



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

Journal:	<i>Annals of Oncology</i>
Manuscript ID	ANNONC-2019-1291.R1
Article Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Haenssle, Holger; UniversitätsKlinikum Heidelberg, Department of Dermatology Fink, Christine; UniversitätsKlinikum Heidelberg, Department of Dermatology Toberer, Ferdinand; UniversitätsKlinikum Heidelberg, Department of Dermatology Winkler, Julia; UniversitätsKlinikum Heidelberg, Department of Dermatology Stolz, Wilhelm; Munich Municipal Hospital Thalkirchner Strasse Branch, Department of Dermatology, Allergology and Environmental Medicine II Deinlein, Teresa; Medical University of Graz, Department of Dermatology and Venerology Hofmann-Wellenhof, Rainer; Medical University of Graz, Department of Dermatology and Venerology Lallas, Aimilios; Aristotle University of Thessaloniki, First Department of Dermatology Emmert, Steffen; University of Rostock, Department of Dermatology Buhl, Timo; Georg-August-Universität Göttingen Universitätsmedizin, Department of Dermatology Zutt, Markus; Klinikum Bremen-Mitte gGmbH, Department of Dermatology and Allergology Blum, Andreas; Hautarzt- und Lehrpraxis Konstanz, Department of Dermatology Abassi, Mohamed Souhayel; University of Passau, Faculty of Computer Science and Mathematics Thomas, Luc; Lyon Sud Hospital Center, Dermatology Tromme, Isabelle; University College Sint-Lukas Brussels, Department of Dermatology, Université Catholique de Louvain

	Tschandl, Philipp; Medical University of Vienna, Department of Dermatology Enk, Alexander; UniversitätsKlinikum Heidelberg, Department of Dermatology Rosenberger, Albert; University of Gottingen, Department of Genetic Epidemiology
Keywords:	deep learning, neural network, Moleanalyzer-pro, skin cancer, melanoma, dermoscopy
Abstract:	Background: Convolutional neural networks (CNN) 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. 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, 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 to 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 of 95.0% [83.5%-98.6%], 76.7% [64.6%-85.6%], and 0.918 [0.866-0.970], respectively. In level-I the dermatologists' management decisions showed a mean sensitivity and specificity of 89.0% [87.4%-90.6%] and 80.7% [78.8%-82.6%]. With level-II information the sensitivity significantly improved to 94.1% ([93.1%-95.1%]; $p<0.001$) while the specificity stayed unchanged (80.4% [78.4%-82.4%]; $p=0.9$). 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 human raters (95% [83.5%-98.6%] versus 94.1% [93.1%-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 solely based on one single case image inadequate.

Original Article

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

H. A. Haenssle¹, C. Fink¹, F. Toberer¹, J. Winkler¹, W. Stolz², T. Deinlein³, R. Hofmann-Wellenhof³, A. Lallas⁴, S. Emmert⁵, T. Buhl⁶, M. Zutt⁷, A. Blum⁸, M.S. Abassi⁹, L. Thomas¹⁰, I. Tromme¹¹, P. Tschandl¹², A. Enk¹, A. Rosenberger¹³; Reader study level-I and level-II Groups

Affiliations:

¹ Department of Dermatology, University of Heidelberg, Heidelberg, Germany

² Department of Dermatology, Allergology and Environmental Medicine II, Hospital Thalkirchner Street, 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, Germany

⁶ Department of Dermatology, University of Göttingen, Göttingen, Germany

⁷ Department of Dermatology and Allergology, Klinikum Bremen-Mitte, Bremen, Germany

⁸ Office based clinic of dermatology, Konstanz, Germany

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

Manuscript word count: 2737, **Abstract:** 296, **Key message:** 438 characters, incl. spaces

Tables: 1, **Figures:** 2, **References:** 20; **Suppl. Material:** 2 Tables, 2 Figures, suppl. methods

Running title: Convolutional neural network for skin cancer classification

Ethics: Reviewed and approved by the ethic committee of the medical faculty of the University of Heidelberg (Approval number S-629/2017)

Corresponding Author:

Prof. Dr. med. Holger A. Haenssle, Department of Dermatology, University of Heidelberg, Im Neuenheimer Feld 440, 69120 Heidelberg, Germany, Holger.Haenssle@med.uni-heidelberg.de, Fon: +49-6221-56-39555, Fax: +49-6221-56-4996

1

2

3 ABSTRACT

4

5

6 **Background:** Convolutional neural networks (CNN) efficiently differentiate skin lesions by

7 image analysis. Studies comparing a market-approved CNN in a broad range of diagnoses to

8 dermatologists working under less artificial conditions are lacking.

9

10

11

12 **Methods:** One hundred cases of pigmented/non-pigmented skin cancers and benign lesions

13 were used for a two-level reader study in 96 dermatologists (level-I: dermoscopy only; level-

14 II: clinical close-up images, dermoscopy, textual information). Additionally, dermoscopic

15 images were classified by a CNN approved for the European market as a medical device

16 (Moleanalyzer-Pro, FotoFinder Systems, Bad Birnbach, Germany). Primary endpoints were

17 the sensitivity and specificity of the CNN’s dichotomous classification in comparison to the

18 dermatologists’ management decisions. Secondary endpoints included the dermatologists’

19 diagnostic decisions, their performance according to their level of experience, and the CNN’s

20 area under the curve (AUC) of receiver operating characteristics (ROC).

21

22 **Results:** The CNN revealed a sensitivity, specificity, and ROC AUC with corresponding 95%

23 confidence intervals of 95.0% [83.5%-98.6%], 76.7% [64.6%-85.6%], and 0.918 [0.866-

24 0.970], respectively. In level-I the dermatologists’ management decisions showed a mean

25 sensitivity and specificity of 89.0% [87.4%-90.6%] and 80.7% [78.8%-82.6%]. With level-II

26 information the sensitivity significantly improved to 94.1% ([93.1%-95.1%]; $p<0.001$) while

27 the specificity stayed unchanged (80.4% [78.4%-82.4%]; $p=0.9$). When fixing the CNN’s

28 specificity at the mean specificity of the dermatologists’ management decision in level-II

29 (80.4%) the CNN’s sensitivity was almost equal to human raters (95% [83.5%-98.6%] versus

30 94.1% [93.1%-95.1%]; $p=0.1$). In contrast, dermatologists were outperformed by the CNN in

31 their level-I management decisions and level-I and -II diagnostic decisions. More experienced

32 dermatologists frequently surpassed the CNN’s performance.

33

34 **Conclusions:** Under less artificial conditions and in a broader spectrum of diagnoses the CNN

35 and most dermatologists performed on the same level. Dermatologists are trained to integrate

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

information from a range of sources rendering comparative studies solely based on one single case image inadequate.

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

For Peer Review

Key message:

The performance of a market-approved convolutional neural network (CNN), which was trained on dermoscopic images, was tested in comparison to dermatologists in a broad range of skin cancers and benign lesions. The average management decisions of 96 dermatologists, who reviewed dermoscopic and clinical images accompanied by textual case information, were on the same level as the CNN’s classifications based on dermoscopic images alone.

Abbreviations:

95% CI	95% confidence interval
AUC	area under the curve
AK	Actinic keratosis
BCC	Basal cell carcinoma
CAD	Computer aided diagnostic
CNN	convolutional neural network
ROC	receiver operating characteristic
SCC	Squamous cell carcinoma

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 lesions²⁻⁵ have been trimmed to attain high sensitivities (>90%), however, mostly at the cost of low specificities.⁶

Recently, some limitations of CAD systems relying on the detection of hand-engineered segmentation features have been overcome by applying convolutional neural networks (CNN). CNN 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 CNN in skin cancer classification tasks have shown a performance on or above the level of dermatologists.⁷⁻¹² However, 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 to be expected in a real-life clinical setting.¹³ With the intention to create a “head-to-head” comparison of CNN and dermatologists, studies granted access to only one dermoscopic or clinical image per case. While CNN were trained to make a classification based on a single image, dermatologists are used to integrate information from various sources (e.g. patient’s risk profile, anamnestic data, lesion’s evolution). Moreover, many earlier studies focused on a limited spectrum of diagnoses (i.e. nevi versus melanomas).^{9;10;12} While 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 imply a commercially or publicly available CNN architecture making it difficult to reproduce reported results.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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, i.e. in a classical store-and-forward teledermatology setting.¹⁴ This meant using dermatologists’ management decision based on the combination of clinical close-up images, dermoscopic images, and textual case information as a comparator.

