ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

ANNALS OF ONCOLOGY
driving innovation in oncology

ORIGINAL ARTICLE

Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions

H. A. Haenssle[1*], C. Fink[1], F. Toberer[1], J. Winkler[1], W. Stolz[2], T. Deinlein[3], R. Hofmann-Wellenhof[3], A. Lallas[4], S. Emmert[5], T. Buhl[6], M. Zutt[7], A. Blum[8], M. S. Abassi[9], L. Thomas[10], I. Tromme[11], P. Tschandl[12], A. Enk[1] & A. Rosenberger[13]; Reader Study Level I and Level II Groups

[1]Department of Dermatology, University of Heidelberg, Heidelberg; [2]Department of Dermatology, Allergology and Environmental Medicine II, Munich, Germany; [3]Department of Dermatology and Venerology, Medical University of Graz, Graz, Austria; [4]First Department of Dermatology, Aristotle University, Thessaloniki, Greece; [5]Department of Dermatology, University of Rostock, Rostock; [6]Department of Dermatology, University of Göttingen, Göttingen; [7]Department of Dermatology and Allergology, Klinikum Bremen-Mitte, Bremen; [8]Office Based Clinic of Dermatology, Konstanz; [9]Faculty of Computer Science and Mathematics, University of Passau, Passau, Germany; [10]Department of Dermatology, Lyons Cancer Research Center, Lyon 1 University, Lyon, France; [11]Department of Dermatology, Université Catholique de Louvain, St Luc University Hospital, Brussels, Belgium; [12]Department of Dermatology, Medical University of Vienna, Vienna, Austria; [13]Department of Genetic Epidemiology, University of Goettingen, Goettingen, Germany

Background: Convolutional neural networks (CNNs) efficiently differentiate skin lesions by image analysis. Studies comparing a market-approved CNN in a broad range of diagnoses to dermatologists working under less artificial conditions are lacking.

Materials and methods: One hundred cases of pigmented/non-pigmented skin cancers and benign lesions were used for a two-level reader study in 96 dermatologists (level I: dermoscopy only; level II: clinical close-up images, dermoscopy, and textual information). Additionally, dermoscopic images were classified by a CNN approved for the European market as a medical device (Moleanalyzer Pro, FotoFinder Systems, Bad Birnbach, Germany). Primary endpoints were the sensitivity and specificity of the CNN's dichotomous classification in comparison with the dermatologists' management decisions. Secondary endpoints included the dermatologists' diagnostic decisions, their performance according to their level of experience, and the CNN's area under the curve (AUC) of receiver operating characteristics (ROC).

Results: The CNN revealed a sensitivity, specificity, and ROC AUC with corresponding 95% confidence intervals (CI) of 95.0% (95% CI 83.5% to 98.6%), 76.7% (95% CI 64.6% to 85.6%), and 0.918 (95% CI 0.866—0.970), respectively. In level I, the dermatologists' management decisions showed a mean sensitivity and specificity of 89.0% (95% CI 87.4% to 90.6%) and 80.7% (95% CI 78.8% to 82.6%). With level II information, the sensitivity significantly improved to 94.1% (95% CI 93.1% to 95.1%; $P < 0.001$), while the specificity remained unchanged at 80.4% (95% CI 78.4% to 82.4%; $P = 0.97$). When fixing the CNN's specificity at the mean specificity of the dermatologists' management decision in level II (80.4%), the CNN's sensitivity was almost equal to that of human raters, at 95% (95% CI 83.5% to 98.6%) versus 94.1% (95% CI 93.1% to 95.1%); $P = 0.1$. In contrast, dermatologists were outperformed by the CNN in their level I management decisions and level I and II diagnostic decisions. More experienced dermatologists frequently surpassed the CNN's performance.

Conclusions: Under less artificial conditions and in a broader spectrum of diagnoses, the CNN and most dermatologists performed on the same level. Dermatologists are trained to integrate information from a range of sources rendering comparative studies that are solely based on one single case image inadequate.

Key words: deep learning, neural network, Moleanalyzer Pro, skin cancer, melanoma, dermoscopy

INTRODUCTION

Computer-aided diagnostic (CAD) systems for the detection of skin cancer have been developed and approved for market access.[1] Understandably, CAD systems for automated classification of lesions[2–5] have been trained to attain high sensitivities (>90%); however, mostly at the cost of low specificities.[6]

*Correspondence to: Prof. Holger A. Haenssle, Department of Dermatology, University of Heidelberg, Im Neuenheimer Feld 440, 69120 Heidelberg, Germany. Tel: +49-6221-56-39555; Fax: +49-6221-56-4996

E-mail: Holger.Haenssle@med.uni-heidelberg.de (H. A. Haenssle).

