ESMO GOOD SCIENCE BETTER MEDICINE BEST PRACTICE

Annals of Oncology 0: 1–7, 2018
doi:10.1093/annonc/mdy166

ORIGINAL ARTICLE

Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists

H. A. Haenssle[1*,†], C. Fink[1†], R. Schneiderbauer[1], F. Toberer[1], T. Buhl[2], A. Blum[3], A. Kalloo[4], A. Ben Hadj Hassen[5], L. Thomas[6], A. Enk[1] & L. Uhlmann[7]

[1]Department of Dermatology, University of Heidelberg, Heidelberg; [2]Department of Dermatology, University of Göttingen, Göttingen; [3]Office Based Clinic of Dermatology, Konstanz, Germany; [4]Dermatology Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, USA; [5]Faculty of Computer Science and Mathematics, University of Passau, Passau, Germany; [6]Department of Dermatology, Lyons Cancer Research Center, Lyon 1 University, Lyon, France; [7]Institute of Medical Biometry and Informatics, University of Heidelberg, Heidelberg, Germany

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

†Both authors contributed equally as co-first authors.

Background: Deep learning convolutional neural networks (CNN) may facilitate melanoma detection, but data comparing a CNN's diagnostic performance to larger groups of dermatologists are lacking.

Methods: Google's Inception v4 CNN architecture was trained and validated using dermoscopic images and corresponding diagnoses. In a comparative cross-sectional reader study a 100-image test-set was used (level-I: dermoscopy only; level-II: dermoscopy plus clinical information and images). Main outcome measures were sensitivity, specificity and area under the curve (AUC) of receiver operating characteristics (ROC) for diagnostic classification (dichotomous) of lesions by the CNN versus an international group of 58 dermatologists during level-I or -II of the reader study. Secondary end points included the dermatologists' diagnostic performance in their management decisions and differences in the diagnostic performance of dermatologists during level-I and -II of the reader study. Additionally, the CNN's performance was compared with the top-five algorithms of the 2016 International Symposium on Biomedical Imaging (ISBI) challenge.

Results: In level-I dermatologists achieved a mean (±standard deviation) sensitivity and specificity for lesion classification of 86.6% (±9.3%) and 71.3% (±11.2%), respectively. More clinical information (level-II) improved the sensitivity to 88.9% (±9.6%, $P = 0.19$) and specificity to 75.7% (±11.7%, $P < 0.05$). The CNN ROC curve revealed a higher specificity of 82.5% when compared with dermatologists in level-I (71.3%, $P < 0.01$) and level-II (75.7%, $P < 0.01$) at their sensitivities of 86.6% and 88.9%, respectively. The CNN ROC AUC was greater than the mean ROC area of dermatologists (0.86 versus 0.79, $P < 0.01$). The CNN scored results close to the top three algorithms of the ISBI 2016 challenge.

Conclusions: For the first time we compared a CNN's diagnostic performance with a large international group of 58 dermatologists, including 30 experts. Most dermatologists were outperformed by the CNN. Irrespective of any physicians' experience, they may benefit from assistance by a CNN's image classification.

Clinical trial number: This study was registered at the German Clinical Trial Register (DRKS-Study-ID: DRKS00013570; https://www.drks.de/drks_web/).

Key words: melanoma, melanocytic nevi, dermoscopy, deep learning convolutional neural network, computer algorithm, automated melanoma detection

Introduction

Over the past few decades, melanoma has emerged as a major challenge in public health [1]. The continuous increase in incidence rates and melanoma mortality have fueled a heightened commitment to early detection and prevention [2]. Several meta-analyses have shown that dermoscopy significantly improves the diagnostic accuracy of the naked eye examination [3–5]. However, dermatologists and medical practitioners formally trained in different dermoscopic algorithms showed an average sensitivity for detecting melanoma of mostly <80% [6, 7]. In recent years, several strategies of automated computer image analysis have been investigated as an aide for physicians to provide a high and widely reproducible diagnostic accuracy for melanoma screening [8–11]. These approaches were limited by using 'man-made' dermoscopic segmentation criteria for the diagnosis of melanoma (e.g. multiple colors, certain morphological structures as streaks/pseudopods, irregular vascular structures) [12]. As a landmark publication, Esteva et al. reported on the training and testing of a deep learning convolutional neural network (CNN) for imaged-based classification in 2017 [13]. In this setting the CNN was not restricted by man-made segmentation criteria, but deconstructed digital images down to the pixel level and eventually created its own diagnostic clues. As in the study reported herein, the authors utilized a pretrained GoogleNet Inception CNN architecture [14] additionally trained with more than 100 000 digital images and corresponding disease labels.

