European Journal of Cancer 169 (2022) 156–165

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

Convolutional neural network assistance significantly improves dermatologists' diagnosis of cutaneous tumours using clinical images

Wei Ba [a,1], Huan Wu [b,1], Wei W. Chen [c,1], Shu H. Wang [d], Zi Y. Zhang [e], Xuan J. Wei [a], Wen J. Wang [a], Lei Yang [f], Dong M. Zhou [c], Yi X. Zhuang [g], Qin Zhong [h], Zhi G. Song [i,**], Cheng X. Li [a,*]

[a] *Department of Dermatology, Chinese PLA General Hospital, Beijing 100853, China*
[b] *Research of Medical Big Data Center, Chinese PLA General Hospital, Beijing 100853, China*
[c] *Department of Dermatology, Beijing Hospital of Traditional Chinese Medicine, Affiliated to the Capital Medical University, Beijing 100010, China*
[d] *Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China*
[e] *Department of Dermatology, Affiliated Hospital of North China University of Science and Technology, Tangshan 063000, China*
[f] *Department of Dermatology, Rocket Force Characteristic Medical Center, Beijing 100088, China*
[g] *Dysion Intelligent, Beijing 100038, China*
[h] *The Medical School of Chinese PLA General Hospital, Beijing 100853, China*
[i] *Department of Pathology, Chinese PLA General Hospital, Beijing 100853, China*

Received 6 February 2022; received in revised form 20 March 2022; accepted 7 April 2022
Available online 12 May 2022

KEYWORDS
Convolutional neural network;
Clinical image;
Cutaneous tumours;
Multi-reader multi-case (MRMC) study;
Fully crossed

Abstract ***Background:*** Convolutional neural networks (CNNs) have demonstrated expert-level performance in cutaneous tumour classification using clinical images, but most previous studies have focused on dermatologist-versus-CNN comparisons rather than their combination. The objective of our study was to evaluate the potential impact of CNN assistance on dermatologists for clinical image interpretation.

Methods: A multi-class CNN was trained and validated using a dataset of 25,773 clinical images comprising 10 categories of cutaneous tumours. The CNN's performance was tested on an independent dataset of 2107 images. A total of 400 images (40 per category) were randomly selected from the test dataset. A fully crossed, self-control, multi-reader multi-case (MRMC) study was conducted to compare the performance of 18 board-certified dermatologists

* *Corresponding author.*
** *Corresponding author.*
E-mail address: songzhg301@139.com (Z.G. Song), chengxinderm@163.com (C.X. Li).

[1] Wei Ba, Huan Wu and Wei W. Chen contribute equally to this work.

https://doi.org/10.1016/j.ejca.2022.04.015

(experience: 13/18 ≤ 10 years; 5/18 > 10 years) in interpreting the 400 clinical images with or without CNN assistance.

Results: The CNN achieved an overall accuracy of 78.45% and kappa of 0.73 in the classification of 10 types of cutaneous tumours on 2107 images. CNN-assisted dermatologists achieved a higher accuracy (76.60% vs. 62.78%, $P < 0.001$) and kappa (0.74 vs. 0.59, $P < 0.001$) than unassisted dermatologists in interpreting the 400 clinical images. Dermatologists with less experience benefited more from CNN assistance. At the binary classification level (malignant or benign), the sensitivity (89.56% vs. 83.21%, $P < 0.001$) and specificity (87.90% vs. 80.92%, $P < 0.001$) of dermatologists with CNN assistance were also significantly improved than those without.

Conclusions: CNN assistance improved dermatologist accuracy in interpreting cutaneous tumours and could further boost the acceptance of this new technique.

1. Introduction

One of the deep learning algorithms, the convolutional neural network (CNN), has been proven to have great potential for classifying images and detecting objects in pictures [1–3]. Due to the advantage of self-regulated learning of CNN and large volumes of cutaneous image accumulation, there has been growing research interest in using CNN to identifying skin tumours [4–7].

Many studies have consistently demonstrated that CNNs could achieve on par or even better performance than human dermatologists in clinical image classification [6,8–11]. It is notable that an accurate CNN will not replace the breadth and contextual knowledge of dermatologist; rather, only through their combination may the CNN benefits be achieved [12,13]. However, most previous studies have focused on dermatologist-versus-CNN comparisons, only few studies investigated the potential of dermatologist-CNN collaboration [13–17]. Han *et al.* evaluated the performance of dermatologists with CNN assistance in clinical image classification by modifying initial diagnoses after being informed of CNN predictions [14]. This process was different from making diagnosis according to CNN prediction in a real assistance modality. Hekler *et al.* developed a fusion algorithm to simulate dermatologist with CNN assistance in identifying dermoscopic image [15]. Lee *et al.* conducted a study to evaluate physicians with CNN assistance in detecting acral lentiginous melanoma using dermoscopic images [17]. Maron *et al.* evaluated dermatologists with CNN assistance in classifying melanoma and nevus using dermoscopic images [16]. However, Lee and Maron studies only included melanocytic lesions. Tschandl *et al.* conducted a seminal research on human–computer collaboration in interpreting dermoscopic images [13]. But Tschandl study didn't clearly define the washout period between two sessions, which may lead to memory bias [18–20]. Furthermore, above studies were conducted firstly without and then with CNN support (sequential design). Compared to cross-over design, reading scenarios (with or without CNN assistance) were not randomised in sequential design, which may lead to bias in performance measures [18,21].

Based on these considerations, we developed a multiclass CNN to classify 10 types of skin tumours and conducted a fully crossed, self-control, multi-reader multi-case (MRMC) study to evaluate the performance of dermatologists in interpreting clinical images with or without CNN assistance.

2. Materials and methods

2.1. Dataset

The study was approved by the institutional review board of the Chinese PLA General Hospital & Medical School (Approval no. 95S2019-123-01). Data on patient demographics and clinical images between 2009 and 2020 were collected via a retrospective chart review, and patient consent was waived by the institutional review board.

Basal cell carcinoma (BCC), squamous cell carcinoma (SCC) and melanoma (MM) are the most common types of skin cancer in populations [22–24]. These malignant entities show increasing incidence rates worldwide [22,25,26]. And mimicking entities regularly create diagnostic dilemmas in clinical practice settings. Clinical images with the following diagnoses were included: BCC, SCC including keratoacanthoma, MM, Bowen disease (BD), actinic keratosis (AK), melanocytic nevus (MN), seborrhoeic keratosis (SK), haemangioma including pyogenic granuloma, cherry haemangioma, sinusoidal haemangioma and angiokeratoma, dermatofibroma (DF) and wart. All of the diagnoses were based on pathological examinations. The clinical images were taken by cameras (Canon EOS 550D, Nikon D610 and

Canon EOS 70D) and smartphones (iPhone 6 s Plus, 7 Plus, 11 Pro, 11 Pro Max and HUAWEI P30, P40, Mate 30Pro).