Recently, some limitations of CAD systems relying on the detection of hand-engineered segmentation features have been overcome by applying convolutional neural networks (CNNs). CNNs are commonly trained by ‘supervised deep learning’, an end-to-end approach using raw image data and corresponding diagnostic labels. Within the network, specialized filters autonomously assess input images on a pixel level for good representations of the true diagnosis. Each additional training image improves the CNN’s ability to assemble and weight features associated with the diagnosis.

Most studies investigating CNNs in skin cancer classification tasks have shown performance at or above the level of dermatologists.[7—12] Prospective studies are still lacking and the experimental design of previous studies was criticized for being highly artificial on the side of dermatologists, thus not reflecting results expected in a real-life clinical setting.[13] With the intention to create a head-to-head comparison of CNNs and dermatologists, studies granted access to only one dermoscopic or clinical image per case. Whereas the CNNs were trained to make a classification based on a single image, dermatologists are used to integrating information from various sources (e.g. patient’s risk profile, anamnestic data, lesion evolution). Moreover, many earlier studies focused on a limited spectrum of diagnoses (i.e. nevi versus melanomas).[9,10,12] Although this approach may be useful for an initial proof of concept, it does not adequately reflect the clinical situation where dermatologists encounter a much broader spectrum of lesions. Finally, most previous publications did not specify a commercially or publicly available CNN architecture making it difficult to reproduce reported results.

The study presented herein was designed to (partly) overcome the aforementioned limitations by including a broad spectrum of pigmented and non-pigmented skin lesions to be classified by the current version of a commercially available and market-approved CNN. Moreover, dermatologists were allowed to work in a more familiar setting, that is, in a classical store-and-forward teledermatology setting.[14] This meant using the dermatologists’ management decision based on the combination of clinical close-up images, dermoscopic images, and textual case information as a comparator.

MATERIALS AND METHODS

The ethics committee of the medical faculty of the University of Heidelberg approved this study (approval number S-629/2017), which was conducted in accordance with the Declaration of Helsinki principles. The CNN used in this study is the current market version of Moleanalyzer Pro® (FotoFinder Systems GmbH, Bad Birnbach, Germany), a CNN architecture based on a modified version of Google’s Inception_v4,[15] specifically trained by dermoscopic images and approved as a medical device in the European Union (Conformité Européenne mark). The prototype of the CNN (supplementary Figure S1, available at *Annals of Oncology* online) was originally developed by a cooperative consortium of industry and academia (with the participation of HAH) and tested in a pivotal study.[9] Details on methods about the CNN architecture and training are in the supplementary methods, available at *Annals of Oncology* online.

Data for test cases

We created a dataset of 100 cases including pigmented/non-pigmented and melanocytic/non-melanocytic skin lesions (supplementary Table S1, available at *Annals of Oncology* online). The images originated from different body sites including special localizations (e.g. face/scalp, mucosa, acral skin) and were manually selected by HAH and CF from a convenience sample collected between 2014 and 2019. Each test case included (i) one clinical close-up image, (ii) one dermoscopic image, (iii) textual case information (patient age, sex, and location of the lesion), and (iv) unequivocal histopathological diagnosis for excised lesions (all malignant lesions, 75% of benign lesions) or unremarkable follow-up data over at least 2 years (25% of benign lesions). Various camera/dermoscope combinations were used for image acquisition. No overlap between datasets for training, validation, and testing was allowed.

The CNN’s performance was also tested in two larger and publicly available datasets (available at https://www.isic-archive.com) containing the full spectrum of diagnoses to confirm the generalizability of the CNN results, namely the MSK-1 dataset (1100 images) and the ISIC-2018 challenge[16] dataset (1511 images). The corresponding diagnoses of the ISIC-2018 challenge dataset will not be released by the organizers, which necessitated external statistical analyses by one of the authors (PT).

Reader study level I and II

Dermatologists were personally invited to participate via a web-based rating application. Participants’ data were de-identified and categorized according to self-reported levels of experience with dermoscopy (beginner, <2 years of experience; skilled, 2—5 years of experience; expert, ≥ 5 years of experience).

Each case included two subsequent computer slides, (i) dermoscopic image (level I information) and (ii) dermoscopic image plus clinical close-up image and textual case information (level II information). Dermatologists were asked to indicate their management decision (treatment/excision, no action, follow-up examination) and dichotomous diagnosis (malignant/premalignant, benign) for each slide.