The aim of the present study was to train, validate, and test a deep learning CNN for the diagnostic classification of dermoscopic images of lesions of melanocytic origin (melanoma, benign nevi) and to compare the results to a large group of 58 dermatologists.

Methods

The study was approved by the local ethics committee and carried out in accordance with the Declaration of Helsinki principles.

Details on methods pertaining to the CNN architecture and CNN training are found in supplementary Methods, available at *Annals of Oncology* online.

We used and specifically trained a modified version of Google's Inception v4 CNN architecture (supplementary Figure S1, available at *Annals of Oncology* online) [14].

Test-set-300

We created a 300-image test-set including 20% melanomas (*in situ* and invasive) of all body sites and of all frequent histotypes, and 80% benign melanocytic nevi of different subtypes and body sites including the so-called 'melanoma simulators' (supplementary Table S1, available at *Annals of Oncology* online). As almost two-third of benign nevi were non-excised lesions validated by follow-up examinations, this dataset represented a spectrum of melanocytic lesions as typically encountered in daily clinical routine. Images of the test-set-300 were retrieved from the high-quality validated image library of the Department of Dermatology, University of Heidelberg, Germany. Various camera/dermoscope combinations were used for image acquisition. No overlap between datasets for training/validation and testing was allowed.

Test-set-100 and reader study level-I and -II

Before CNN testing two experienced dermatologists prospectively selected 100 images of set-300 for an increased diagnostic difficulty (supplementary Table S2, available at *Annals of Oncology* online). Set-100 was used for CNN testing in comparison to dermatologists in a global reader study. Readers ($n = 172$) were invited via mailing lists of the International Dermoscopy Society, and 58 (33.7%) returned their completed voting sheets. Participants indicated their level of experience in dermoscopy ('Beginner' <2 years of experience, 'Skilled' 2–5 years of experience, 'Expert' ≥ 5 years of experience).

In level-I of the reader study, dermatologists were presented solely the dermoscopic image and asked to indicate their dichotomous diagnosis (melanoma, benign nevus) and their management decision (excision, short-term follow-up, send away/no action needed). After an interval of 4 weeks, the same participants indicated their diagnosis and management decision in level-II of the reader study, which included dermoscopic images supplemented by additional clinical information and close-up images of the same 100 cases.

International Symposium on Biomedical Imaging challenge dataset

We used another 100-image dataset created by the International Skin Imaging Collaboration (ISIC) melanoma project for the occasion of the 2016 International Symposium on Biomedical Imaging (ISBI) challenge. This dataset enabled the direct comparison of our CNN to the internationally top-five ranked algorithms [15].

Statistical analysis

The primary outcome measures were sensitivity, specificity, and area under the curve (AUC) of receiver operating characteristics (ROC) for the diagnostic classification (dichotomous) of lesions by the CNN versus dermatologists during level-I or -II of the reader study. Secondary end points included the assessment of the dermatologists' diagnostic performance in their management decisions and the differences in the diagnostic performance of dermatologists between level-I and II of the reader study. For management decisions the option of a 'short-term follow-up' was positively accounted for both sensitivity and specificity calculations. The mean number (percentage) of all lesions and all melanomas indicated for follow-up, the benign nevus excision rate (number of excised nevi/number of all nevi), and the number needed to excise (NNE; number of excised lesions/number of excised melanomas) were calculated.

The CNN put out a 'malignancy score' ranging from 0 to 1 with a cut-off of > 0.5 for the dichotomous classification of malignant versus benign lesions. For comparison of the CNN to dermatologists a two-sided, one-sample *t*-test was applied and the specificity at the level of the average dermatologist sensitivity and the ROC AUC of the CNN versus the mean ROC area of dermatologists was calculated. For dermatologists' dichotomous predictions, area under ROC curves is equivalent to the average of sensitivity and specificity. Descriptive statistics as frequency, mean, range, and standard deviation were used. Two-sided *t*-tests were used to assess differences in the dermatologists' diagnostic performance between level-I and II of the reader study. Results were considered statistically significant at the $P < 0.05$ level. All analyses were carried out using SPSS Version 24 (IBM, SPSS; Chicago, IL).

