Augmented Intelligence Dermatology: Deep Neural Networks Empower Medical Professionals in Diagnosing Skin Cancer and Predicting Treatment Options for 134 Skin Disorders

Seung Seog Han[1,8], Ilwoo Park[2,8], Sung Eun Chang[3,8], Woohyung Lim[4], Myoung Shin Kim[5], Gyeong Hun Park[6], Je Byeong Chae[7], Chang Hun Huh[7] and Jung-Im Na[7]

Although deep learning algorithms have demonstrated expert-level performance, previous efforts were mostly binary classifications of limited disorders. We trained an algorithm with 220,680 images of 174 disorders and validated it using Edinburgh (1,300 images; 10 disorders) and SNU datasets (2,201 images; 134 disorders). The algorithm could accurately predict malignancy, suggest primary treatment options, render multi-class classification among 134 disorders, and improve the performance of medical professionals. The area under the curves for malignancy detection were 0.928 ± 0.002 (Edinburgh) and 0.937 ± 0.004 (SNU). The area under the curves of primary treatment suggestion (SNU) were 0.828 ± 0.012, 0.885 ± 0.006, 0.885 ± 0.006, and 0.918 ± 0.006 for steroids, antibiotics, antivirals, and antifungals, respectively. For multi-class classification, the mean top-1 and top-5 accuracies were 56.7 ± 1.6% and 92.0 ± 1.1% (Edinburgh) and 44.8 ± 1.2% and 78.1 ± 0.3% (SNU), respectively. With the assistance of our algorithm, the sensitivity and specificity of 47 clinicians (21 dermatologists and 26 dermatology residents) for malignancy prediction (SNU; 240 images) were improved by 12.1% ($P < 0.0001$) and 1.1% ($P < 0.0001$), respectively. The malignancy prediction sensitivity of 23 non-medical professionals was significantly increased by 83.8% ($P < 0.0001$). The top-1 and top-3 accuracies of four doctors in the multi-class classification of 134 diseases (SNU; 2,201 images) were increased by 7.0% ($P = 0.045$) and 10.1% ($P = 0.0020$), respectively. The results suggest that our algorithm may serve as augmented intelligence that can empower medical professionals in diagnostic dermatology.

Journal of Investigative Dermatology (2020) ■, ■–■; doi:10.1016/j.jid.2020.01.019

INTRODUCTION

One of the most successful deep learning architectures—convolutional neural network (CNN) has consistently demonstrated its outstanding performance in medical research (Chilamkurthy et al., 2018; Gulshan et al., 2016; Rajpurkar et al., 2018). Several recent studies have used CNNs in diagnostic dermatology and reported dermatologist-level performance in diagnosing skin cancer. However, most studies have been limited to specific binary tasks, such as differentiating melanoma from nevus (Brinker et al., 2019; Haenssle et al., 2018; Tschandl et al., 2019b; Yu et al., 2018), or multi-class classification of limited number of skin tumors (Han et al., 2018a; Maron et al., 2019; Tschandl et al., 2019a). The performance of CNN needs to be tested in an environment similar to real practice, which requires distinguishing skin cancer from numerous other skin disorders including inflammatory and infectious conditions. In addition, the robustness and repeatability of this approach require further validations (Brinker et al., 2018; Narla et al., 2018; Topol, 2019).

This study utilized CNN architectures that were trained with 220,680 images consisting of 174 disease classes (Table 1). Both binary classification (predicting malignancy and suggesting treatment option) and multi-class classification of 134 skin disorders were performed with the algorithm, and the performance was compared with that of clinicians and algorithm-assisted clinicians.

[1]I Dermatology Clinic, Seoul, Korea; [2]Department of Radiology, Chonnam National University Medical School and Hospital, Gwangju, Korea; [3]Department of Dermatology, Asan Medical Center, Ulsan University College of Medicine, Seoul, Korea; [4]LG Sciencepark, Seoul, Korea; [5]Department of Dermatology, Sanggye Paik Hospital, Inje University College of Medicine, Seoul, Korea; [6]Department of Dermatology, Dongtan Sacred Heart Hospital, Hallym University College of Medicine, Dongtan, Korea; and [7]Department of Dermatology, Seoul National University Bundang Hospital, Seongnam, Korea

[8]These authors contributed equally to this work as co-first authors.

Correspondence: Jung-Im Na, Department of Dermatology, Seoul National University Bundang Hospital, 82 Gumi-Ro 173 Beon-Gil, Seongnam, Gyeonggi 463-707, Korea. E-mail: jina1@snu.ac.kr

Abbreviations: AUC, area under the curve; CNN, convolutional neural network

Received 3 September 2019; revised 2 December 2019; accepted 13 January 2020; accepted manuscript published online XXXX; corrected proof published online XXXX

Table 1. Summary of the Training and Validation Data and the Corresponding Demographic Information

Variable/Dataset	ASAN[1]	Normal[2]	MED-NODE[2]	Web[2]	SNU[3]	Edinburgh[3]
No. of images	120,780	48,271	170	51,459	2,201	1,300
No. of unique individuals	20,765	5,849	–	–	1,608	–
Age (mean ± SD)	41.8 ± 21.6	39.4 ± 21.3	–	–	42.4 ± 23.6	–
% of male	44.4%	43.4%	–	–	45.9%	–
Race	>99%, Asian	>99%, Asian	Mainly Caucasian	Mainly Caucasian	>99%, Asian	Mainly Caucasian
No. of disorders	174	Normal, or nonspecific	Melanoma and nevus	174	134	10
Diagnosis method	Clinical diagnosis and/or biopsy	Image finding	Biopsy	Image finding	Clinical diagnosis and/or biopsy	Biopsy
Usage	Training	Training	Training	Training	Validation	Validation

[1]Patient demographics, including age and sex, were available from a retrospective chart review for 96.8% of the ASAN dataset and 65.5% of the normal dataset. Clinical images were collected after the chart review, and all data were fully anonymized.
[2]Demographic information was not available for the Edinburgh, MED-NODE, and Web datasets.
[3]The SNU dataset consisted of data from three university hospitals (Seoul National University Bundang Hospital, Inje University Sanggye Paik Hospital, and Hallym University Dongtan Hospital).

