JAMA Dermatology | Original Investigation

Assessment of Diagnostic Performance of Dermatologists Cooperating With a Convolutional Neural Network in a Prospective Clinical Study
Human With Machine

Julia K. Winkler, MD; Andreas Blum, MD; Katharina Kommoss, MD; Alexander Enk, MD; Ferdinand Toberer, MD; Albert Rosenberger, MSc; Holger A. Haenssle, MD

Supplemental content

IMPORTANCE Studies suggest that convolutional neural networks (CNNs) perform equally to trained dermatologists in skin lesion classification tasks. Despite the approval of the first neural networks for clinical use, prospective studies demonstrating benefits of human with machine cooperation are lacking.

OBJECTIVE To assess whether dermatologists benefit from cooperation with a market-approved CNN in classifying melanocytic lesions.

DESIGN, SETTING, AND PARTICIPANTS In this prospective diagnostic 2-center study, dermatologists performed skin cancer screenings using naked-eye examination and dermoscopy. Dermatologists graded suspect melanocytic lesions by the probability of malignancy (range 0-1, threshold for malignancy ≥0.5) and indicated management decisions (no action, follow-up, excision). Next, dermoscopic images of suspect lesions were assessed by a market-approved CNN, Moleanalyzer Pro (FotoFinder Systems). The CNN malignancy scores (range 0-1, threshold for malignancy ≥0.5) were transferred to dermatologists with the request to re-evaluate lesions and revise initial decisions in consideration of CNN results. Reference diagnoses were based on histopathologic examination in 125 (54.8%) lesions or, in the case of nonexcised lesions, on clinical follow-up data and expert consensus. Data were collected from October 2020 to October 2021.

MAIN OUTCOMES AND MEASURES Primary outcome measures were diagnostic sensitivity and specificity of dermatologists alone and dermatologists cooperating with the CNN. Accuracy and receiver operator characteristic area under the curve (ROC AUC) were considered as additional measures.

RESULTS A total of 22 dermatologists detected 228 suspect melanocytic lesions (190 nevi, 38 melanomas) in 188 patients (mean [range] age, 53.4 [19-91] years; 97 [51.6%] male patients). Diagnostic sensitivity and specificity significantly improved when dermatologists additionally integrated CNN results into decision-making (mean sensitivity from 84.2% [95% CI, 69.6%-92.6%] to 100.0% [95% CI, 90.8%-100.0%]; *P* = .03; mean specificity from 72.1% [95% CI, 65.3%-78.0%] to 83.7% [95% CI, 77.8%-88.3%]; *P* < .001; mean accuracy from 74.1% [95% CI, 68.1%-79.4%] to 86.4% [95% CI, 81.3%-90.3%]; *P* < .001; and mean ROC AUC from 0.895 [95% CI, 0.836-0.954] to 0.968 [95% CI, 0.948-0.988]; *P* = .005). In addition, the CNN alone achieved a comparable sensitivity, higher specificity, and higher diagnostic accuracy compared with dermatologists alone in classifying melanocytic lesions. Moreover, unnecessary excisions of benign nevi were reduced by 19.2%, from 104 (54.7%) of 190 benign nevi to 84 nevi when dermatologists cooperated with the CNN (*P* < .001). Most lesions were examined by dermatologists with 2 to 5 years (96, 42.1%) or less than 2 years of experience (78, 34.2%); others (54, 23.7%) were evaluated by dermatologists with more than 5 years of experience. Dermatologists with less dermoscopy experience cooperating with the CNN had the most diagnostic improvement compared with more experienced dermatologists.

CONCLUSIONS AND RELEVANCE In this prospective diagnostic study, these findings suggest that dermatologists may improve their performance when they cooperate with the market-approved CNN and that a broader application of this human with machine approach could be beneficial for dermatologists and patients.

JAMA Dermatol. doi:10.1001/jamadermatol.2023.0905
Published online May 3, 2023.

Author Affiliations: Department of Dermatology, University of Heidelberg, Heidelberg, Germany (Winkler, Kommoss, Enk, Toberer, Haenssle); Public, Private and Teaching Practice of Dermatology, Konstanz, Germany (Blum); Institute of Genetic Epidemiology, University Medical Center, Georg-August University of Goettingen, Goettingen, Germany (Rosenberger).

Corresponding Author: Holger A. Haenssle, MD, Department of Dermatology, University of Heidelberg, Im Neuenheimer Feld 440, 69120 Heidelberg, Germany (holger.haenssle@med.uni-heidelberg.de).

The incidence of skin cancer remains high around the globe, and early diagnosis is relevant to patients' prognosis.[1,2] In skin cancer classification tasks, convolutional neural networks (CNN) achieved diagnostic accuracies similar to trained dermatologists.[3-6] Most previous studies applied a retrospective setting, using images of skin lesions with validated diagnoses.[7] In one initial competition of human against machine, the diagnostic performance of dermatologists was compared with a CNN that classified dermoscopic images of skin lesions.[4,5] Since then, numerous studies have confirmed the high-level diagnostic performance of different classifiers but also unraveled important limitations, particularly increased numbers of false diagnoses in images including artifacts such as scale bars and skin markings[8,9] or in rare lesions found at mucosal or subungual sites.[10] Despite a European market approval of a number of neural networks for skin lesion classification, prospective clinical studies investigating the integration of CNN results into daily clinical decision-making after live patient examinations are lacking.

