European Journal of Cancer 144 (2021) 192–199

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

Skin lesions of face and scalp – Classification by a market-approved convolutional neural network in comparison with 64 dermatologists

Holger Andreas Haenssle [a,*], Julia Katharina Winkler [a], Christine Fink [a], Ferdinand Toberer [a], Alexander Enk [a], Wilhelm Stolz [b], Teresa Deinlein [c], Rainer Hofmann-Wellenhof [c], Harald Kittler [d], Philipp Tschandl [d], Cliff Rosendahl [e], Aimilios Lallas [f], Andreas Blum [g], Mohamed Souhayel Abassi [h], Luc Thomas [i], Isabelle Tromme [j], Albert Rosenberger [k], Reader study level-I and level-II Groups Christina Alt

[a] *Department of Dermatology, University of Heidelberg, Heidelberg, Germany*
[b] *Department of Dermatology, Allergology and Environmental Medicine II, Hospital Thalkirchner Street, Munich, Germany*
[c] *Department of Dermatology and Venerology, Medical University of Graz, Graz, Austria*
[d] *Department of Dermatology, Medical University of Vienna, Vienna, Austria*
[e] *School of Medicine, The University of Queensland, Queensland, Australia*
[f] *First Department of Dermatology, Aristotle University, Thessaloniki, Greece*
[g] *Office Based Clinic of Dermatology, Konstanz, Germany*
[h] *Faculty of Computer Science and Mathematics, University of Passau, Passau, Germany*
[i] *Department of Dermatology, Lyons Cancer Research Center, Lyon 1 University, Lyon, France*
[j] *Department of Dermatology, Université Catholique de Louvain, St. Luc University Hospital, Brussels, Belgium*
[k] *Department of Genetic Epidemiology, University of Goettingen, Goettingen, Germany*

Received 16 September 2020; received in revised form 2 November 2020; accepted 22 November 2020

KEYWORDS
Deep learning;
Neural network;
Moleanalyzer-pro;
Dermoscopy;
Skin cancer;

Abstract ***Background:*** The clinical differentiation of face and scalp lesions (FSLs) is challenging even for trained dermatologists. Studies comparing the diagnostic performance of a convolutional neural network (CNN) with dermatologists in FSL are lacking.
Methods: A market-approved CNN (Moleanalyzer-Pro, FotoFinder Systems) was used for binary classifications of 100 dermoscopic images of FSL. The same lesions were used in a two-level reader study including 64 dermatologists (level I: dermoscopy only; level II: dermoscopy,

* *Corresponding author*: Department of Dermatology, University of Heidelberg, Im Neuenheimer Feld 440, 69120, Heidelberg, Germany. Fax: +49 6221 56 4996.
E-mail address: Holger.Haenssle@med.uni-heidelberg.de (H.A. Haenssle).

https://doi.org/10.1016/j.ejca.2020.11.034

Melanoma;
Lentigo maligna;
Solar lentigo;
Actinic keratosis;
Seborrheic keratosis;
Basal cell carcinoma

clinical close-up images, textual information). Primary endpoints were the CNN's sensitivity and specificity in comparison with the dermatologists' management decisions in level II. Generalizability of the CNN results was tested by using four additional external data sets.

Results: The CNN's sensitivity, specificity and ROC AUC were 96.2% [87.0%–98.9%], 68.8% [54.7%–80.1%] and 0.929 [0.880–0.978], respectively. In level II, the dermatologists' management decisions showed a mean sensitivity of 84.2% [82.2%–86.2%] and specificity of 69.4% [66.0%–72.8%]. When fixing the CNN's specificity at the dermatologists' mean specificity (69.4%), the CNN's sensitivity (96.2% [87.0%–98.9%]) was significantly higher than that of dermatologists (84.2% [82.2%–86.2%]; $p < 0.001$). Dermatologists of all training levels were outperformed by the CNN (all $p < 0.001$). In confirmation, the CNN's accuracy (83.0%) was significantly higher than dermatologists' accuracies in level II management decisions (all $p < 0.001$). The CNN's performance was largely confirmed in three additional external data sets but particularly showed a reduced specificity in one Australian data set including FSL on severely sun-damaged skin.