Results

Dermatologists' diagnostic accuracy

Seventeen (29.3%) out of the 58 participating dermatologists from 17 countries indicated being a 'beginner' in dermoscopy (< 2 years of experience) while 11 (19%) and 30 (51.7%) declared to be 'skilled' (2–5 years of experience) or an 'expert' (> 5 years of

Table 1. Results of reader study level-I and -II

	Classification			Management decision		
Dermatologists	**Sensitivity (%)**	**Specificity (%)**	**ROC area**	**Sensitivity (%)**	**Specificity (%)**	**ROC area**
Level-I						
All (n=58)	86.6	71.3	0.79	98.8	64.6	0.82
'Expert' (n=30)	89.0	74.5	0.82	98.8	68.1	0.83
'Skilled' (n=11)	85.9	68.5	0.77	98.6	61.6	0.80
'Beginner' (n=17)	82.9	67.6	0.75	98.8	60.7	0.80
Level-II						
All (n=58)	88.9	75.7	0.82	98.6	66.7	0.83
'Expert' (n=30)	89.5	77.7	0.84	99.1	69.0	0.84
'Skilled' (n=11)	90.9	77.2	0.84	98.2	68.4	0.83
'Beginner' (n=17)	86.6	71.2	0.79	98.1	61.3	0.80

ROC, receiver operating characteristic.
Level-I, readers were solely provided with dermoscopic images.
Level-II, readers were additionally provided with clinical information close-up images.
'Expert', the reader indicated to have >5 years of experience in dermoscopy.
'Skilled', the reader indicated to have 2–5 years of experience in dermoscopy.
'Beginner', the reader indicated to have <2 years of experience in dermoscopy.

experience), respectively. Due to reasons of feasibility dermatologists were asked to read only test-set-100.

Diagnostic classification in reader study level-I (dermoscopy only). The mean [±standard deviation (SD)] sensitivity and specificity of the 58 dermatologists for the dichotomous classification of set-100 lesions during study level-I was 86.6% (±9.3%) and 71.3% (±11.2%), respectively (Table 1). This translated into an average (±SD) ROC area of 0.79 (±0.06). Experts in dermoscopy showed a significantly higher mean sensitivity, specificity, and ROC area than beginners [89% (±9.2%), 74.5% (±12.6%), 0.82 (±0.06) versus 82.9% (±7.1%), 67.6% (±6.3%), 0.75 (±0.04), respectively; all $P < 0.02$; Table 1].

Management decisions in reader study level-I (dermoscopy only). Participants were offered (i) excision, (ii) short-term follow-up, or (iii) send away/no action needed as management decisions. In this setting, the average (±SD) sensitivity and ROC area significantly increased to 98.8% (±2.9%, $P < 0.01$) and 0.82 (±0.07, $P = 0.03$), respectively (Table 1). In contrast, the specificity significantly decreased from 71.3% to 64.6% (±13.6%, $P < 0.01$). Similar changes were observed across all levels of experience. Among all dermatologists the average (±SD) benign nevus excision rate was 35.4% (±13.6%) and the lesion follow-up rate was 33.5% (±11.7%). Dermatologists included an average number (±SD) of 1.9 (±1.6) melanomas in follow-up and attained a NNE of 2.3 (±0.6). Higher experience was associated with a significant reduction of the benign nevus excision rate, the lesion follow-up rate, and the number of melanomas under follow-up (all $P < 0.05$). The NNE also slightly improved with experience, however, without reaching statistical significance.

Diagnostic classification in reader study level-II (dermoscopy and clinical information). The addition of clinical information (age, sex, and body site) and close-up images improved the dermatologists' mean (±SD) sensitivity, specificity, and ROC area to 88.9% (±9.6%, $P = 0.19$), 75.7% (±11.7%, $P < 0.05$), and 0.82 (±0.06, $P < 0.01$), respectively (Table 1). These changes were solely based on significant improvements of 'beginners' and 'skilled' dermatologists, while 'experts' in dermoscopy showed no relevant benefit from supplemented clinical information and images.

Management decisions in reader study level-II (dermoscopy and clinical information). When asked for their management decisions during level-II of the study, dermatologists improved their level-II results of the dichotomous classification to a mean (±SD) sensitivity, specificity, and ROC area of 98.6% (±2.8%, $P < 0.01$), 66.7% (±12.4%, $P < 0.01$), and 0.83 (±0.06, $P = 0.76$) (Table 1). However, we found no significant differences between these results and management decision of study level-I. The average (±SD) number of melanomas included into short-term follow-up dropped from 1.9 (±1.6) to 1.3 (±1.5) melanomas ($P = 0.03$) and the NNE remained unchanged at 2.3 benign nevi excised for the detection of one melanoma. For management decisions in study level-II a higher level of experience ('experts' versus 'beginners') was associated with a significantly better mean (±SD) ROC area [0.84 (±0.06) versus 0.79 (±0.06), $P = 0.03$], whereas other parameters of management decisions in study level-II showed no significant differences in relation to the level of experience.