Most previous studies were limited to the assessment of human and machine collaborations in a retrospective setting where dermatologists were asked to review images of skin lesions with or without the availability of CNN classification results.[11,12] However, there are important differences between retrospective and prospective studies involving the diagnosis of skin lesions by dermatologists and a CNN. First, in a prospective study, dermatologists may directly interview and examine patients (live examinations); whereas in most retrospective studies, they could review only a single dermoscopic image. Second, clinical decisions in a prospective study have a direct association with patient well-being; whereas retrospective studies lack the harsh consequences of missing any malignant lesions. Therefore, it remains to be assessed whether and how dermatologists may incorporate CNN recommendations into their clinical decision-making process.

With the present study, we aimed to elucidate the cooperation of dermatologists with a market-approved CNN in a prospective clinical setting. Moreover, we used a validated questionnaire measuring patient acceptance and trust toward the tested CNN.

Methods

This prospective diagnostic study was approved by the ethics committee of the medical faculty of the University of Heidelberg (approval number: S-836/2020) and performed in accordance with the Declaration of Helsinki principles. All patients gave written informed consent before study-related procedures. The Standards for Reporting of Diagnostic Accuracy (STARD) reporting guideline was followed.

Study Settings

This prospective, explorative, observational, clinical study aimed to investigate the cooperation of dermatologists with a CNN approved for clinical use, Moleanalyzer-Pro (FotoFinder Systems), and to measure possible changes in diagnostic performance. Due to statistical considerations, this study included only melanocytic lesions. The study was performed at 2 sites: a university department of dermatology (University of Heidelberg, Germany) and at a private practice of dermatology (Public, Private and Teaching Practice of Dermatology, Andreas Blum, Konstanz, Germany).

All study procedures are listed in the study flowchart (eFigure 1 in Supplement 1). Briefly, dermatologists with different levels of experience in using dermoscopy for skin cancer screening (<2 years, 2-5 years, >5 years) performed full-body examinations using the unaided eye and dermoscopy. Dermatologists were asked to indicate the probability of malignancy on a visual analog scale between 0 and 1 for suspect melanocytic lesions. They were informed to apply a threshold of at least 0.5 for suspected melanoma. Management decisions of dermatologists were recorded (no action, follow-up examination, excision). Next, patients were sent to another room for CNN assessment of 1 dermoscopic image per suspect lesion. The CNN put out a malignancy score between 0 and 1 with a threshold for malignancy of at least 0.5.[4,5] The CNN malignancy scores were then forwarded to the examining dermatologists, who were asked to revise their diagnoses and management decisions in consideration of the CNN results. Finally, dermatologists were asked whether or not they judged the CNN scores to be helpful and/or reassuring. Patients were provided a questionnaire to assess their trust and acceptance toward CNN-based assistant systems. The questionnaire included 10 statements based on the validated "trust in medical technology" instrument.[13,14] Histopathologic examination was performed by experienced board-certified histopathologists using a consensus conference for difficult-to-diagnose cases.

Statistical Analysis

A sample size of 183 patients was found necessary to detect an improved specificity of 90% (lower threshold >75%) for dermatologists using CNN support vs unsupported dermatologists with a power of 90% at a level of significance of 5%. Primary outcome measures were the sensitivity and specificity of melanoma detection. Accuracy and receiver operating characteristic area under the curve (ROC AUC) were considered as additional

Key Points

Question Do dermatologists benefit from cooperation with a market-approved convolutional neural network (CNN) in classifying melanocytic lesions?

Findings In this prospective diagnostic study, the CNN achieved a comparable sensitivity, higher specificity, and higher diagnostic accuracy compared with dermatologists alone in classifying melanocytic lesions, and dermatologists cooperating with the CNN significantly improved their diagnostic performance. The CNN's higher specificity guided dermatologists to excise significantly fewer benign nevi; additionally, dermatologists with less dermoscopy experience cooperating with the CNN had the most diagnostic improvement.

Meaning Such human with machine cooperation approaches should be further evaluated and potentially implemented in clinical settings, such as classifying melanocytic lesions in dermatology.

measures supporting interpretation. Secondary outcome measures included the CNN's performance itself and the patients' trust and acceptance as determined by the questionnaire.

The reference diagnoses of examined lesions (ground truth) were based on histopathologic examination reports for excised lesions or clinical follow-up data and expert consensus for nonexcised lesions. For statistical calculations dichotomous diagnostic classifications of melanocytic lesions (benign/malignant) and dermatologists' management decisions were used. Management "excision" in melanomas and "follow-up" or "no action" in melanocytic nevi were considered true-positive or true-negative, respectively. Whenever dermatologists recommended excision, they were asked to indicate reason(s) for excision (clinical appearance, dermoscopic appearance, anamnestic information/patient concern).