For Peer Review

METHODS

The study was approved by local ethics committees and performed in accordance with the Declaration of Helsinki principles. The CNN used in this study was 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**) was originally developed by a cooperative consortium of industry and academia (under participation of HAH) and tested in a pivotal study.⁹ Details on methods pertaining to the CNN architecture and training are found in 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**). 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 additionally 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's results, namely the MSK-1 dataset (1100 images) and the ISIC-2018 challenge¹⁶ dataset (1511 images). Ground truth of ISIC-2018 challenge

dataset will not be release by the organizers, which necessitated external statistical analyses (Tschandl P., Department of Dermatology, University of Vienna, Austria).

Reader study level-I and -II

Dermatologists were personally invited to participated via a web-based rating application.

Participants' data was de-identified and categorized according to self-reported level 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), ii) dermoscopic image plus clinical close-up image and textual case information (level-II information). For each slide dermatologists were asked to indicate their management decision (treatment/excision, no action, follow-up examination) and dichotomous diagnosis (malignant/premalignant, benign).

Statistical analysis

Primary outcome measures were the CNN's sensitivity and specificity in comparison to 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 '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 put out 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 to 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 to the average sensitivity of dermatologists by a 2-sided one sample *t*-test. We further applied the non-parametric Kruskal-Wallis test as omnibus test of heterogeneity between dermatologists with different levels of experience, and performed post-hoc comparisons of any pair of levels using 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 Wilcoxon paired signed-rank test with observations related to readers. 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, Illinois, USA) or SAS/STAT software, Version 9.4 (SAS Institute Inc., Cary, NC, USA.).

RESULTS

CNN's diagnostic performance

In its current market configuration (June 2019), the CNN showed a sensitivity and specificity of 95.0% (95% CI [83.5%-98.6%]) and 76.7% (95% CI [64.6%-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 of >0.5 the percentage of correct classifications in malignant lesions was 100% in AK, 100% in Bowen's disease, 100% in melanoma, 100% in BCCs, and 60% in 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**), 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 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 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%-90.6%]) and 80.7% (95% CI [78.8%-82.6%]), respectively (**Table 1**). With additional case information in level-II the sensitivity significantly improved to 94.1% (95% CI [93.1%-95.1%]; $p<0.001$) while the specificity stayed largely unchanged (80.4%, 95% CI [78.4%-82.4%]; $p=0.97$).

Expectedly, the performance of dermatologists improved with more experience (**supplementary Table S2**). In level-I the percentage of correct management decisions (accuracy) increased from 79.9% in beginners (95% CI [77.7%-82.1%]) to 83.3% in skilled (95% CI [80.1%-85.6%]) and 86.9% in experts (95% CI [85.5%-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 to 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%-84.7%]), skilled: 85.4% (95% CI [83.0%-87.8%]), experts: 88.5% (95% CI [87.0%-90.0%])). However, in level-II the differences between experts and beginners also were significant for specificity ($p=0.029$), while significance for sensitivity was still missed ($p=0.225$).

Dichotomous diagnostic classification

When viewing solely 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%-85.8%]) and 77.6% (95% CI [75.2%-80.0%]), respectively. With more information in level-II the sensitivity significantly improved to 90.6% (95% CI [89.3%-

92.0%]; $p<0.001$). In contrast to management decisions, the specificity also significantly increased to 82.4% (95% CI [80.5%-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%-77.6%]) to 79.3% in skilled (95% CI [76.3%-82.3%]) to 84.2% in experts (95% CI [82.0%-86.4%]) (all $p<0.01$; **supplementary Table S2**). The same observation was made in level-II (beginners: 81.2% (95% CI [78.0%-84.4%]), skilled: 85.1 (95% CI [82.5%-87.7%]), experts: 88.7% (95% CI [87.0%-90.4%])). As shown for 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%-98.6%]), which was similar to the mean sensitivity of dermatologists (94.1% (95% CI [93.1%-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%-90.6%]) *versus* 95.0% (95% CI [83.5%-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%-85.8%]); $p<0.001$) or -II (90.6% (95% CI [89.3%-92.0%]); $p<0.001$). **Figure 1** depicts the performance of dermatologists during level-I and -II in comparison to the CNN.

1
2
3 Additionally, we compared the CNN's accuracy (percentage of correct classifications) to
4 dermatologists (**Table 1**). When the CNN's accuracy (84.0% (95% CI [75.6%-89.9%]) was
5
6 compared to the dermatologists' mean accuracy in level-II management decisions (85.9%
7
8 (95% CI [84.7%-87.1%]) dermatologists performed slightly but significantly better ($p=0.003$).
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Peer Review

DISCUSSION

The incidence rates of melanoma and non-melanoma skin cancers are globally rising 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 a 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 clinical routine to confirm or rule out malignancy and brings the setting closer to a real-life clinical situation. Moreover, our results clearly demonstrate that the amount of case information and the sort of 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 teledermatology¹⁴ and entails dermatologists to integrate different levels of information. Interestingly, this observations also holds true 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, first reports of deep learning models using clinical non-imaging information have shown promising results in predicting skin cancer¹⁹ and may be included in neural networks. In daily routine dermatologists make management decisions rather than clear cut classifications whether it be more simple dichotomous classifications (benign vs. 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 offering physicians enough data for review and to use their management decisions as primary outcome. Of note, as the

dermatologists' experience will strongly impact 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 questions of 'who will benefit the most from applying a CNN' is still a subject of much debate. By 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.¹³ However, 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 the setting of our study we can safely rule out overfitting, because training images were collected from multiple sources around the world and test cases were derived from a source 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, lack of interpretability prevents taking interventions for improvement, which leads to repeating the same mistakes in the future. Second, to safeguard the feasibility of the reader study our test-set included 100 lesions, thus leaving only a few cases for certain

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

diagnoses. The CNN’s results attained in these diagnoses should, therefore, be interpreted with caution. Third, 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 attain comparable results in a population with divergent 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 availability of 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 a CNN’s classification results to take research to a next level.

Acknowledgements:

We would like to thank all dermatologists, who actively and voluntarily spend much time to participate in the reader study level-I and -II

Funding: This research project received funding from a public, non-for-profit domain, namely the Skin Cancer Council Germany (Nationale Versorgungskonferenz Hautkrebs (NVKH) e.V.), www.nvkh.de/.

Grant/Award Number: AD-LEARN DERMOSCOPY (HF01-Z02-P05).

Disclosures: A Blum received honoraria and/or travel expenses from FotoFinder Systems GmbH. C Fink received travel expenses from Magnosco GmbH. HA Haenssle 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. P Tschandl has received honoraria from Silverchair, and an unrestricted research grant from MetaOptima Technology Inc. All other authors declared no conflict of interest.

Reference List

1. Fink C, Haenssle HA. Non-invasive tools for the diagnosis of cutaneous melanoma. Skin Res Technol 2017;23:261-71.
2. Forschner A, Keim U, Hofmann M, Spankuch I, Lomberg D, Weide B 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;179:478-85.
3. Lui H, Zhao J, McLean D, Zeng H. Real-time Raman spectroscopy for in vivo skin cancer diagnosis. Cancer Res 2012;72:2491-500.

4. Malvehy J, Hauschild A, Curiel-Lewandrowski C, Mohr P, Hofmann-Wellenhof R, Motley R 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;171:1099-107.

5. Monheit G, Cognetta AB, Ferris L, Rabinovitz H, Gross K, Martini M et al. The performance of MelaFind: a prospective multicenter study. *Arch Dermatol* 2011;147:188-94.

