specifications too broad to guide clinical next steps (for example, hair loss without further details)[34]. As a result, the utility of these works in actual clinical settings is unclear. By contrast, the images in our data were taken by medical assistants across 17 sites, representing a wide variety of lighting conditions, perspectives and backgrounds. Our dataset is also representative of cases that require dermatology consultations, and the conditions that our DLS predicts are specific enough to guide a clinician to the next steps in clinical care. However, due to the impracticality of performing exhaustive tests or biopsies for all skin conditions, there exists inevitable diagnostic uncertainty in actual clinical settings. To help resolve this, our DLS learns to predict a differential diagnosis instead of a single diagnosis, enabling a decision support tool that surfaces potential diagnoses for clinicians to consider.

Our DLS can potentially augment the current clinical workflow in a primary-care setting in several ways. First, the DLS can prompt clinicians to include on their differential a diagnosis that they would not have considered previously, particularly for difficult cases that they might have otherwise referred. Our data (Fig. 2c,d) suggest that clinicians could identify low-confidence cases and leverage the DLS's higher accuracy on those cases. This applies to most NPs in this study, regardless of confidence level, while all PCPs would have benefited at confidence ≤4. The DLS may thus prevent misdiagnosis, delay to care, and improper treatment, which can lead to poor clinical outcomes, a bad patient experience and increased costs of care. Second, by helping to improve the accuracy of non-dermatologists, the referrals to dermatologists may be more appropriate. With challenges to access, it is important to identify referred cases as urgent versus non-urgent. If the non-dermatologist clinician provides a more accurate diagnostic assessment at the time of referral, the patient can be more appropriately triaged for an appointment.

From a technical aspect, while most previous work used a single image as input, our DLS integrates information from both metadata and multiple images. We further quantify the magnitude of improvement as metadata or more images are provided for each case. Similarly, dermatologists in a teledermatology setting look at multiple images to better appreciate the three-dimensional and textural aspects of the skin findings. We also show that visual features alone enable reasonable diagnostic accuracy by the DLS, and accuracy improves with more images, albeit with diminishing returns. This has implications for broader real-world usage: a single image is likely to be suboptimal, but more than five provides marginal benefit. Adding metadata provides another 4–5% improvement, with most of the benefit coming from a handful of features. Thus, a few questions may be sufficient to capture most of the diagnostic accuracy benefits. Moreover, even the most 'important' metadata (the type of self-reported skin problem) caused a small 0.8% reduction in top-1 accuracy when permuted (see Methods), suggesting robustness of the DLS to metadata error.

Our study has limitations. First, we lacked a completely external dataset for validation, but instead split the data temporally. This mimics developing a DLS using retrospective data and then validating that DLS at the same sites on data collected over the next year. To aid generalization beyond the specific metadata in this dataset, we also trained a version of the DLS that uses only images as input (Fig. 4b), which may be more easily applicable to practices without or with different metadata. Second, our data did not include additional testing, and only a subset of suspected malignancies had biopsy confirmation. Instead, our reference standard for each case was based on aggregating the differential diagnoses of a panel

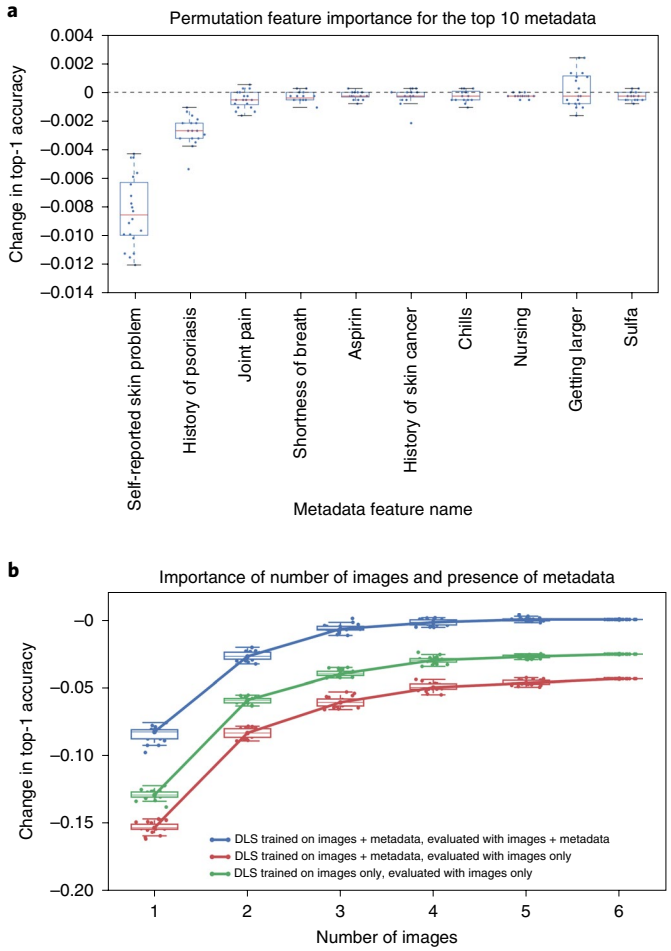


Fig. 4 | Importance of different inputs to the DLS. a, Impact on the top-1 accuracy of permuting each of the top 10 most important clinical metadata across validation set A examples, using the same trained DLS (for results for all metadata see Supplementary Fig. 3). **b**, The blue line illustrates the impact on the top-1 accuracy of different numbers of input images for the same DLS (that was trained using all images and metadata). The red line illustrates a similar trend when the clinical metadata are absent from this same DLS. Finally, the green line illustrates the trend, but for a DLS retrained without using clinical metadata (so that the DLS cannot depend on the presence of clinical metadata). All trends were the average of 20 different runs to reduce the effects of stochasticity from the permutation, image sampling and/or training process. In both panels, results are presented as boxplots with raw points ($n = 20$) indicated by dots. Edges of boxes indicate quartiles, red lines indicate the median, whiskers represent the ranges and outliers are defined by 1.5 times the interquartile range.