CNN's diagnostic accuracy

Boxplots in Figure 1 show the distribution of melanoma probability scores for benign nevi, *in situ* melanomas, and invasive melanomas. When the aforementioned settings were applied to

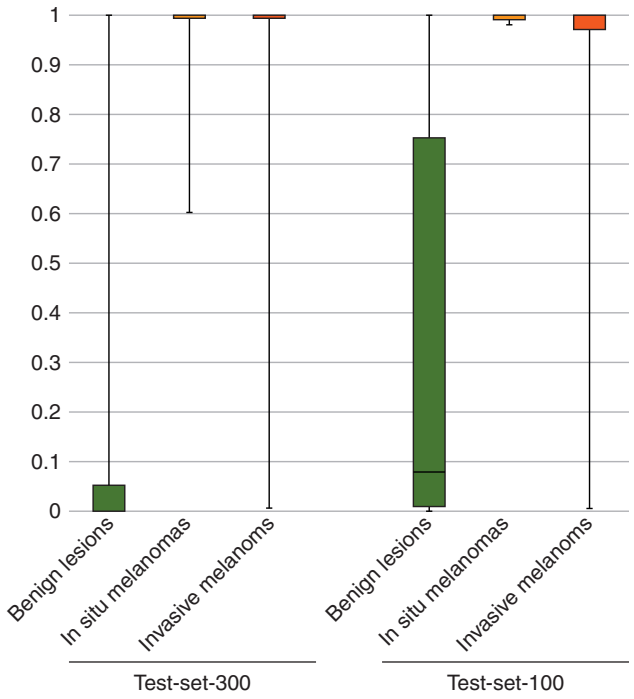

Figure 1. The CNN's melanoma probability scores (range 0–1) for benign nevi (green online) in comparison to *in situ* (orange online) or invasive melanomas (red online) are depicted as boxplots for test-set-300 and test-set-100. Scores closer to 1 indicated a higher probability of melanoma. The upper 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. Statistical analyses revealed significantly different melanoma probability scores when comparing benign lesions to *in situ* or invasive melanomas ($P < 0.001$). However, melanoma probability scores for *in situ* and invasive melanomas showed no significant differences (set-300 $P = 0.84$, set-100 $P = 0.24$).

test-set-100, the sensitivity, specificity, and ROC AUC were 95%, 63.8%, and 0.86, respectively. For the larger test-set-300 including less difficult-to-diagnose lesions the sensitivity, specificity, and ROC AUC were 95%, 80%, and 0.95, respectively. Both ROC curves are depicted in Figure 2A and B.

Diagnostic accuracy of CNN versus dermatologists

We used the dermatologists' mean sensitivity of 86.6% for the diagnostic classification in study level-I as the benchmark for comparison to the CNN (Figure 2A). At this sensitivity the CNN's specificity was higher (82.5%) than the mean specificity of dermatologists (71.3%, $P < 0.01$). Moreover, in level-I the CNN ROC AUC (0.86) was greater than the mean ROC area of dermatologists (0.79, $P < 0.01$).

When dermatologists received more clinical information and images (study level-II) their diagnostic performance improved. Using the dermatologists' level-II mean sensitivity of 88.9% as the operating point on the CNN ROC curve, the CNN specificity was 82.5%, which was significantly higher than the dermatologists' mean specificity of 75.7% ($P < 0.01$). Again, the CNN ROC AUC (0.86) was greater than the mean ROC area of dermatologists (0.82, $P < 0.01$).

CNN comparison to top-five algorithms of ISBI challenge

The head-to-head comparison of ROC curves of our CNN to the international top-five ranked individual algorithms of the ISBI 2016 challenge [15] is shown in Figure 3. With an ROC AUC of 0.79 the CNN presented herein was among the three top algorithms of the ISBI 2016 challenge with almost overlaying ROC curves.