Fig. 1 shows the data flow of the study. The overall data collection was split into training/validation and testing datasets. Of the 29,280 images, 25,773 images were allocated for training and validation. In the remaining 3507 images, 2107 eligible images (excluding 1400 unqualified images) were used for testing (Table 1). The excluded images contained two components. (1) Blurry and distant images were deleted from the test set but were still used for training. (2) In the test set, each image corresponds to one lesion (no overlap derived from multiple viewpoints) to maintain the diversity of the test set and to ensure that the test set reflected the real performance of the CNN.

For a number of the patients, multiple images were from the same lesion, including images taken from different distances and angles. These images were divided into training or test dataset according to the specific lesion to prevent images from the same lesion being used in both training and testing.

2.2. Multi-class CNN development

We trained our CNN based on the EfficientNet-B3 architecture, which was pre-trained on the ImageNet dataset. The model has 27 layers and 12 million parameters, which is suitable for real-time calculation with high accuracy using mobile phones. A total of 25,773 clinical images were used to train the model. An 80/20 training/validation split was conducted to perform cross-validation. The data splitting was conducted according to the lesions to avoid images from the same lesion being used in both training and validation.

During training, each image was trimmed to 720 × 720 pixels to set the lesion at the centre of the image. For each image, we used a 300 × 300 pixel sliding window to sample ordered, same-sized patches as the CNN input. In the test phase, the CNN produced 10 outputs for a given patch. The image prediction was obtained by averaging all the patches' probabilities. Colour constancy was performed on all images during training and validation using the Shades of Grey method with Minkowski norm $p = 6$ [27].

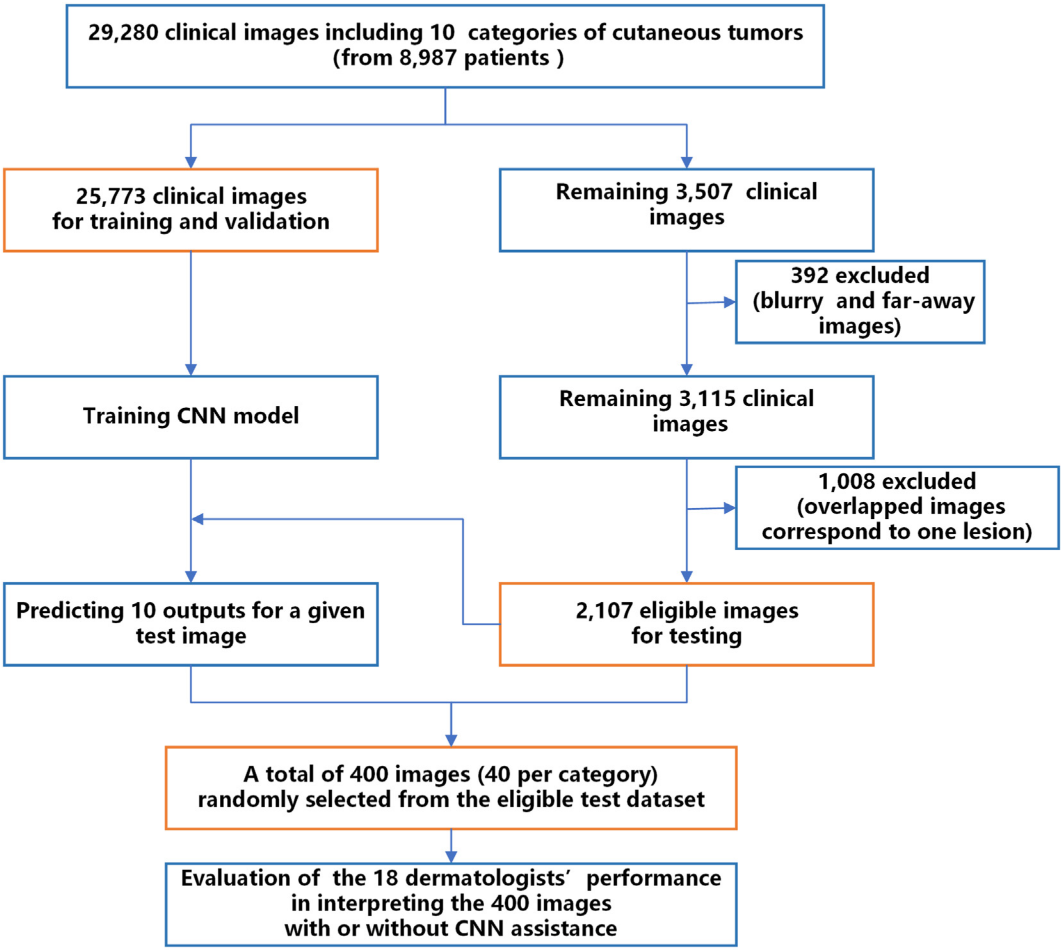

29,280 clinical images including 10 categories of cutaneous tumors (from 8,987 patients)
25,773 clinical images for training and validation
Remaining 3,507 clinical images
392 excluded (blurry and far-away images)
Training CNN model
Remaining 3,115 clinical images
1,008 excluded (overlapped images correspond to one lesion)
Predicting 10 outputs for a given test image
2,107 eligible images for testing
A total of 400 images (40 per category) randomly selected from the eligible test dataset
Evaluation of the 18 dermatologists' performance in interpreting the 400 images with or without CNN assistance

Fig. 1. Data flow of the CNN assistance study. The contents in the orange box represent the data sources of the training (validation), testing and evaluation study.

Table 1
Clinical images in the dataset.

Diagnosis of dataset	Training dataset [a] (n)	Testing dataset [b] (n)	Total (n)
Patient demographics			
Unique individuals (n)	8987		
Age in years, mean ± SD	42.77 ± 21.68		
Basal cell carcinoma	2053	125	2178
Squamous cell carcinoma	1031	77	1108
Melanoma	931	99	1030
Bowen	480	82	562
Actinic keratosis	748	104	852
Nevus	6065	849	6914
Seborrhoeic keratosis	7613	360	7973
Haemangioma	2276	142	2418
Dermatofibroma	2991	151	3142
Wart	1585	118	1703
Total	25,773	2107	27,880

[a] In the training dataset, multiple images from the same lesion (taken at different distances and angles) and blurry images and far-away images could be included.

[b] In the testing dataset, each image corresponded to one lesion (no overlap existing from multiple viewpoints) to maintain the diversity of the test set.

Data augmentation in our study included random rotations and random flips (horizontal and vertical) [28]. To address the images unbalanced in different tumours, we trained CNN with weighted loss functions where weights were determined using the inverse frequency of classes in the training data. Specifically, the weight of the class with the largest images (Nmax) was 1.0, and the image number of class k was Nk. Then, the weight of the image from class k was represented as Nmax/Nk. The developed CNN can be accessed at the following URL (http://ai.skinreader.cn/) and has an online prediction function.