We investigated differences in diagnostic performance of the CNN and dermatologists, making decisions with and without knowledge of CNN results. We used McNemar test to detect differences in proportions of categorical variables and Wilcoxon signed rank test to assess continuous data. Moreover, a pairwise statistical comparison of ROC AUCs was performed.[15] Bonferroni corrections were used to adjust for multiple testing. A significance threshold of *P* < .05 was used. Significance tests were 2-tailed. For all analyses, SPSS statistical software, version 25 (IBM) was used.

Results

Patient and Lesion Characteristics

In this study, 22 dermatologists examined 188 patients (mean [range] age, 53.4 [19-91] years; 97 [51.6%] male patients; **Table 1**) and detected 228 suspect melanocytic lesions. A total of 166 (88.3%) patients were included at the Department of Dermatology at the University of Heidelberg , and 22 (11.7%) patients were included at the Public, Private and Teaching Practice of Dermatology in Konstanz, Germany. The majority of patients showed skin type (Fitzpatrick classification) 2 (33.5%) or 3 (56.4%). A high total body nevus count per patient (>50) was documented in 30.9% of patients (51-100 in 33 patients, >100 in 25 patients). The study included 51 (27.1%) patients with previous melanoma and 25 (13.3%) patients with multiple (>5) atypical nevi. A family history of melanoma was reported by 13 (6.9%) patients. Nonmelanoma skin cancer had previously been diagnosed in 29 (15.4%) patients. Overall, 111 (59.0%) patients had no personal or family history of any skin cancer.

Out of the 228 suspect lesions, there were 190 (83.3%) nevi and 38 (16.7%) melanomas (Table 1). Most lesions were localized on the trunk (n = 148, 64.9%) or lower (n =35, 15.4%) and upper extremities (n = 22, 9.6%). Of note, this study also included lesions from special localizations, such as 18 from the head and neck area, 3 from acral skin, or 2 from the nail unit. For 125 (54.8%) lesions, the reference diagnosis was based on histopathologic examination reports. Histopathologic examination identified 44 dysplastic nevi without suggesting melanoma or recommending re-excision. The diagnosis of the remaining nonexcised lesions was validated by clinical follow-up and/or expert consensus.

Table 1. Characteristics of the Patients and Lesions Included in the Study

Characteristic	No. (%)
Patients (n = 188)	
Gender	
Female	91 (48.4)
Male	97 (51.6)
Skin type according to Fitzpatrick	
Type 1	5 (2.7)
Type 2	63 (33.5)
Type 3	106 (56.4)
Type 4	12 (6.4)
Type 5	2 (1.1)
Type 6	0 (0)
No. of nevi	
0-15 Nevi	42 (22.3)
16-50 Nevi	88 (46.8)
51-100 Nevi	33 (17.6)
>100 Nevi	25 (13.3)
No. of atypical nevi	
0-5 Nevi	163 (86.7)
>5 Nevi	25 (13.3)
Personal/family history	
Previous melanoma	51 (27.1)
Previous nonmelanoma skin cancer	29 (15.4)
Positive family history for melanoma	13 (6.9)
Previous skin cancer screening	74 (39.6)
Lesions (n = 228)	
Validated diagnosis	
Melanocytic nevus	190 (83.3)
Melanoma	38 (16.7)
In situ melanoma	12 (5.3)
Invasive melanoma (median thickness, 1.0 mm)	26 (11.4)
Localization	
Head/neck	18 (7.9)
Trunk	148 (64.9)
Upper extremities	22 (9.6)
Lower extremities	35 (15.4)
Acral	3 (1.3)
Nail	2 (0.9)
Type of diagnostic validation	
Histopathologic examination report	125 (54.8)
Follow-up/expert opinion	103 (45.2)

Diagnostic Classifications

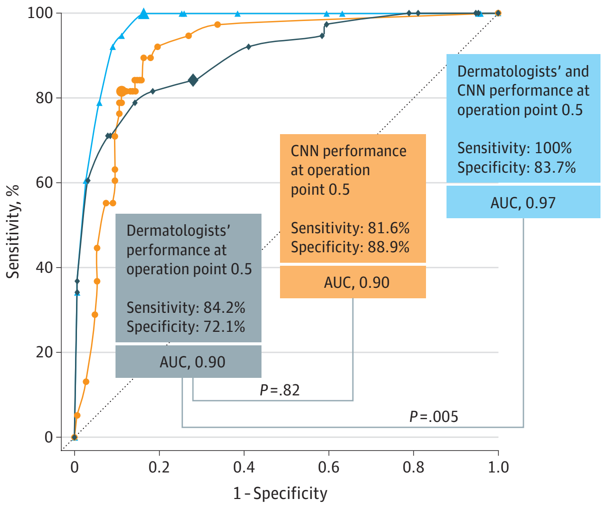
Dermatologists who prospectively examined patients and lesions (live examination, no access to CNN results) achieved a mean diagnostic sensitivity of 84.2% (95% CI, 69.6%-92.6%) and specificity of 72.1% (95% CI, 65.3%-78.0%) (**Table 2**). After receiving and integrating CNN results, dermatologists significantly improved their mean sensitivity and specificity to 100% (95% CI, 90.8%-100.0%; *P* = .03) and 83.7% (95% CI, 77.8%-88.3%; *P* < .001), respectively. The CNN itself, that solely assessed 1 dermoscopic image per lesion,