Discussion

Melanoma incidence rates are rising steadily in most fair-skinned populations and were predicted to further increase [2]. Notwithstanding the different levels of training and experience of physicians engaged in early melanoma detection, a reproducible high diagnostic accuracy would be desirable. To this end, we trained and tested a convolutional deep learning CNN for differentiating dermoscopic images of melanoma and benign nevi. For the first time we compared the diagnostic performance of a CNN with a large international group of 58 dermatologists from 17 countries, including 30 experts with more than 5 years of dermoscopic experience. When dermatologists were provided with dermoscopic images only (study level-I) their dichotomous classification of lesions was significantly outperformed by the CNN. However, in a real-life clinical setting dermatologists will incorporate more clinical information into decision-making. Therefore, we investigated the effect of additional clinical information and close-up images and found a much-improved diagnostic performance of dermatologists (study level-II). However, at their improved mean sensitivity (88.9%) dermatologists still showed a specificity inferior to the CNN (75.7% versus 82.5%, $P < 0.01$). Our data clearly show that a CNN algorithm may be a suitable tool to aid physicians in melanoma detection irrespective of their individual level of experience and training. Of note, in study level-I thirteen (22.4%) of 58 dermatologists showed a slightly higher diagnostic performance than the CNN.

We deliberately chose the dermatologists' dichotomous classification of lesions in set-100 as the primary outcome measure for comparison to the CNN. However, it may be argued that 'management decisions' rather than 'diagnostic classifications' represent more the dermatologists' everyday task in skin cancer screenings. Besides 'excision' and 'send away/no action needed' management decisions implied a 'third way', namely the option of a short-term follow-up examination, which was introduced and validated for single lesions with a higher grade of atypia (e.g. variegated tonalities of color, asymmetry in shape, or prominent network) that do not warrant immediate excision for a suspicion of melanoma [16]. The statistical assessment of the follow-up option introduces some difficulties. On the one hand short-term follow-up was shown to be an effective measure to differentiate early melanomas from benign nevi by unmasking dynamic changes [17–19], on the other hand excessive use of the follow-up 'wild-card' (i) may be used to conceal a lack of dermoscopic expertise, (ii) may be largely impracticable in daily clinical routine, and (iii) may delay melanoma excision. Therefore, we positively included the choice to follow-up a lesion into sensitivity (melanomas under follow-up: 'true positives') and specificity calculations (nevi under follow-up: 'true negatives'). However, we also measured details about the use of the