Statistical analysis

The primary outcome measures were the CNN’s sensitivity and specificity in comparison with the dermatologists’ management decisions in study level I and II. Secondary endpoints included the dermatologists’ diagnostic decisions, their performance according to their experience, and the CNN’s area under the curve (AUC) of receiver operating characteristics (ROC). Estimates are provided along with 95% confidence intervals.

Management decisions of ‘no action’ and ‘follow-up examination’ were considered true-negative for benign lesions. Actinic keratosis (AK) shows limited potential to progress to invasive carcinoma and ‘excision/treatment’ and ‘follow-up examination’ were considered as true-positive.

The CNN’s softmax layer gave a malignancy score ranging from 0 to 1 with the a priori cutoff of >0.5 for classifying a lesion as being malignant.

The CNN’s performance was compared with dermatologists’ by using their mean specificity to determine the corresponding cutoff in CNN malignancy scores within a 400-image validation set. This cutoff was applied to the test set and the resulting CNN’s sensitivity was then compared with the average sensitivity of dermatologists by a two-sided one-sample *t*-test. We further applied the non-parametric Kruskal—Wallis test as an omnibus test of heterogeneity between dermatologists with different levels of experience and carried out *post hoc* comparisons of any pair of levels using the Dunn-Nemenyi procedure to adjust for multiple comparisons.[17] Changes in dermatologists’ diagnostic performance after receiving level I or level II information were tested by the Wilcoxon paired signed-rank test with observations related to readers. The results were considered statistically significant at the $P < 0.05$ level due to the observational nature of the investigation. All analyses were carried out using SPSS version 24 (IBM, SPSS, Chicago, IL) or SAS/STAT software, version 9.4 (SAS Institute Inc., Cary, NC).

RESULTS

CNN’s diagnostic performance

In its current market configuration (June 2019), the CNN showed a sensitivity and specificity of 95.0% [95% confidence interval (CI) 83.5% to 98.6%] and 76.7% (95% CI 64.6% to 85.6%), respectively. The ROC AUC was 0.918 (95% CI 0.866—0.970) (Figure 1). Boxplots in Figure 2 show the distribution of malignancy scores in relation to diagnostic categories. With the a priori malignancy cutoff at >0.5, the percentage of correct classifications in malignant lesions was 100% in AK, 100% in Bowen’s disease, 100% in melanoma, 100% in basal cell carcinomas (BCCs), and 60% in squamous cell carcinomas (SCCs). In benign lesions, the percentage of correct classifications was 90% in nevi, 80% in angioma/angiokeratoma, 70% in seborrheic keratoses, 60% in dermatofibroma, and 50% in solar lentigo.

In order to rule out overfitting and to confirm the generalizability of our results, two larger external datasets were used for testing (supplementary Figure S2, available at *Annals of Oncology* online), namely MSK-1 (1100 images) and ISIC-2018 challenge (1511 images). In the MSK-1 datasets the CNN attained an almost identical performance in comparison to our test set (sensitivity 94.2%, specificity 73.8%, ROC AUC 0.939). In the ISIC-2018 challenge dataset the CNN showed a lower sensitivity of 84.7% at a higher specificity of 84.1% and a comparable ROC AUC of 0.926. The pairwise comparison of ROC AUCs attained by the CNN in all three datasets revealed no significant differences (all $P > 0.527$).

Diagnostic performance of dermatologists

Dermatologists ($N = 96$) were categorized into beginners ($n = 17$, <2 years of experience), skilled ($n = 29$, 2—5 years of experience), and experts ($n = 40$, >5 years of experience). Ten participants did not provide information. The mean diagnostic performance of the dermatologists was assessed for their management decisions and dichotomous diagnostic classification of lesions in level I and II (Table 1).

Management decisions. The mean sensitivity and specificity of dermatologists for management decisions during study level I (dermoscopy only) was 89.0% (95% CI 87.4% to 90.6%) and 80.7% (95% CI 78.8% to 82.6%), respectively (Table 1). With additional case information in level II the sensitivity significantly improved to 94.1% (95% CI 93.1% to 95.1%; $P < 0.001$) while the specificity stayed largely unchanged (80.4%, 95% CI 78.4% to 82.4%; $P = 0.97$).