Table 2. Sensitivity, Specificity, and Accuracy of Diagnostic Classifications and Management Decisions of Dermatologists, Convolutional Neural Network (CNN), and Dermatologists Cooperating With CNN[a]

	Sensitivity	Specificity	Accuracy
Diagnostic classifications			
Dermatologists alone	84.2 (69.9-92.6)	72.1 (65.3-78.0)	74.1 (68.1-79.4)
CNN	81.6 (66.6-90.8)	88.9 (83.7-92.7)	87.7 (82.8-91.4)
Dermatologists with CNN	100.0 (90.8-100.0)	83.7 (77.8-88.3)	86.4 (81.3-90.3)
Management decisions			
Dermatologists alone	97.4 (86.5-99.5)	45.3 (38.3-52.4)	53.9 (47.5-60.3)
Dermatologists with CNN	100.0 (90.8-100.0)	55.8 (48.7-62.7)	63.2 (56.7-69.2)

[a] All data reported as mean (95% CI) percentages.

Figure 1. Receiver Operator Characteristic (ROC) Curves for Binary Classifications of Nevi vs Melanomas by Dermatologists Alone, Convolutional Neural Network (CNN), and Dermatologists Cooperating With CNN



The CNN used was Moleanalyzer Pro (FotoFinder Systems). Sensitivities and specificities at the a priori operation point (cutoff for malignancy, ≥0.5) are depicted as larger symbols on the corresponding ROC curves. Abbreviation: AUC, area under the curve.

obtained a mean sensitivity of 81.6% (95% CI, 66.6%-90.8%) and specificity of 88.9% (95% CI, 83.7%-92.7%). Without access to CNN results, dermatologists achieved a mean diagnostic accuracy of 74.1% (95% CI, 68.1%-79.4%), which significantly improved to 86.4% (95% CI, 81.3%-90.3%) when cooperating with the CNN (*P* < .001). The mean ROC AUC of dermatologists alone was 0.895 (95% CI, 0.836-0.954) and increased to 0.968 (95% CI, 0.948-0.988) when dermatologists integrated CNN results (*P* = .005, **Figure 1**).

Of note, the mean sensitivity of dermatologists alone and CNN was comparable (84.2% [95% CI, 69.6%-92.6%] vs 81.6% [95% CI, 66.6%-90.8%]; *P* > .99); whereas the specificity of the CNN was significantly higher compared with dermatologists (72.1% [95% CI, 65.3%-78.0%] vs 88.9% [95% CI, 83.7%-92.7%]; *P* < .001). As a result, the mean percentage of correct diagnoses (accuracy) was significantly better for the CNN compared with dermatologists (87.7% [95% CI, 82.8%-91.4%] vs 74.1% [95% CI, 68.1%-79.4%]; *P* < .001). The mean ROC AUC of the CNN (0.904 [95% CI, 0.856-0.951]) was slightly but not significantly higher compared with dermatologists (0.895 [95% CI, 0.836-0.954]; *P* = .82) (Figure 1).

Management Decisions

Besides diagnostic classifications, dermatologists' management decisions were recorded (Table 2). Here, the mean sensitivity of dermatologists' management decisions was 97.4% (95% CI, 86.5%-99.5%) and increased to 100% (95% CI, 90.8%-100.0%) when dermatologists cooperated with the CNN (*P* > .99). With access to CNN results, the dermatologists mean specificity of 45.3% (95% CI, 38.3%-52.4%) significantly improved to 55.8% (95% CI, 48.7%-62.7%; *P* < .001). Dermatologists originally recommended the excision of 104 of 190 (54.7%) benign nevi. After reviewing and integrating CNN results into decision-making, the rate of unnecessary excisions was significantly reduced by 19.2% from 104 to 84 nevi (*P* < .001; **Figure 2**A). At the same time the excision rate of malignant lesions was not significantly altered by including CNN results (*P* > .99). The percentage of nevi managed by follow-up examinations was slightly increased after receiving CNN results (from 37.9% to 44.7%), but the differences missed statistical significance (*P* = .053).