RESULTS

Binary classification – malignancy prediction and treatment suggestion

Using both the SNU dataset, which consisted of 2,201 images representing 134 diseases (5 malignancies and 129 non-malignancies), and the Edinburgh dataset, which consisted of 1,300 images representing 10 disorders (four malignancies and six non-malignancies), the ability of our algorithm for malignancy diagnosis was validated in a situation that was representative of a real clinical practice, where clinicians are required to differentiate malignancy from several types of other skin diseases. The area under the curves (AUCs) for predicting malignancy were 0.937 ± 0.004 (SNU) and 0.928 ± 0.002 (Edinburgh), which were determined by using the malignancy output (Table 2). The test for predicting a treatment option among four primary medications (steroids, antibiotics, antivirals, and antifungals) using the SNU dataset demonstrated the potential of our algorithm to be applied to treatment suggestion with the mean AUCs of 0.893 ± 0.006 (Table 2).

The performance of our model for predicting malignancy and treatment option among 134 diseases was compared with those of 47 medical professionals (21 board-certified dermatologists and 26 dermatology residents) using randomly selected 240 images from the SNU dataset. Overall, in both the malignancy and treatment predictions, the algorithm showed the similar performance as that of the dermatology residents but slightly lower than those of the dermatologists (Figure 1).

The efficacy of our algorithm to augment the diagnostic accuracy of test participants was evaluated by using the 240 images from the SNU dataset. After the initial test, the test participants were informed of the result of the algorithm for each test image and their answers were modified. For the detection of malignancy, the individual performance of 47 medical professionals assessed by F1 score showed a significant improvement from 0.66 ± 0.08 to 0.75 ± 0.06 ($P < 0.0001$) with the assistance of the model (Figure 1a). The sensitivity and specificity of the malignancy diagnosis of the 47 medical professionals improved from 77.4 ± 10.7% to 86.8 ± 8.7% ($P < 0.0001$) and from 92.9 ± 2.4% to 93.9 ± 2.3% ($P < 0.0001$), respectively. Similarly, the sensitivity of the malignancy diagnosis of 23 non-medical professionals improved markedly from 47.6 ± 33.1% to 87.5 ± 17.2% ($P < 0.0001$) without a loss in specificity. Similar to the malignancy test, the F1 score of the 47 medical professionals in predicting a primary treatment method significantly improved from 0.50 ± 0.08 to 0.61 ± 0.05 for steroids, from 0.47 ± 0.10 to 0.55 ± 0.08 for antibiotics, from 0.70 ± 0.09 to 0.75 ± 0.08 for antifungals, and from 0.48 ± 0.11 to 0.54 ± 0.09 for antivirals with the assistance of the algorithm ($P < 0.0001$ for four treatments, Figure 1b–e).

In the prediction of malignancy, medical professionals modified 5.1 ± 2.7% of their answers after reviewing the result of the algorithm. A total of 74.0 ± 15.0% of their modified answers became correct (26.0 ± 15.0% became incorrect). The non-medical professionals modified 19.6 ± 10.2% of their answers. As a result, 72.0 ± 16.2% of their modified answers became correct, whereas 28.0 ± 16.2% of them were incorrect. For the treatment prediction test, the medical professionals modified 8.8 ± 4.9%, 7.3 ± 4.2%, 4.0 ± 2.4%, and 4.4 ± 2.4% of their answers for predicting steroids, antibiotics, antivirals, and antifungals, respectively. As a result, 70.1 ± 14.9% (steroids), 64.2 ± 15.0% (antibiotics), 66.5 ± 20.8% (antivirals), and 67.0 ± 20.6% (antifungals) of the modified answers became correct. Overall, the diagnostic accuracy of the test participants for both the malignancy and treatment prediction improved with the assistance of the algorithm.

Multi-class classification of 134 diseases

Using the SNU dataset (2,201 images), the mean top-1, 3, and 5 accuracies for the classification of 134 diseases were 44.8 ± 1.2%, 69.0 ± 0.9%, and 78.1 ± 0.3%, respectively (Table 2). The classification performance for each of the 134 diseases are listed in Supplementary Table S1. The algorithm was able to differentiate between eczematous and infectious conditions, as well as classify very rare skin lesions, such as

Table 2. Summary of Binary Classification (Malignancy Prediction and Treatment Suggestion) and Multi-Class Classification (Diagnosis of 134 Disorders)

	Statistical Parameter	Analysis Method	Result
Malignancy prediction (binary classification)			
SNU (2,201 images)	AUC	Malignancy output	0.937 ± 0.004
Edinburgh (1,300 images)	AUC	Malignancy output	0.928 ± 0.002
Treatment prediction (binary classification)			
SNU (2,201 images)	AUC	Steroids output	0.828 ± 0.012
SNU (2,201 images)	AUC	Antibiotics output	0.885 ± 0.006
SNU (2,201 images)	AUC	Antivirals output	0.944 ± 0.006
SNU (2,201 images)	AUC	Antifungals output	0.918 ± 0.006
Disease classification (multi-class classification)			
SNU (2,201 images)	Top-1 accuracy	Target-class output	44.8 ± 1.2%
	Top-3 accuracy	Target-class output	69.0 ± 0.9%
	Top-5 accuracy	Target-class output	78.1 ± 0.3%
	AUC[1]	Target-class output	0.978 ± 0.001
Edinburgh (1,300 images)	Top-1 accuracy	Target-class output	56.7 ± 1.6%
	Top-3 accuracy	Target-class output	83.6 ± 0.9%
	Top-5 accuracy	Target-class output	92.0 ± 1.1%
	AUC[1]	Target-class output	0.939 ± 0.003

The SNU dataset consisted of 2,201 images of 134 disorders and the Edinburgh dataset consisted of 1,300 images of 10 tumorous skin diseases. As an analytic metric, the following parameters were used:

malignancy output = a sum of model outputs for five malignant tumorous disorders

steroids output = a sum of model outputs for 59 disorders requiring a steroids treatment

antibiotics output = a sum of model outputs for 21 disorders requiring an antibiotics treatment

antivirals output = a sum of model outputs for five disorders requiring an antivirals treatment

antifungals output = a sum of model outputs for eight disorders requiring an antifungals treatment.

Abbreviation: AUC, area under the curve.

[1]The AUCs for each disease were calculated by converting the multi-class model outputs into a binary classification: one vs rest of the 173 diseases.

lichen amyloidosis (Figure 2f). For the same SNU dataset (2,201 images), the mean top-1 and 3 accuracies of four doctors (two dermatologists and two dermatology residents) were 49.9 ± 7.0% and 67.2 ± 5.4%, respectively (Supplementary Table S1).