As expected, the performance of dermatologists improved with more experience (supplementary Table S2, available at *Annals of Oncology* online). In level I, the percentage of correct management decisions (accuracy) increased from 79.9% in beginners (95% CI 77.7% to 82.1%) to 83.3% in skilled (95% CI 80.1% to 85.6%) and 86.9% in experts (95% CI 85.5% to 88.3%). Similar observations were made for differences in sensitivity and specificity. The comparison of all three groups was significant for accuracy ($P < 0.001$) and specificity ($P = 0.005$). When applying pairwise comparisons, significant differences in diagnostic performance were only observed for accuracy when comparing experts with beginners ($P = 0.006$). The sample was too small to attain significance for the observed trend regarding sensitivity ($P = 0.108$, all group comparison). The same observations were made for accuracy in level II [beginners: 82.0% (95% CI 79.3% to 84.7%), skilled: 85.4% (95% CI 83.0% to 87.8%), experts: 88.5% (95% CI 87.0% to 90.0%)]. However, in level II, the differences between experts and beginners were also significant for specificity ($P = 0.029$), while significance for sensitivity was still missed ($P = 0.225$).

Dichotomous diagnostic classification. When viewing one dermoscopic image per case (study level I), the 96 dermatologists achieved a mean sensitivity and specificity for the dichotomous diagnostic classification of 83.8% (95% CI 81.8% to 85.8%) and 77.6% (95% CI 75.2% to 80.0%), respectively. With more information in level II, the sensitivity significantly improved to 90.6% (95% CI 89.3% to 92.0%; $P < 0.001$). In contrast with management decisions, the specificity also significantly increased to 82.4% (95% CI 80.5% to 84.3%; $P < 0.001$).

In level I, the percentage of correct dichotomous classifications increased with more experience from 72.6% in beginners (95% CI 67.6% to 77.6%) to 79.3% in skilled (95% CI 76.3% to 82.3%) to 84.2% in experts (95% CI 82.0% to 86.4%) (all $P < 0.01$; supplementary Table S2, available at *Annals of Oncology* online). The same observation was made in level II [beginners: 81.2% (95% CI 78.0% to 84.4%), skilled: 85.1 (95% CI 82.5% to 87.7%), experts: 88.7% (95% CI 87.0% to 90.4%)]. As shown for