Conclusions: When applied as an assistant system, the CNN's higher sensitivity at an equivalent specificity may result in an improved early detection of face and scalp skin cancers.

Abbreviations

95% CI	95% confidence interval
AUC	area under the curve
CNN	convolutional neural network
ROC	receiver operating characteristic

1. Introduction

The skin of face and scalp differs from skin of other anatomic sites because of rather flat rete ridges and high densities of follicular openings, sweat glands, and sebaceous glands [1]. Therefore, the clinical and dermoscopic examination of face and scalp lesions (FSLs) requires additional knowledge on specific patterns at these sites [2]. Moreover, a straightforward clinical diagnosis is often challenging as FSL of benign (e.g. solar lentigo, lichen planus like keratosis) or malignant biology (e.g. lentigo maligna, pigmented actinic keratosis [AK]) may show overlapping features [3–5]. The appearance of FSL on severely sun damaged skin or collisions of two or more different lesions may additionally hamper a correct diagnosis [6,7].

Recently, computer vision systems were reported to attain a high level of diagnostic accuracy in the classification of skin lesions [8–12]. These systems often use deep learning convolutional neural networks (CNNs) trained by images and corresponding diagnostic labels. In this end-to-end approach software filters autonomously scan training images on a pixel level for 'good representations' of the specific diagnostic label. Each additional training image increases the CNN's 'knowledge' on what the diagnostic label 'looks like'. Well-trained CNNs achieved a diagnostic accuracy on or above the level of dermatologists. However, the methodology of many earlier studies made it difficult to transfer results into a real-life clinical setting [13]. Some studies applied a disadvantageous setting on the side of dermatologists by asking them to make a diagnosis by only one image. Moreover, the differentiation of only two diagnoses (e.g. nevi versus melanomas) did not adequately reflect the clinical situation with a much broader spectrum of lesions. With few exceptions [14,15] most studies did not include lesions of special anatomic sites (e.g. lesions of acral skin, mucosa, face, scalp). Finally, most previous publications did not imply a commercially or publicly available CNN architecture making it difficult to reproduce reported results.

The aim of the present study was to investigate a broad spectrum of FSL by a market-approved CNN [9,14–16]. As a main comparator, we chose the management decisions of dermatologists after reviewing full case information, comparable with the setting of a classical store-and-forward teledermatology consultation [17].

2. Methods

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

2.1. Data for test cases

2.1.1. Test-set of this study

We compiled a test-set of 100 FSL by randomly drawing image IDs created via a web-based application (Random.org, Randomness and Integrity Services Ltd., Dublin, Ireland). Included images were provided by five participating institutions, namely 1) Department of

Dermatology, University of Heidelberg; 2) Department of Dermatology, Hospital Thalkirchner Street, Munich; 3) Department of Dermatology, Medical University of Graz; 4) First Department of Dermatology, Aristotle University, Thessaloniki; 5) Dermatology Office based clinic of Dermatology, Konstanz. The test-set was 'fully external', because none of the institutions providing test cases was involved in CNN training or validation. The final set included many difficult-to-differentiate lesions, namely flat/macular lesions (n = 77) of melanocytic or non-melanocytic origin (Table 1). Most malignant cases were lentigo maligna and lentigo maligna melanoma, and false negative classifications in these diagnoses may significantly worsen the prognosis. Raised/nodular cases (n = 23) targeted the differential diagnoses of (pigmented) basal cell carcinoma, nodular melanoma, papillomatous (dermal) nevus, or seborrheic keratosis. All cases included close-up images, dermoscopic images, and textual information (age, sex, anatomic location).

Table 1
Characteristics of patients and lesions (n = 100 face and scalp cases).