2.3. Test set enrolment for evaluation

To make a statistically adequate comparison for each category of cutaneous tumour, with or without CNN assistance, while keeping a feasible total number of clinical images, a total of 400 images (40 images per category) were randomly selected from the test dataset.

2.4. Dermatologists

A total of 18 board-certified dermatologists from nine different hospitals participated as readers in this study. Their years of clinical work experience ranged from 1 to 20 years. The majority of them were not experts (experience: 13/18 ≤ 10 years; 5/18 > 10 years). All dermatologists voluntarily participated, understood and agreed with the basic principles and objectives of this study.

2.5. Study design

To evaluate dermatologists' performance in both assisted and unassisted conditions, the study was designed as a fully crossed, self-control, MRMC study (Fig. 2). Dermatologists interpreted all the clinical images in both modalities (with or without CNN assistance), which was separated by a washout period of 2 months. In each modality, dermatologists were only offered one clinical image for their diagnosis and no additional information was supplied.

To reduce possible performance differences at the beginning versus the end of interpreting the test set, 400 clinical images were divided into blocks of 40 images with each block containing a roughly similar distribution of tumour categories with a random order. Eighteen dermatologists were randomised into two groups, either of which began with (mode 1) or without (mode 2) CNN assistance first. In either mode, the order and images interpreted were identical; the difference was solely in modality (assisted or not assisted).

The dermatologists independently assessed 400 clinical images on a 24" LED monitor (Lenovo Q24i-1L) at a self-controlled pace. When the dermatologists reviewed the images, modalities (with or without CNN assistance) switched every 40 image intervals. If the image had CNN assistance, the top 3 predictions of the CNN appeared in the upper right corner of the corresponding image. If the image was without CNN assistance, only a clinical image was displayed. The participants gave the most likely diagnosis to the corresponding image (Supplement Fig. 1).

2.6. Statistical analysis

The predicted results of the multi-class CNN on 2107 test images were shown with a confusion matrix. The performance of the CNN model was evaluated by accuracy and kappa coefficient. In the performance evaluation of dermatologists with or without CNN assistance, the primary endpoint was the overall accuracy of dermatologists with pathological diagnosis as the gold standard. The sensitivity and specificity of binary classification was the secondary evaluation endpoint, and kappa consistency analysis was used to evaluate the consistency between dermatologist's diagnoses (with or without CNN assisted) and pathological results. Dermatologists' accuracy and kappa with or without assistance were compared by a two-sided paired *t* test or Wilcoxon signed-rank test, as appropriate. On the binary classification level, the sensitivity and specificity between assisted and unassisted modes were also analysed using a two-sided paired *t* test. And $P < 0.05$ was regarded as statistically significant. Statistical analyses were performed in R (version 4.1) software.

3. Results

3.1. Performance of the multi-class CNN

The performance of the multi-class CNN was evaluated using 2107 test images. The CNN produced 10 outputs

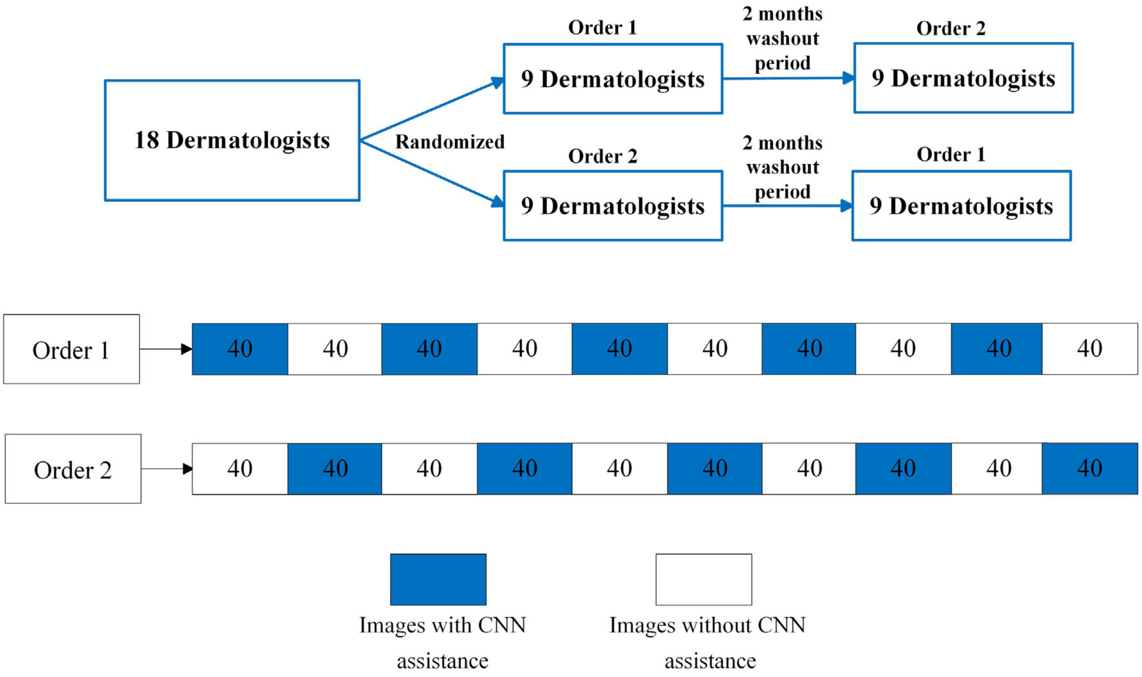


Fig. 2. Study design. The 18 dermatologists reviewed the same clinical images in the same sequence but with different modalities: with or without CNN assistance after a washout period of 2 months. Eighteen dermatologists were randomised to one of the two assistance "orders." Each rectangle included 40 images in random order. The colour of the rectangle indicates the modality. In either mode, the order and images interpreted were identical; the difference was solely in modality (assisted or not assisted).