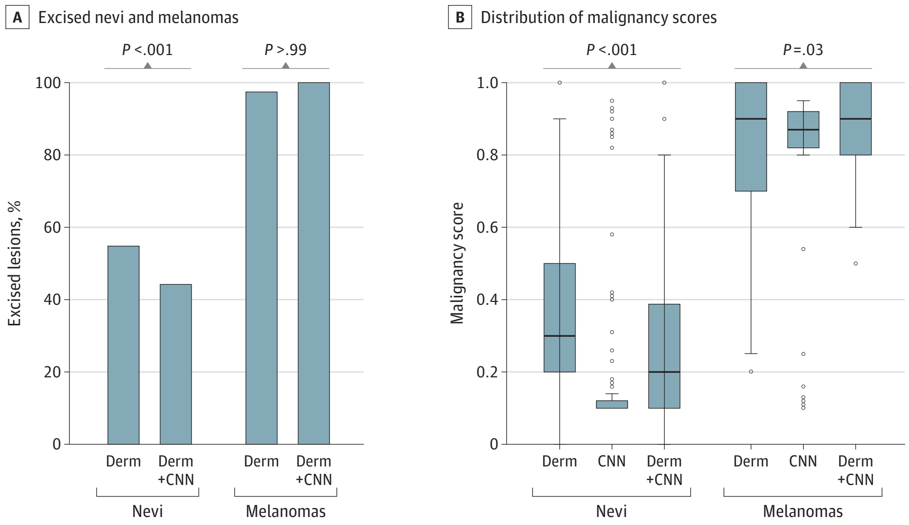
Distribution of Malignancy Scores

Malignancy scores for suspect lesions were provided by dermatologists, the CNN, and dermatologists cooperating with the CNN (Figure 2B). Mean malignancy scores of nevi were 0.35 (95% CI, 0.31-0.38) when given by dermatologists, 0.20 (95% CI, 0.16-0.23) by the CNN, and 0.27 (95% CI, 0.24-0.30) by dermatologists cooperating with the CNN. Dermatologists indicated significantly lower (and thus improved) malignancy scores for nevi after receiving CNN results (*P* < .001). Mean malignancy scores of melanomas were 0.80 (95% CI, 0.72-0.88) when given by dermatologists, 0.74 (95% CI, 0.64-0.84) by the CNN, and 0.86 (95% CI, 0.81-0.91) by dermatologists cooperating with the CNN. Again, scores given by dermatologists cooperating with the CNN were significantly higher (and thus improved) compared with those given by dermatologists alone (*P* = .03). Generally, box plots of malignancy scores by the CNN revealed smaller boxes because of less dispersion (variability) of results in comparison with dermatologists (Figure 2B).

Dermatologists' Performance Related to Their Clinical Experience

All 22 participating dermatologists indicated their level of experience with dermoscopy. Most lesions were examined

Figure 2. Bar Chart and Box Plot for Excised Nevi and Melanomas and Distribution of Malignancy Scores



A, This bar chart depicts the number of excised nevi and melanomas according to management decisions by dermatologists alone vs dermatologists cooperating with the convolutional neural network (CNN). B, This box plot shows the distribution of malignancy scores (range, 0-1) for nevi and melanomas according to dermatologists alone, CNN, and dermatologists cooperating with CNN. The upper and lower bounds of boxes represent the 25th and 75th percentiles while the medians are illustrated by the horizontal lines within the boxes. The whiskers represent the largest and smallest values that are not more than 1.5 interquartile ranges away from the box. Abbreviations: DERM, dermatologists alone; DERM+CNN, dermatologists cooperating with convolution neural network.

by dermatologists with 2 to 5 years (96, 42.1%) or less than 2 years of experience (78, 34.2%). All remaining examinations (54, 23.7%) were performed by dermatologists with more than 5 years of experience.

Of 78 lesions assessed, dermatologists with less than 2 years of experience showed a significant increase in the percentage of correct diagnoses (accuracy) from 70.5% to 87.2% after receiving CNN malignancy scores (*P* < .01) (eTable 1 in Supplement 1). Similarly, of 96 lesions assessed, the accuracy of dermatologists with 2 to 5 years of experience improved from 77.1% to 91.7% (*P* < .01). In contrast, of 54 lesions assessed, the accuracy of dermatologists with more than 5 years of experience showed only a small increase without reaching statistical significance (74.1% vs 75.9%, *P* > .99).

Dermatologists' and Patients' Perspectives Toward CNN Assistance

For each of the 228 evaluated lesions, dermatologists were asked about their personal opinion toward CNN assistance. For most lesions, dermatologists consented that CNN scores were reassuring (77.6%, 159 of 205 replies) and/or helpful (84.4%, 173 of 205 replies).

Moreover, in this study written questionnaires were collected from 152 patients measuring their acceptance and trust toward CNN support (eFigure 2 in Supplement 1). Most patients consented that the CNN might improve physicians' performance (85 [56.3%] strongly agree, 48 [31.8%] agree). The majority of patients was willing to accept longer examination times for an additional CNN-assisted diagnosis (50 [33.3%] strongly agree, 60 [40%] agree). Yet, only very few agreed that CNN classifications might completely replace physicians (8 [5.3%] strongly agree, 20 [13.2%] agree). Most patients demanded an expert physician's opinion for interpretation of CNN results (120 [79.5%] strongly agree, 28 [18.5%] agree).