	Face and scalp cases (*n* = 100)	
Age (mean, ±SD)	59	±22.8
Sex (n, %)		
Female	42	42.0%
Male	58	58.0%
Confirmation of diagnosis (n, %)		
Histopathology	98	98.0%
Unremarkable follow-up (>2 years)	2	2.0%
Localization (n, %)		
Cheek	49	49.0%
Chin	1	1.0%
Ear	4	4.0%
Eye lid	4	4.0%
Forehead	15	15.0%
Nose	15	15.0%
Philtrum	3	3.0%
Scalp	5	5.0%
Temple	4	4.0%
Malignant diagnosis (n, %)		
Lentigo maligna	22	22.0%
Invasive melanoma	19	19.0%
Lentigo maligna melanoma	*16*	*16.0%*
Nodular melanoma	*2*	*2.0%*
Superficial spreading melanoma	*1*	*1.0%*
(Pigmented) basal cell carcinoma	5	5.0%
(Pigmented) actinic keratosis	6	6.0%
Benign diagnosis (n, %)		
Nevus	18	18.0%
Solar lentigo/seborrheic keratosis	30	30.0%
Breslow thickness of invasive melanomas (mean, ±SD)	0.75	±0.91
Pigmentation (n, %)		
Pigmented	87	87.0%
Non-pigmented	13	13.0%
Origin of pigmentation (n, %)		
Melanocytic	59	59.0%
Non-melanocytic	41	41.0%
Clinical morphology (n, %)		
Flat/macular	77	77.0%
Raised/nodular	23	23.0%

Melanomas included hypomelanotic/amelanotic forms, actinic keratoses included pigmented forms, basal cell carcinomas included pigmented forms.

2.1.2. Additional external test-sets

To test for generalizability, we additionally used four other external test-sets comprising a total of 4832 dermoscopic images (Supplementary Table S1). First, the 'Australian data set' included solely FSL (n = 240) and was collected by C. Rosendahl between 2007 and 2012 in a primary skin cancer practice in Queensland, Australia [2]. Authors were fully blinded to diagnoses and lesions were often located on severely sun damaged skin and included many small and early lentigo maligna lesions. All except one melanoma (Breslow 0.3 mm) were *in situ*. Second, the 'ISIC2018 data set' including 1511 lesions with full blinding to diagnoses [11]. Third, the 'MSK-1 data set' including 1100 lesions available at https://www.isic-archive.com. Finally, the 'Prospective data set' a real-world dermoscopic data set of 1981 lesions prospectively acquired during 15 years of follow-up examinations in 435 patients at increased melanoma risk [18]. The CNN performance in data sets with blinding was kindly assessed by P. Tschandl.

2.2. The convolutional neural network

The CNN was approved as a medical device in the European Union (Moleanalyzer-Pro®, FotoFinder Systems, Bad Birnbach, Germany), and uses a modified architecture of Google's Inception_v4 [19], specifically trained with dermoscopic images. The prototype of the CNN was tested in a pivotal study [9] and details pertaining to the CNN architecture and training were reported earlier [9,14–16]. The CNN put out a 'malignancy score' ranging from 0 to 1 (higher scores indicating higher probability for malignancy) with the a priori cut-off of >0.5 for classifying a lesion as malignant.

2.3. Reader study level I and II

Dermatologists were personally invited to participate via a Web-based rating application. Participants were categorised 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). Dermatologists were not limited in the time spend per case and could discontinue their work at their own convenience for a 6-month period.

Each case was presented via two subsequent computer slides: i) dermoscopic image (level I information) and ii) dermoscopic image plus close-up image and textual information (level II information). For each slide dermatologists had to indicate management decisions (treatment/excision, follow-up examination, no action)

and binary diagnostic decisions (benign, malignant). Diagnostically unclear case could be marked as 'unclear'. All study cases are shown in Supplementary Fig. S1.

2.4. Statistical analysis

Primary outcome measures were the CNN's sensitivity, specificity and accuracy in comparison with the dermatologists' management decisions in study level 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 (ROCs). Estimates are provided along with 95% confidence intervals.

Management decisions 'no action' and 'follow-up examination' were considered true-negative for benign lesions. AK shows limited potential to progress to invasive carcinoma and 'excision/treatment' and 'follow-up examination' were considered true-positive.