of board-certified dermatologists ('collective intelligence'; see Methods and Supplementary Methods). Ambiguities in diagnosis do exist in clinical practice, which makes it challenging to evaluate the accuracy of clinicians and DLS, especially for conditions like rashes, which are not typically biopsied. Third, as our dataset was de-identified, only structured metadata were available. Although useful, it is less rich than free text clinical notes or an in-person examination. With regard to the top-3 metrics, although instructed to, some clinicians provided fewer than three diagnoses when sufficiently confident in their first few diagnoses. The clinicians may thus have higher top-3 metrics if forced to provide at least three diagnoses. Additionally, actual clinical cases may include multiple concurrent conditions, which were excluded from this study (Table 1). In principle, multiple conditions may be handled as several single-condition diagnoses, although treatment plans may be more complex. Importantly, our validation data were limited with respect to the uncommon skin types[46] (0.2% type I, 2.7% type V and 0% type VI); further validation on these skin types will be needed to complement the race and ethnicity analysis in Supplementary Table 6. Future work will also need to assess the generalizability of the DLS to the full spectrum of cases seen in primary care by including cases that were not referred, data from additional sites and settings spanning more countries and states, and cases imaged on a greater variety of devices. Finally, additional studies are needed to better understand the optimal implementation of such tools and assess their impact in clinical practice.

To conclude, we have developed a DLS to identify 26 common skin conditions at a level comparable to board-certified dermatologists and more accurate than general practitioners. Our approach could be directly applied to store-and-forward teledermatology by assisting clinicians in triaging cases, thus shortening wait times for specialty care and reducing morbidity from skin diseases. Within (in-person) primary care, our algorithm could help improve the accuracy of non-dermatologists for cases that might otherwise have been referred, thus allowing the treatment to be initiated instead of waiting for referrals.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-020-0842-3.

Received: 11 September 2019; Accepted: 19 March 2020;
Published online: 18 May 2020

References

1. Hay, R. J. et al. The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions. *J. Invest. Dermatol.* **134**, 1527–1534 (2014).
2. Lowell, B. A., Froelich, C. W., Federman, D. G. & Kirsner, R. S. Dermatology in primary care: prevalence and patient disposition. *J. Am. Acad. Dermatol.* **45**, 250–255 (2001).
3. Awadalla, F., Rosenbaum, D. A., Camacho, F., Fleischer, A. B. Jr & Feldman, S. R. Dermatologic disease in family medicine. *Fam. Med.* **40**, 507–511 (2008).
4. Feng, H., Berk-Krauss, J., Feng, P. W. & Stein, J. A. Comparison of dermatologist density between urban and rural counties in the United States. *JAMA Dermatol.* **154**, 1265–1271 (2018).
5. Resneck, J. & Kimball, A. B. The dermatology workforce shortage. *J. Am. Acad. Dermatol.* **50**, 50–54 (2004).
6. Johnson, M. L. On teaching dermatology to nondermatologists. *Arch. Dermatol.* **130**, 850–852 (1994).
7. Ramsay, D. L. & Weary, P. E. Primary care in dermatology: whose role should it be? *J. Am. Acad. Dermatol.* **35**, 1005–1008 (1996).
8. *The Distribution of the US Primary Care Workforce* (Agency for Healthcare Research & Quality, 2012); https://www.ahrq.gov/research/findings/factsheets/primary/pcwork3/index.html
9. Seth, D., Cheldize, K., Brown, D. & Freeman, E. F. Global burden of skin disease: inequities and innovations. *Curr. Dermatol. Rep.* **6**, 204–210 (2017).
10. Federman, D. G., Concato, J. & Kirsner, R. S. Comparison of dermatologic diagnoses by primary care practitioners and dermatologists. A review of the literature. *Arch. Fam. Med.* **8**, 170–172 (1999).
11. Moreno, G., Tran, H., Chia, A. L. K., Lim, A. & Shumack, S. Prospective study to assess general practitioners' dermatological diagnostic skills in a referral setting. *Australas. J. Dermatol.* **48**, 77–82 (2007).
12. Tran, H., Chen, K., Lim, A. C., Jabbour, J. & Shumack, S. Assessing diagnostic skill in dermatology: a comparison between general practitioners and dermatologists. *Australas. J. Dermatol.* **46**, 230–234 (2005).
13. Federman, D. G. & Kirsner, R. S. The abilities of primary care physicians in dermatology: implications for quality of care. *Am. J. Manag. Care* **3**, 1487–1492 (1997).
14. *UpToDate* https://www.uptodate.com/home
15. Cutrone, M. & Grimalt, R. Dermatological image search engines on the Internet: do they work? *J. Eur. Acad. Dermatol. Venereol.* **21**, 175–177 (2007).