Discussion

Early diagnosis of melanoma is crucial with regard to prognosis. Physicians, however, may show various levels of training and experience with a direct association with their success rates. Therefore, it has been of interest to design tools that support clinicians (1) to not miss melanoma, (2) to limit unnecessary excisions, and (3) to reduce the number of lesions that are considered diagnostically unclear and trigger time-consuming follow-up examinations. In this context, neural networks trained for skin cancer detection were shown to achieve a performance on par or even superior to dermatologists.[3-6] For various clinical tasks, studies have investigated approaches for a human with machine collaboration,[16-18] and strategies for an optimized collaboration in skin cancer screening were discussed.[11] Yet, to our knowledge, this is the first prospective clinical study under clinical settings to evaluate whether or not dermatologists benefit from integrating CNN classifications into their decision-making process.

This prospective diagnostic study was performed at a university hospital and office-based dermatology service to enable a representative spectrum of patients and lesions. Of note, lesions from special localizations (eg, acral sites, nail unit) were also included being particularly challenging for clinicians and the CNN, most probably due to a lack of sufficient numbers of training cases.[10] Expectedly, many included patients revealed an increased melanoma risk (history of previous melanoma, multiple atypical nevi), as high-risk patients often seek screening examinations at specialized centers and at regular intervals.[19,20]

Dermatologists of this study performed live examinations allowing inclusion of the clinical and dermoscopic appearance of the melanocytic lesions as well as anamnestic information and each patient's risk profile into decision-making. In contrast, CNN classifications were based solely on a single dermoscopic image. Nevertheless, in line with results of previous retrospective studies, the CNN achieved a high-level diagnostic performance showing a similar sensitivity but significantly higher specificity (88.9%) than dermatologists (72.1%).[4,5]

Irrespective of these convincing CNN results, the main outcome measures of the present study were the differences in sensitivity, specificity, accuracy, and ROC AUC of dermatologists before and after access to CNN results. To our knowledge, there have been no data showing to what extent dermatologists would apply CNN recommendations and revise their original decisions in a prospective clinical situation to date. Interestingly, within this study, all the previously mentioned main outcome measures significantly improved after dermatologists gained access to CNN results (improvements roughly between 10% to 15%). The results of this prospective study largely confirm data of retrospective studies using lesion images instead of live examinations. In a reader study, Tschandl et al[11] found that the accuracy of human raters with artificial intelligence–based multiclass probabilities increased from 63.6% to 77% in a broader spectrum of skin lesions, which is in line with the 10% to 15% increase that was found in our study. Hekler et al[12] reported tremendous improvements for physicians using CNN support (accuracy from 43% to 83%, sensitivity from 66% to 89%, specificity from 62% to 84%), yet the performance of unaided dermatologists in this study was apparently low. Moreover, in a retrospective study by Maron et al,[21] dermatologists with CNN support showed a significant increase in their mean sensitivity (from 59.4% to 74.6%) and accuracy (from 65.0% to 73.6%) at a largely unchanged specificity (70.6% vs 72.4%).

In the prospective setting of the present study, dermatologists with CNN support did not miss any melanomas, although sensitivities of the CNN as well as dermatologists alone were considerably lower. Hence, the cooperation of dermatologists with the CNN outperformed either modality on its own, which is an essential finding of this study. Obviously, whenever the CNN provided a malignancy score discordant with the dermatologist's original assessment, this provided an incentive to critically re-evaluate decisions. The overall outcome of this study is best illustrated by ROC curves graphically depicting side-by-side the performance of dermatologists, the CNN, and both working in collaboration. Here, ROC AUCs for dermatologists (0.895) and CNN (0.904) were not significantly different and close to previously published results (same CNN in melanocytic lesions: 0.86, same CNN in a broader spectrum of lesions: 0.918).[4,5] Yet, the cooperation of dermatologists with the CNN helped to significantly increase the AUC to a markedly higher level of 0.968. Not surprisingly, we found that less experienced clinicians showed the largest benefit from CNN support.[11] This observation may further boost the vision of establishing routine CNN support at less specialized institutions.

In daily clinical practice, management decisions are even more relevant than diagnostic decisions. Dermatologists without CNN support here already achieved a high sensitivity of 97.4%; however, at a low specificity of only 45.3%. This underlines that in a clinical setting, dermatologists tend to be cautious and excise more lesions to not miss melanoma. Here, cooperation with the CNN reduced unnecessary excisions of benign nevi by 19.2% and thereby significantly improved the specificity to 55.8%. These results are in line with a study by Tschandl et al,[11] who asked dermatologists to re-evaluate their face-to-face management decisions after gaining access to CNN support and found that dermatologists switched from excision to follow-up in 15.5% of benign lesions, without additionally missing excisions of malignant lesions.