For a head-to-head comparison of the CNN's and the dermatologists' sensitivity at the same specificity we determined the cut-off in CNN malignancy scores corresponding to the dermatologists' mean specificity within a 1100-image validation-set. This cut-off was transferred to the test-set and the CNN's *versus* dermatologists' sensitivity could be compared by a non-parametric one sample Wilcoxon signed-rank test. Changes in dermatologists' diagnostic performance after receiving additional level II information were tested by Wilcoxon paired signed-rank test with observations related to readers. Bonferroni corrections were used to adjust for multiple testing. Results were considered statistically significant at the $p < 0.05$ level. All analyses were carried out using SPSS Version 24 (IBM, SPSS; Chicago, Illinois, USA).

3. Results

3.1. Characteristics of face and scalp cases ($n = 100$)

The average age (±SD) of patients was 59 (±22) years (range 4–92 years) and 42% were female (Table 1). Face and scalp cases included 48 benign lesions and 52 malignant/premalignant lesions (Supplementary Fig. S1).

3.2. Performance of convolutional neural network

3.2.1. Performance in test-set of this study

The CNN showed a sensitivity across all malignant diagnoses and specificity across all benign diagnoses of 96.2% [87.0%–98.9%] and 68.8% [54.7%–80.1%], respectively. The ROC AUC was 0.929 [0.880–0.978] (Fig. 1). Boxplots in Fig. 2 show the distribution of malignancy scores in relation to diagnoses. The percentage of correct classifications in malignant lesions was 100% in basal cell carcinomas, 100% in actinic keratoses, 95.5% in invasive melanomas and 94.7% in lentigo maligna, whereas in benign lesions 88.9% of nevi and 56.7% of solar lentigo/seborrheic keratosis were classified correctly (83.0% overall accuracy).

3.2.2. Performance in additional external test-sets

We used four external data sets (two with full blinding to diagnoses) to test for generalizability of the CNN performance (Supplementary Table S1). The 'Australian data set' included 240 FSL collected in Queensland, Australia [2]. Here, the CNN's sensitivity was 90.4% [83.5%–94.5%], however at a lower specificity of 48.4% [39.9%–57.1%] resulting in a significantly lower ROC AUC in comparison with our study data set (0.814 [0.760–0.868] versus 0.929 [0.880–0.978], $p = 0.014$) (Fig. 3). As a possible explanation for the reduced specificity we found a frequent co-localization of benign lesions with severely sun-damaged adjacent skin. In contrast, the CNN's ROC-AUC in our study data set was comparable with that attained in the much larger 'ISIC2018 data set' (0.926 [0.912–0.940])[11], 'MSK-1 data set' (0.939 [0.923–0.954]) and 'Prospective data set' [18] (0.945 [0.930–0.961], all $p > 0.585$).

3.3. Dermatologists' management decisions

The mean sensitivity and specificity of dermatologists for management decisions during study level II (full case information) was 84.2% [82.2%–86.2%] and 69.4%

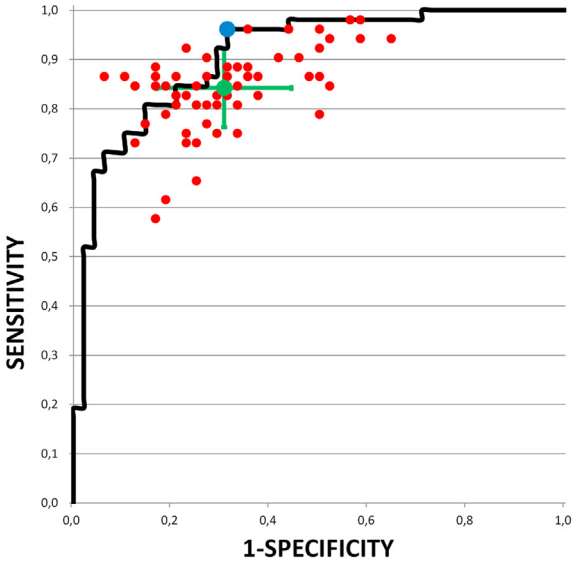

Fig. 1. Receiver operating characteristic curve of the convolutional neural network (CNN) (black curve) in relation to the results of all dermatologists (n = 64, red dots) in their level II management decisions (full case information). The average (±SD) sensitivity and specificity of all dermatologists (mean: green circle; ± SD: green error bars) and the CNN's point of operation (blue circle, sensitivity: 96.2%, specificity: 68.8%) are depicted.