16. Yim, K. M., Florek, A. G., Oh, D. H., McKoy, K. & Armstrong, A. W. Teledermatology in the United States: an update in a dynamic era. *Telemed. e-Health* **24**, 691–697 (2018).
17. Whited, J. D. et al. Clinical course outcomes for store and forward teledermatology versus conventional consultation: a randomized trial. *J. Telemed. Telecare* **19**, 197–204 (2013).
18. Mounessa, J. S. et al. A systematic review of satisfaction with teledermatology. *J. Telemed. Telecare* **24**, 263–270 (2018).
19. Cruz-Roa, A. A., Arevalo Ovalle, J. E., Madabhushi, A. & González Osorio, F. A. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. *Med. Image Comput. Comput. Assist. Inter.* **16**, 403–410 (2013).
20. Codella, N. C. F. 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* (IEEE, 2018); https://doi.org/10.1109/isbi.2018.8363547
21. Yuan, Y., Chao, M. & Lo, Y.-C. Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance. *IEEE Trans. Med. Imaging* **36**, 1876–1886 (2017).
22. Haenssle, H. 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.* **29**, 1836–1842 (2018).
23. Brinker, T. J. et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *Eur. J. Cancer* **113**, 47–54 (2019).
24. Maron, R. C. et al. Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. *Eur. J. Cancer* **119**, 57–65 (2019).
25. Okuboyejo, D. A., Olugbara, O. O. & Odunaike, S. A. Automating skin disease diagnosis using image classification. In *Proceedings of the World Congress on Engineering and Computer Science* Vol. 2, 850–854 (International Association of Engineers, 2013).
26. Tschandl, P. 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.* **20**, 938–947 (2019).
27. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
28. Han, S. S. 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* **13**, e0191493 (2018).
29. Sun, X., Yang, J., Sun, M. & Wang, K. A benchmark for automatic visual classification of clinical skin disease images. *Proceedings of the European Conference on Computer Vision (ECCV) 2016* 206–222 (Springer, 2016); https://doi.org/10.1007/978-3-319-46466-4_13
30. Boer, A. & Nischal, K.C. www.derm101.com: a growing online resource for learning dermatology and dermatopathology. *Indian J. Dermatol. Venereol. Leprol.* **73**, 138–140 (2007).
31. Wilmer, E. N. et al. Most common dermatologic conditions encountered by dermatologists and nondermatologists. *Cutis* **94**, 285–292 (2014).
32. Yang, J., Sun, X., Liang, J. & Rosin, P. L. Clinical skin lesion diagnosis using representations inspired by dermatologist criteria. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2018); https://doi.org/10.1109/cvpr.2018.00137
33. Okuboyejo, D. A. Towards automation of skin disease diagnosis using image classification. In *Proceedings of the World Congress on Engineering and Computer Science* Vol. 2, 850–854 (International Association of Engineers, 2013).
34. Mishra, S., Imaizumi, H. & Yamasaki, T. Interpreting fine-grained dermatological classification by deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (IEEE, 2019).
35. Guyatt, G. *Users' Guides to the Medical Literature: Essentials of Evidence-Based Clinical Practice* 3rd edn (McGraw-Hill Education/Medical, 2015).
36. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br. J. Surg.* **102**, 148–158 (2015).
37. Webber, W., Moffat, A. & Zobel, J. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.* **28**, 1–38 (2010).
38. Krauss, J. C., Boonstra, P. S., Vantsevich, A. V. & Friedman, C. P. Is the problem list in the eye of the beholder? An exploration of consistency across physicians. *J. Am. Med. Inform. Assoc.* **23**, 859–865 (2016).
39. Eng, C., Liu, Y. & Bhatnagar, R. Measuring clinician–machine agreement in differential diagnoses for dermatology. *Br. J. Dermatol.* https://doi.org/10.1111/bjd.18609 (2019).
40. Sundararajan, M., Taly, A., & Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning vol.* 70, 3319–3328 (2017).
41. Karimkhani, C. et al. Global skin disease morbidity and mortality: an update from the global burden of disease study 2013. *JAMA Dermatol.* **153**, 406–412 (2017).
42. Stern, R. S. & Nelson, C. The diminishing role of the dermatologist in the office-based care of cutaneous diseases. *J. Am. Acad. Dermatol.* **29**, 773–777 (1993).
43. *Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2017* (*GBD 2017*) *Results* (Institute for Health Metrics and Evaluation (IHME), 2018); http://ghdx.healthdata.org/gbd-results-tool
44. Romano, C., Maritati, E. & Gianni, C. Tinea incognito in Italy: a 15-year survey. *Mycoses* **49**, 383–387 (2006).
45. Prabhu, V. et al. Prototypical clustering networks for dermatological disease diagnosis. In *Proceedings of the 4th Conference on Machine Learning for Health Care* (MLHC, 2019).
46. He, S. Y. et al. Self-reported pigmentary phenotypes and race are significant but incomplete predictors of Fitzpatrick skin phototype in an ethnically diverse population. *J. Am. Acad. Dermatol.* **71**, 731–737 (2014).

Methods

Dataset. The dataset for this study consisted of retrospective consecutive adult cases from a teledermatology service serving 17 primary-care and specialist sites from two states in the United States. Cases were predominantly referred by medical doctors, doctors of osteopathic medicine, NPs and physician assistants. Each case contained between one and six clinical photographs of the affected skin areas taken by medical assistants or trained nurses (~75% of cases had six or fewer images; for cases with more images, six images were randomly selected). Images were taken on a mix of devices (Canon point-and-shoot cameras and Apple iPad Minis).