For a deeper insight into dermatologists' and the CNN's assessments, we evaluated the distribution of malignancy scores. Box plots illustrated that the CNN mostly provided scores at the upper (for melanomas) or lower (for nevi) end of the scale, while dermatologists tended to provide intermediate scores close to the threshold for malignancy (0.4-0.6), particularly when not feeling confident with their diagnosis. In cases of nevi unclear to dermatologists, low CNN malignancy scores frequently had a reassuring effect to leave lesions unexcised and, in some cases rather, switch to follow-up examinations. Interestingly, it has previously been suggested that providing a CNN's level of confidence with its classification result (eg, by a risk-aware Bayesian deep learning model) could further improve human computer collaboration.[22]

Besides the main outcome measures, we also assessed the clinicians' and patients' attitudes toward a CNN-based support system. For a majority of lesions dermatologists consented that CNN support was reassuring and/or helpful. In general, these data confirm an overall optimistic attitude of dermatologists toward CNN support.[23] Similarly, results of the study's questionnaire indicated that patients were open-minded toward a CNN-based support system.[24] Nevertheless, most patients still wished an interpretation of results by an expert clinician and rejected a full replacement of clinicians by neural networks.[25]

Limitations

First, we only included melanocytic lesions to enable a closer look on the human-machine collaboration in a prospective but well-controlled setting, rather than an investigation into fully generalizable performance data. Therefore, the results of this study are not representative for settings with a larger spectrum of lesions and less frequent diagnoses. Second, the prospective nature of the study made it impossible to calculate the true sensitivity across all patients' skin lesions. Instead, our statistical considerations of sensitivity apply to those 228 lesions prospectively deemed suspect by participating dermatologists. Third, most patients of the present study showed light-colored skin (skin types 2 or 3). Hence, performance data of humans, CNN, and human-CNN collaboration may largely differ in very light (skin type 1) or darker skin types (skin types 4 or higher) warranting further investigations.[26] Finally, many patients included in the present study were at an increased risk to develop melanoma, which forbids a direct transfer of results to the general population.

Conclusions

To our knowledge, we herein present the first prospective diagnostic study investigating the collaboration of dermatologists with a market-approved CNN in a melanoma screening task. In this study, dermatologists significantly improved their diagnostic performance when cooperating with the tested CNN. These results indicate that a broader application of this human with machine approach, particularly in nonspecialized institutions, could be beneficial to clinicians and patients.

ARTICLE INFORMATION

Accepted for Publication: March 5, 2023.

Published Online: May 3, 2023. doi:10.1001/jamadermatol.2023.0905

Author Contributions: Dr Winkler and Prof Haenssle had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: Winkler, Kommoss, Enk, Haenssle.
Acquisition, analysis, or interpretation of data: Winkler, Blum, Toberer, Rosenberger, Haenssle.
Drafting of the manuscript: Winkler, Haenssle.
Critical revision of the manuscript for important intellectual content: All authors.
Statistical analysis: Winkler, Rosenberger, Haenssle.
Administrative, technical, or material support: Toberer, Haenssle.
Supervision: Enk, Haenssle.

Conflict of Interest Disclosures: Dr Winkler reported personal fees from FotoFinder Systems, Amgen, Bristol Myers Squibb, MSD, Philochem, and Roche outside the submitted work. Prof Blum reported personal fees from HEINE Optotechnik and FotoFinder Systems. Prof Haenssle reported personal fees from SciBase, FotoFinder Systems, HEINE Optotechnik, and Magnosco outside the submitted work. No other disclosures were reported.

Additional Contributions: We would like to thank and acknowledge all dermatologists, who were not compensated, including Ines Bertlich, Olivia Bochnig, Therezia Bokor-Billmann, Jana Burghaus-Chang, Christiane Cussigh, Carmen Dietrich, Lara-Elena Hakim-Meibodi, Susanne Hanner, Julia Hartmann, Katharina Kälber, Aric Keller, Kai Linse, Pawel Majenka, Silvia Mihalceanu, Christina Naoum, Alexander Wald, Priscila Wölbing, and others who remain anonymous for actively and voluntarily spending time to recruit patients for this study.

Data Sharing Statement: See Supplement 2.