We also tested whether our algorithm could be used to augment the diagnostic accuracy of the test participants in the multi-class diagnosis task of 134 disorders by using the SNU dataset (2,201 images). The mean top-1 and 3 accuracies of four doctors showed 7.0 ± 4.5% ($P = 0.045$) and 10.1 ± 2.5% ($P = 0.0020$) improvements, respectively, with the reference of the top-1, 2, and 3 diagnoses predicted by the algorithm (Supplementary Table S1).

DISCUSSION

Several studies have compared the performance of artificial intelligence with those of dermatologists in terms of classifying skin cancer and other skin lesions by using a digital camera and dermoscopy images. Using the Edinburgh and Stanford Hospital datasets for validation, Esteva et al. (2017) showed that a CNN can diagnose carcinoma and melanoma with an AUC of 0.96, which was on par with the performance of dermatologists. Haenssle et al. (2018) compared a CNN's diagnostic performance with 58 dermatologists by using dermoscopic images of melanoma and nevi, and showed that most dermatologists were outperformed by the CNN. Fujisawa et al. (2019) trained and validated a CNN by using Tsukuba Hospital's dataset and showed a 96.3% sensitivity and a 89.5% specificity for melanoma detection, which was superior to the performance of dermatologists. Tschandl et al. (2019b) showed that the algorithm had an accuracy similar to 62 human experts for a combined evaluation of both dermoscopic and close-up images of nonpigmented lesions . Tschandl et al. (2019a) also showed that machine learning classifiers outperformed human experts in the diagnosis of pigmented skin lesions of dermoscopy. Brinker et al. (2019) reported the CNN trained with open-source dermoscopic images performed on par with 145 dermatologists on a clinical image classification task. Recently, Cho et al. (2019) showed that CNN was equivalent to the dermatologists when classifying malignancy with unified composition images of the lip.

For the CNN models to be practically useful, it is important to develop an algorithm with both specific binary task and multi-classification capabilities. A binary classifier alone, such as one trained for melanoma versus nevi, could lead to wrong conclusions if inputs other than melanoma or nevi, such as nail hematoma or melanonychia, are given. Our model renders results for the multi-classification and the binary task (malignancy or not) at the same time. In the specific binary classification of melanoma versus nevus, the performance of the model (AUC = 0.971 ± 0.003; 76 melanoma and 331 nevi of the Edinburgh dataset; Supplementary Figure S1) was superior to that of dermatologists. However, the model performance showed a relative decline in the 134 multi-class diagnosis, in which the