Each case also contained metadata such as patient demographic information and medical history (for a complete list see Supplementary Table 1), which were available to both the DLS and all clinicians in this study. Histology (biopsy) was only done in a small number of patients, reflecting the daily routine in dermatology. All images and metadata were de-identified according to the Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor before transfer to study investigators. The protocol was reviewed by Advarra IRB, which determined that it was exempt from further review under 45 CFR 46.

The dataset was split in a 80:20 ratio based on the submission date of the case: the development set contained cases from 2010–2017, while the validation set contained cases from 2017–2018 (Table 1). Such a temporal split strategy helps simulate a study where a model is developed on past data and validated on future cases, and is arguably a form of external validation[36]. An anonymous patient ID was used to ensure that no patients overlapped between the development set and the validation set, thus preventing any potential label leakage due to the presence of cases from previous visits (Supplementary Fig. 1). This validation set A was further subsampled to reduce class imbalance among the skin conditions of interest to obtain validation set B (Table 1). Selection of skin conditions is described in the section 'Labeling tool and skin condition mapping'.

Reference standard labeling validation set. Because of the impracticality of pathologic confirmation of all diagnoses (for example, rashes are rarely biopsied), each case's differential diagnosis for the validation set was provided by a rotating panel of three dermatologists from a pool of 14 US board-certified dermatologists. The dermatologists practiced in multiple states (Colorado, Hawaii, Iowa, Massachusetts, New York, South Carolina, Tennessee and Texas) and participated in the study via Advanced Clinical. None of the dermatologists were actively practicing teledermatology, although some may have had prior experience or training (for example, during residency); nine were in private or group practice; three were in private practice with concurrent academic appointments; two had primary academic affiliations. They had 5–30 years of experience (average 9.1 years, median 6.5 years) and were actively seeing patients in clinic. The dermatologists also passed a certification test on a small number of cases to ensure that they were comfortable with grading cases using the labeling tool (Supplementary Table 9 and Supplementary Fig. 7). Every dermatologist graded each case (clinical photographs, demographic information and medical history) independently for the presence of multiple skin conditions, diagnosability (for example, due to poor image quality, minimal visible pathology or limited field of view) and up to three differential diagnoses using a custom annotation tool (see 'Labeling tool and skin condition mapping' and Supplementary Figs. 8 and 9). Cases labeled as containing multiple skin conditions or as undiagnosable by the majority of the dermatologists were excluded from the study.

Because grades from individual graders can demonstrate substantial variability, to determine the reference standard we aggregated the differential diagnoses of the three dermatologists that reviewed each case based on a previously proposed 'voting' procedure[47] (see Supplementary Methods for details and Supplementary Fig. 10 for an example). Briefly, for each grader, each diagnosis was first mapped to one of 419 conditions (see 'Labeling tool and skin condition mapping'), then duplicate mapped conditions were removed. Votes for each of these mapped conditions were summed across the three dermatologists based on the relative position of each diagnosis within each dermatologist's differential. The final differential was thus based on the aggregated votes across three board-certified dermatologists.

We verified that this procedure provides substantially higher reproducibility in differential diagnoses than between individual dermatologists (0.77 versus 0.66; for more details see Supplementary Methods). The distribution of the top differential diagnoses is presented in Table 1.

Reference standard labeling development set. The development set was further split into a training set to 'learn' the neural network weights and a tuning set to select hyperparameters for the training process. To maximize the amount of training data, more dermatologists labeled the development set: 1–39 dermatologists (from a cohort of 37 US board-certified and five Indian board-certified dermatologists) labeled each case (mean number of dermatologists per case 4.0, standard deviation 2.6). This range was a result of dispatching reviews on a rolling basis to accelerate reviews and does not reflect the difficulty of cases. Only cases considered by all of the dermatologists grading that case as having multiple skin conditions or undiagnosable were discarded. Reference standard differential diagnosis was established in the same way as for the validation set.

Labeling tool and skin condition mapping. Our labeling tool provided a search-as-you-type interface (Supplementary Table 9 and Supplementary Fig. 7) based on the standardized Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT)[48], within which more than 20,000 terms were related to cutaneous disease. If the dermatologist could not find a matching SNOMED-CT term, the diagnosis could be entered as free text.

Because SNOMED-CT contains terms at varying granularities and there are complex and incompletely specified relationships between terms[49], three board-certified dermatologists mapped these terms and free text diagnoses entries to a list. The list was initially populated with dermatologic conditions that were common or high-acuity, and more conditions were added as needed. Considerations during this mapping were a granularity that would (1) allow a non-dermatology clinician to reasonably determine the next steps in clinical care, (2) enable clear and concise communication with another healthcare provider and (3) exclude superfluous information for most purposes (for example, specific site of the condition). For example, a diagnosis such as 'alopecia' would be too broad, but 'alopecia areata' and 'androgenetic alopecia' would allow a non-dermatologist to engage in the next steps in clinical care.