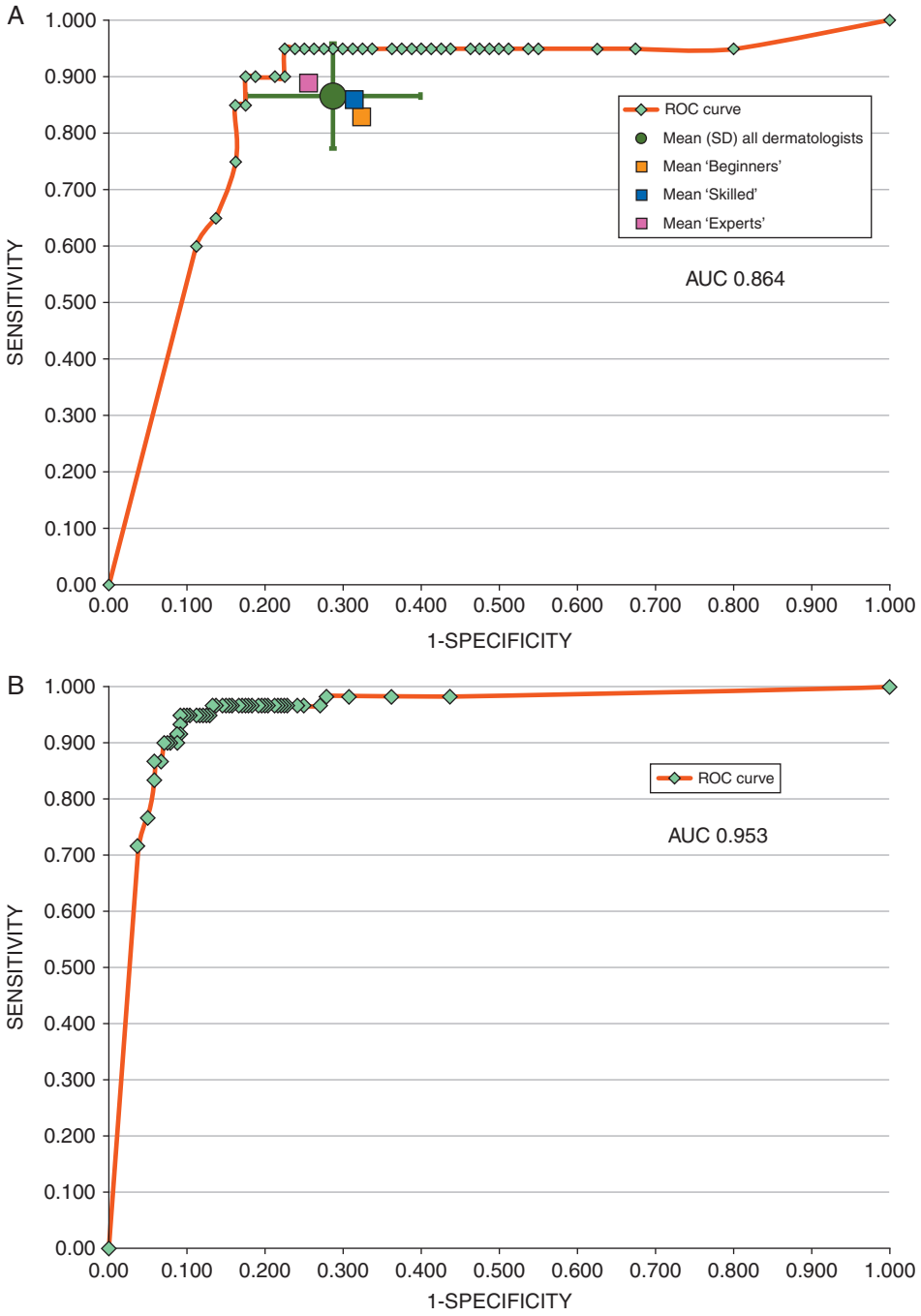

Figure 2. (A) ROC curve of the CNN in relation to the average (±SD) sensitivity and specificity of all dermatologists [mean: green (online) circle; ±SD: green (online) error bars] in set-100 (dichotomous classification, study level-I) and the dermatologists' mean sensitivity and specificity in relation to their level of experience. (B) ROC curve of the CNN in set-300.

follow-up option and found that dermatologists selected approximately one-third of lesions for follow-up, while the mean absolute number of melanomas under follow-up was in the range of 1.3–1.9. As expected, a higher level of experience and more clinical information were associated with reduced follow-up rates.

Important to mention, that differences in the level of difficulty inherent to any image test-set will directly impact the diagnostic performance of algorithms and physicians. In order to generate comparability of different computer algorithms it is therefore of utmost importance to include a large group of dermatologists with various levels of experience as well as to create and use open source datasets as provided by the ISIC [15]. In contrast to Marchetti et al. [15] other authors have not used 'benchmark' image datasets, and only a few studies included a small number of readers for comparison with their designed computer algorithms [13, 20]. Moreover, wherever possible datasets should include lesions of different anatomical sites and histotypes. As shown in supplementary Tables S1 and S2, available at *Annals of Oncology* online, both set-100 and set-300 met these requirements in order to create a less artificial study setting.