https://doi.org/10.1016/j.annonc.2019.10.013

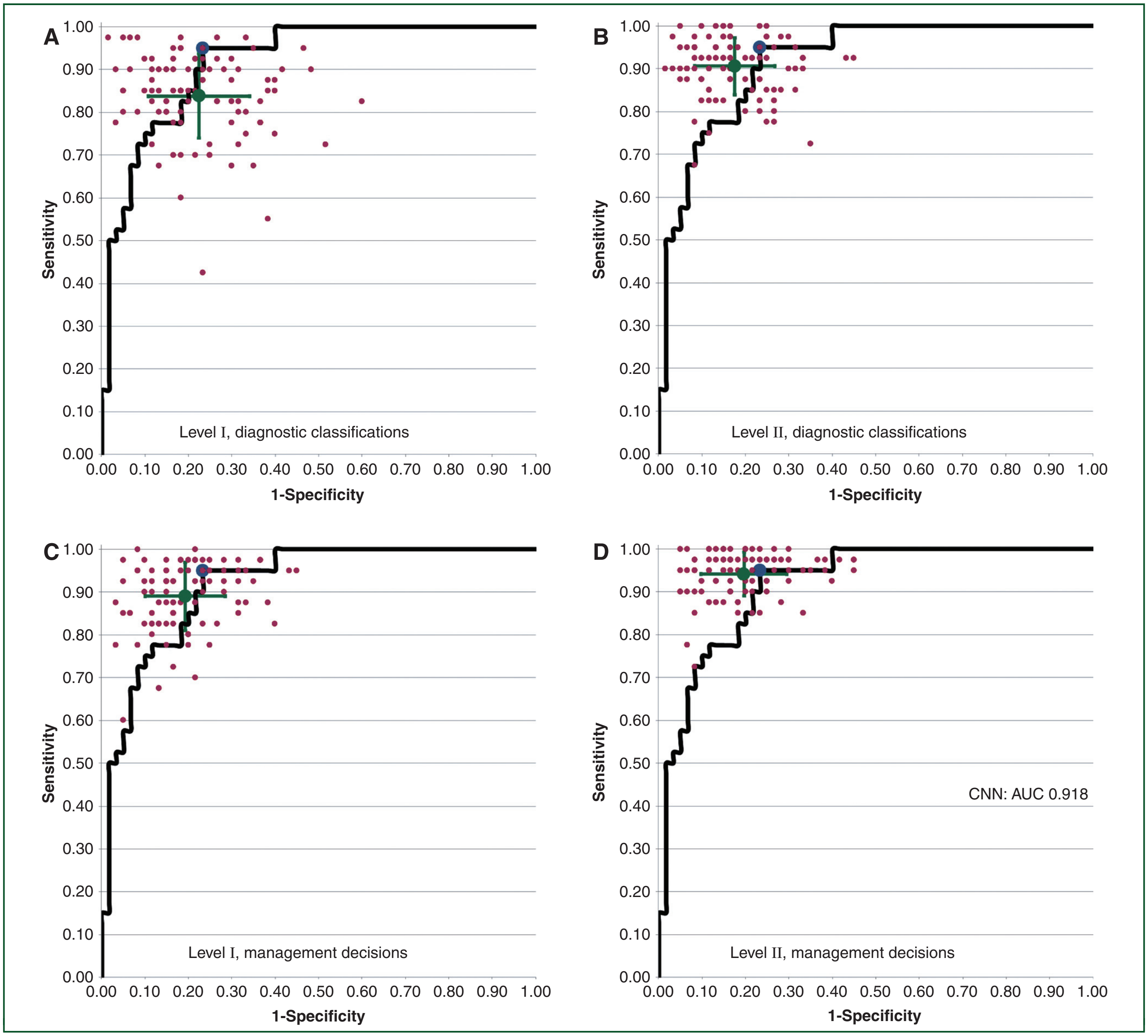

Figure 1. **ROC curve of the CNN (black curve) in relation to the results of all dermatologists ($N = 96$, red dots) in their dichotomous classifications (A: level I, B: level II) and their management decisions (C: level I, D: level II). The average (± SD) sensitivity and specificity of all dermatologists (mean: green circle; ± SD: green error bars) and the CNN's point of operation (blue circle, sensitivity: 95.0%, specificity: 76.7%) is depicted. Dermatologists performed best when allowed to review more case information (B better than A, D better than C) and when deciding on the management of cases (D better than B).**

management decisions, significance was attained in the comparison between groups in terms of accuracy and specificity (all $P < 0.001$) driven by differences between experts and beginners.

Diagnostic performance of CNN versus dermatologists

We used the mean specificity of all dermatologists' management decisions in level II (80.4%) as the benchmark for comparison to the CNN (Figure 1D). To this end, the specificity of 80.4% was used to attain the CNN's corresponding a priori malignancy score cutoff in a 400-image validation set. At this cutoff, the CNN's sensitivity in the test set was 95.0% (95% CI 83.5% to 98.6%), which was similar to the mean sensitivity of dermatologists [94.1% (95% CI 93.1% to 95.1%); $P = 0.1$].

When management decisions were solely based on one dermoscopic image per case (level I), the dermatologists' sensitivity was significantly lower than the CNN's [89.0% (95% CI 87.4% to 90.6%) versus 95.0% (95% CI 83.5% to 98.6%); $P < 0.001$]. Similarly, the CNN showed a superior sensitivity when compared with the dermatologists' dichotomous classifications in level I [83.8% (95% CI 81.8% to 85.8%); $P < 0.001$] or II [90.6% (95% CI 89.3% to 92.0%); $P < 0.001$]. Figure 1 depicts the performance of dermatologists during level I and II in comparison with the CNN.

Additionally, we compared the CNN's accuracy (percentage of correct classifications) with the dermatologists' accuracy (Table 1). When the CNN's accuracy (84.0% (95% CI 75.6% to 89.9%) was compared with the dermatologists' mean accuracy in level II management decisions [85.9%