REFERENCES

1. Arnold M, Singh D, Laversanne M, et al. Global burden of cutaneous melanoma in 2020 and projections to 2040. *JAMA Dermatol*. 2022;158(5):495-503. doi:10.1001/jamadermatol.2022.0160

2. Barreiro-Capurro A, Andrés-Lencina JJ, Podlipnik S, et al. Differences in cutaneous melanoma survival between the 7th and 8th edition of the American Joint Committee on Cancer (AJCC): a multicentric population-based study. *Eur J Cancer*. 2021;145:29-37. doi:10.1016/j.ejca.2020.11.036

3. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. doi:10.1038/nature21056

4. 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(8):1836-1842. doi:10.1093/annonc/mdy166

5. Haenssle HA, Fink C, Toberer F, et al. 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. *Ann Oncol*. 2020;31(1):137-143. doi:10.1016/j.annonc.2019.10.013

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(7):938-947. doi:10.1016/S1470-2045(19)30333-X

7. Haggenmüller S, Maron RC, Hekler A, et al. Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. *Eur J Cancer*. 2021;156:202-216. doi:10.1016/j.ejca.2021.06.049

8. Winkler JK, Sies K, Fink C, et al. Association between different scale bars in dermoscopic images and diagnostic performance of a market-approved deep learning convolutional neural network for melanoma recognition. *Eur J Cancer*. 2021;145:146-154. doi:10.1016/j.ejca.2020.12.010

9. Winkler JK, Fink C, Toberer F, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol*. 2019;155(10):1135-1141. doi:10.1001/jamadermatol.2019.1735

10. Winkler JK, Sies K, Fink C, et al. Melanoma recognition by a deep learning convolutional neural network—performance in different melanoma subtypes and localisations. *Eur J Cancer*. 2020;127:21-29. doi:10.1016/j.ejca.2019.11.020

11. Tschandl P, Rinner C, Apalla Z, et al. Human-computer collaboration for skin cancer recognition. *Nat Med*. 2020;26(8):1229-1234. doi:10.1038/s41591-020-0942-0

12. Hekler A, Utikal JS, Enk AH, et al; Collaborators. Superior skin cancer classification by the combination of human and artificial intelligence. *Eur J Cancer*. 2019;120:114-121. doi:10.1016/j.ejca.2019.07.019

13. Montague E. Validation of a trust in medical technology instrument. *Appl Ergon*. 2010;41(6):812-821. doi:10.1016/j.apergo.2010.01.009

14. Fink C, Uhlmann L, Hofmann M, et al. Patient acceptance and trust in automated computer-assisted diagnosis of melanoma with dermatofluoroscopy. *J Dtsch Dermatol Ges*. 2018;16(7):854-859. doi:10.1111/ddg.13562

15. Vergara IA, Norambuena T, Ferrada E, Slater AW, Melo F. StAR: a simple tool for the statistical comparison of ROC curves. *BMC Bioinformatics*. 2008;9(1):265. doi:10.1186/1471-2105-9-265

16. Garg AX, Adhikari NK, McDonald H, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA*. 2005;293(10):1223-1238. doi:10.1001/jama.293.10.1223

17. Codella NC, Lin C-C, Halpern A, Hind M, Feris R, Smith JR. Collaborative Human-AI (CHAI): evidence-based interpretable melanoma classification in dermoscopic images. In: Stoyanov D, Taylor Z, Kia SM, et al, eds. *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Springer; 2018:97-105.

18. Bien N, Rajpurkar P, Ball RL, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS Med*. 2018;15(11):e1002699. doi:10.1371/journal.pmed.1002699

19. Winkler JK, Tschandl P, Toberer F, et al. Monitoring patients at risk for melanoma: may convolutional neural networks replace the strategy of sequential digital dermoscopy? *Eur J Cancer*. 2022;160:180-188. doi:10.1016/j.ejca.2021.10.030

20. Tschandl P, Hofmann L, Fink C, Kittler H, Haenssle HA. Melanomas vs nevi in high-risk patients under long-term monitoring with digital dermatoscopy: do melanomas and nevi already differ at baseline? *J Eur Acad Dermatol Venereol*. 2017;31(6):972-977. doi:10.1111/jdv.14065

21. Maron RC, Utikal JS, Hekler A, et al. Artificial intelligence and its effect on dermatologists' accuracy in dermoscopic melanoma image classification: web-based survey study. *J Med Internet Res*. 2020;22(9):e18091. doi:10.2196/18091

22. Mobiny A, Singh A, Van Nguyen H. Risk-aware machine learning classifier for skin lesion diagnosis. *J Clin Med*. 2019;8(8):1241. doi:10.3390/jcm8081241

23. Polesie S, Gillstedt M, Kittler H, et al. Attitudes towards artificial intelligence within dermatology: an international online survey. *Br J Dermatol*. 2020;183(1):159-161. doi:10.1111/bjd.18875

24. Jutzi TB, Krieghoff-Henning EI, Holland-Letz T, et al. Artificial intelligence in skin cancer diagnostics: the patients' perspective. *Front Med (Lausanne)*. 2020;7:233. doi:10.3389/fmed.2020.00233

25. Nelson CA, Pérez-Chada LM, Creadore A, et al. Patient perspectives on the use of artificial intelligence for skin cancer screening: a qualitative study. *JAMA Dermatol*. 2020;156(5):501-512. doi:10.1001/jamadermatol.2019.5014

26. Goyal M, Knackstedt T, Yan S, Hassanpour S. Artificial intelligence-based image classification methods for diagnosis of skin cancer: challenges and opportunities. *Comput Biol Med*. 2020;127:104065. doi:10.1016/j.compbiomed.2020.104065