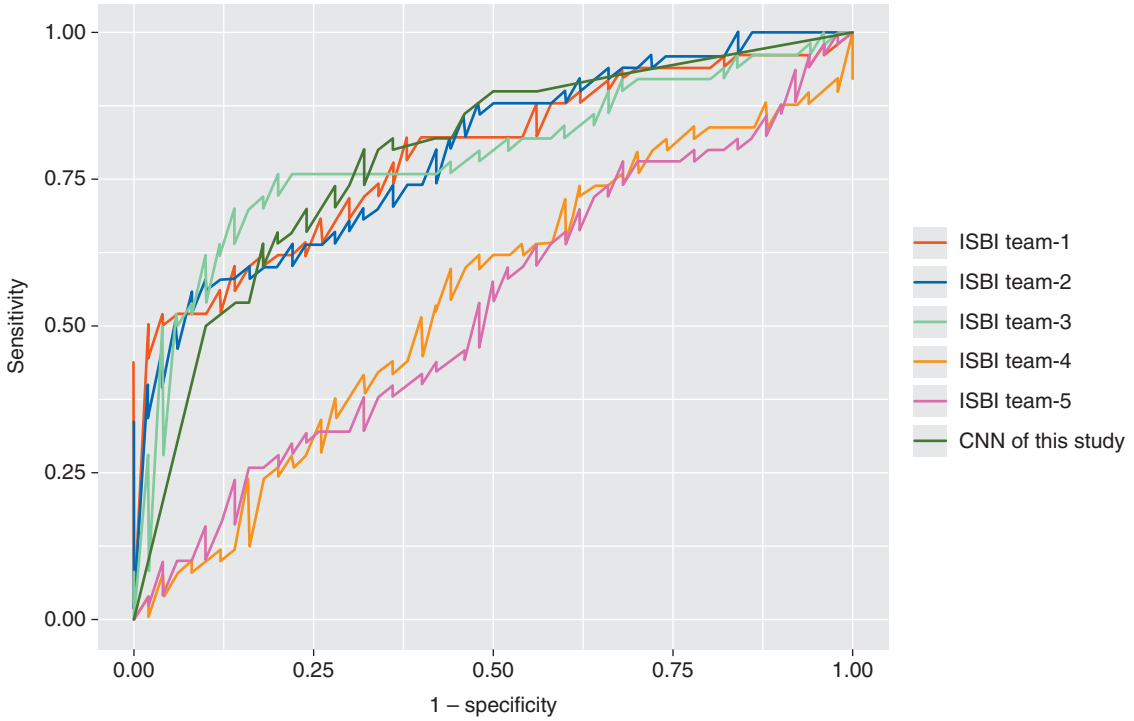

Figure 3. Comparison of ROC curves of the CNN described in this study (dark green (online) line) to the top-five ranked individual algorithms of the 2016 International Symposium on Biomedical Imaging (ISBI) challenge [18]. ROC AUCs in descending order were as follows: ISBI team-2: 0.7956; ISBI team-1: 0.7928; ISBI team-3: 0.7892; CNN of this study: 0.7868; ISBI team-4: 0.5460; ISBI team-5: 0.5324.

Our study shows a number of limitations that may impede a broader generalization. First, as for all reader studies, the setting for testing the dermatologists' diagnostic performance was artificial as they did not need to fear the harsh consequences of missing melanoma. Second, the test-sets of our study did not display the full range of lesions (e.g. pigmented basal cell carcinoma or seborrheic keratosis). Third, the poor availability of validated images led to a shortage of melanocytic lesions from other skin types and genetic backgrounds. Fourth, as shown in earlier reader studies, operating physicians may not follow the recommendations of a CNN they not fully trust, which may diminish the reported diagnostic performance [21]. Besides confirmation of our results with the help of larger and more diverse test-sets, prospective studies are needed that also address the acceptance of patients and physicians involved with screening for skin cancer.

In conclusion, the results of our study demonstrate that an adequately trained deep learning CNN is capable of a highly accurate diagnostic classification of dermoscopic images of melanocytic origin. In conjunction with results from the reader study level-I and -II we could show, that the CNN's diagnostic performance was superior to most but not all dermatologists. While a CNN's architecture is difficult to set up and train, its implementation on digital dermoscopy systems or smart phone applications may easily be deployed. Therefore, physicians of all different levels of training and experience may benefit from assistance by a CNN's image classification.

Acknowledgements

We would like to thank and acknowledge the following dermatologists in alphabetical order, who actively and voluntarily spend much time to participate in the reader study level-I and reader study level-II: Christina Alt, Monika Arenbergerova, Renato Bakos, Anne Baltzer, Ines Bertlich, Andreas Blum, Therezia Bokor-Billmann, Jonathan Bowling, Naira Braghiroli, Ralph Braun, Kristina Buder-Bakhaya, Timo Buhl, Horacio Cabo, Leo Cabrijan, Naciye Cevic, Anna Classen, David Deltgen, Christine Fink, Ivelina Georgieva, Lara-Elena Hakim-Meibodi, Susanne Hanner, Franziska Hartmann, Julia Hartmann, Georg Haus, Elti Hoxha, Raimonds Karls, Hiroshi Koga, Jürgen Kreusch, Aimilios Lallas, Pawel Majenka, Ash Marghoob, Cesare Massone, Lali Mekokishvili, Dominik Mestel, Volker Meyer, Anna Neuberger, Kari Nielsen, Margaret Oliviero, Riccardo Pampena, John Paoli, Erika Pawlik, Barbar Rao, Adriana Rendon, Teresa Russo, Ahmed Sadek, Kinga Samhaber, Roland Schneiderbauer, Anissa Schweizer, Ferdinand Toberer, Lukas Trennheuser, Lyobomira Vlahova, Alexander Wald, Julia Winkler, Priscila Wölbing, Iris Zalaudek. Some participants asked to remain anonymous and we also thank these colleagues for their commitment. Moreover, we thank the International Dermoscopy Society (IDS) for providing the mailing list that enabled the invitation of dermatologists to participate in the study.

Ethical approval

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

Funding

This research received no specific grant from any public, commercial or not-for-profit sector.