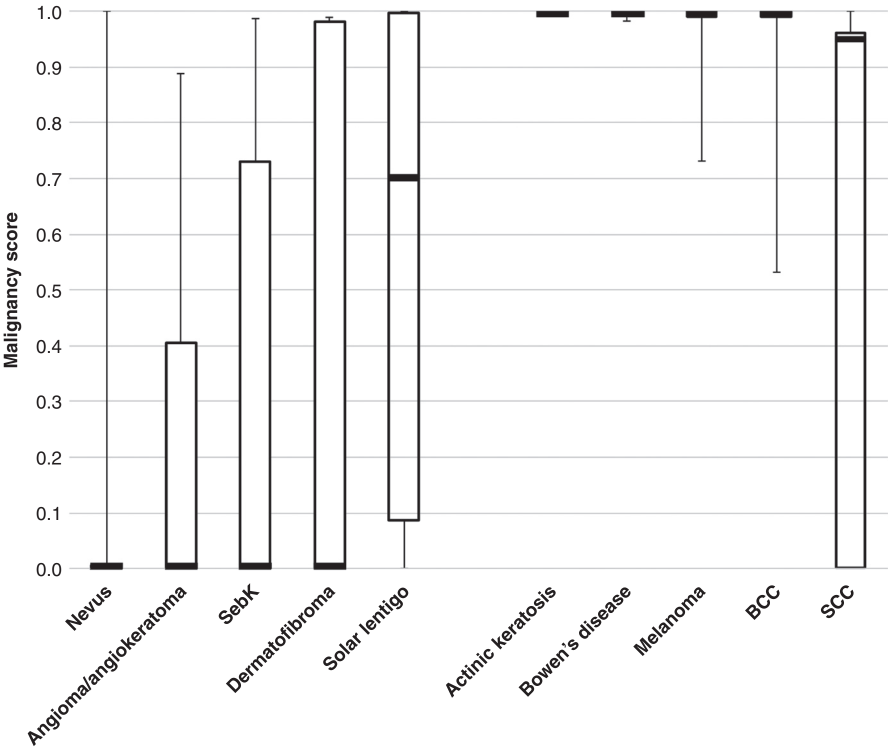

Figure 2. The CNN's melanoma probability scores (range 0—1) for major benign and malignant diagnostic categories are depicted as boxplots. Scores closer to 1 indicated a higher probability of malignancy. The upper and lower bounds of boxes indicate the 25th and 75th percentiles, and the median is indicated by the line intersection of the upper and lower box. Whiskers indicate the full range of probability scores.

BCC, basal cell carcinoma; SCC, squamous cell carcinoma; SebK, seborrheic keratosis.

(95% CI 84.7% to 87.1%)] dermatologists performed slightly but significantly better ($P = 0.003$).

DISCUSSION

The incidence rates of melanoma and non-melanoma skin cancers are rising globally in most fair-skinned populations.[18] Additional efforts in primary and secondary prevention are necessary to contain and possibly reverse these trends. Previous reports on the application of CNN in the diagnosis of skin cancer demonstrated performance on or above the level of dermatologists but were legitimately criticized for highly artificial study settings.[13] Unlike earlier studies,[9,10,16] our test cases included malignant and benign, melanocytic and non-melanocytic, and pigmented and non-pigmented skin lesions. This broad range comprised the vast majority of skin lesions biopsied in the

Table 1. Reader study level I and II results in comparison with CNN

	Management decision			Binary classification		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Raters level I						
All ($N = 96$)	89.0%	80.7%	84.0%	83.8%	77.6%	**80.1%**[a]
Beginner ($n = 17$)	85.7%	76.1%	**79.9%**[a]	80.0%	67.7%	**72.6%**[a]
Skilled ($n = 29$)	89.7%	79.1%	83.3%	84.3%	75.9%	**79.3%**[a]
Expert ($n = 40$)	91.1%	84.1%	**86.9%**[b]	86.2%	82.9%	84.2%
Raters level II						
All ($N = 96$)	94.1%	80.4%	**85.9%**[b]	90.6%	82.4%	**85.7%**[b]
Beginner ($n = 17$)	92.9%	74.7%	82.0%	89.0%	76.0%	81.2%
Skilled ($n = 29$)	94.7%	79.2%	85.4%	90.9%	81.2%	85.1%
Expert ($n = 40$)	94.8%	84.4%	**88.5%**[b]	91.8%	86.6%	**88.7%**[b]
CNN	95.0%	76.7%	84.0%	95.0%	76.7%	84.0%

Level I: readers were provided with dermoscopic images only.

Level II: readers were provided with close-up images and textual case information in addition to dermoscopic images.

Accuracy: calculated as the percentage of correct classifications/decisions ([true-positive + true-negative]/all cases).

Self-reported level of experience was categorized into expert: >5 years of experience; skilled: 2—5 years of experience; and beginner: <2 years of experience. Ten participants did not report their level of experience.

[a] The CNN's accuracy was significantly higher than the mean accuracy of dermatologists.

