Diagnostic performance of a deep learning convolutional neural network in the differentiation of combined naevi and melanomas

C. Fink,1 A. Blum,2 T. Buhi,3 C. Mitteldorf,3 R. Hofmann-Wellenhof,4 T. Deinlein,4 W. Stolz,5 L. Trennheuser,1 C. Cussigh,1 D. Deltgen,1 J.K. Winkler,1 F. Toberer,1 A. Enk,1 A. Rosenberger,6 H.A. Haenssle1,*

1Department of Dermatology, University of Heidelberg, Heidelberg, Germany
2Public, Private and Teaching Practice, Konstanz, Germany
3Department of Dermatology and Allergology, University Medical Center Göttingen, Göttingen, Germany
4Department of Dermatology and Venerology, Medical University of Graz, Graz, Austria
5Department of Dermatology, Allergology and Environmental Medicine II, Hospital Thalkirchner Street, Munich, Germany
6Department of Genetic Epidemiology, University Medical Center of Goettingen, Goettingen, Germany
*Correspondence: H.A. Haenssle. E-mail: holger.haenssle@med.uni-heidelberg.de

Linked Commentary: M.V. Heppt et al. J Eur Acad Dermatol Venereol 2020; 34: 1134–1135. https://doi.org/10.1111/jdv.16577.

Abstract
Background  Deep learning convolutional neural networks (CNN) may assist physicians in the diagnosis of melanoma. The capacity of a CNN to differentiate melanomas from combined naevi, the latter representing well-known melanoma simulators, has not been investigated.

Objective  To assess the diagnostic performance of a CNN when used to differentiate melanomas from combined naevi in comparison with dermatologists.

Methods  In this study, a CNN with regulatory approval for the European market (MoleAnalyzer-Pro, FotoFinder Systems GmbH, Bad Birnbach, Germany) was used. We attained a dichotomous classification (benign, malignant) in dermoscopic images of 36 combined naevi and 36 melanomas with a mean Breslow thickness of 1.3 mm. Primary outcome measures were the CNN’s sensitivity, specificity and the diagnostic odds ratio (DOR) in comparison with 11 dermatologists with different levels of experience.

Results  The CNN revealed a sensitivity, specificity and DOR of 97.1% (95% CI [82.7–99.6]), 78.8% (95% CI [62.8–89.1]) and 34 (95% CI [4.8–239]), respectively. Dermatologists showed a lower mean sensitivity, specificity and DOR of 90.6% (95% CI [84.1–94.7]; P = 0.092), 71.0% (95% CI [62.6–78.1]; P = 0.256) and 24 (95% CI [11.6–48.4]; P = 0.1114). Under the assumption that dermatologists use the CNN to verify their (initial) melanoma diagnosis, dermatologists achieve an increased specificity of 90.3% (95% CI [79.8–95.6]) at an almost unchanged sensitivity. The largest benefit was observed in ‘beginners’, who performed worst without CNN verification (DOR = 12) but best with CNN verification (DOR = 98).

Conclusion  The tested CNN more accurately classified combined naevi and melanomas in comparison with trained dermatologists. Their diagnostic performance could be improved if the CNN was used to confirm/overrule an initial melanoma diagnosis. Application of a CNN may therefore be of benefit to clinicians.

Conflict of interest
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. C Fink received travel expenses from Magnosco GmbH. All other authors declared no conflict of interest.

Funding source
This research received no specific grant from any public, commercial or not-for-profit sector.
Introduction
Early detection of malignant melanoma is critical to decrease the mortality rate which is still rising in many countries.1,2 As for now, visual inspection and dermoscopy are the most frequently applied methods for melanoma detection.3 However, concerns may arise from a heterogeneous level of training and experience of physicians involved in skin cancer screening. Accordingly, a reproducible, standardized and objective method with a high diagnostic accuracy to assist physicians in the decision-making process would be desirable. One promising approach is the application of deep learning algorithms, which are directly trained by raw pixels of input images and corresponding diagnostic labels. Deep learning is part of machine learning methods and uses large datasets and classification labels that are fed into a neural network. Deep learning neural networks are composed of many sequential layers that feed forward the training image while searching on a pixel level for ‘good representation’ of the input classification. Lower layers within the network are responsible for the detection of simple features within small areas (e.g. edges and colours), while higher layers assemble these information into more complex representations of larger areas. Therefore, deep learning could be best described as ‘hierarchical feature learning’. During training with dermoscopic images and the task to differentiate naevi from melanomas, the network collects and weights features that are good representations of either diagnosis. A convolutional neural network (CNN) as used in our study is a subclass of deep neural networks.4,5