As labels for cases were collected, additional conditions were added to the list, as appropriate, based on the discussion of at least two of the three dermatologists. Some diagnoses were marked as invalid if they were too broad, non-skin entries, reflected multiple skin conditions (such as a syndrome with multiple skin findings) or were semantically unclear (for example, tooth abrasion). All mappings and labeling were performed while blinded to DLS predictions. The final list contained 419 conditions (Supplementary Table 10).

Selection of the 26 skin conditions. As in actual clinical practice, the prevalence of different skin conditions was heavily skewed in our dataset, ranging from skin conditions with >10% prevalence like acne, eczema and psoriasis to those with sub-1% prevalence like lentigo, melanoma and stasis dermatitis. To ensure that there were sufficient data to develop and evaluate the DLS, we filtered the 419 conditions to the top 26 with the highest prevalence based on the development set (when the labeling was ~80% complete). Specifically, this ensured that, for each of these conditions, there were at least approximately 100 cases in the development dataset (for DLS training purposes) and a projected 25 cases in the validation set (for DLS evaluation). The remaining conditions were aggregated into an 'Other' category (which comprised 21% of the cases in validation dataset A).

DLS development. The DLS has two main components, an image-processing deep convolutional neural network and a shallow network that processes clinical metadata (demographic information and medical history). The image-processing component consisted of a variable number (1–6, depending on the number of images in each case) of Inception-v4[50] modules with shared weights. All images were resized to 459×459 pixels, the default size of this network architecture. The clinical metadata were featurized using the one-hot encoding for all categorical features. Age was used as a number normalized to [0,1] based on the range in the development set. These two components were joined at the top using a fully connected layer (that is, late fusion[51]). The DLS did not have access to any other clinician-provided diagnoses as input.

To help the DLS learn to predict a differential diagnosis (as opposed to a pure classification to predict a single label), the target label of the DLS was based on each case's reference standard differential diagnosis. Specifically, the summed 'votes' of each condition in the differential were normalized (to sum to 1), and the DLS was trained using a softmax cross-entropy loss to learn these 'soft' target labels. To account for class imbalance, when calculating cross-entropy loss, each class was weighted as a function of its frequency, so that cases of rare conditions would contribute more to the loss function. The network weights were optimized using a distributed stochastic gradient descent implementation[52] to predict both the full list of 419 conditions and the shorter list of 27 conditions (26 conditions plus 'Other'). To speed up the training and improve training performance, batch normalization[53] and pre-initialization from ImageNet dataset were used[54]. Training was stopped after a fixed number of steps (100,000) with a batch size of 8.

To train the DLS, the development set was partitioned into a training set to learn DLS's parameters and a tuning set to tune hyperparameters. Because of the severe class imbalance, we created the tune set via stratified sampling (of up to 50 cases per condition). To ensure a clean split with respect to patients, all cases from the patients represented in this sampling were moved to the tune set.

Data augmentation was applied to improve generalization (random flipping, rotating, cropping and color perturbation). The random cropping was parameterized to ensure that the crops had a minimum overlap of 20% with the pathologic skin region (a separately collected label for every case in the training set). Random dropout was applied to metadata features (assigned to unknown), to help improve robustness to missing values or potential data errors. Six networks were trained with the same input and hyperparameters (see Supplementary Table 11 for a complete list of hyperparameters) and ensembled[55] to provide the final prediction.

DLS evaluation. To evaluate the DLS performance, we compared its predicted differential diagnosis with the 'voted' reference standard differential diagnosis

using the top-*k* accuracy and the average top-*k* sensitivity. The top-*k* accuracy measures how frequently the top *k* predictions capture any of the primary diagnoses in the reference standard (that is, ranked first in the differential). The top-*k* sensitivity assesses this for each of the 26 conditions separately, whereas the final average top-*k* sensitivity is the average across the 26 conditions. Averaging across the 26 conditions avoids biasing towards more common conditions, particularly in validation set A. We use both the top-1 and top-3 metrics in this study.

In addition to comparing both the DLS and clinicians against the voting-based reference standard differential diagnoses, we also evaluated against a reference standard based on agreeing with 'at least one' of the three board-certified dermatologists comprising the reference standard ('$\text{Accuracy}_{\text{any}}$'; Supplementary Tables 8, 12 and 13).

Finally, we also measured the agreements in the full differential diagnosis between the DLS and reference standard using the AO metric[37–39]. Because the clinicians were instructed to provide up to three diagnoses, we similarly filtered the DLS's predictions to retain the top-3. Next, unlikely diagnoses lower than a predicted likelihood of 0.1 (selected based on the AO computed on the tune dataset) were filtered to produce the final DLS-predicted differential (up to three diagnoses in ranked order).

Comparison to clinicians. To compare the DLS performance with clinicians, a group of 18 clinicians (who did not participate in earlier parts of this study) provided differential diagnoses for validation set B. These clinicians were composed of three groups of six US board-certified clinicians (dermatologists, PCPs and NPs) and participated in the study via Advanced Clinical. The dermatologists included both academic and private-practice physicians, and all PCPs and NPs were employed in a private practice or medical group and recruited through Advanced Clinical or referrals. In terms of state of practice, the dermatologists practiced in the states of California, Georgia, Maryland, New York and Tennessee and the PCPs and NPs in Arizona, California, Colorado, Connecticut, Florida, Illinois, New York, South Dakota, Vermont and Washington. All clinicians were compensated at an hourly rate comparable to market rates. One dermatologist had prior teledermatology experience, and the others may have had prior training (for example, as part of residency). The NPs were selected from those who were practicing independently as primary-care providers without physician supervision. Every clinician graded a random one-third of the cases, and each case was graded by two random clinicians from each group (six clinicians in total). These clinicians used the same labeling tool as the dermatologists involved in determining the reference standard, and their diagnoses were mapped and processed similarly. There was no time constraint. In case of ties, the top-*k* diagnoses were determined by randomly selecting the diagnosis from the tied candidates. This tie-breaking affected the top-1 analyses for 13% of dermatologist-provided, 24% of PCP-provided and 14% of NP-provided diagnoses. The top-3 analysis was minimally affected, with no ties from dermatologists and NPs and 0.6% ties from PCPs. This tie-breaking avoided confounding the analysis by biasing towards clinicians who provided tied differential diagnoses (which indicates uncertainty).