[b] The mean accuracy of dermatologists was significantly higher than the CNN's accuracy.

routine clinical setting to confirm or rule out malignancy and brings the setting closer to a real-life clinical situation. Moreover, our results demonstrate that the amount of case information and the decisions asked of dermatologists had a major impact on their performance. Dermatologists performed best when allowed to review clinical plus dermoscopic images accompanied by textual metadata. This setting largely reflects the presentation of cases in classical store-and-forward teledermatology[14] and allows dermatologists to integrate different levels of information. Interestingly, these observations also hold for neural networks, as the fusion of two separate CNNs, one assessing clinical close-up images and the other dermoscopic images of the same cases, achieved better results than either modality alone.[11] Moreover, the first reports of deep learning models using clinical non-imaging information have shown promising results in predicting skin cancer[19] and may be included in neural networks. In their daily routine, dermatologists make management decisions rather than clear-cut classifications whether it be more simple dichotomous classifications (benign versus malignant) or naming specific diagnoses. For each probable diagnosis there will be differential diagnoses and dermatologists are not used to ranking these by a probability score. It may be assumed that conflicting options lead to random decisions. Therefore, for future studies we recommend giving physicians enough data for review and to use their management decisions as the primary outcome. Of note, as the dermatologists' experience will strongly affect the results of a comparison to a CNN, we further recommend reporting their performance ranked by experience.

While the CNN in our present study undoubtedly performed very well, the question of who will benefit the most from applying a CNN is still a subject of much debate. On leaving the framework of a clinical study, 'man against machine' becomes 'man with a machine' and physicians will need to incorporate the CNN's classification into their decision-making process. Unfortunately, there are no prospective studies available to show the impact of a CNN assisting dermatologists in their daily clinical work.[13] Nevertheless, our summarized data in Table 1 suggest that less experienced physicians will benefit the most. For management decisions of highly trained experts with all case information at hand, our data indicate a possible decrease in specificity at an unchanged sensitivity should they strictly follow each of the CNN's classifications.

Overfitting is an important limitation not well recognized by many previous CNN-based studies. Overfitting may occur when using an image dataset collected from a few sources that is randomly split into training versus validation and testing. A similarly narrow distribution of lesions for training and testing may result in an overestimation of the CNN's performance and lack of generalizability.[16] In our study setting we can safely rule out overfitting because training images were collected from multiple sources around the world and test cases were derived from sources that did not provide training images. Moreover, we gave evidence for the generalizability of the CNN's results by including two larger external datasets (one with full blinding of authors to true diagnoses) in which the tested CNN showed comparable ROC AUCs.

There are several limitations to this study. First, current deep learning algorithms lack interpretability;[20] therefore, we are unable to name specific causes for false classifications. Unfortunately, this lack of interpretability prevents making changes for improvement, which leads to repeating the same mistakes in the future. Secondly, to safeguard the feasibility of the reader study our test set included 100 lesions, thus leaving only a few cases for certain diagnoses. The CNN results attained in these diagnoses should, therefore, be interpreted with caution. Thirdly, the compilation of our test set did not include some other benign (e.g. viral warts), malignant (e.g. Merkel cell carcinoma), or inflammatory skin lesions (e.g. clear cell acanthoma). Therefore, our results should not be generalized to a large prospective patient population. Finally, the CNN was mostly trained with dermoscopic images of patients with a Caucasian genetic background and may not provide comparable results in a population of non-white skin types.

In conclusion, the results of our study demonstrate that the tested CNN is capable of classifying a broad spectrum of skin tumors at a high level of sensitivity and specificity. Under less artificial conditions with clinical close-up images, dermoscopic images, and textual case information for review, the management decisions of most dermatologists were either on or slightly above the level of the CNN. Experts in dermoscopy with access to relevant case information commonly outperformed the CNN's specificity at a comparable sensitivity. There is a need for prospective studies and an improved interpretability of the CNN's classification results to take this research to the next level.

ACKNOWLEDGEMENTS

We would like to thank all dermatologists who actively and voluntarily spent a lot of time participating in the reader study level I and II.

The following dermatologists, in alphabetical order, agreed to be acknowledged for their participation in the reader study level I and II: Christina Alt, Marie Bachelerie, Sonali Bajaj, Alise Balcere, Sophie Baricault, Clément Barthaux, Yvonne Beckenbauer, Ines Bertlich, Andreas Blum, Marie-France Bouthenet, Sophie Brassat, Philipp Marcel Buck, Kristina Buder-Bakhaya, Maria-Letizia Cappelletti, Cécile Chabbert, Julie De Labarthe, Eveline DeCoster, Teresa Deinlein, Michèle Dobler, Daphnée Dumon, Steffen Emmert, Julie Gachon-Buffet, Mikhail Gusarov, Franziska Hartmann, Julia Hartmann, Anke Herrmann, Isabelle Hoorens, Eva Hulstaert, Raimonds Karls, Andreea Kolonte, Christian Kromer, Aimilios Lallas, Céline Le Blanc Vasseux, Annabelle Levy-Roy, Pawel Majenka, Marine Marc, Veronique Martin Bourret, Nadège Michelet-Brunacci, Christina Mitteldorf, Jean Paroissien, Camille Picard, Diana Plise, Valérie Reymann, Fabrice Ribeaudeau, Pauline Richez, Hélène Roche Plaine, Deborah Salik, Elke Sattler, Sarah Schäfer, Roland Schneiderbauer, Thierry Secchi, Karen Talour, Lukas Trennheuser, Alexander Wald, Priscila Wölbing, and Pascale Zukervar.