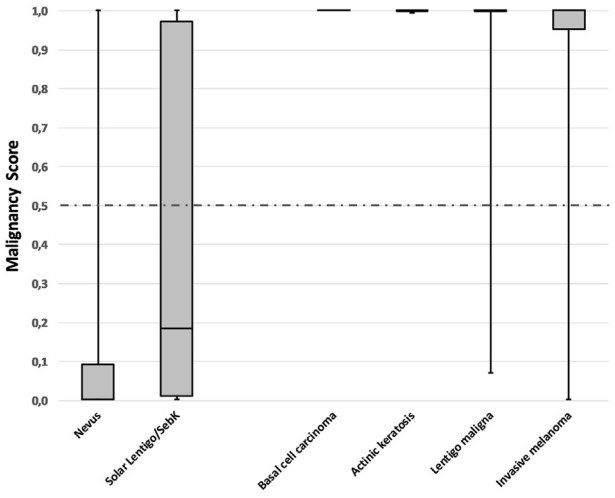

Fig. 2. The convolutional neural network's (CNN) malignancy scores (range 0–1) for major benign and malignant diagnostic categories are depicted as boxplots. Scores closer to 1 indicate a higher probability of malignancy. The upper and lower bounds of boxes indicate the 25th and 75th percentiles while the median is indicated by the line intersection the boxes. Whiskers indicate the full range of scores. SebK, seborrheic keratosis.

[66.0%–72.8%], respectively. With only one dermoscopic image at hand (study level I) dermatologists were less sensitive (77.1% [74.0%–80.2%]) at a similar specificity (69.5% [66.3%–72.7%]). Expectedly, the performance of dermatologists improved with more experience. The percentage of correct management decisions (accuracy) was lowest in beginners and highest in experts (Supplementary Table S2). Full details on the dermatologists' performance are given in the supplement and Supplementary Fig. S2.

3.4. Performance of convolutional neural network versus dermatologists

The benchmark analysis was the difference in sensitivity between the CNN and dermatologists when compared at the mean specificity of dermatologists' management decisions in level-II. The CNN's a priori cut-off at this specificity (69.4%) was attained in a 1100-image validation-set and then applied to the test-set. The resulting CNN's sensitivity was 96.2% [87.0%–98.9%], which was significantly higher than the mean sensitivity of dermatologists (84.2% [82.2%–86.2%]; $p < 0.001$).

For management decisions based on one dermoscopic image (level I) the dermatologists' sensitivity was lower and more inferior to the CNN (77.1% [74.0%–80.2%]; $p < 0.001$). Moreover, accuracies (percentage of correct diagnoses) were used for comparing dermatologist with the CNN. In this statistical approach the CNN also outperformed dermatologists in their management decisions (Supplementary Table S2).

4. Discussion

Chronically sun exposed skin of the face and scalp is particularly prone to develop skin lesions of benign and/or malignant biology. Clinically, diagnostic uncertainties concerning FSL arise from overlapping morphologies of benign and malignant lesions [2,3].