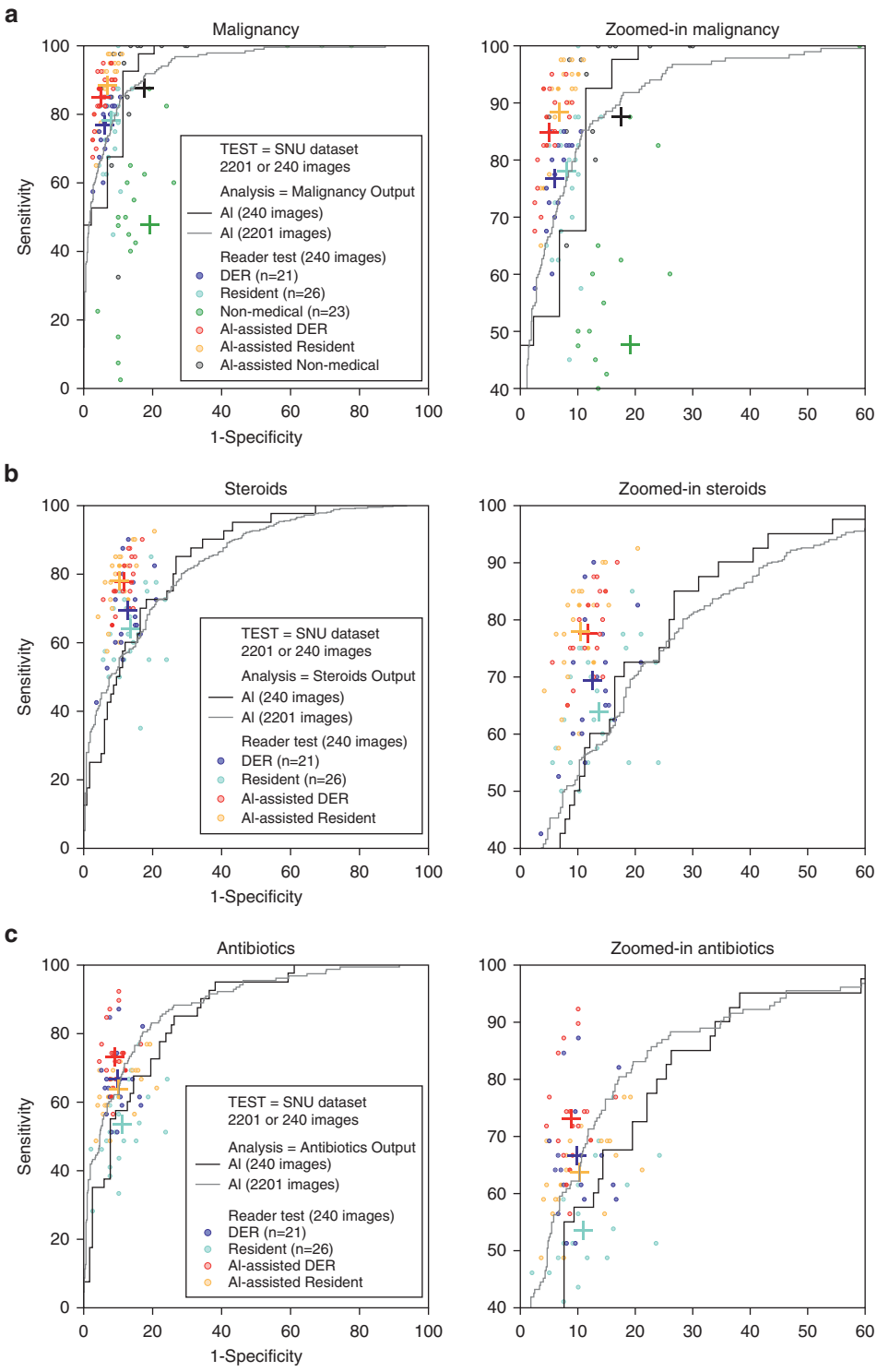

Figure 1. Performance of the algorithm and comparison with human tests for malignancy prediction among 134 diseases and four treatment options test. The SNU dataset, which consisted of 2,201 images accounting for 134 diseases, was used for the prediction of (**a**) malignancy and treatment selection tests for (**b**–**e**) four primary medications. The algorithm was validated using both 2,201 and 240 test images, whereas human tests were performed using the latter 240 images. Steroids, antibiotics, antivirals, and antifungals output were defined as the sums of outputs for the diseases whose primary treatment of choice belongs to one of the four treatment choices (Supplementary Table 1; Category I). The performance of the algorithm, which is represented by ROC curves, was tested against the results obtained from 21 dermatologists and 26 dermatology residents. In addition, 23 non-medical professionals participated in the malignancy test. For each test image, they were asked if the lesion was malignant, and which of the four medications they would use to treat the lesion with. Overall, the performance of the algorithm was similar to that of the dermatologic residents but slightly lower than that of the dermatologists. The dots represent individual performance of test participants, whereas crosses (+) indicate the average performance of them. The zoomed-in ROC curves for each graph is located in the right column. AI, our algorithm; DER, dermatologists; ROC, receiver operator characteristic.

number of classes was increased dramatically and the background of images contained various structures, such as ones in Figure 2a.

Although there have been a few efforts (De Fauw et al., 2018; Esteva et al., 2017; Han et al., 2018a; Rajpurkar et al., 2018) in comparing the performance between CNN models and human experts in multi-classification medical diagnosis, the number of disease classes in medical research were limited to 14 (Rajpurkar et al., 2018) or less(De Fauw et al., 2018; Esteva et al., 2017; Han et al., 2018a; Maron et al., 2019). In nine disease-groups classification, Esteva et. al.(2017) demonstrated a 53.3 and 55.0% top-1 accuracy of two dermatologists, whereas their algorithm showed a 55.4% top-1 accuracy (Esteva et al., 2017). For clinical images of 10 tumorous diseases, Han et. al. (2018a) showed 57.3% and 55.7% top-1 accuracies for their internal dataset

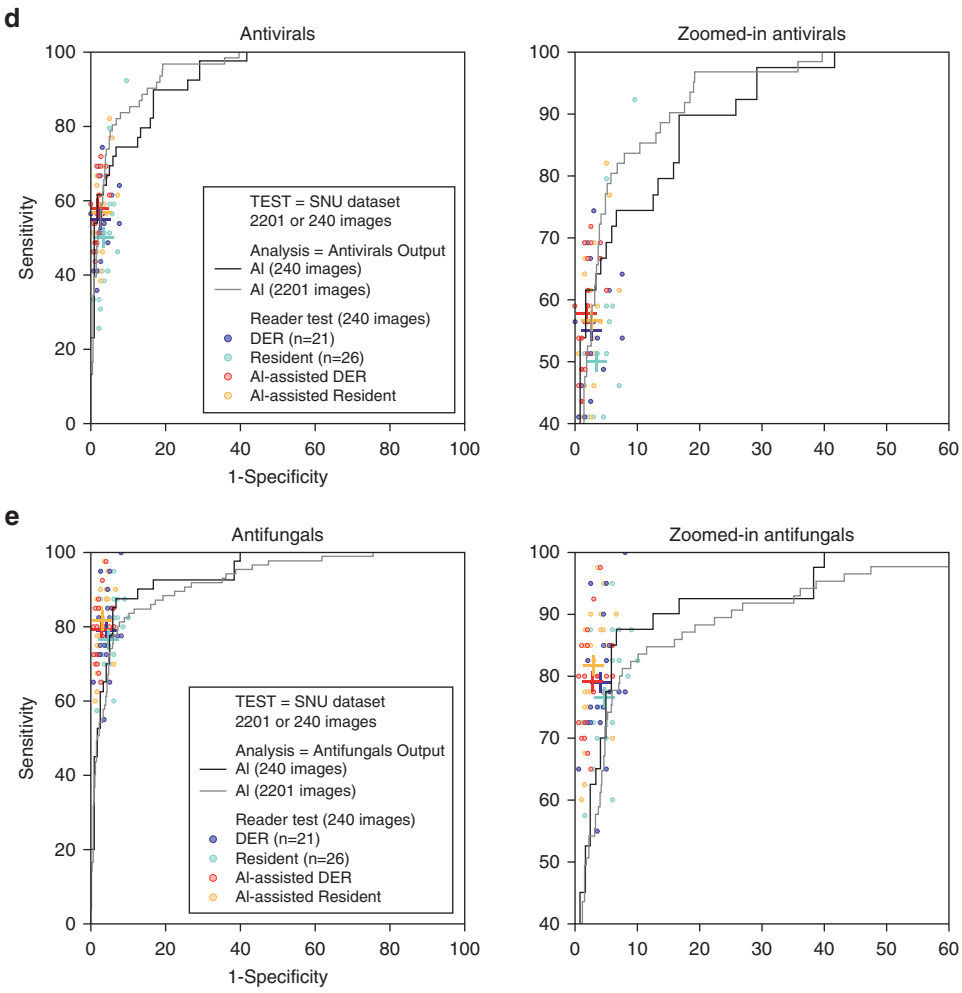

Figure 1. Continued

and the Edinburgh dataset, respectively. Our study represents an unprecedented challenge with the multi-class classification task of 134 diseases and demonstrated the top-1 and 3 accuracy of 44.8% and 69.0% for the algorithm, respectively, while the mean top-1 and 3 accuracy of four doctors was 49.9% and 67.2%, respectively (Supplementary Tables S2 and S3). These multi-classification results may have been underestimated because the 134 diseases included the classes that could have been lumped together. For example, tinea cruris and tinea corporis were categorized as separate classes, and pigmented basal cell carcinoma was regarded as an incorrect answer for the case of melanoma.

One of the most interesting findings of our study was that the algorithm can be used to provide an appropriate treatment strategy. Some skin lesions are very similar under clinical and visual evaluations, making it difficult to determine a proper treatment plan. For example, because eczematous and infectious conditions are similar in appearance (Figure 2a–d), it is very challenging to determine appropriate medications for these diseases. When we trained the algorithm to perform the classification into four categories (steroids, antibiotics, antifungals, and antivirals), the algorithm produced accurate treatment predictions with the mean AUC of 0.893 ± 0.006 (Table 2). Although further validation is warranted, our preliminary results suggest that our algorithm may be utilized for the selection of appropriate treatment plans.

Based on the algorithms used in this study, we have created a website, Model Dermatology (http://modelderm.com), that is accessible using both PCs and smartphones. It reports both top accuracies for the classification among the 174 disorders and binary results for predicting malignancy (i.e., requiring biopsy or not) and four treatment options.