6. Cukras AR. On the comparison of diagnosis and management of melanoma between dermatologists and MelaFind. *JAMA Dermatol* 2013;149:622-3.

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.

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. 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.

10. Marchetti MA, Codella NCF, Dusza SW, Gutman DA, Helba B, Kalloo A et al. Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. *J Am Acad Dermatol* 2018;78:270-7.

11. 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* 2019;155:58-65.
12. 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.
13. Lallas A, Argenziano G. Artificial intelligence and melanoma diagnosis: ignoring human nature may lead to false predictions. *Dermatol Pract Concept* 2018;8:249-51.
14. Finnane A, Dallest K, Janda M, Soyer HP. Teledermatology for the Diagnosis and Management of Skin Cancer: A Systematic Review. *JAMA Dermatol* 2017;153:319-27.
15. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Reprint at <https://arxiv.org/abs/1512.00567> 2015.
16. 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.
17. Elliott AC, Hynan LS. A SAS((R)) 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, Castellsague X. Skin Cancer: Epidemiology, Disease Burden, Pathophysiology, Diagnosis, and Therapeutic Approaches. *Dermatol Ther (Heidelb)* 2017;7:5-19.

19. Wang HH, Wang YH, Liang CW, Li YC. Assessment of Deep Learning Using Nonimaging Information and Sequential Medical Records to Develop a Prediction Model for Nonmelanoma Skin Cancer. JAMA Dermatol 2019.

20. Pereira S, Meier R, McKinley R, Wiest R, Alves V, Silva CA 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;44:228-44.

APPENDIX:

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-Bufferet, 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 Ribeau, Pauline Richez, Hélène Roche Plaine, Deborah Salik, Elke Sattler, Sarah Schäfer, Roland Schneiderbauer, Thierry Secchi, Karen Talour, Lukas Trennheuser, Alexander Wald, Priscila Wölbing, Pascale Zukervar.

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

FIGURE LEGENDS

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 in their management decisions (C: level-I, D: level-II). The average ($\pm$ SD) sensitivity and specificity of all dermatologists (mean: green circle; $\pm$ 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).

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 while the median is indicated by the line intersection the upper and lower box. Whiskers indicate the full range of probability scores.

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

Original Article

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

H. A. Haenssle¹, C. Fink¹, F. Toberer¹, J. Winkler¹, W. Stolz², T. Deinlein³, R. Hofmann-Wellenhof³, A. Lallas⁴, S. Emmert⁵, T. Buhl⁶, M. Zutt⁷, A. Blum⁸, M.S. Abassi⁹, L. Thomas¹⁰, I. Tromme¹¹, P. Tschandl¹², A. Enk¹, A. Rosenberger¹³; Reader study level-I and level-II Groups

Affiliations:

- ¹ Department of Dermatology, University of Heidelberg, Heidelberg, Germany
- ² Department of Dermatology, Allergology and Environmental Medicine II, Hospital Thalkirchner Street, 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, Germany
- ⁶ Department of Dermatology, University of Göttingen, Göttingen, Germany
- ⁷ Department of Dermatology and Allergology, Klinikum Bremen-Mitte, Bremen, Germany
- ⁸ Office based clinic of dermatology, Konstanz, Germany
- 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

Manuscript word count: 2737, **Abstract:** 296, **Key message:** 438 characters, incl. spaces
Tables: 1, **Figures:** 2, **References:** 20; **Suppl. Material:** 2 Tables, 2 Figures, suppl. methods

Running title: Convolutional neural network for skin cancer classification

Ethics: Reviewed and approved by the ethic committee of the medical faculty of the University of Heidelberg (Approval number S-629/2017)

Corresponding Author:

Prof. Dr. med. Holger A. Haenssle, Department of Dermatology, University of Heidelberg, Im Neuenheimer Feld 440, 69120 Heidelberg, Germany, Holger.Haenssle@med.uni-heidelberg.de, Fon: +49-6221-56-39555, Fax: +49-6221-56-4996

ABSTRACT

Background: Convolutional neural networks (CNN) 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.

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, 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 to 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 of 95.0% [83.5%-98.6%], 76.7% [64.6%-85.6%], and 0.918 [0.866-0.970], respectively. In level-I the dermatologists' management decisions showed a mean sensitivity and specificity of 89.0% [87.4%-90.6%] and 80.7% [78.8%-82.6%]. With level-II information the sensitivity significantly improved to 94.1% ([93.1%-95.1%]; $p<0.001$) while the specificity stayed unchanged (80.4% [78.4%-82.4%]; $p=0.9$). 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 human raters (95% [83.5%-98.6%] versus 94.1% [93.1%-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

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

information from a range of sources rendering comparative studies solely based on one single case image inadequate.

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

For Peer Review

Key message:

The performance of a market-approved convolutional neural network (CNN), [which was trained on dermoscopic images](#), was tested in comparison to dermatologists in a broad range of skin cancers and benign lesions. The average management decisions of 96 dermatologists, who reviewed dermoscopic and clinical images accompanied by textual case information, were on the same level as the CNN's classifications based on dermoscopic images alone.

~~**Key words:** Melanoma, melanocytic nevi, squamous cell carcinoma, basal cell carcinoma, early detection, Dermoscopy, atypical mole syndrome, deep learning convolutional neural network, computer algorithm, automated skin cancer detection, classification~~

Abbreviations:

95% CI	95% confidence interval
AUC	area under the curve
AK	Actinic keratosis
BCC	Basal cell carcinoma
CAD	Computer aided diagnostic
CNN	convolutional neural network
ROC	receiver operating characteristic
SCC	Squamous cell carcinoma

INTRODUCTION

Computer aided diagnostic (CAD) systems for the detection of skin cancer have been developed and approved for market access ~~by the U.S. food and drug administration and/or national competent authorities under European Union legislation.~~¹ Understandably, CAD systems for automated classification of lesions²⁻⁵ have been trimmed to attain high sensitivities (>90%), however, mostly at the cost of low specificities.⁶ ~~Low specificities raised concerns about some CAD systems⁶ and a high rate of unnecessary biopsies was found when using a multispectral imaging system under real-life clinical conditions.~~⁷

Recently, some limitations of CAD systems relying on the detection of hand-engineered segmentation features have been overcome by applying convolutional neural networks (CNN). CNN are commonly trained by “supervised deep learning”, an end-to-end approach using raw image data and corresponding diagnostic labels ~~that are fed into the network.~~

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 CNN in skin cancer classification tasks have shown a performance on or above the level of dermatologists.⁷⁻¹² However, 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 to be expected in a real-life clinical setting.¹³ With the intention to create a “head-to-head” comparison of CNN and dermatologists, studies granted access to only one dermoscopic or clinical image per case.

While CNN were trained to make a classification based on a single image, dermatologists are used to integrate information from various sources (e.g. patient’s risk profile, anamnestic data, lesion’s evolution~~morphology and evolution of the lesion~~). Moreover, many earlier studies focused on a limited spectrum of diagnoses (i.e. nevi versus melanomas).^{9;10;12} While 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 ~~pigmented and non-pigmented skin~~ lesions. Finally, most previous publications did not imply 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, i.e. in a classical store-and-forward teledermatology setting.¹⁴ This meant using dermatologists' management decision based on the combination of clinical close-up images, dermoscopic images, and textual case information as a comparator.

METHODS

The study was approved by local ethics committees and performed in accordance with the Declaration of Helsinki principles. The ~~neural network~~CNN used in this study was the current market version of Moleanalyzer-Pro® (FotoFinder Systems GmbH, Bad Birnbach, Germany), a ~~specifically trained~~ CNN architecture based on a modified version of Google’s Inception_v4¹⁵, 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) was originally developed by a cooperative consortium of industry and academia (under participation of HAH) and tested in a pivotal study.⁹ Details on methods pertaining to the CNN architecture and training are found in 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). Images originated from different body sites including special localizations (e.g. face/scalp, mucosa, acral skin) and ~~Lesions of the nail apparatus were excluded as well as lesions non-compliant with skin imaging standards, e.g. blurred or out of focus images. As 15 out of 60 benign lesions (25%) were non-exeised but validated by follow-up examinations, this dataset represented a spectrum that may typically be encountered in daily clinical routine.~~ All cases were manually selected by HAH and CF from a convenience sample collected between 2014 and 2019, ~~at the university department of dermatology in Heidelberg for educational reasons. CNN training did not comprise any images from the Heidelberg department of dermatology, thus ruling out overfitting, a common problem of overestimating a CNN’s performance when using split samples of the same source for training and testing.~~

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.