Feature importance. Additionally, we investigated the relative importance of different types of input on DLS performance. To study the effect of the number of images, we selected a random subset of the images for each case and measured DLS performance on this subsampled dataset. For the clinical metadata, we used a permutation procedure ('permutation feature importance'[56]). Briefly, for a metadata variable of interest, this procedure randomly permutes its assignment across cases in validation set A. Next, the performance of the DLS was measured using the perturbed dataset. To understand the importance of all the metadata collectively, we 'dropped out' all the metadata by assigning all their values to unknown. Because the network could have been dependent on metadata in this analysis, thus over-representing the importance of metadata, we further trained a DLS using only images and evaluated its performance. Finally, we used integrated gradients[40] to highlight the parts of each image that have the biggest effect on the prediction.

Statistical analysis. To compute the confidence intervals, we used a non-parametric bootstrap procedure[57] with 1,000 samples. Because of the intensive computation required to re-run DLS inference, confidence intervals for the feature importance analyses were calculated using the normal approximation with 20 runs (1.96 × standard error, with each run performed on the entire validation set A). To compare the DLS performance to clinicians, a standard one-tailed permutation test[57] was used. Briefly, in each of the 10,000 trials, the DLS's score was randomly swapped with itself or a comparator clinician's score for each case, yielding a DLS–human difference in top-1 accuracy sampled from the null distribution. To perform the non-inferiority test, the empirical *P* value was computed by adding the 5% margin to the observed difference and comparing this number to its empirical quantiles[38,57]. Non-inferiority compared to dermatologists in top-1 accuracy was documented in an institutional mailing list as our pre-specified primary endpoint prior to evaluating the DLS on the validation dataset. See the Life Sciences Reporting Summary for a summary.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this Article.

Data availability

The de-identified teledermatology data used in this study are not publicly available due to restrictions in the data-sharing agreement.

Code availability

The deep learning framework (TensorFlow) used in this study is available at https://www.tensorflow.org/. The training framework (Estimator) is available at https://www.tensorflow.org/guide/estimators. The deep learning architecture (Inception-v4) is available at https://github.com/tensorflow/models/blob/master/research/slim/nets/inception_v4.py.

References

47. Barnett, M. L., Boddupalli, D., Nundy, S. & Bates, D. W. Comparative accuracy of diagnosis by collective intelligence of multiple physicians vs individual physicians. *JAMA Netw. Open* **2**, e190096 (2019).
48. SNOMED home page. *SNOMED* http://www.snomed.org/
49. Simpson, C. R., Anandan, C., Fischbacher, C., Lefevre, K. & Sheikh, A. Will systematized nomenclature of medicine-clinical terms improve our understanding of the disease burden posed by allergic disorders? *Clin. Exp. Allergy* **37**, 1586–1593 (2007).
50. Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. Inception-v4, inception-ResNet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence* 4278–4284 (AAAI, 2017).
51. Snoek, C. G. M., Worring, M. & Smeulders, A. W. M. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia* 399–402 (ACM, 2005); https://doi.org/10.1145/1101149.1101236
52. Dean, J. et al. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems* 1223–1231 (NIPS, 2012).
53. Ioffe, S. & Szegedy, C. Batch normalization: accelerating deep network training by reducing internal covariate shift. Preprint at https://arxiv.org/pdf/1502.03167.pdf (2015).
54. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
55. Opitz, D. & Maclin, R. Popular ensemble methods: an empirical study. *J. Artif. Intell. Res.* **11**, 169–198 (1999).
56. Permutation feature importance. *Azure Machine Learning Studio* https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/permutation-feature-importance.
57. Chihara, L. M. & Hesterberg, T. C. *Mathematical Statistics with Resampling and R* (Wiley, 2018).
58. Hahn, S. Understanding noninferiority trials. *Korean J. Pediatr.* **55**, 403–407 (2012).

Acknowledgements

We thank W. Chen, J. Yoshimi, X. Ji and Q. Duong for software infrastructure support for data collection. Thanks also go to G. Foti, K. Su, T. Saensuksopa, D. Wang, Y. Gao and L. Tran. We also appreciate the input of C. Chen, M. Howell and A. Paller for their feedback on the manuscript. Last, but not least, this work would not have been possible without the participation of the dermatologists, primary care physicians and nurse practitioners who reviewed cases for this study, and S. Bis who helped to establish the skin condition mapping.