All other participants asked to remain anonymous and we also thank these colleagues for their commitment.

FUNDING

This research project received funding from a public, non-for-profit domain, namely the Skin Cancer Council of Germany (Nationale Versorgungskonferenz Hautkrebs (NVKH) e.V.), www.nvkh.de/. Grant/Award Number: AD-LEARN DERMOSCOPY (HF01-Z02-P05).

DISCLOSURE

AB received honoraria and/or travel expenses from FotoFinder Systems GmbH. CF received travel expenses from Magnosco GmbH. HAH received honoraria and/or travel expenses from companies involved in the development of devices for skin cancer screening: Scibase AB, FotoFinder Systems GmbH, Heine Optotechnik GmbH, Magnosco GmbH. PT received honoraria from Silverchair and an unrestricted research grant from MetaOptima Technology Inc. All other authors declared no conflict of interest.

REFERENCES

1. Fink C, Haenssle HA. Non-invasive tools for the diagnosis of cutaneous melanoma. *Skin Res Technol*. 2017;23:261—271.
2. Forschner A, Keim U, Hofmann M, et al. Diagnostic accuracy of dermatofluoroscopy in cutaneous melanoma detection: results of a prospective multicentre clinical study in 476 pigmented lesions. *Br J Dermatol*. 2018;179:478—485.
3. Lui H, Zhao J, McLean D, et al. Real-time Raman spectroscopy for in vivo skin cancer diagnosis. *Cancer Res*. 2012;72:2491—2500.
4. Malvehy J, Hauschild A, Curiel-Lewandrowski C, et al. Clinical performance of the Nevisense system in cutaneous melanoma detection: an international, multicentre, prospective and blinded clinical trial on efficacy and safety. *Br J Dermatol*. 2014;171:1099—1107.
5. Monheit G, Cognetta AB, Ferris L, et al. The performance of MelaFind: a prospective multicenter study. *Arch Dermatol*. 2011;147:188—194.
6. Cukras AR. On the comparison of diagnosis and management of melanoma between dermatologists and MelaFind. *JAMA Dermatol*. 2013;149:622—623.
7. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115—118.
8. Fujisawa Y, Otomo Y, Ogata 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—3781.
9. Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol*. 2018;29:1836—1842.
10. Marchetti MA, Codella NCF, Dusza SW, et al. Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. *J Am Acad Dermatol*. 2018;78:270—277.
11. Tschandl P, Rosendahl C, Akay BN, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. *JAMA Dermatol*. 2019;155:58—65.
12. Yu C, Yang S, Kim W, et al. Acral melanoma detection using a convolutional neural network for dermoscopy images. *PLoS One*. 2018;13:e0193321.
13. Lallas A, Argenziano G. Artificial intelligence and melanoma diagnosis: ignoring human nature may lead to false predictions. *Dermatol Pract Concept*. 2018;8:249—251.
14. Finnane A, Dallest K, Janda M, et al. Teledermatology for the diagnosis and management of skin cancer: a systematic review. *JAMA Dermatol*. 2017;153:319—327.
15. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. Available at https://arxiv.org/abs/1512.00567. Accessed November 19, 2019.
16. Tschandl P, Codella N, Akay BN, 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—947.
17. Elliott AC, Hynan LS. A SAS$^{®}$ macro implementation of a multiple comparison post hoc test for a Kruskal-Wallis analysis. *Comput Methods Programs Biomed*. 2011;102:75—80.
18. Apalla Z, Nashan D, Weller RB, et al. Skin cancer: epidemiology, disease burden, pathophysiology, diagnosis, and therapeutic approaches. *Dermatol Ther (Heidelb)*. 2017;7:5—19.
19. Wang HH, Wang YH, Liang CW, et al. Assessment of deep learning using nonimaging information and sequential medical records to develop a prediction model for nonmelanoma skin cancer. *JAMA Dermatol*. 2019;155:1277—1283.
20. Pereira S, Meier R, McKinley R, et al. Enhancing interpretability of automatically extracted machine learning features: application to a RBM-random forest system on brain lesion segmentation. *Med Image Anal*. 2018;44:228—244.