~~Additionally~~The CNN's performance was additionally tested in ~~we used~~ two larger and publicly available datasets ([available at https://www.isic-archive.com](https://www.isic-archive.com)) containing the full spectrum of ~~diagnosis diagnoses~~ to confirm the generalizability of the CNN's results, namely the ~~Memorial Sloan Kettering MSK-1~~ dataset (~~MSK-1~~, 1100 images, ~~supplementary Table S1~~) and the ISIC-2018 challenge¹⁶ dataset (1511 images). ~~Authors were kept blinded to ground~~ ~~Ground~~ truth of ISIC-2018 challenge dataset ~~will not be release by the organizers to keep~~ ~~researchers blinded~~, ~~which~~ necessitated ~~eding~~ external statistical analyses (Tschandl P., Department of Dermatology, University of Vienna, Austria).

Reader study level-I and -II

Dermatologists were personally invited ~~to and~~ participated ~~in the study~~ via a web-based rating application. Participants' data was de-identified and categorized according to self-reported level of experience with dermoscopy ('beginner' <2 years of experience, 'skilled' 2-5 years of experience, 'expert' ≥5 years of experience).

Each case ~~included was presented to dermatologists by three two~~ subsequent computer slides, ~~i) The first slide showed the~~ dermoscopic image (level-I information), ~~ii) the second slide showed the~~ dermoscopic image ~~plus a~~ clinical close-up image; and textual case information (level-II information). For each slide dermatologists were asked to indicate their management decision (treatment/excision, no action, follow-up examination) and dichotomous diagnosis (malignant/premalignant, benign). ~~The third slide depicted the correct answers and closed the case.~~

Statistical analysis

Primary outcome measures were the CNN's sensitivity and specificity ~~of the CNN's~~ ~~dichotomous classification (malignant versus benign)~~ in comparison to 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 'no action' and 'follow-up examination' were considered true-negative for benign lesions. Actinic keratosis (AK) ~~shows is an early form of intraepithelial~~ ~~carcinoma with~~ limited potential to progress to invasive carcinoma and. ~~Therefore, we considered~~ 'excision/treatment' and 'follow-up examination' were considered as true-positive ~~for AK.~~

The CNN's softmax layer put out 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 to dermatologists' by using their mean specificity ~~of dermatologists~~ 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 to the average sensitivity of dermatologists by a 2-sided one sample *t*-test. We further applied the non-parametric Kruskal-Wallis test as omnibus test of heterogeneity between dermatologists with different levels of experience, and performed post-hoc comparisons of any pair of levels using Dunn-Nemenyi-procedure to adjust for multiple comparisons.¹⁷ ~~For comparisons among dermatologists 2-sided Mann-Whitney U tests were used.~~ Changes in dermatologists' diagnostic performance after receiving level-I or level-II information were tested by Wilcoxon paired signed-rank test with observations related to readers. 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, Illinois, USA) or SAS/STAT software, Version 9.4 (SAS Institute Inc., Cary, NC, USA.).

For Peer Review

1
2
3 **RESULTS**
4

5
6 **CNN’s diagnostic performance**
7

8 In its current market configuration (June 2019), the CNN showed a sensitivity and specificity
9 of 95.0% (95% CI [83.5%-98.6%]) and 76.7% (95% CI [64.6%-85.6%]), respectively. The
10 ROC AUC was 0.918 (95% CI [0.866-0.970]) (**Figure 1**). Boxplots in **Figure 2** show the
11 distribution of malignancy scores in relation to diagnostic categories. With the a priori
12 malignancy cutoff of >0.5 the percentage of correct classifications in malignant lesions was
13 100% in AK, 100% in Bowen’s disease, 100% in melanoma, 100% in BCCs, and 60% in
14 SCCs. In benign lesions the percentage of correct classifications was 90% in nevi, 80% in
15 angioma/angiokeratoma, 70% in seborrheic keratoses, 60% in dermatofibroma, and 50% in
16 solar lentigo.
17

18 In order to rule out overfitting and to confirm the generalizability of our results two larger
19 external datasets were used for testing (**supplementary Figure S2**), namely MSK-1 (1100
20 images) and ISIC-2018 challenge (1511 images). In the MSK-1 datasets the CNN attained an
21 almost identical performance in comparison to our test-set (sensitivity 94.2%, specificity
22 73.8%, ROC AUC 0.939). In the ISIC-2018 challenge dataset the CNN showed a lower
23 sensitivity of 84.7% at a higher specificity 84.1% and a comparable ROC AUC of 0.926. The
24 pairwise comparison of ROC AUCs attained by the CNN in all three datasets revealed no
25 significant differences (all $p>0.527$).
26
27

28
29 **Diagnostic performance of dermatologists**
30

31 Dermatologists (n=96) were categorized into ‘beginners’ (n=17, <2 years of experience),
32 ‘skilled’ (n=29, 2-5 years of experience), and ‘experts’ (n=40, >5 years of experience). Ten
33 participants did not provide information. The mean diagnostic performance of dermatologists
34 was assessed for their management decisions and dichotomous diagnostic classification of
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

lesions in level-I and -II ~~Moreover, differences according to the level of experience were investigated (Table 21).~~

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%-90.6%]) and 80.7% (95% CI [78.8%-82.6%]), respectively (Table 21). With additional case information in level-II the sensitivity significantly improved to 94.1% (95% CI [93.1%-95.1%]; $p < 0.001$) while the specificity stayed largely unchanged (80.4%, 95% CI [78.4%-82.4%]; $p = 0.97$).

Expectedly, the performance of dermatologists improved with more experience (supplementary Table S2). In level-I the percentage of correct management decisions (accuracy) increased from 79.9% in beginners (95% CI [77.7%-82.1%]) to 83.3% in skilled (95% CI [80.1%-85.6%]) and 86.9% in experts (95% CI [85.5%-88.3%]) ~~(all $p < 0.03$)~~. 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 to 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 ~~was were~~ made for accuracy in level-II (beginners: 82.0% (95% CI [79.3%-84.7%]), skilled: 85.4% (95% CI [83.0%-87.8%]), experts: 88.5% (95% CI [87.0%-90.0%])). However, ~~differences~~ in level-II the differences between experts and beginners also were significant for specificity ($p = 0.029$), while significance for sensitivity was still missed ($p = 0.225$). ~~were only significant between beginners and skilled ($p = 0.04$) and between beginners and experts ($p < 0.001$).~~

Dichotomous diagnostic classification

When viewing solely 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%-85.8%]) and 77.6% (95% CI [75.2%-80.0%]), respectively. With more information in level-II the sensitivity significantly improved to 90.6% (95% CI [89.3%-92.0%]; $p<0.001$). In contrast to management decisions, At the same time the specificity also significantly increased to 82.4% (95% CI [80.5%-84.3%]; ~~$p=0.003$~~ $p<0.001$).

In level-I the percentage of correct dichotomous classifications increased with more experience from 72.6% in beginners (95% CI [67.6%-77.6%]) to 79.3% in skilled (95% CI [76.3%-82.3%]) to 84.2% in experts (95% CI [82.0%-86.4%]) (all $p<0.01$; supplementary Table S2). The same observation was made in level-II (beginners: 81.2% (95% CI [78.0%-84.4%]), skilled: 85.1 (95% CI [82.5%-87.7%]), experts: 88.7% (95% CI [87.0%-90.4%])). As shown for 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. However, differences were only significant between beginners and experts ($p<0.001$) and between skilled and experts ($p=0.036$).