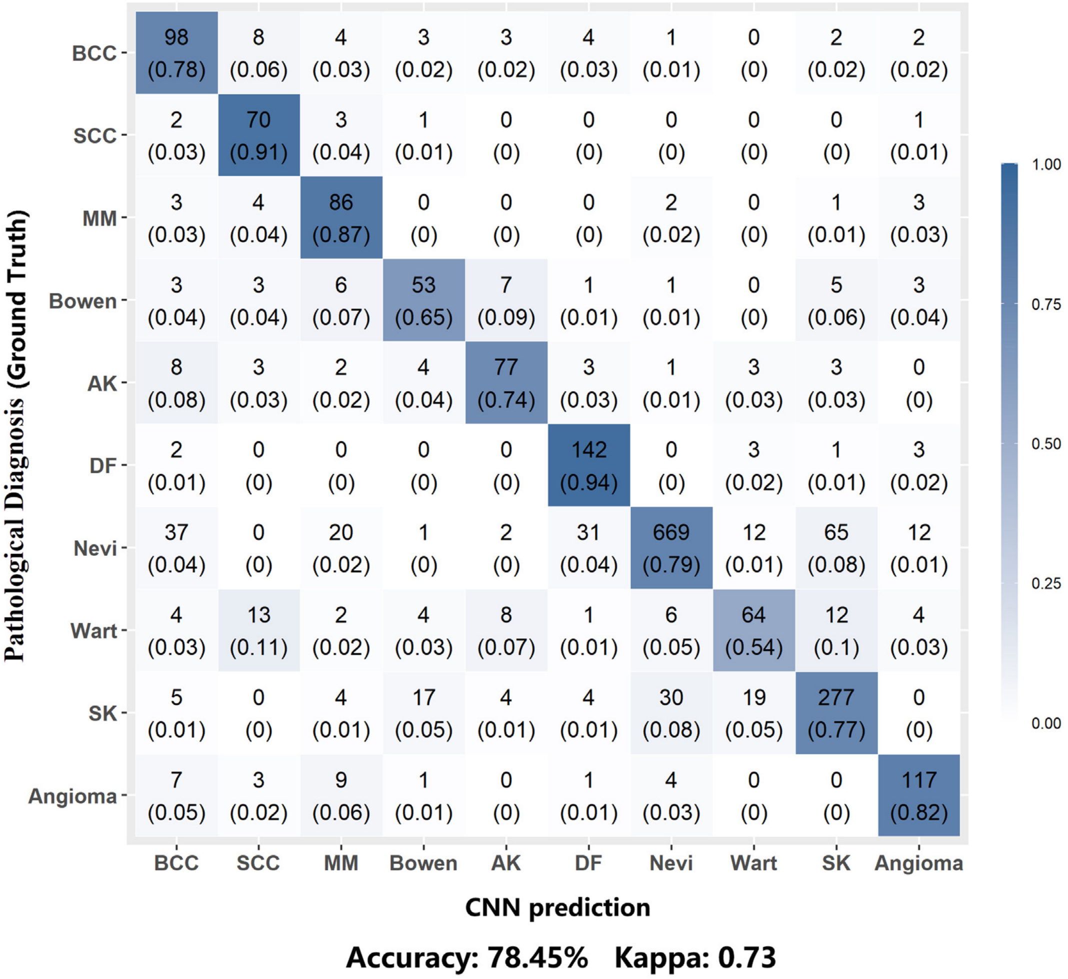
for a given image. The top 1 prediction was used to evaluate the performance of the model. The results were presented with a confusion matrix (Fig. 3). The overall accuracy of the CNN on the top 1 diagnosis level was 78.45%. The accuracies for BCC, SCC, MM, Bowen, AK, DF, Nevi, Wart, SK and angioma were 78.40%, 90.91%, 86.87%, 64.63%, 74.04%, 94.04%, 78.80%, 54.24%, 76.94% and 82.39%, respectively. In addition to accuracy, the kappa coefficient is an indicator for the consistency test and can be used to measure the effect of a multi-class model, especially in the situation of imbalanced sample numbers. The kappa index of the CNN on 2107 test images was 0.73, which demonstrates that the predictions of CNN are moderately consistent with the gold standard diagnoses.

3.2. Performance of dermatologists with or without CNN assistance

A total of 400 images (40 images per category) were randomly selected from the test set to evaluate the performance of dermatologists with or without CNN assistance. The participants gave the most likely diagnosis to each image in each modality (Supplement Fig. 1). For each category, 720 predictions (40 images × 18 dermatologists) were collected with or without CNN assistance. The CNN's detailed performance and each dermatologist's results on the 400 images were shown with confusion matrixs (Supplement Fig. 2). To reflect the overall performance of 18 dermatologists with or without CNN assistance, 18 confusion matrices were accumulated under two different modalities (Fig. 4a). The overall accuracy and kappa of dermatologists were 62.78% and 0.59 without assistance, while 76.60% and 0.74 with assistance. Compared with no assistance, CNN assistance improved the accuracy of dermatologists by 13.82% ($P < 0.001$, 95% confidence interval [CI]: 11.83%–15.80%) and kappa by 0.15 ($P < 0.001$, 95% CI: 0.131–0.176) (Fig. 4b), which demonstrated that CNN improved the diagnostic performance of the dermatologists. The diagnostic improvement with CNN assistance depended on the experience of the dermatologists. The majority of dermatologists were not experts (experience: 13/18 ≤ 10 years; 5/18 > 10 years). From Fig. 4c, we could conclude that dermatologists with less experience benefited more from CNN assistance.

For each category of cutaneous tumour, except for wart, the accuracies of BCC, SCC, MM, Bowen, AK, DF, Nevi, SK and angioma with CNN assistance were significantly improved compared with those without CNN assistance. A summary of the results is shown in Fig. 5a (BCC, SCC, MM and Bowen) and Supplemental Fig. 3 (AK, DF, Nevi, wart, SK and angioma). Fig. 5b shows six representative examples indicating that the performance of dermatologists significantly improved with CNN assistance.

According to the dermatologist's diagnosis, BCC, MM, SCC and Bowen were clustered as malignant, and AK, SK, MN, DF, haemangioma and wart were



	BCC	SCC	MM	Bowen	AK	DF	Nevi	Wart	SK	Angioma
BCC	98 (0.78)	8 (0.06)	4 (0.03)	3 (0.02)	3 (0.02)	4 (0.03)	1 (0.01)	0 (0)	2 (0.02)	2 (0.02)
SCC	2 (0.03)	70 (0.91)	3 (0.04)	1 (0.01)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.01)
MM	3 (0.03)	4 (0.04)	86 (0.87)	0 (0)	0 (0)	0 (0)	2 (0.02)	0 (0)	1 (0.01)	3 (0.03)
Bowen	3 (0.04)	3 (0.04)	6 (0.07)	53 (0.65)	7 (0.09)	1 (0.01)	1 (0.01)	0 (0)	5 (0.06)	3 (0.04)
AK	8 (0.08)	3 (0.03)	2 (0.02)	4 (0.04)	77 (0.74)	3 (0.03)	1 (0.01)	3 (0.03)	3 (0.03)	0 (0)
DF	2 (0.01)	0 (0)	0 (0)	0 (0)	0 (0)	142 (0.94)	0 (0)	3 (0.02)	1 (0.01)	3 (0.02)
Nevi	37 (0.04)	0 (0)	20 (0.02)	1 (0)	2 (0)	31 (0.04)	669 (0.79)	12 (0.01)	65 (0.08)	12 (0.01)
Wart	4 (0.03)	13 (0.11)	2 (0.02)	4 (0.03)	8 (0.07)	1 (0.01)	6 (0.05)	64 (0.54)	12 (0.1)	4 (0.03)
SK	5 (0.01)	0 (0)	4 (0.01)	17 (0.05)	4 (0.01)	4 (0.01)	30 (0.08)	19 (0.05)	277 (0.77)	0 (0)
Angioma	7 (0.05)	3 (0.02)	9 (0.06)	1 (0.01)	0 (0)	1 (0.01)	4 (0.03)	0 (0)	0 (0)	117 (0.82)

Fig. 3. Performance of the convolutional neural network (CNN) in top 1 classification on 2107 test images.