Augmented intelligence to empower the diagnostic performance of medical professionals

We showed that the performance of dermatologists improved with the assistance of the algorithm for the predictions of malignancy and treatment options as well as multi-disease classification tasks. In the prediction of malignancy, the mean sensitivity and specificity of 21 dermatologists significantly improved from 76.7 ± 8.1% to 84.9 ± 8.0% ($P < 0.0001$) and from 93.9 ± 2.0% to 94.9 ± 1.9% ($P = 0.0062$), respectively (Figure 1a). The F-score of the 21 dermatologists also exhibited a significant improvement from 0.66 ± 0.06 to 0.75 ± 0.06 ($P < 0.0001$) with the assistance of the algorithm. The same trends were observed in the treatment prediction tasks where the algorithm significantly improved the F1 score of 47 doctors ($P < 0.0001$; Figure 1b–e). In the multi-disease classification test using

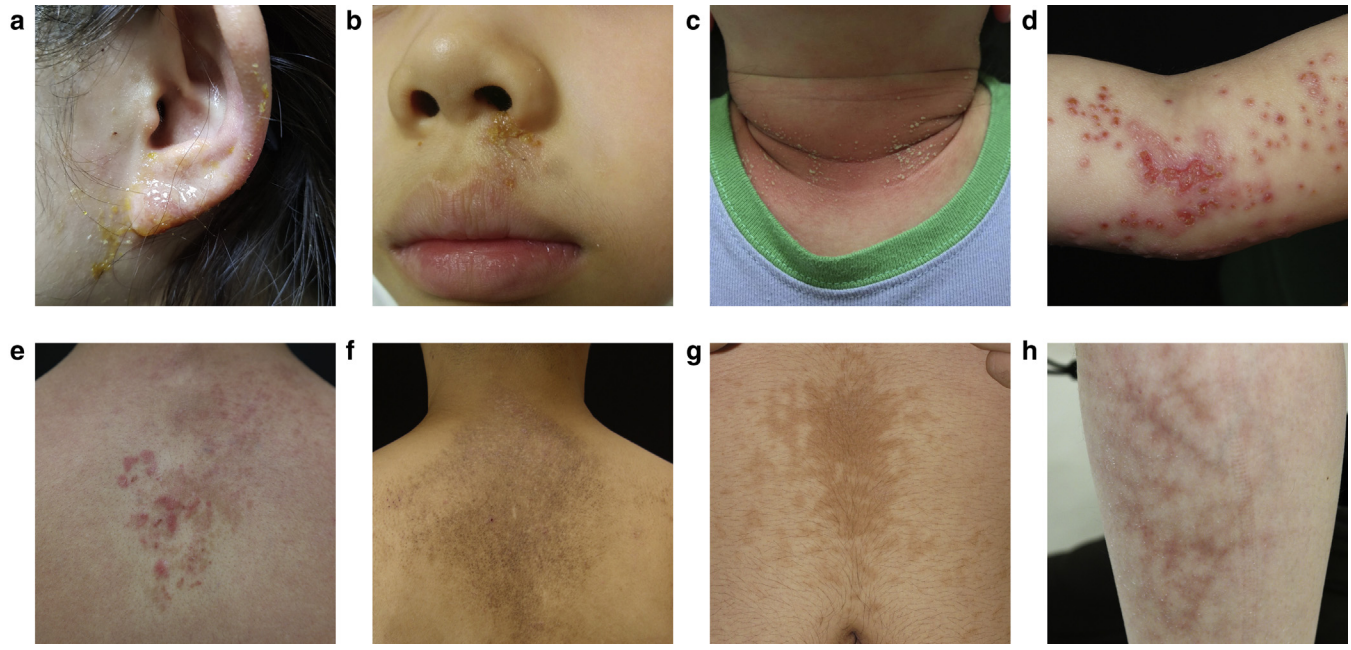

Figure 2. Representative examples of multi-class diagnosis using the algorithm. The results from the algorithm indicated that the convolutional neural network trained in this study can be used for multi-class diagnosis (134 disorders) of various cutaneous diseases. The confirmed diagnosis of each disease, the top-1, 2, and 3 predictions and the corresponding outputs from the model are demonstrated as follows: (a) contact dermatitis, prediction = seborrheic dermatitis (0.39)/squamous cell carcinoma (0.09)/lupus erythematosus (0.08); (b) impetigo, prediction = impetigo (0.92)/herpes simplex (0.04)/wart (0.01); (c) staphylococcal scalded skin syndrome, prediction = pustular psoriasis (0.43)/staphylococcal scalded skin syndrome (0.35)/contact dermatitis (0.12); (d) eczema herpeticum, prediction = eczema herpeticum (0.93)/nummular eczema (0.02)/ impetigo (0.01); (e) prurigo pigmentosa, prediction = prurigo pigmentosa (0.19)/psoriasis (0.18)/bullous disease (0.17); (f) lichen amyloidosis, prediction = lichen amyloidosis (0.96)/Rhiel melanosis (0.01)/prurigo pigmentosa (0.01); (g) confluent reticulated papillomatosis, prediction = confluent reticulated papillomatosis (0.72)/Becker nevus (0.13)/vitiligo (0.04); (h) erythema ab igne, prediction = erythema ab igne (0.54)/hemangioma (0.20)/portwine stain (0.07). (**a**) Contact dermatitis and (**b**) impetigo are lesions with a similar appearance with an oozing erythematous patch. Although the algorithm did not match the right answer in (a), eczema of the ear is easily confused with other infectious conditions, and the algorithm preferentially predicted the condition as seborrheic dermatitis, which is a common eczematous condition around ear. In contrast, the lesions in (**b**–**d**) represent infectious conditions that are frequently misdiagnosed as contact dermatitis; however, the algorithm diagnosed them correctly as (**b**) impetigo and (**d**) eczema herpeticum. In the case of (**c**), staphylococcal scalded skin syndrome was not predicted as a top-1, but a top-2 choice; however, all three predictions would be plausible diagnoses by a clinician for the given image. The images shown in (**e**–**h**) represent the examples of uncommon disorders. Pruritus is a chief complaint in both (**e**) prurigo pigmentosa and (**f**) lichen amyloidosis. In the absence of an examination by dermatologists, however, only contact dermatitis is usually considered for pruritus. (**g**) Confluent reticulated papillomatosis and (**h**) erythema ab igne represent other rare disorders that exhibit nonpruritic reticulated patches. The algorithm produced a correct differential diagnosis for these rare dermatologic disorders based on the characteristic findings from these rare lesions.