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%-98.6%]), which was ~~almost equals~~similar to the mean sensitivity of dermatologists (94.1% (95% CI [93.1%-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%-90.6%]) *versus* 95.0% (95% CI [83.5%-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%-85.8%]); $p<0.001$) or -II (90.6% (95% CI [89.3%-92.0%]; $p<0.001$)). **Figure 3-1** depicts the performance of dermatologists during level-I and -II in comparison to the CNN.

Additionally, we compared the CNN's accuracy (percentage of correct classifications) to dermatologists (**Table 21**). When the CNN's accuracy (84.0% (95% CI [75.6%-89.9%]) was compared to the dermatologists' mean accuracy in level-II management decisions (85.9% (95% CI [84.7%-87.1%]) dermatologists performed slightly but significantly better ($p=0.003$).

DISCUSSION

The incidence rates of melanoma and non-melanoma skin cancers are globally rising in most fair-skinned populations.¹⁸ Additional efforts in primary and secondary prevention are necessary to contain and possibly reverse these trends. ~~While P~~previous reports on the application of CNN in the diagnosis of skin cancer demonstrated a performance ~~either on or~~ above the level of dermatologists, ~~but the were setting of these studies was~~ legitimately criticized for ~~being~~ highly artificial study settings.¹³ Unlike ~~most~~ 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 clinical routine to confirm or rule out malignancy and brings the setting closer to a real-life clinical situation. Moreover, our results clearly demonstrate that the amount of case information and the sort of decisions asked of dermatologists had a major impact on their performance. Dermatologists performed best when allowed to review ~~a clinical~~ plus dermoscopic images accompanied by -close-up image, a dermoscopic image, and textual metadata ~~about the patient and the lesion~~. This setting largely reflects the presentation of cases in classical store-and-forward teledermatology¹⁴ and entails dermatologists to integrate different levels of information ~~before arriving at a decision~~. Interestingly, this observations also holds true 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, first reports of deep learning models using clinical non-imaging information have shown promising results in predicting skin cancer¹⁹ and may be included in neural networks. In daily routine dermatologists make management decisions rather than clear cut classifications whether it be more simple dichotomous classifications (benign vs. malignant) or naming specific diagnoses. ~~This is mostly due to the fact that ff~~ or each probable diagnosis there will be ~~one or more~~ differential diagnoses and .

~~D~~dermatologists are not used to ranking differential diagnoses ~~these~~ by a 'probability score'.

~~and it~~ It may be assumed that conflicting options lead to random decisions. Therefore, for future studies we recommend offering physicians enough data for review and to use their management decisions as primary outcome. Of note, ~~we also found significant differences in the performance of dermatologists according to their self-reported level of experience with dermoscopy. As as~~ the dermatologists' experience will strongly impact the results of a comparison to a CNN, we further recommend reporting ~~their physicians'~~ performance ranked by experience.

While the CNN ~~tested~~ in our present study undoubtedly performed very well, the questions of 'who will benefit the most from applying a CNN' is still a subject of much debate. By leaving the framework of a clinical study, 'man against machine' becomes 'man with a machine' and physicians will need to incorporate the ~~additional information of the~~ CNN's classification into their decision-making process. Unfortunately, there are no prospective studies available to show the impact.¹³ However, our summarized data in Table ~~2-1~~ suggest that ~~dermatologists less experienced physicians (or physicians of other specialties) who are less experienced with dermoscopy~~ 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 the setting of our study we can safely rule out overfitting, because training images were collected from ~~hundreds of multiple~~ sources around the world and test cases were derived from a source that did not provide training images. Moreover, we gave evidence for the generalizability of the CNN's results by including two larger external

1
2
3 datasets (one with full blinding of authors to true diagnoses) in which the tested CNN showed
4
5 comparable ROC AUCs.
6

7
8 There are several limitations to this study. First, (current) deep learning algorithms lack
9
10 interpretability.²⁰ Therefore, we are unable to name specific causes for ~~a false classification~~
11 ~~of a certain input image~~. Unfortunately, lack of interpretability prevents taking interventions
12
13 for improvement, which leads to repeating the same mistake in the future. ~~Frequent~~
14
15 ~~misclassifications in a specific diagnostic category are often explained by a 'lack of training~~
16
17 ~~images'. However, this is mostly speculative, and misclassifications may persist with more~~
18
19 ~~training for other reasons~~. Second, to safeguard the feasibility of the reader study our test-set
20
21 included 100 lesions, thus leaving only a few cases for certain diagnoses. The CNN's results
22
23 attained in these diagnoses should, therefore, be interpreted with caution. Third, the
24
25 compilation of our test-set did not include some other benign (e.g. viral warts), malignant (e.g.
26
27 Merkel cell carcinoma), or inflammatory skin lesions (e.g. clear cell acanthoma) ~~that may be~~
28
29 ~~encountered in daily clinical routine~~. Therefore, our results should not be generalized to a
30
31 large prospective patient population. Finally, the CNN was mostly trained with dermoscopic
32
33 images of patients with a Caucasian genetic background ~~(Fitzpatrick skin types 1-3)~~ and may
34
35 not attain comparable results in a population with divergent skin types.
36
37
38
39
40
41
42

43 In conclusion, the results of our study demonstrate that the tested CNN is capable of
44
45 classifying a broad spectrum of skin tumors at a high level of sensitivity and specificity.
46
47 Under less artificial conditions, with availability of clinical close-up images, dermoscopic
48
49 images, and textual case information for review, the management decisions of most
50
51 dermatologists were either on or slightly above the level of the CNN. Experts in dermoscopy
52
53 with access to relevant case information ~~(study level-II)~~ commonly outperformed the CNN's
54
55 specificity at a comparable sensitivity. There is a need for prospective studies and an
56
57 improved interpretability of a CNN's classification results to take research to a next level.
58
59
60

Acknowledgements:

We would like to thank ~~all the following~~ dermatologists ~~in alphabetical order~~, who ~~agreed to be acknowledged and who~~ actively and voluntarily spend much time to participate 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-Buffer, 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 Ribeau, Pauline Richez, Hélène Roche-Plaine, Deborah Salik, Elke Sattler, Sarah Schäfer, Roland Schneiderbauer, Thierry Secchi, Karen Talour, Lukas Trennheuser, Alexander Wald, Priscila Wölbing, Pascale Zukervar.~~

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

Funding-statement: This research project received funding from a public, non-for-profit domain, namely the Skin Cancer Council Germany (Nationale Versorgungskonferenz Hautkrebs (NVKH) e.V.), www.nvkh.de/.

Grant/Award Number: AD-LEARN DERMOSCOPY (HF01-Z02-P05).

Conflict-of-interestDisclosures: A Blum received honoraria and/or travel expenses from FotoFinder Systems GmbH. C Fink received travel expenses from Magnosco GmbH. HA Haenssle 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. P Tschandl has received honoraria from Silverchair, and an unrestricted research grant from MetaOptima Technology Inc. All other authors declared no conflict of interest.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Financial Disclosure: None reported

Reference List

1. Fink C, Haenssle HA. Non-invasive tools for the diagnosis of cutaneous melanoma. *Skin Res Technol* 2017;23:261-71.
2. Forschner A, Keim U, Hofmann M, Spankuch I, Lomberg D, Weide B 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;179:478-85.
3. Lui H, Zhao J, McLean D, Zeng H. Real-time Raman spectroscopy for in vivo skin cancer diagnosis. *Cancer Res* 2012;72:2491-500.
4. Malvehy J, Hauschild A, Curiel-Lewandrowski C, Mohr P, Hofmann-Wellenhof R, Motley R 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;171:1099-107.
5. Monheit G, Cognetta AB, Ferris L, Rabinovitz H, Gross K, Martini M et al. The performance of MelaFind: a prospective multicenter study. *Arch Dermatol* 2011;147:188-94.
6. Cukras AR. On the comparison of diagnosis and management of melanoma between dermatologists and MelaFind. *JAMA Dermatol* 2013;149:622-3.
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.