To create a setting closer to a real-life clinical situation, test cases in this study included a broad range of malignant and benign, melanocytic and non-melanocytic, pigmented and non-pigmented, flat/macular and raised/nodular skin lesions. These lesions comprised the vast majority of FSL frequently biopsied in clinical routine to confirm or rule out malignancy. Most lesions (61%) were 'difficult to differentiate' lesions, namely lentigo maligna (22%), pigmented solar lentigo/seborrheic keratosis (28%), pigmented AK (6%) and pigmented basal cell carcinoma (5%). Moreover, we aimed to apply a fair comparator to the CNN's performance [13]. Our data clearly show that dermatologists' management results improved after having reviewed clinical plus dermoscopic images accompanied by textual case information. Of note, gaining better results by integrating different levels of information into the diagnostic process also works for CNNs. It was shown, that the fusion of one CNN assessing clinical images to one classifying dermoscopic images achieved better results than either CNN alone [20]. Moreover, textual non-imaging information may also be included into the analysis by neural networks for successfully predicting skin cancer [21].

The CNN of our study significantly outperformed the dermatologists' sensitivity by approximately 12% when operating at a threshold equivalent to the dermatologists' mean specificity. This difference in sensitivity translated to an average of 6.2 of 52 malignant lesions more frequently missed by dermatologists than by the CNN (CNN: 2/52 malignant lesions missed, dermatologists: mean of 8.2/52 malignant lesions missed). Of note, the CNN evaluated dermoscopic images only, whereas dermatologists reviewed the full case information. Unsurprisingly, when grouping dermatologists by their self-reported experience, more experienced dermatologists attained better results. This was reflected by an improved rate of correct decisions in their diagnostic and management decisions. When comparing the management decisions of dermatologist subgroups (experts, skilled, beginners) after reviewing full information with the CNN we found a significantly higher accuracy for the CNN (only 8 of 30 experts surpassed the CNN's performance). Additionally, when setting the CNN's operation point to the specificity of individual dermatologist subgroups, the sensitivity of the CNN was significantly higher across all levels of experience.

Most of the CNN's outperformance was related to a higher sensitivity at an equivalent specificity. In a

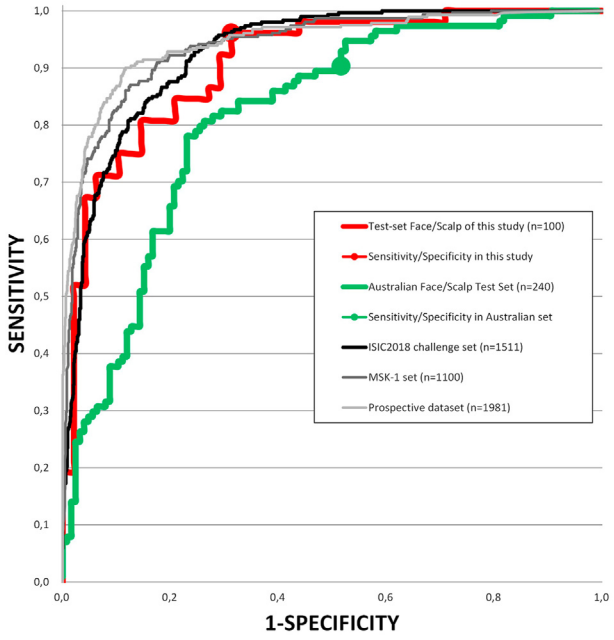

Fig. 3. The convolutional neural network's (CNN) receiver operating characteristic (ROC) curve in the data set of this study (red curve) is depicted in comparison with ROC curves resulting from additional external data sets. The 'Australian data set [2]' solely included face and scalp lesion (FSL) (green curve). The point of operation of the CNN within both FSL data sets is depicted as a circle on the curve in red (this study: sensitivity 96.2%, specificity 68.8%) and in green ('Australian data set': sensitivity 90.4%, specificity 48.4%). Moreover, ROC curves of three other external data set including more than 4500 images are shown. The ROC AUCs were 0.929 [0.880–0.978] in this study's data set, 0.814 [0.760–0.868] in 'Australian data set', 0.926 [0.912–0.940] in 'ISIC2018 data set', 0.939 [0.923–0.954] in 'MSK-1 data set', and 0.945 [0.930–0.961] in 'Prospective data set'.

clinical routine application of the CNN, this should result in less missed malignant lesions but not necessarily in less biopsied benign lesions. Yet, in a prospective real-life situation, the CNN would be applied as an assistance system and its classifications would be incorporated into the decision-making process. Unfortunately, there are no prospective studies available to show in which situation physicians would decide to follow a CNN's recommendation or not.