2,201 images consisting of 134 disorders, the top-1 and 3 accuracies of four doctors improved significantly with the assistance of the algorithm ($P = 0.045$ for top-1 and $P = 0.0020$ for top-3).

Although the algorithm demonstrated a performance comparable to that of dermatologic residents, the dermatologists also benefited from the assistance from the algorithm, which may be explained by the difference in diagnostic and error profiles between the algorithm and human experts. As shown in Figure 3a and b, the results from the algorithm may have assisted doctors in evaluating the ambiguous images resembling eczema, which led to an improvement in performance. In contrast, the errors by the algorithm owing to suboptimal image qualities, such as blurry images or shadows in Figure 2a, may have been avoided by clinicians. In our previous study on onychomycosis, the CNN algorithm was more efficient in analyzing ambiguous, difficult-to-answer images than dermatologists (Han et al., 2018). Similarly, the difference in the patterns of errors that were more susceptible to the algorithm and human has been reported in the ImageNet studies (Dodge and Karam, 2017; Russakovsky et al., 2015).

This study demonstrated that the algorithm plus dermatologists produced the maximal effectiveness in predicting malignancy as well as deciding on the treatment options. This observation suggests that the deep learning algorithm developed in this study may represent an "Augmented Intelligence," thereby serving as an ancillary tool to enhance the diagnostic accuracy of clinicians. Similar efforts to enhance human performance with computer supports have been demonstrated by using machine learning techniques in dermatology (Cho et al., 2019; Farberg et al., 2017) or deep learning algorithms in pathology (Litjens et al., 2016; Wang et al., 2016a). To maximize the efficiency of using the algorithm as an ancillary tool, further research is needed to analyze the algorithm's diagnostic patterns and identify the features that the model uses to derive its outcome.

Lastly, we expect that our algorithm may be able to encourage the public to visit specialists for cancerous lesions that might have been otherwise neglected. Figure 3c represents the typical case of pigmented basal cell carcinoma, which is often misidentified as nevus and, thus, neglected by the public. A total of 73.9% (17 out of 23) non-medical

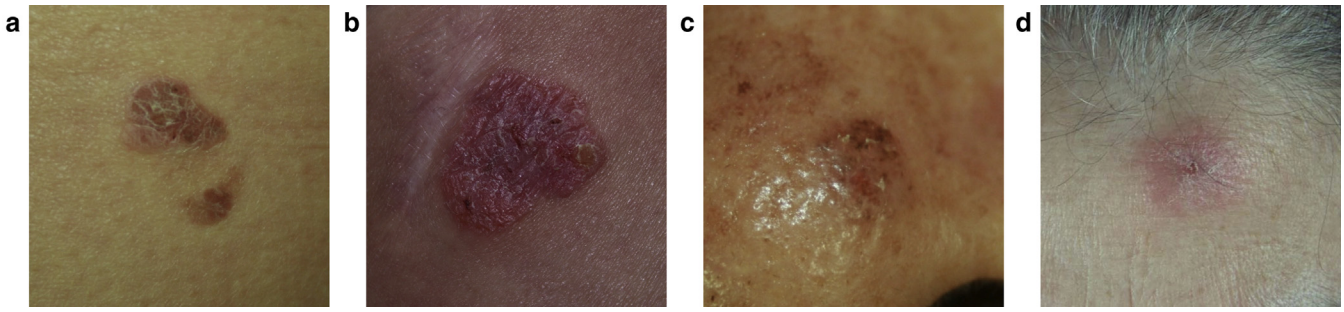

Figure 3. Augmented intelligence as an ancillary tool: representative examples of how the assistance of algorithm resulted in correct or incorrect answers. (**a**) Intraepithelial carcinoma; model' malignancy output = 1.00; (**b**) intraepithelial carcinoma; model' malignancy output = 0.91; (**c**) basal cell carcinoma; model' malignancy output = 1.00; (**d**) insect bite; model' malignancy output = 0.77. A total of 21 dermatologists, 26 residents, and 23 non-medical professionals were tested with 240 test images and their performances were compared with the results of our convolutional neural network model for malignancy prediction. In addition, after reviewing the prediction of the model for the same test set, they were given an opportunity to modify their answers. With assistance from the algorithm, 46.8% (22 out of 47) and 31.9% (15 out of 47) of the doctors modified their answers to the correct diagnosis for (a) and (b), respectively. A total of 73.9% (17 out of 23) of the non-professionals benefited from the assistance of the algorithm for the image in (c). Figure (d) represents the case where the assistance from the algorithm directed the test participants toward incorrect choices. In this instance, the model arrived at the diagnosis of malignancy and 27.7% (13 out of 47) of doctors and 65.2% (15 out of 23) of non-medical professionals followed the algorithm's suggestion and made an incorrect prediction. This illustrates one of the current limitations of our algorithm and signifies that it needs to be used with proper guidance from clinicians.

professionals modified their answers from non-malignancy to malignancy for this image with assistance from the algorithm.

Limitations of this study

Although our study demonstrated the potential of the deep learning models to be used for screening melanoma and other dermatologic disorders, our algorithm still presents several limitations, especially to be used as a standalone diagnostic tool. First, the outcome of a deep learning algorithm can be significantly affected by the composition of the input images (Navarrete-Dechent et al., 2018). Skin disorders have their own characteristic locations and compositions for optimal diagnosis, which are recognized by specialists, but not by non-medical professionals. For example, scabies is characteristically found between finger webs or under breasts, and the images of scabies in our training data echoed this characteristic because they were collected by dermatologists. However, the images submitted by nonexperts may contain the regions other than these typical locations because when scabies develops, it is also found on other parts of body.

Second, because our algorithm was trained and tested by using high-quality images, its performance is generally suboptimal if the input images are of low quality (Dodge and Karam, 2017). Current artificial intelligence technologies may not be able to distinguish between blurry shadows on normal skin and irregular borders in a malignancy unless it is trained by using data with such substandard qualities. As shown in Figure 2a, where the squamous cell carcinoma output was increased owing to the presence of an ear in the image, the algorithm often produced a misdiagnosis as malignancy for images that contained ears and noses presumably because of a shading or curved shape on these anatomical structures.