- 1
2
3 8. Fujisawa Y, Otomo Y, Ogata Y, Nakamura Y, Fujita R, Ishitsuka Y et al. Deep-learning-based,
4
5 computer-aided classifier developed with a small dataset of clinical images surpasses board-
6
7 certified dermatologists in skin tumour diagnosis. *Br J Dermatol* 2019;180:373-81.
8
9
10
11
12 9. Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A et al. Man against machine:
13
14 diagnostic performance of a deep learning convolutional neural network for dermoscopic
15
16 melanoma recognition in comparison to 58 dermatologists. *Ann Oncol* 2018;29:1836-42.
17
18
19
20
21 10. Marchetti MA, Codella NCF, Dusza SW, Gutman DA, Helba B, Kalloo A et al. Results of the 2016
22
23 International Skin Imaging Collaboration International Symposium on Biomedical Imaging
24
25 challenge: Comparison of the accuracy of computer algorithms to dermatologists for the
26
27 diagnosis of melanoma from dermoscopic images. *J Am Acad Dermatol* 2018;78:270-7.
28
29
30
31
32 11. Tschandl P, Rosendahl C, Akay BN, Argenziano G, Blum A, Braun RP et al. Expert-Level Diagnosis
33
34 of Nonpigmented Skin Cancer by Combined Convolutional Neural Networks. *JAMA Dermatol*
35
36 2019;155:58-65.
37
38
39
40
41 12. Yu C, Yang S, Kim W, Jung J, Chung KY, Lee SW et al. Acral melanoma detection using a
42
43 convolutional neural network for dermoscopy images. *PLoS One* 2018;13:e0193321.
44
45
46
47
48 13. Lallas A, Argenziano G. Artificial intelligence and melanoma diagnosis: ignoring human nature
49
50 may lead to false predictions. *Dermatol Pract Concept* 2018;8:249-51.
51
52
53
54 14. Finnane A, Dallest K, Janda M, Soyer HP. Teledermatology for the Diagnosis and Management
55
56 of Skin Cancer: A Systematic Review. *JAMA Dermatol* 2017;153:319-27.
57
58
59
60

15. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Reprint at <https://arxiv.org/abs/1512.00567> 2015.

16. 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.

17. Elliott AC, Hynan LS. A SAS((R)) 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, Castellsague X. Skin Cancer: Epidemiology, Disease Burden, Pathophysiology, Diagnosis, and Therapeutic Approaches. Dermatol Ther (Heidelb) 2017;7:5-19.

19. Wang HH, Wang YH, Liang CW, Li YC. Assessment of Deep Learning Using Nonimaging Information and Sequential Medical Records to Develop a Prediction Model for Nonmelanoma Skin Cancer. JAMA Dermatol 2019.

20. Pereira S, Meier R, McKinley R, Wiest R, Alves V, Silva CA 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;44:228-44.

APPENDIX:

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-Buffer, 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 Ribeau, Pauline Richez, Hélène Roche Plaine, Deborah Salik, Elke
Sattler, Sarah Schäfer, Roland Schneiderbauer, Thierry Secchi, Karen Talour, Lukas
Trennheuser, Alexander Wald, Priscila Wölbing, Pascale Zukervar.
All other participants asked to remain anonymous and we also thank these colleagues for their
commitment.

FIGURE LEGENDS

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 in their management decisions (C: level-I, D: level-II). The average ($\pm$ SD) sensitivity and specificity of all dermatologists (mean: green circle; $\pm$ 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).

~~ROC curve of the CNN (black curve) in relation to the results of all dermatologists (n=96, red dots) in their level-II management decisions. The average ($\pm$ SD) sensitivity and specificity of all dermatologists (green, Sensitivity: 94.1%, Specificity: 80.4%) in their level-II management decisions is depicted. The point of operation of the CNN is indicated by a blue circle (Sensitivity: 95.0%, Specificity: 76.7%).~~

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 while the median is indicated by the line intersection the upper and lower box. Whiskers indicate the full range of probability scores.

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

Figure 3.

~~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 in their management~~

1
2
3 decisions (C: level-I, D: level-II). The average ($\pm$ SD) sensitivity and specificity of all
4
5 dermatologists (mean: green circle; $\pm$ SD: green error bars) and the CNN's point of operation
6
7 (blue circle) is depicted. Dermatologists performed best when allowed to review more case
8
9 information (B better than A, D better than C) and when deciding on the management of cases
10
11 (D better than B).
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

For Peer Review

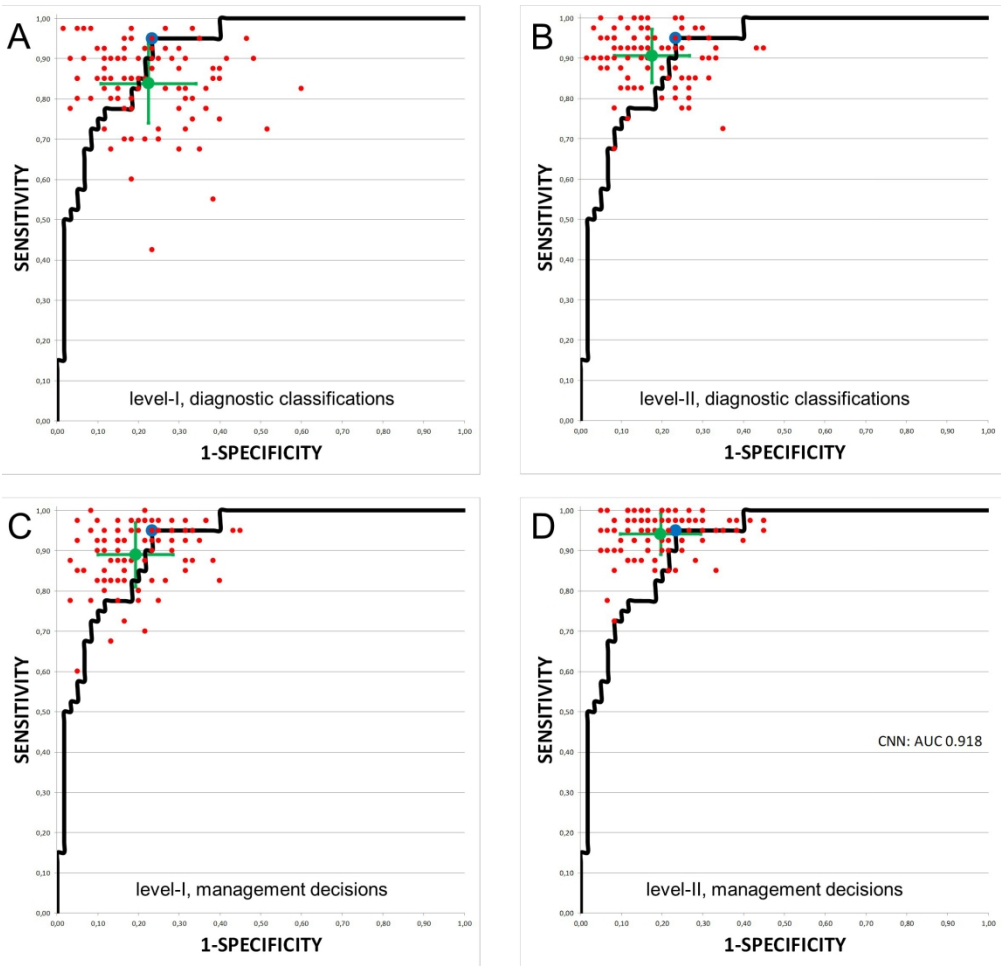


Figure 1.
ROC curve of the CNN (black curve) in relation to the results of all dermatologists (n=96, red dots) in their dichotomous classifications (A: level-I, B: level-II) and in their management decisions (C: level-I, D: level-II). The average (□ SD) sensitivity and specificity of all dermatologists (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).