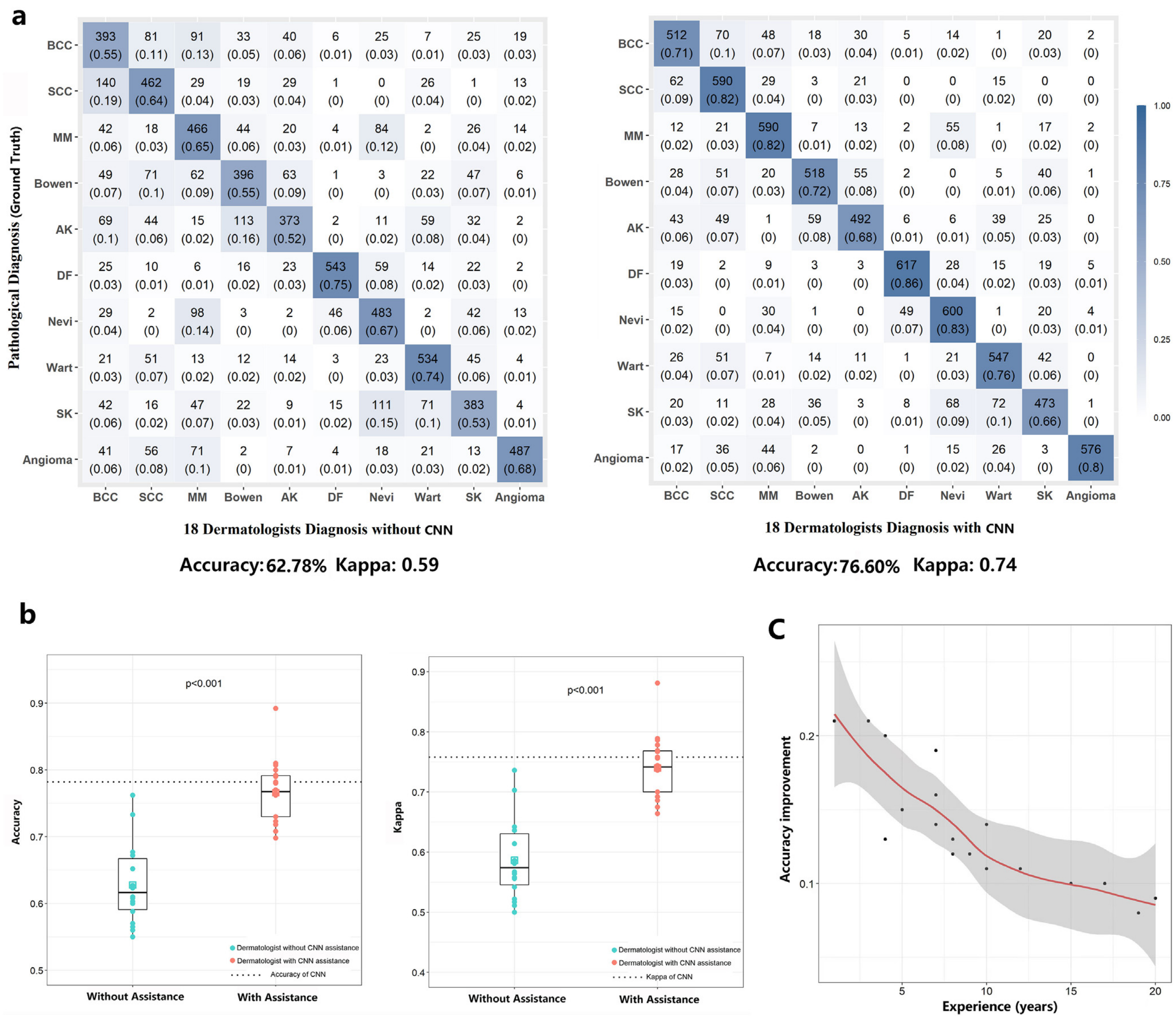
clustered as benign. On the binary classification level, the average sensitivities of the dermatologists without and with CNN assistance were 83.21% and 89.56%. The sensitivity improved by 6.35% ($P < 0.001$, 95% CI: 4.61%–8.10%) with assistance compared with that without assistance. The average specificities of the dermatologists without and with CNN assistance were 80.92% and 87.90%. The specificity improved by 6.98% ($P < 0.001$, 95% CI: 4.80%–9.20%) with assistance than that without assistance (Fig. 6).

4. Discussion

Studies have demonstrated that CNNs could achieve comparable or better performance than board-certified dermatologists in clinical image classification [6,29–31]. However, CNN with full automation with no dermatologist backup is not the objective. Rather, it is through their combination that the significant benefits of CNN may be achieved [4,12,13]. Therefore, we designed a fully crossed MRMC study to investigate the potential of CNN assistance for dermatologists in interpreting clinical images. Our results demonstrated that CNN assistance clearly increased the accuracy of dermatologists in identifying cutaneous tumours.

Our multi-class' CNN was developed on 25,773 clinical images and achieved a top 1 accuracy of 78.45% and kappa of 0.73 in classifying 10 types of skin tumours on 2107 images. Fujisawa *et al.* trained a CNN using 4867 clinical images and tested its performance on 1142 images including 14 diagnoses, which achieved an overall accuracy of 76.5% [8]. Esteva *et al.* trained a CNN using a dataset of 129,450 clinical images consisting of 2032 different diseases and tested its performance on 1942 images, achieving an overall accuracy of 55% in classifying nine types of skin diseases [31]. Our CNN achieved a comparable or better performance than previous studies.

However, a well-developed CNN with high accuracy alone was not sufficient, as our objective was to conduct further evaluation of CNN-dermatologist collaboration. The overall accuracy of dermatologists improved significantly with CNN assistance compared to that without CNN assistance. Furthermore, dermatologist