Overfitting and lack of generalizability are important limitations not well recognised by many earlier studies [22]. Overfitting occurs whenever an image data set is collected from a few sources resulting in a narrow distribution of lesions and also of unrecognised image features. A similarly narrow distribution of lesions and unrecognised image features in training and testing can easily result in a major overestimation of the CNN's performance and lack of generalizability. For the CNN of our study, overfitting was largely avoided by using >150,000 training images from hundreds of cooperating clinics and dermatologists around the world. Moreover, all test cases were truly external cases, meaning that sources providing test cases did not provide training images. We also tested for generalizability by four larger external data sets, one focused solely on FSL from Australia with full blinding to the true diagnoses [2]. It was reassuring to still find a CNN's sensitivity of 90% in the Australian data set [11], however, at a reduced specificity and ROC AUC. The lower specificity in this setting may be explained by appearance of many benign lesions on severely sun-damaged skin. Conceivably, a CNN that had 'learned' to classify actinic keratoses as 'malignant' during training could be misled by actinic damage when present on surrounding skin. We are currently investigating appropriate solutions to this limitation in another study. In the remaining data sets, which were not exclusively focused on FSL and included more than 4500 dermoscopic lesions, the CNN's performance was comparable with that observed in our study data set as shown by non-significant differences in ROC AUCs.

There are several limitations to this study. First, our test-set included only 100 lesions to guarantee recruitment of enough dermatologists willing to review the data, thus leaving only a few cases for certain diagnoses. The CNN's and the dermatologists' results in these diagnoses should, therefore, be interpreted with caution. Moreover, dermatologists did not receive textual information on the history of lesions, e.g. recent change in size or colour. The skin of face and scalp may bring forward many more lesions (e.g. inflammatory lesions) than were included in our test-set so that results should not be transferred to a general population. Finally, (current) deep learning algorithms lack interpretability [23], rendering specific explanations for false classifications impossible.

In conclusion, our study demonstrates that the tested CNN may classify dermoscopic images of FSL at a high level of sensitivity and specificity. The CNN's performance showed superiority in comparison with management decisions of dermatologists after they had reviewed the full case information. If applied in clinical routine, the CNN's higher sensitivity at an equivalent specificity in comparison with dermatologists may result in less missed malignant lesions but probably not in a reduction of biopsied benign lesions.

Ethics

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

Author contribution section

Holger Andreas Haenssle: study concepts and design, data acquisition, quality control, data analysis and

interpretation, statistical analysis, manuscript preparation, -editing, -review; **Julia Katharina Winkler**: study concepts, data acquisition, manuscript editing and -review; **Christine Fink**: study concepts, data acquisition, manuscript editing and -review; **Ferdinand Toberer**: data acquisition, manuscript editing and -review; **Alexander Enk**: data acquisition, manuscript editing and -review; **Wilhelm Stolz**: data acquisition, manuscript editing and -review; **Teresa Deinlein**: data acquisition, manuscript editing and -review; **Rainer Hofmann-Wellenhof**: data acquisition, manuscript editing and -review; **Harald Kittler**: data acquisition, data analysis and interpretation, statistical analysis, manuscript editing and -review; **Philipp Tschandl**: data acquisition, data analysis and interpretation, statistical analysis, manuscript editing and -review; **Cliff Rosendahl**: data acquisition, manuscript editing and -review; **Aimilios Lallas**: data acquisition, manuscript editing and -review; **Andreas Blum**: data acquisition, manuscript editing and -review; **Mohamed Souhayel Abassi**: data acquisition, manuscript editing and -review; **Luc Thomas**: data acquisition, manuscript editing and -review; **Isabelle Tromme**: data acquisition, manuscript editing and -review; **Albert Rosenberger**: data acquisition, quality control, data analysis and interpretation, statistical analysis, manuscript editing and -review.