198x190mm (300 x 300 DPI)

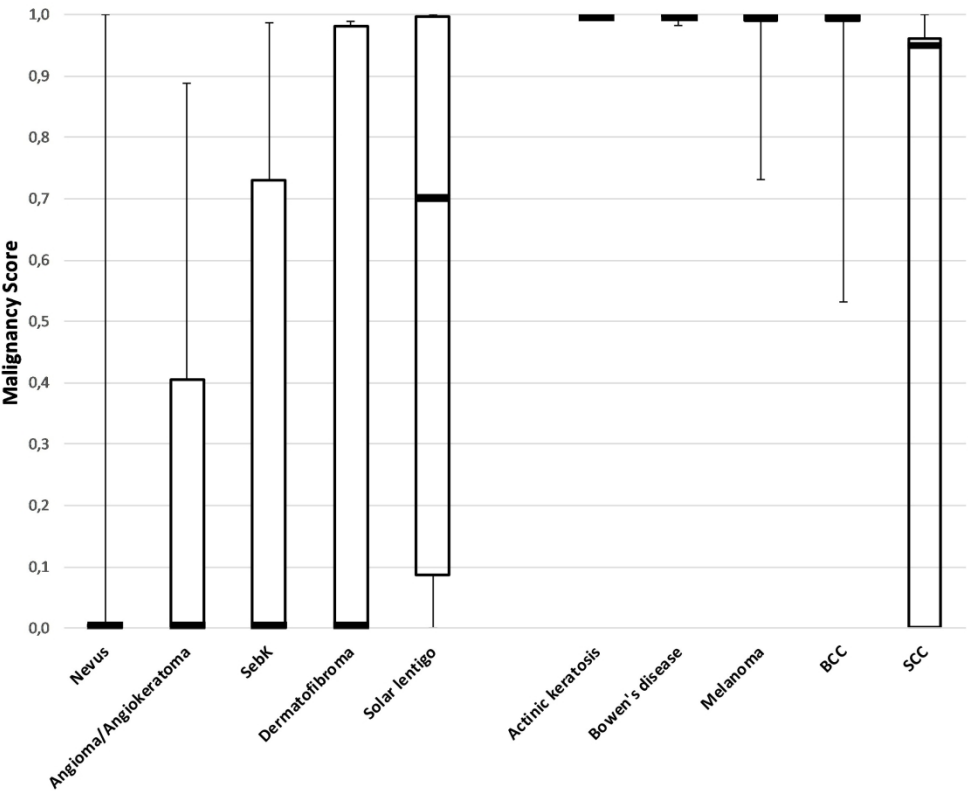


Figure 2.
The CNN's melanoma probability scores (range 0-1) for major benign and malignant diagnostic categories are depicted as boxplots. Scores closer to 1 indicated a higher probability of malignancy. The upper and lower bounds of boxes indicate the 25th and 75th percentiles while the median is indicated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

235x189mm (300 x 300 DPI)

Table 1. Reader study level-I and -II results in comparison to 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%¹
‘Beginner’ (n=17)	85.7%	76.1%	79.9%¹	80.0%	67.7%	72.6%¹
‘Skilled’ (n=29)	89.7%	79.1%	83.3%	84.3%	75.9%	79.3%¹
‘Expert’ (n=40)	91.1%	84.1%	86.9%²	86.2%	82.9%	84.2%
Raters Level-II						
All (n=96)	94.1%	80.4%	85.9%²	90.6%	82.4%	85.7%²
‘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%²	91.8%	86.6%	88.7%²
CNN	95.0%	76.7%	84.0%	95.0%	76.7%	84.0%

Level-I: readers were solely provided with dermoscopic images
Level-II: readers were additionally provided with close-up images and textual case information
Accuracy: Was 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; ‘*Beginner*’: <2 years of experience. Ten participants did not report level of experience
¹The CNN’s accuracy was significantly higher than the mean accuracy of dermatologists
²The mean accuracy of dermatologists was significantly higher than the CNN’s accuracy

SUPPLEMENTARY METHODS

Convolutional neural network architecture

We used Google's Inception_v4 CNN architecture, pretrained across 1,000 object classes and composed of 27 layers.¹ Known by its high predictive performance, Inception_v4 architecture is a pure Inception variant without residual connections that has roughly the same recognition performance as Inception-ResNet-v2. The global Inception_v4 network architecture is displayed in **supplementary figure S1**. The network starts with the input of a 299x299 pixels image that is fed to the stem module containing 3 inception modules. The output is fed to the 'Inception A' module four times. 'Inception A' contains combination of average pooling, 1x1 convolutions and 3x3 convolutions. The concatenated output is run through a reduction module. Next, the data is fed to the 'Inception B' module seven times. The 'Inception B' module is the concatenation of average pooling, 1x1 convolutions and 7x1 and 1x7 convolutions. Resulting data is fed to another reduction module, followed by an 'Inception C' module three times. 'Inception C' is composed of average pooling, 1x1 convolution, and 1x3 and 3x1 convolutions. The result is run through an average pooling and a dropout layer and finally through a softmax layer that puts out the final diagnostic classification.

Data collection

Dermoscopic images and corresponding diagnostic labels of skin lesions of different anatomic localizations were collected from multiple sources, including many cooperating dermatologists and the International Skin Imaging Collaboration (ISIC) dermoscopic archive.² Diagnostic labels were validated by histopathology, or, in case of non-excised lesions, by experienced dermatologists and/or inconspicuous sequential digital dermoscopy follow-up examinations. Each collected dataset was split at a 9:1 ratio for training and validation/testing. Additional

image sets for validation and testing were manually compiled and contained difficult-to-diagnoses cases (from a human perspective) to determine the model’s performance.

Data pre-processing

Because of imbalances in the ratio of (more frequent) benign and (less frequent) malignant lesions, an oversampling of the data was needed. We used an oversampling by several geometric image transformations (e.g. random brightness, reflection, rotation, minimal color distortion) which left the original classification unchanged. In our model, we randomly applied (using a normal distribution) one or more of these transformations on images of the original dataset to finally attain a ratio of 1:1 for images of benign versus malignant lesions. For each epoch, images were randomly selected from the batches, and none, one, or several of the transformations were applied. As a result, the same original data with slight variations were re-entered from one epoch to the other.

Training Process

We initialized Google’s Inception_v4 CNN by using transfer learning,³ i.e. the neurons’ weights were initialized kept as previously set in the model pretrained across multiple object classes, instead of using random initiation. By using this approach, we were able to finish the training with the limited specific dataset (dataset of dermoscopic images). This process allowed us to shorten the training time as well as improve the results and minimizes the risk of the gradient descent to get blocked in a local minimum. Due to the design of the network, input images derived from a range of different capturing devices had to be resized to 299x299 pixels by padding and cropping. Training with the limited specific dataset was attained by using a variation of the stochastic gradient descent algorithm called Adaptive Moment estimation (AdaM) optimizer⁴ and Google’s deep learning library TensorFlow.⁵ A fixed learning rate at 0.001 with 0.9 decay factor, a momentum of 0.999 and $\epsilon = 10^{-8}$ were used. The training batch

size was 16, and the number of epochs was 80. No early stopping criteria were used during all the training instances.

Overfitting and generalizability

In order to rule out overfitting and to better confirm the generalizability of the results attained by the tested CNN in the dataset of the reader study, we additionally assessed the images of two larger, publicly available, external datasets. The Memorial Sloan Kettering dataset (MSK-1) includes 1100 images of a similar broad spectrum of diagnoses and was downloaded with accompanying metadata from the ISIC-Archive web interface (<https://www.isic-archive.com>). The ISIC-2018 challenge dataset⁶ includes 1511 dermoscopic images of a broad spectrum of pigmented skin lesions and was also downloaded from the ISIC-Archive web interface (<https://submission.challenge.isic-archive.com/#phase/5bee4312c5eaea4f24b5ec0c>). In contrast to the MSK-1 dataset the diagnoses for the ISIC-2018 challenge images will not be released by the challenge organizers thus necessitating external evaluation of the CNN's classifications in our study (P. Tschandl, Department of Dermatology, University of Vienna, Austria). Supplementary figure S2 depicts the ROC curves attained within the dataset of our reader study in comparison to the MSK-1 and ISIC-2018 challenge datasets.