Author contributions

Yuan Liu, A.J., C.E., D.H.W., K.L. and D.C. prepared the dataset for usage. S.J.H., K.K. and R.H.-W. provided clinical expertise and guidance for the study. Yuan Liu, A.J., C.E., K.L., P.B., G.d.O.M., J.G., D.A., S.J.H. and K.K. worked on the technical, logistical and quality control aspects of label collection. S.J.H. and K.K. established the skin condition mapping. Yuan Liu, K.L., V.G. and D.C. developed the model. Yuan Liu, A.J., N.S. and V.N. performed statistical analysis and additional analysis. Yun Liu guided study design, analysis of the results and statistical analysis. S.G. studied the potential utility of the model. R.C.D. and D.C. initiated the project and led the overall development, with strategic guidance and executive support from G.S.C., L.H.P. and D.R.W. Yuan Liu, Yun Liu and S.J.H. prepared the manuscript with the assistance and feedback from all other co-authors. K.K. and S.J.H. performed the work at Google Health via Advanced Clinical. G.d.O.M. performed the work at Google Health via Adecco Staffing. N.S. performed the work at Google Health.

Competing interests

K.K. and S.J.H. were consultants of Google LLC. R.H.-W. is an employee of the Medical University of Graz. G.d.O.M. is an employee of Adecco Staffing supporting Google LLC. This study was funded by Google LLC. The remaining authors are employees of Google

LLC and own Alphabet stock as part of the standard compensation package. Yuan Liu, A.J., C.E., D.H.W., K.L., P.B., J.G., V.G., D.A., Yun Liu, R.C.D. and D.C. are inventors on a filed patent related to this work. The authors declare no other competing interests.

Additional information

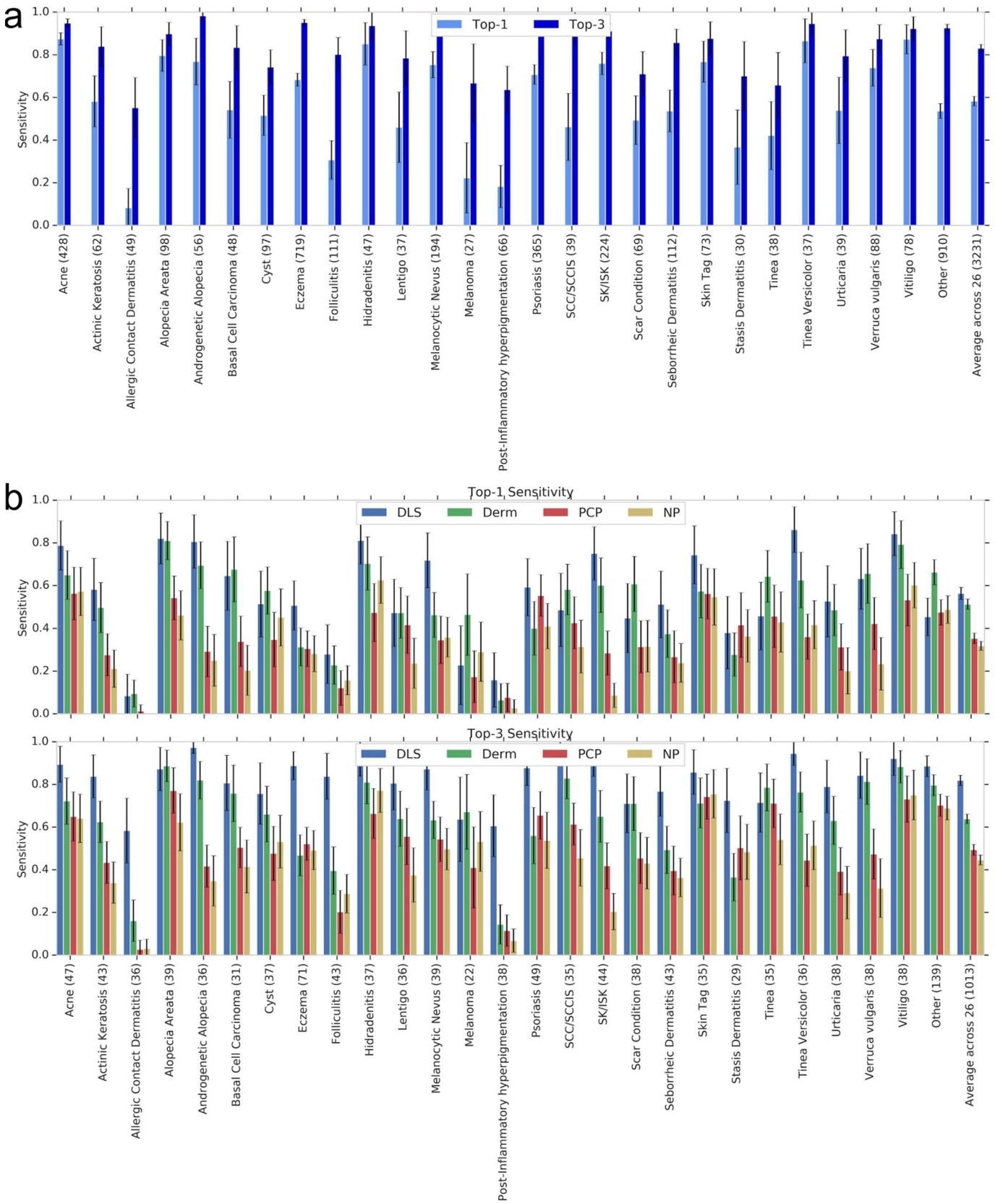
Extended data is available for this paper at https://doi.org/10.1038/s41591-020-0842-3.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-020-0842-3.