Conflict of interest statement

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: A Blum received honoraria and/or travel expenses from Heine Optotechnik GmbH and 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.

R Hofmann-Wellenhof received honoraria and/or travel expenses from FotoFinder Systems GmbH and is founder and shareholder of e-derm-consult GmbH.

All other authors declared no conflict of interest.

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.

Appendix A

The following dermatologists in alphabetical order agreed to be acknowledged for their participation in the reader study level I and II:

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

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

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejca.2020.11.034.

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).

References

[1] Arda O, Goksugur N, Tuzun Y. Basic histological structure and functions of facial skin. Clin Dermatol 2014;32(1):3–13.

[2] Tschandl P, Rosendahl C, Kittler H. Dermatoscopy of flat pigmented facial lesions. J Eur Acad Dermatol Venereol 2015;29(1):120–7.

[3] Akay BN, Kocyigit P, Heper AO, et al. Dermatoscopy of flat pigmented facial lesions: diagnostic challenge between pigmented actinic keratosis and lentigo maligna. Br J Dermatol 2010;163(6):1212–7.

[4] Tanaka M, Sawada M, Kobayashi K. Key points in dermoscopic differentiation between lentigo maligna and solar lentigo. J Dermatol 2011;38(1):53–8.

[5] Lallas A, Tschandl P, Kyrgidis A, et al. Dermoscopic clues to differentiate facial lentigo maligna from pigmented actinic keratosis. Br J Dermatol 2016;174(5):1079–85.
[6] Chung HJ, McGuigan KL, Osley KL, et al. Pigmented solar (actinic) keratosis: an underrecognized collision lesion. J Am Acad Dermatol 2013;68(4):647–53.
[7] Blum A, Siggs G, Marghoob AA, et al. Collision skin lesions-results of a multicenter study of the International Dermoscopy Society (IDS). Dermatol Pract Concept 2017;7(4):51–62.
[8] Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542(7639):115–8.
[9] Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018;29(8):1836–42.
[10] Marchetti MA, Codella NCF, Dusza SW, et al. Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. J Am Acad Dermatol 2018;78(2):270–7.
[11] 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):938–47.
[12] Yu C, Yang S, Kim W, et al. Acral melanoma detection using a convolutional neural network for dermoscopy images. PLoS One 2018;13(3):e0193321.
[13] Lallas A, Argenziano G. Artificial intelligence and melanoma diagnosis: ignoring human nature may lead to false predictions. Dermatol Pract Concept 2018;8(4):249–51.
[14] 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):137–43.
[15] 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 Canc 2020: 12721–9.
[16] Fink C, Blum A, Buhl T, et al. Diagnostic performance of a deep learning convolutional neural network in the differentiation of combined naevi and melanomas. J Eur Acad Dermatol Venereol 2020 Jun;34 (6):1355–61.
[17] Finnane A, Dallest K, Janda M, et al. Teledermatology for the diagnosis and management of skin cancer: a systematic review. JAMA Dermatol 2017;153(3):319–27.
[18] Sies K, Winkler JK, Fink C, et al. Past and present of computer-assisted dermoscopic diagnosis: performance of a conventional image analyser versus a convolutional neural network in a prospective data set of 1,981 skin lesions. Eur J Canc 2020:13539–46.
[19] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision. 2015. Reprint at: https://arxiv.org/abs/1512.00567.
[20] Tschandl P, Rosendahl C, Akay BN, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. JAMA Dermatol 2019;155(1):58–65.
[21] Wang HH, Wang YH, Liang CW, et al. Assessment of deep learning using nonimaging information and sequential medical records to develop a prediction model for nonmelanoma skin cancer. JAMA Dermatol 2019 Sep 4;155(11):1277–83.
[22] Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence Technology for medical diagnosis and prediction. Radiology 2018;286(3): 800–9.
[23] Pereira S, Meier R, McKinley R, et al. Enhancing interpretability of automatically extracted machine learning features: application to a RBM-Random Forest system on brain lesion segmentation. Med Image Anal 2018:44228–44.