Third, a diagnosis made with only one image with the most optimal composition may present inherent limitations compared to diagnoses made in a clinical setting. In a real practice, a dermatological diagnosis is made based on the combination of multiple sources of information including past medical history, symptoms, appearance compared to other lesions on the patient, and the texture of the lesion assessed by physical contact.

Finally, although 174 disorders were used for training and 134 disorders for validation in this study, there still exist numerous skin diseases that were omitted. Our training dataset lacked the images of mild inflammatory lesions that were not of clinical interest. Given that the prevalence of these conditions compared to skin cancer is high, frequent false positives attributed to these lesions could potentially cause unnecessary visits to the physician's office if our algorithm is used without a proper guidance from clinicians. The collection of broader spectrum of training data should be pursued to overcome this issue. The coordinated efforts to archive publicly available datasets such as ISIC (https://isic-archive.com), HAM10000, and Seven-Point Checklist Dermatology dataset are expected to contribute to the accumulation and sharing of dermatologic data (Codella et al., 2018; Giotis et al., 2015; Kawahara et al., 2018; Tschandl et al., 2018).

CONCLUSIONS

We have demonstrated that deep learning algorithms trained with the substantial numbers of both Asian and Caucasian populations can be used for malignancy diagnosis, treatment suggestion, and classification of 134 diseases with a performance that is comparable to that of the experts. In addition, we demonstrated that the CNN model can serve as an ancillary tool that empowers the performance of medical professionals in diagnosing cutaneous skin diseases.

For optimal results, the submission of an image with adequate quality is required to minimize the possibility of potential false positives, especially those caused by shaded or blurry images. Future studies are warranted to evaluate the utility and performance of our algorithms in a clinical setting (Topol, 2019).

MATERIALS AND METHODS

A total of 224,181 clinical photographs were obtained with approvals of Asan Institutional Review Board (No. 2017-0087) and

Seoul National University Bundang Hospital Institutional Review Board (No. B-1802-451-005). In total, 220,680 images from four datasets (ASAN, Web, MED-NODE [Giotis et al., 2015], and Normal) were used for training the CNNs (Table 1).

For the validation of our CNN models, two datasets (Edinburgh and SNU) were used. The Edinburgh and SNU datasets were used to represent the dataset consisting of Caucasian and Asian populations, respectively. The former consisted of 1,300 images of 10 tumorous skin diseases and is commercially available (https://licensing.eri.ed.ac.uk/i/software/dermofit-image-library.html), whereas the latter consisted of clinical photographs obtained in the Department of Dermatology at Seoul National University Bundang Hospital, Inje University Sanggye Paik Hospital, and Hallym University Dongtan Hospital. Out of 174 skin diseases used for training the model, 134 general skin disorders were selected for validation from the SNU dataset. The excluded 40 diseases were either a sign or rare diseases such as ulcer, purpura, dilated pore, and senile gluteal dermatosis. At least 10 images per disease were obtained in the SNU dataset. A total of 10 tumorous disorders in the SNU dataset contained at least 30 biopsy-confirmed images per class. In total, 2,201 images with either pathological confirmation (1,057 images) or clinical diagnosis identified on medical charts (1,144 images) were included in the SNU dataset.

With Berkeley Vision and Learning Center (BVLC) Caffe (Jia et al., 2014), we fine-tuned the ImageNet pretrained models of SENet (He et al., 2015, Hu et al., 2018), SE-ResNet-50 (He et al., 2015, Hu et al., 2018), and visual geometry group (VGG)-19 (Simonyan and Zisserman, 2014) (Supplementary Methods). The algorithm was trained with 174 disease classes and produced 174 outputs for a given test image.

To appropriately address a specific binary classification in multi-class output models, we created the clusters of outputs that shared the same traits by using the disease categorization based on the classification listed in Supplementary Table S1, Category I:

- malignancy output = a sum of model outputs for five malignant tumorous disorders (melanoma, basal cell carcinoma, squamous cell carcinoma, intraepithelial carcinoma, and keratoacanthoma)
- benign output = a sum of model outputs for 40 benign tumorous disorders (seborrheic keratosis, wart, etc.)
- steroids output = a sum of model outputs for 59 disorders requiring a steroids treatment (atopic dermatitis, contact dermatitis, etc.)
- antibiotics output = a sum of model outputs for 21 disorders requiring an antibiotics treatment (cellulitis, impetigo, etc.)
- antivirals output = a sum of model outputs for five disorders requiring an antivirals treatment (herpes simplex, herpes zoster, etc.)
- antifungals output = a sum of model outputs for eight disorders requiring an antifungals treatment (tinea pedis, tinea versicolor, etc.)

These outputs were used for the analysis of binary classifications. For example, we used steroids output as a target-class output for the binary classification of steroid prediction. In the multi-class classification, we used the individual disease class output as a target-class output. In order to draw a receiver operating characteristic curve, a threshold for the target class was varied from 0 to 1 by a very small increment and the outputs that were greater than the threshold were considered as positive predictions.

For the binary classification of malignancy and treatment determination, a total of 240 images (40 images per each of the 6 group) randomly selected from the SNU dataset were used for the human tests. The test images were categorized into one of the following six groups: malignant nodule, benign lesions requiring antifungals, those requiring antivirals, those requiring antibiotics, or those requiring steroids treatment, and other benign lesions. A total of 21 dermatologists, 26 residents, and 23 non-medical professionals participated in the reader tests. After the initial test, the test participants were informed of the results obtained by using the algorithm and the test was repeated.

For the multi-disease classification of 134 disorders, two dermatologists and two dermatology residents participated in the reader test, in which they were asked to provide their top-1 and 3 choices among 134 diseases for the 2,201 images from the SNU dataset. The list of 134 diseases was given to the test participants. Following the completion of the entire test set, the test participants were informed of the algorithms' top-1, 2, and 3 diagnoses for each test image and they repeated the entire test set.

A two-tailed paired *t*-test was performed to determine if there were significant differences in the test performance before and after the assistance from the algorithm (Supplementary Table S4). We used R version 3.5.3 (pROC package version 1.14) for calculating the result of *t*-test and the AUC of receiver operating characteristic curve. The AUCs for each disease were calculated by converting the multi-class model outputs into a binary classification: one versus rest of the 173 diseases. In calculating top accuracy, if the top prediction of algorithm did not belong the 134 classes in the validation, then the prediction was counted as incorrect. To calculate mean accuracy, a macro-average was computed independently for each class and then the average was taken.