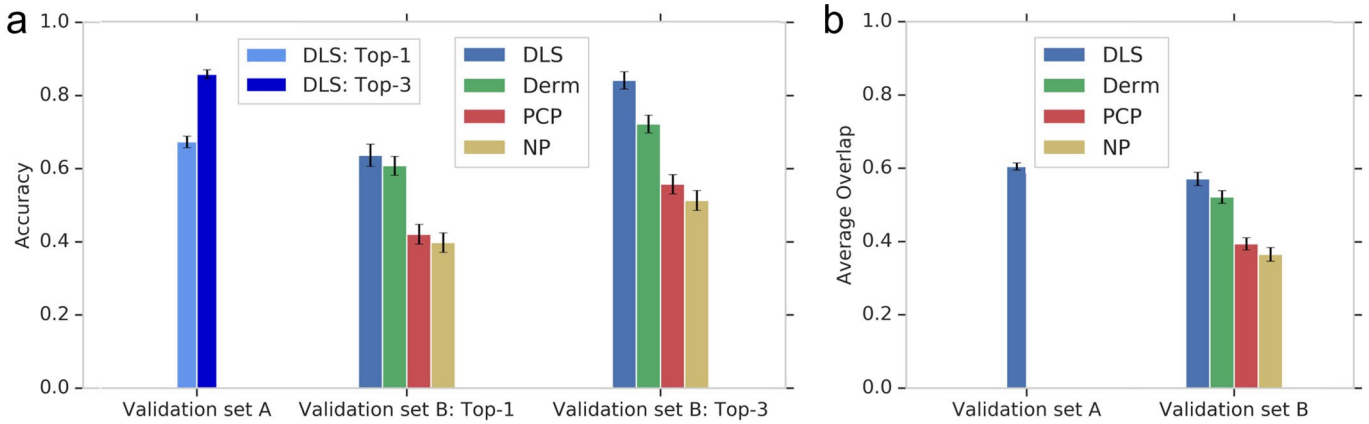
Correspondence and requests for materials should be addressed to Y.L.

Peer review information Javier Carmona was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | Performance of the deep learning system (DLS) and clinicians, broken down for each of the 26 categories of skin conditions and 'other'. a, Top-1 and top-3 sensitivity of the DLS on validation set A (n=3,756). **b**, Top-1 and top-3 sensitivity of the DLS and three types of clinicians: dermatologists (Derm), primary care physicians (PCP), and nurse practitioners (NP) on validation set B (n=963). Numbers in parentheses in the x-axes indicate the number of cases. Detailed breakdown of each clinician and the DLS performance on the subset of cases graded by each clinician are in Supplementary Table 8. Error bars indicate 95% CI (see Statistical Analysis).



Extended Data Fig. 2 | Performance of the deep learning system (DLS) and the clinicians on the 419-way classification: dermatologists (Derm), primary care physicians (PCP), and nurse practitioners (NP) on validation set A (n=3,756) and validation set B (n=963). a, Top-1 and top-3 accuracy for the DLS and clinicians across all cases and 419 categories of skin conditions. b, Average overlap (to assess the full differential diagnosis) of the DLS and clinicians. Error bars indicate 95% confidence intervals (see Statistical Analysis).

nature research

Corresponding author(s): Yun Liu

Last updated by author(s): Mar 9, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data from a teledermatology service were de-identified prior to transfer to study investigators.
Data analysis	The deep learning framework (TensorFlow v1.15.0) used in this study is available at https://www.tensorflow.org/; the training framework (Estimator) is available at https://www.tensorflow.org/guide/estimators; the deep learning architecture (Inception-v4) is available at: https://github.com/tensorflow/models/blob/master/research/slim/nets/inception_v4.py.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The de-identified teledermatology data used in this study are not publicly available due to restrictions in the data-sharing agreement.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The 80/20 development/validation split was determined a priori without sample size calculations. The number of skin conditions (26) was chosen such that at least 100 examples per skin condition (based on preliminary labels) would be available for model development, and a projected 25 examples per skin condition for model validation.
Data exclusions	Several pre-determined exclusion criteria were applied. First, cases with multiple skin conditions or not diagnosable were excluded from the study. Second, from the validation sets, cases from repeat visits of a patient in the development dataset were excluded to avoid bias.
Replication	The deep learning system is an ensemble of 6 trained models. The spread in tuning set performance between the 6 trained models was measured at 1-3%.
Randomization	Validation set B cases were randomly subselected from validation set A using stratified random sampling to enrich for rarer skin conditions.
Blinding	The skin condition mapping were established by dermatologists who were blinded to the predictions of the deep learning system.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	De-identified data from cases submitted to a teledermatology service. Baseline variables are reported in Table 1, but briefly on the development and validation sets, patients were aged 40-43 on average, with 62-64% female, and the most common diagnoses were eczema, psoriasis, and acne. Post-visit treatment data were not available.
Recruitment	All consecutive adult cases in the time period (2010-2018) were used, excluding those with multiple skin conditions or not diagnosable (as in exclusion criteria above).
Ethics oversight	The protocol was reviewed by Advarra IRB (Columbia, MD), which determined that it was exempt from further review under 45 CFR 46.

Note that full information on the approval of the study protocol must also be provided in the manuscript.