18 Dermatologists Diagnosis without CNN

	BCC	SCC	MM	Bowen	AK	DF	Nevi	Wart	SK	Angioma
BCC	393 (0.55)	81 (0.11)	91 (0.13)	33 (0.05)	40 (0.06)	6 (0.01)	25 (0.03)	7 (0.01)	25 (0.03)	19 (0.03)
SCC	140 (0.19)	462 (0.64)	29 (0.04)	19 (0.03)	29 (0.04)	1 (0)	0 (0)	26 (0.04)	1 (0)	13 (0.02)
MM	42 (0.06)	18 (0.03)	466 (0.65)	44 (0.06)	20 (0.03)	4 (0.01)	84 (0.12)	2 (0)	26 (0.04)	14 (0.02)
Bowen	49 (0.07)	71 (0.1)	62 (0.09)	396 (0.55)	63 (0.09)	1 (0)	3 (0)	22 (0.03)	47 (0.07)	6 (0.01)
AK	69 (0.1)	44 (0.06)	15 (0.02)	113 (0.16)	373 (0.52)	2 (0)	11 (0.02)	59 (0.08)	32 (0.04)	2 (0)
DF	25 (0.03)	10 (0.01)	6 (0.01)	16 (0.02)	23 (0.03)	543 (0.75)	59 (0.08)	14 (0.02)	22 (0.03)	2 (0)
Nevi	29 (0.04)	2 (0)	98 (0.14)	3 (0)	2 (0)	46 (0.06)	483 (0.67)	2 (0)	42 (0.06)	13 (0.02)
Wart	21 (0.03)	51 (0.07)	13 (0.02)	12 (0.02)	14 (0.02)	3 (0)	23 (0.03)	534 (0.74)	45 (0.06)	4 (0.01)
SK	42 (0.06)	16 (0.02)	47 (0.07)	22 (0.03)	9 (0.01)	15 (0.02)	111 (0.15)	71 (0.1)	383 (0.53)	4 (0.01)
Angioma	41 (0.06)	56 (0.08)	71 (0.1)	2 (0)	7 (0.01)	4 (0.01)	18 (0.03)	21 (0.03)	13 (0.02)	487 (0.68)

18 Dermatologists Diagnosis with CNN

	BCC	SCC	MM	Bowen	AK	DF	Nevi	Wart	SK	Angioma
BCC	512 (0.71)	70 (0.1)	48 (0.07)	18 (0.03)	30 (0.04)	5 (0.01)	14 (0.02)	1 (0)	20 (0.03)	2 (0)
SCC	62 (0.09)	590 (0.82)	29 (0.04)	3 (0)	21 (0.03)	0 (0)	0 (0)	15 (0.02)	0 (0)	0 (0)
MM	12 (0.02)	21 (0.03)	590 (0.82)	7 (0.01)	13 (0.02)	2 (0)	55 (0.08)	1 (0)	17 (0.02)	2 (0)
Bowen	28 (0.04)	51 (0.07)	20 (0.03)	518 (0.72)	55 (0.08)	2 (0)	0 (0)	5 (0.01)	40 (0.06)	1 (0)
AK	43 (0.06)	49 (0.07)	1 (0)	59 (0.08)	492 (0.68)	6 (0.01)	6 (0.01)	39 (0.05)	25 (0.03)	0 (0)
DF	19 (0.03)	2 (0)	9 (0.01)	3 (0)	3 (0)	617 (0.86)	28 (0.04)	15 (0.02)	19 (0.03)	5 (0.01)
Nevi	15 (0.02)	0 (0)	30 (0.04)	1 (0)	0 (0)	49 (0.07)	600 (0.83)	1 (0)	20 (0.03)	4 (0.01)
Wart	26 (0.04)	51 (0.07)	7 (0.01)	14 (0.02)	11 (0.02)	1 (0)	21 (0.03)	547 (0.76)	42 (0.06)	0 (0)
SK	20 (0.03)	11 (0.02)	28 (0.04)	36 (0.05)	3 (0)	8 (0.01)	68 (0.09)	72 (0.1)	473 (0.66)	1 (0)
Angioma	17 (0.02)	36 (0.05)	44 (0.06)	2 (0)	0 (0)	1 (0)	15 (0.02)	26 (0.04)	3 (0)	576 (0.8)

Fig. 4. (a) The accumulated classification of 18 dermatologists without and with CNN assistance. (b) The average accuracy of dermatologists with CNN assistance was higher than that of without assistance (76.60% vs. 62.78%, P < 0.001). The average kappa of the dermatologists was improved with assistance compared to that without assistance (0.74 vs. 0.59, P < 0.001). The circles represent the value of each dermatologist with different assistance modalities, the squares indicate the average in that modality, and the vertical lines of the box represent quartiles. (C) Dermatologist with less experience tended to get larger accuracy improvement from CNN assistance (The red line representing fitting curve, shaded area representing 95% confidence intervals).

with less experience tended to get larger accuracy improvement from CNN assistance. Nevertheless, we noticed that, although the average accuracy of dermatologists with CNN assistance significantly improved, it was still lower than the accuracy of CNN alone (Fig. 4b). This may due to dermatologists selectively trusting the predictions of the CNN based on their clinical experience, their confidence about the diagnosis, and the relative difference between the multi-class probabilities. This assumption was also confirmed by a previous study [13].

The CNN output was not only possible diagnoses but calculated probabilities for each diagnosis. The specific value of probability also has an impact on dermatologists. In assisted modality, the top 3 predictions of each image were presented to dermatologists. For example, even if an MM was predicted as a benign nevus with the highest probability at 0.528, while the probability of MM was calculated as 0.231 and BCC was 0.079. Dermatologists understand the implications of false positive and false negative for patients; they may still perform a biopsy to avoid a false negative result, even though the probability was 0.528 (nevus) vs. 0.231 (melanoma).

For each category of cutaneous tumours, except for wart, the accuracy of the remaining nine tumours significantly improved with CNN assistance compared