Reference List

1. Szegedy C, Vanhoucke V, Ioffe S *et al.* Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016; 2818-26.

2. Finnane A, Curiel-Lewandrowski C, Wimberley G *et al.* Proposed Technical Guidelines for the Acquisition of Clinical Images of Skin-Related Conditions. *JAMA dermatology* 2017; **153**: 453-7.

3. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 2010; **22**: 1345-59.

4. Ruder S. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747* 2016.

5. Abadi M, Barham P, Chen J *et al.* Tensorflow: A system for large-scale machine learning. In: *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16).* 2016; 265-83.

6. 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;20:938-47.

Supplementary Figure S1

The global Inception_v4 network architecture

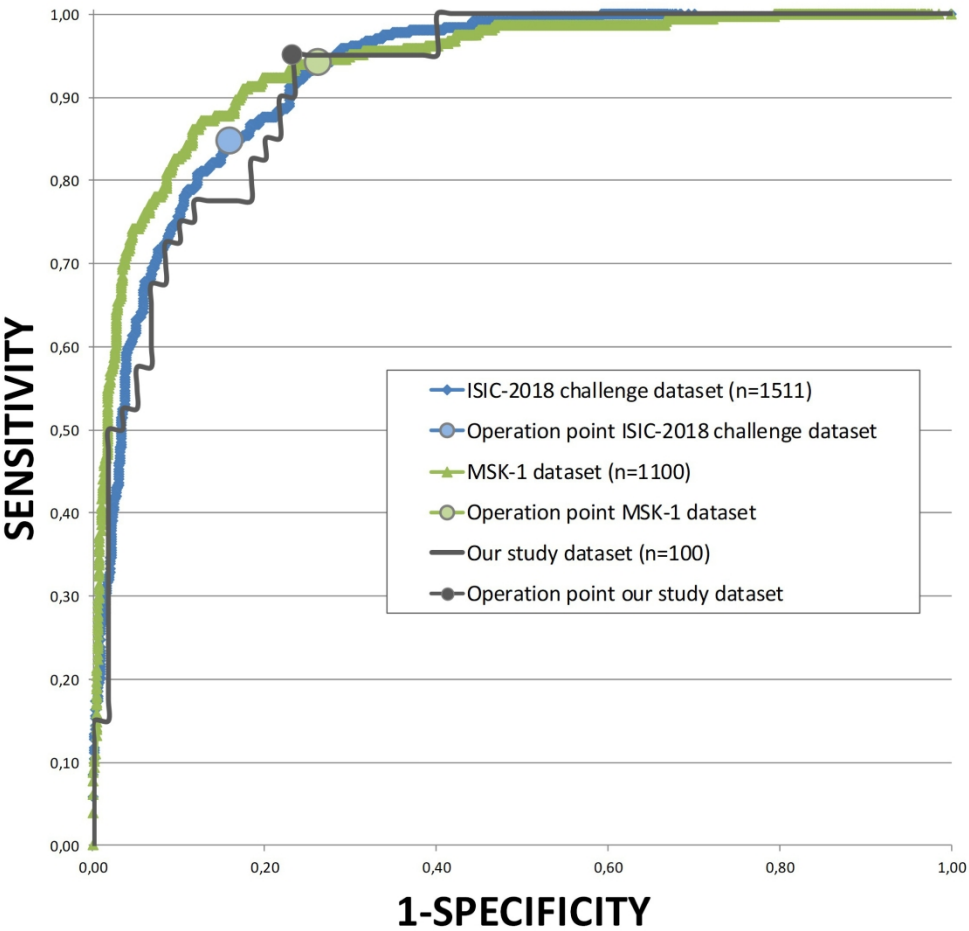
Supplement Figure S2.

ROC curve of the CNN attained in the dataset of our reader study (grey curve) in comparison to ROC curves that resulted from the diagnostic classification of the MSK-1 dataset (green curve) and the ISIC-2018 challenge dataset (blue curve). The point of operation of the CNN within each dataset is depicted as a circle on the curve (our study: sensitivity 95.0%, specificity 76.7%; MSK-1 dataset sensitivity 94.2%, specificity 73.8%; ISIC-2018 challenge dataset sensitivity 84.7%, specificity 84.1%). The ROC AUCs in descending order were 0.939 in the MSK-1 dataset (95% CI [0.923-0.954]), 0.926 in the ISIC-2018 challenge dataset (95% CI [0.912-0.940]), and 0.918 in our study test-set (95% CI [0.866-0.970]) and no significant differences were found in statistical pairwise comparisons (all $p>0.527$).

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



338x50mm (300 x 300 DPI)



194x188mm (300 x 300 DPI)

Supplementary Table S1: Characteristics of test-set lesions (n=100)

	Benign lesions (n=60)		Malignant lesions (n=40)	
	n	(%)	n	(%)
Localization				
Non-glabrous common skin	37	(61.7%)	22	(55.0%)
Palmoplantar skin	6	(10.0%)	1	(2.5%)
Facial skin	14	(23.3%)	12	(30.0%)
Hairbearing scalp	1	(1.7%)	5	(12.5%)
Mucosa	2	(3.3%)	-	-
Ground truth				
Histopathology	45	(75.0%)	40	(100%)
Unremarkable follow-up (>2 years)	15	(25.0%)		
Melanoma				
			15	(37.5%)
In situ melanoma			3	(7.5%)
Invasive melanoma			12	(30.0%)
Superficial spreading melanoma			9	(22.5%)
Lentigo maligna melanoma			1	(2.5%)
Acrolentiginous melanoma			1	(2.5%)
Nodular melanoma			1	(2.5%)
Basal cell carcinoma				
			10	(25%)
Pigmented BCC			5	(12.5%)
Ulcerated BCC			2	(5.0%)
Superficial BCC			1	(2.5%)
Squamous cell carcinoma				
			5	(12.5%)
Actinic keratosis				
			5	(12.5%)
Bowen's disease				
			5	(12.5%)
Melanocytic nevus				
	30	(50.0%)		
Congenital type nevus	8	(13.3%)		
Compound nevus	6	(10.0%)		
Junctional nevus	2	(3.3%)		
Blue nevus	2	(3.3%)		
Spitz nevus	4	(6.7%)		
Acral nevus	4	(6.7%)		
Combined nevus	2	(3.3%)		
Mucosal nevus	1	(1.7%)		
Speckled lentiginous nevus (spilus)	1	(1.7%)		
Seborrheic keratosis				
	10	(16.7%)		
Solar lentigo				
	10	(16.7%)		
Angioma/Angiokeratoma				
	5	(8.3%)		
Dermatofibroma				
	5	(8.3%)		

Mean thickness (min-max) of invasive melanomas was 2.2mm (0.7mm-3.8mm). Melanomas included hypomelanotic/amelanotic forms, actinic keratoses included pigmented forms, Bowen's disease included pigmented forms, angioma/angiokeratoma included (partially) thrombosed forms.

Supplementary Table S2: Comparisons of dermatologists' diagnostic performance with regard to different levels of experience (p-values)

p-values	Management decision			Binary classification		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Raters Level-I						
Beginner vs Expert	0.295	0.078	0.006	0.452	0.014	0.009
Beginner vs Skilled	0.579	0.646	0.246	0.784	0.456	0.330
Expert vs Skilled	0.860	0.316	0.229	0.827	0.157	0.204
All groups KW	0.108	0.005	<0.001	0.230	<0.001	<0.001
Raters Level-II						
Beginner vs Expert	0.539	0.029	0.013	0.512	0.019	0.019
Beginner vs Skilled	0.477	0.448	0.272	0.677	0.435	0.371
Expert vs Skilled	0.980	0.281	0.326	0.963	0.223	0.275
All groups KW	0.225	0.001	<0.001	0.290	<0.001	<0.001

Level-I: readers were solely provided with dermoscopic images

Level-II: readers were additionally provided with close-up images and textual case information

Accuracy: Was 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; '*Beginner*': <2 years of experience. Ten participants did not report level of experience

All groups KW: Kruskal-Wallis omnibus test of heterogeneity between groups, pairwise comparisons were corrected for multiple comparisons by Dunn-Nemenyi-procedure.