Data availability statement

The images used to train and test the neural networks described in the manuscript are subject to privacy regulations and cannot be made available in totality. We provided images in the validation datasets (SNU) as thumbnails. The 240 test images used for the human test are available for download as full-size files (https://doi.org/10.6084/m9.figshare.6454973). The test subset may be available upon a reasonable request and an approval from the originating university hospitals.

ORCIDs

Seung Seog Han: https://orcid.org/0000-0002-0500-3628
Ilwoo Park: https://orcid.org/0000-0001-6022-8363
Woohyung Lim: https://orcid.org/0000-0003-0525-9065
Myoung Shin Kim: https://orcid.org/0000-0002-0660-8098
Gyeong Hun Park: https://orcid.org/0000-0001-8890-8678
Je Byeong Chae: https://orcid.org/0000-0002-0968-3819
Chang Hun Huh: https://orcid.org/0000-0003-3944-7777
Sung Eun Chang: https://orcid.org/0000-0003-4225-0414
Jung-Im Na: https://orcid.org/0000-0002-5717-2490

CONFLICT OF INTEREST

One of the study authors (Woohyung Lim) is employed by LG Sciencepark. However, the company did not have any role in the study design, data collection and analysis, the decision to publish, or the preparation of this manuscript.

ACKNOWLEDGMENTS

The authors would like to thank the professors and clinicians who participated in the tests. The authors also thank Kim Sohyun for the assistance with the

survey part of the investigation. The authors are grateful to Park Jeoong Sung for his efforts to improve the stability of the web DEMO.

The co-first author (Ilwoo Park) was supported by the National Research Foundation of Korea grant funded by the Ministry of Science and ICT (No. 2017R1C1B5018396) along with grants from the Chonnam National University Hospital Biomedical Research Institute (CRI18019-1 and CRI18094-2). For any additional correspondence queries please contact Sung Eun Chang (csesnumd@gmail.com).

AUTHOR CONTRIBUTIONS

Conceptualization: SSH, IP, JN; Data Curation: MSK, GHP, JBC, CHH, SEC, JN; Formal Analysis: SSH, GHP; Funding Acquisition: IP; Investigation: SSH, JBC, JN; Methodology: SSH, IP, SEC, JN; Project Administration: SSH, SEC, JN; Resources: MSK, GHP, JBC, CHH, SEC, JN; Software: SSH, WL; Supervision: SSH, SEC, JN; Validation: SSH, IP, MSK, GHP, JBC, SEC, JN; Visualization: SSH, IP, WL, JBC; Writing - Original Draft Preparation: SSH, IP; Writing - Review and Editing: SSH, IP, SEC, JN

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at www.jidonline.org, and at https://doi.org/10.1016/j.jid.2020.01.019.

REFERENCES

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;111:148–54.

Brinker TJ, Hekler A, Utikal JS, Grabe N, Schadendorf D, Klode J, et al. Skin cancer classification using convolutional neural networks: systematic review. J Med Internet Res 2018;20:e11936.

Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 2018;392:2388–96.

Cho SI, Sun S, Mun JH, Kim C, Kim SY, Cho S, et al. Dermatologist-level classification of malignant lip diseases using a deep convolutional neural network. Br J Dermatol 2019. https://doi.org/10.1111/bjd.18459 (accessed September 1, 2019).

Codella NC, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza SW, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). Piscataway, NJ: IEEE; 2018. p. 168–72.

De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018;24:1342–50.

Dodge S, Karam L. A study and comparison of human and deep learning recognition performance under visual distortions. 2017 26th international conference on computer communication and networks (ICCCN). Piscataway, NJ: IEEE; 2017. p. 1–7.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8.

Farberg AS, Winkelmann RR, Tucker N, White R, Rigel DS. The impact of quantitative data provided by a multi-spectral digital skin lesion analysis device on dermatologists' decisions to biopsy pigmented lesions. J Clin Aesthet Dermatol 2017;10:24–6.

Fujisawa Y, Otomo Y, Ogata Y, Nakamura Y, Fujita R, Ishitsuka Y, et al. Deep learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumor diagnosis. Br J Dermatol 2019;180:373–81.

Giotis I, Molders N, Land S, Biehl M, Jonkman MF, Petkov N, et al. MED-NODE: a computer-assisted melanoma diagnosis system using non-dermoscopic images. Expert Syst Appl 2015;42:6578–85.

Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402–10.

Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018;29:1836–42.

Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermatol 2018a;138:1529–38.

Han SS, Park GH, Lim W, Kim MS, Na JI, Park I, et al. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: automatic construction of onychomycosis datasets by region-based convolutional deep neural network. PLoS one 2018b;13:e0191493.

He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision 2015. Piscataway, NJ: IEEE; 2015. p. 1026–34.

Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2018. Piscataway, NJ: IEEE; 2018. p. 7132–41.

Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, et al. Caffe: convolutional architecture for fast feature embedding. In: MM' 14: Proceedings of the 22nd ACM international conference on Multimedia. New York, NY: ACM; 2014. p. 675–8.

Kawahara J, Daneshvar S, Argenziano G, Hamarneh G. 7-Point Checklist and Skin Lesion Classification using multi-task multi-modal neural nets. IEEE J Biomed Health Inform 2019;23:538–46.

Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep 2016;6:26286.

Maron RC, Weichenthal M, Utikal JS, Hekler A, Berking C, Hauschild A, et al. Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. Eur J Cancer 2019;119:57–65.

Narla A, Kuprel B, Sarin K, Novoa R, Ko J. Automated classification of skin lesions: from pixels to practice. J Invest Dermatol 2018;138:2108–10.

Navarrete-Dechent C, Dusza SW, Liopyris K, Marghoob AA, Halpern AC, Marchetti MA. Automated dermatological diagnosis: hype or reality? J Invest Dermatol 2018;138:2277–9.

Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. Deep learning for chest radiograph diagnosis: a retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018;15:e1002686.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. Int J Comput Vis 2015;115:211–52.

Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Paper presented at: ICLR 2015. 7–9 May 2015; San Diego, CA.

Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44–56.

Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. Lancet Oncol 2019a;20:938–47.

Tschandl P, Rosendahl C, Akay BN, Argenziano G, Blum A, Braun RP, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. JAMA Dermatol 2019b;155:58–65.

Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 2018;5:180161.

Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep learning for identifying metastatic breast cancer. arXiv preprint 2016a;arXiv:1606.05718.

Yu C, Yang S, Kim W, Jung J, Chung KY, Lee SW, et al. Acral melanoma detection using a convolutional neural network for dermoscopy images. PLoS one 2018;13:e0193321.