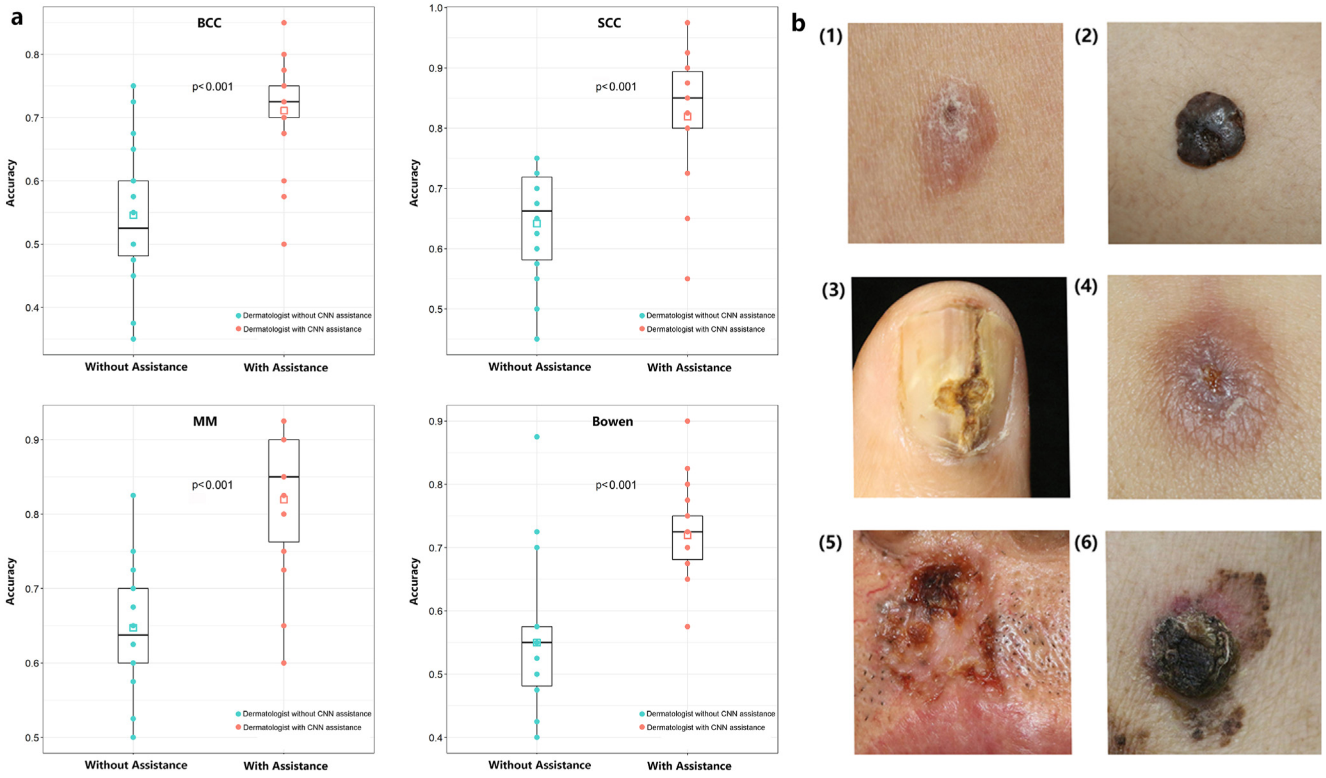


Fig. 5. (a) The accuracies of dermatologists in classifying BCC, SCC, MM and Bowen significantly improved with CNN assistance compared to those without assistance. (b) Representative examples showing that the accuracy of dermatologists significantly improved after CNN assistance. The pathological diagnosis, top 3 output of CNN and number of dermatologists with correct diagnoses were described as follows. (1) Bowen; Top 3 prediction: Bowen (0.603), DF (0.362), SK (0.015); improvement: 3/18 → 10/18. (2) BCC; Top-3 prediction: BCC (0.414), MM (0.231), SK (0.211); improvement: 7/18 → 13/18. (3) MM; Top-3 prediction: MM (0.993), SCC (0.002), Haemangioma (0.002); improvement: 10/18 → 15/18. (4) DF; Top-3 prediction: DF (0.976), Nevus (0.007), BCC (0.005); improvement: 12/18 → 17/18. (5) BCC; Top-3 prediction: BCC (0.679), SCC (0.253), SK (0.023); improvement: 9/18 → 14/18. (6) Bowen; Top-3 prediction: Bowen (0.397), MM (0.325), SCC (0.121); Improvement: 8/18 → 14/18.

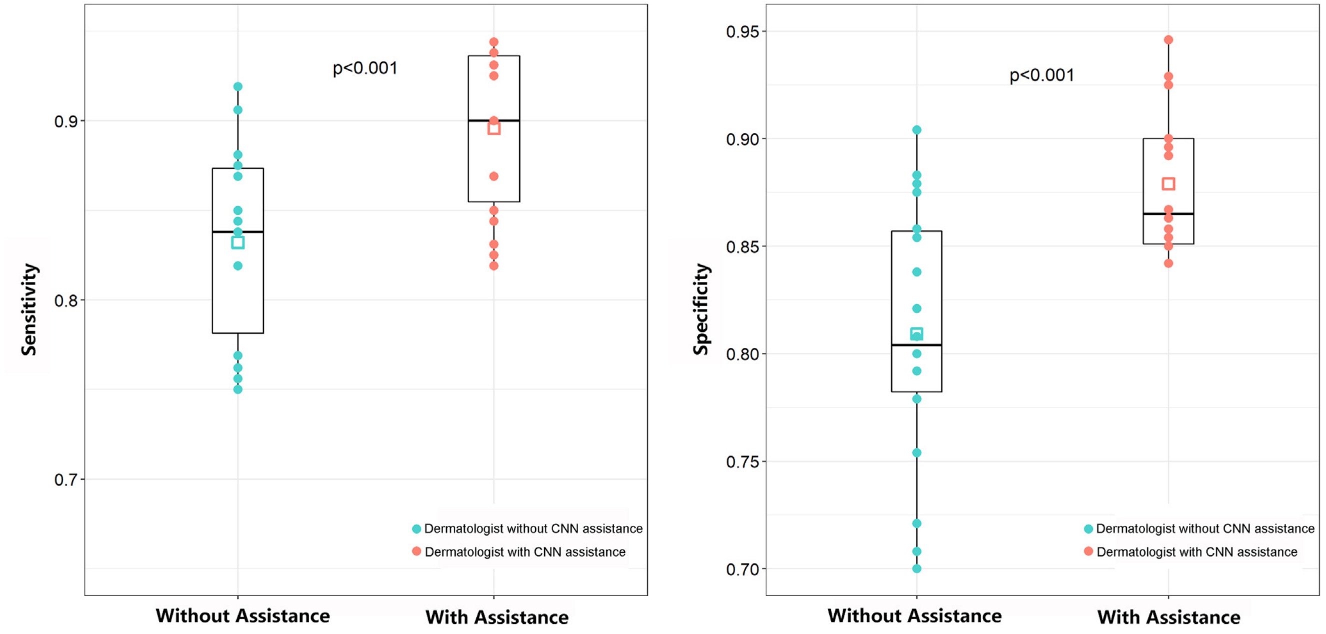


Fig. 6. At the binary classification level, the sensitivity (89.56% vs. 83.21%, $P < 0.001$) and specificity (87.90% vs. 80.92%, $P < 0.001$) were significantly improved with CNN assistance compared with those without assistance.

to that without CNN assistance. This was mainly because morphological differences were obvious between wart and the other nine tumours, leading to limited effect of CNN assistance on dermatologists. For cases with similar visual appearances (examples in Fig. 5b), CNN assistance significantly improved the diagnostic accuracy. This may be that the data-driven CNN could learn a variety of clinical manifestations of the tumour and discover distinguishing features automatically that might not be apparent to humans [32–35]. We assume that these kinds of situations often occur when dermatologists do not have extensive clinical experience or work in haste, such as being overloaded with work or evaluating the last several patients of the day in the clinical workflow. CNN, as an analogue to a second opinion from a fellow dermatologist, predicts a possible diagnosis, which may alert dermatologists to reconsider the potential diagnosis and improve patient care.