Disclosure

HAH received honoraria and/or travel expenses from companies involved in the development of devices for skin cancer screening: Scibase AB, FotoFinder Systems GmbH, Heine Optotechnik GmbH, Magnosco GmbH. CF received travel expenses from Magnosco GmbH. The remaining authors declared no conflicts of interest.

References

1. Koh HK. Melanoma screening: focusing the public health journey. Arch Dermatol 2007; 143(1): 101–103.
2. Nikolaou V, Stratigos AJ. Emerging trends in the epidemiology of melanoma. Br J Dermatol 2014; 170(1): 11–19.
3. Vestergaard ME, Macaskill P, Holt PE, Menzies SW. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. Br J Dermatol 2008; 159: 669–676.
4. Bafounta ML, Beauchet A, Aegerter P, Saiag P. Is dermoscopy (epiluminescence microscopy) useful for the diagnosis of melanoma? Results of a meta-analysis using techniques adapted to the evaluation of diagnostic tests. Arch Dermatol 2001; 137(10): 1343–1350.
5. Salerni G, Teran T, Puig S et al. Meta-analysis of digital dermoscopy follow-up of melanocytic skin lesions: a study on behalf of the International Dermoscopy Society. J Eur Acad Dermatol Venereol 2013; 27(7): 805–814.
6. Dolianitis C, Kelly J, Wolfe R, Simpson P. Comparative performance of 4 dermoscopic algorithms by nonexperts for the diagnosis of melanocytic lesions. Arch Dermatol 2005; 141(8): 1008–1014.
7. Carli P, Quercioli E, Sestini S et al. Pattern analysis, not simplified algorithms, is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology. Br J Dermatol 2003; 148(5): 981–984.
8. Barata C, Celebi ME, Marques JS. Improving dermoscopy image classification using color constancy. IEEE J Biomed Health Inform 2015; 19: 1–52.
9. Glaister J, Wong A, Clausi DA. Segmentation of skin lesions from digital images using joint statistical texture distinctiveness. IEEE Trans Biomed Eng 2014; 61(4): 1220–1230.
10. Garnavi R, Aldeen M, Bailey J. Computer-aided diagnosis of melanoma using border and wavelet-based texture analysis. IEEE Trans Inform Technol Biomed 2012; 16(6): 1239–1252.
11. Kaya S, Bayraktar M, Kockara S et al. Abrupt skin lesion border cutoff measurement for malignancy detection in dermoscopy images. BMC Bioinformatics 2016; 17(S13): 367.
12. Pehamberger H, Steiner A, Wolff K. In vivo epiluminescence microscopy of pigmented skin lesions. I. Pattern analysis of pigmented skin lesions. J Am Acad Dermatol 1987; 17(4): 571–583.
13. Esteva A, Kuprel B, Novoa RA et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542(7639): 115–118.
14. Szegedy C, Vanhoucke V, Ioffe S et al. Rethinking the inception architecture for computer vision 2015. https://arxiv.org/abs/1512.00567 (9 May 2018, date last accessed).
15. Marchetti MA, Codella NCF, Dusza SW et al. Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. J Am Acad Dermatol 2018; 78(2): 270–277.
16. Menzies SW, Gutenev A, Avramidis M et al. Short-term digital surface microscopic monitoring of atypical or changing melanocytic lesions. Arch Dermatol 2001; 137(12): 1583–1589.
17. Altamura D, Avramidis M, Menzies SW. Assessment of the optimal interval for and sensitivity of short-term sequential digital dermoscopy monitoring for the diagnosis of melanoma. Arch Dermatol 2008; 144(4): 502–506.
18. Menzies SW, Emery J, Staples M et al. Impact of dermoscopy and short-term sequential digital dermoscopy imaging for the management of pigmented lesions in primary care: a sequential intervention trial. Br J Dermatol 2009; 161(6): 1270–1277.
19. Menzies SW, Stevenson ML, Altamura D, Byth K. Variables predicting change in benign melanocytic nevi undergoing short-term dermoscopic imaging. Arch Dermatol 2011; 147(6): 655–659.
20. Ferris LK, Harkes JA, Gilbert B et al. Computer-aided classification of melanocytic lesions using dermoscopic images. J Am Acad Dermatol 2015; 73(5): 769–776.
21. Hauschild A, Chen SC, Weichenthal M et al. To excise or not: impact of MelaFind on German dermatologists' decisions to biopsy atypical lesions. J Dtsch Dermatol Ges 2014; 12(7): 606–614.