There are also several limitations to our study, mainly stemming from the assessment being performed as a simulation rather than in actual clinical practice. The test set was enriched for cutaneous tumours, which is not directly comparable to the mixed cases during the clinical workflow. In our study, each dermatologist was given one image to make the diagnosis. In a real clinical setting, dermatologists could access additional clinical data, such as the clinical history, size, location, tactility or dermoscopic image, which might improve diagnostic performance. In addition, our training data are mainly from Asian patients. The clinical features of BCC in Asians differ from those in Caucasians [36,37]. The most common melanoma subtype in Asians is acral melanoma, accounting for more than 50% of the total melanoma incidence, which is different from Caucasians [38–40]. It is possible that the CNN is likely to perform less effectively on external data. Furthermore, the images in the training and test sets only include lesions with a histopathology report, which may lead to a reduced CNN performance in completely banal skin lesions. Future studies could use images from more diverse datasets, making the system universally useful, and prospective studies are needed to further verify the feasibility of this auxiliary tool.

In summary, our study demonstrated that the combination of CNN and human dermatologists has the potential to improve the diagnostic accuracy of cutaneous tumours using clinical images. This research is a useful attempt to understand how CNN improves the performance of dermatologists. Therefore, it could further boost the dermatologists' acceptance of this new technique.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Data availability

The data that support the findings of this study may be available upon a reasonable request and an approval from the PLA general hospital.

Model availability

The multi-class CNN model is available from the following website (http://ai.skinreader.cn/).

Author contributions

Concept and design: Wei Ba, Huan Wu, Weiwen Chen, Zhigang Song and Chengxin Li, **Search and collection of the data:** all authors, **Experiment conduction:** Wei Ba, Huan Wu, Weiwen Chen and Chengxin Li, **Analysis of data and interpretation:** all authors, **Statistical analysis:** Wei Ba, Huan Wu and Weiwen Chen **Manuscript writing and review:** Wei Ba, Huan Wu, Zhigang Song and Chengxin Li.

Conflict of interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejca.2022.04.015.

References

[1] Rawat W, Wang Z. Deep convolutional neural networks for image classification: a comprehensive review. Neural Comput 2017;29:2352–449.

[2] He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019;25:30–6.

[3] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med 2019;25:24–9.

[4] Han SS, Moon IJ, Lim W, Suh IS, Lee SY, Na JI, et al. Keratinocytic skin cancer detection on the face using region-based convolutional neural network. JAMA Dermatol 2020;156:29–37.

[5] 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.

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

[7] Young AT, Xiong M, Pfau J, Keiser MJ, Wei ML. Artificial intelligence in dermatology: a primer. J Invest Dermatol 2020;140: 1504–12.

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

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

[10] 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.

[11] Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. Nat Med 2020;26:900–8.

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

[13] Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human-computer collaboration for skin cancer recognition. Nat Med 2020;26:1229–34.

[14] Han SS, Park I, Eun Chang S, Lim W, Kim MS, Park GH, et al. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. J Invest Dermatol 2020;140:1753–61.

[15] Hekler A, Utikal JS, Enk AH, Hauschild A, Brinker TJ. Superior skin cancer classification by the combination of human and artificial intelligence. Eur J Cancer 2019;120:114–21.

[16] Maron RC, Utikal JS, Hekler A, Hauschild A, Sattler E, Sondermann W, et al. Artificial intelligence and its effect on dermatologists' accuracy in dermoscopic melanoma image classification: web-based survey study. J Med Internet Res 2020;22:e18091.

[17] Lee S, Chu YS, Yoo SK, Choi S, Choe SJ, Koh SB, et al. Augmented decision-making for acral lentiginous melanoma detection using deep convolutional neural networks. J Eur Acad Dermatol Venereol 2020;34:1842–50.

[18] Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data in-Premarket Notification (510(k)) https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-performance-assessment-considerations-computer-assisted-detection-devices-applied-radiology.

[19] Cesana BM, Antonelli P, Chiumello D. Statistical methods for evidence-based medicine: the diagnostic test. Part I. Minerva Anestesiol 2008;74:431–7.

[20] Rudolfer SM. Statistical methods in diagnostic medicine. In: Biometrics. vol. 59. New York: Wiley; 2002. p. 203–4. 2015.

[21] Obuchowski NA, Meziane M, Dachman AH, Lieber ML, Mazzone PJ. What's the control in studies measuring the effect of computer-aided detection (CAD) on observer performance? Acad Radiol 2010;17:761–7.

[22] Leiter U, Keim U, Garbe C. Epidemiology of skin cancer: update 2019. Adv Exp Med Biol 2020;1268:123–39.

[23] Lomas A, Leonardi-Bee J, Bath-Hextall F. A systematic review of worldwide incidence of nonmelanoma skin cancer. Br J Dermatol 2012;166:1069–80.

[24] Siegel RL, Miller KD, Fedewa SA, Ahnen DJ, Meester RGS, Barzi A, et al. Colorectal cancer statistics. CA Cancer J Clin 2017; 67:177–93. 2017.

[25] Rogers HW, Weinstock MA, Harris AR, Hinckley MR, Feldman SR, Fleischer AB, et al. Incidence estimate of non-melanoma skin cancer in the United States, 2006. Arch Dermatol 2010;146:283–7.

[26] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68:394–424.

[27] Finlayson GD, Trezzi E. Shades of gray and colour constancy. Color and imaging conference. 2004.

[28] Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: challenges and opportunities. Med Image Anal 2016;33:170–5.

[29] Arevalo J, Cruz-Roa A, Arias V, Romero E, Gonzalez FA. An unsupervised feature learning framework for basal cell carcinoma image analysis. Artif Intell Med 2015;64:131–45.

[30] Masood A, Al-Jumaily AA. Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms. Int J Biomed Imag 2013;2013:323268.

[31] 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.

[32] Singh S, Okun A, Jackson A. Artificial intelligence: learning to play Go from scratch. Nature 2017;550:336–7.

[33] Madani A, Ong JR, Tibrewal A, Mofrad MRK. Deep echocardiography: data-efficient supervised and semi-supervised deep learning towards automated diagnosis of cardiac disease. NPJ Digit Med 2018;1:59.

[34] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529:484–9.

[35] Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 2020;368:m689.

[36] McLoone P, McLoone P, Imanbayev K, Norval M. The incidence and body site of skin cancers in the population groups of Astana. Kazakhstan. Health Sci Rep. 2018;1:e51.

[37] Kim GK, Del Rosso JQ, Bellew S. Skin cancer in asians: part 1: nonmelanoma skin cancer. J Clin Aesthet Dermatol 2009;2:39–42.

[38] Korir A, Yu Wang E, Sasieni P, Okerosi N, Ronoh V, Maxwell Parkin D. Cancer risks in Nairobi (2000-2014) by ethnic group. Int J Cancer 2017;140:788–97.

[39] Hogue L, Harvey VM. Basal cell carcinoma, squamous cell carcinoma, and cutaneous melanoma in skin of color patients. Dermatol Clin 2019;37:519–26.

[40] Higgins S, Nazemi A, Feinstein S, Chow M, Wysong A. Clinical presentations of melanoma in African Americans, Hispanics, and Asians. Dermatol Surg 2019;45:791–801.