Pioneering publications reported an expert-level performance of deep learning CNNs in the task of skin cancer detection from dermoscopic images.4–7 More recently, a first deep learning CNN for the classification of skin neoplasms has gained market access as a medical device in Europe (Molealyzer-Pro®, FotoFinder GmbH, Bad Birnbach, Germany). The developmental prototype of this CNN, that is based on a pretrained GoogleNet Inception architecture additionally trained with large numbers of dermoscopic images and corresponding disease labels, showed an expert-level performance in an international cross-sectional reader study.4 While these results appear promising, potential limitations need to be identified. One critical aspect could be benign lesions mimicking cutaneous melanomas. Such ‘melanoma simulators’ may potentially decrease the specificity of a CNN leading to increased numbers of unnecessary excisions. In this context, combined naevi represent an important group of melanoma simulators that share many clinical and dermoscopic features of melanomas such as asymmetry, multi-colour, and/or alternating flat and raised areas.8 Combined naevi are defined as benign naevi that histologically show two or more different types of naevus components within one lesion.9,10 According to an analysis by Scolyer et al.9 the most frequent combination triggering excision is a common naevus and a blue naevus followed by a common naevus and a Spitz naevus. Our study aimed at investigating the diagnostic performance of a CNN in differentiating combined naevi and melanomas in comparison with a group of trained dermatologists with different levels of experience.

Material and methods

Dermoscopic images
For this study, we created an image set including 72 melanocytic lesions (36 combined naevi, 36 melanomas) from 72 different patients. The study was approved by the local ethics committee and performed in accordance with the Declaration of Helsinki principles.11 Combined naevi (n = 36) were collected from the departments of dermatology of the university medical centres of Heidelberg, Göttingen, and from the medical centre Thalkirchner Straße, Munich. Cases retrieved from histopathological databases (n = 34) were included whenever a corresponding dermoscopic image of sufficient quality was available. In non-excised cases of combined naevi (n = 2), the diagnosis was based on expert consensus and an unremarkable follow-up over at least 2 years. Histopathology showed 32 (94.1%) combinations of compound naevi with blue naevi, one compound naevus with Spitz naevus (2.9%) and one compound naevus with deep penetrating naevus (2.9%). Melanoma cases were retrieved from the annotated dermoscopic image archive of the department of dermatology in Heidelberg. Melanomas were manually selected to meet the requirement of blue-grey areas, which were also present in all CN. Histopathology of melanoma cases revealed 31 (86.1%) superficial spreading and 5 (13.9%) nodular melanomas. All melanomas were invasive with a mean (min–max) thickness of 1.3 mm (0.2–2.5 mm; Table 1). Representative lesions of the study image set are shown in Fig. 1.

Convolutional neural network and dermatologists
For this study, the current version (accessed June 2019) of a CNN with regulatory approval as a medical device for the European market was used (Molealyzer-Pro®; FotoFinder Systems GmbH, Bad Birnbach, Germany). The CNN is based on a pretrained GoogleNet Inception_v4 architecture and was additionally trained with >120.000 dermoscopic images and labels.4 The CNN architecture and training are described in detail elsewhere (Fig. S1).4 The softmax output layer of the CNN classified each of the 72 dermoscopic images by a ‘melanoma probability’ score ranging from 0 to 1. The dichotomous classification (benign vs. malignant) was based on a priori validated cut-off of >0.5. The same set of dermoscopic images was assessed by a group of 11 dermatologists. First, dermatologists indicated their level of experience in the use of dermoscopy (‘beginner’ <2 years of experience, ‘skilled’ 2–5 years of experience and ‘expert’ ≥5 years of experience). Next, dermatologists were presented the dermoscopic images on a screen and asked to indicate their
dichotomous diagnosis (melanoma vs. benign melanocytic lesion) and their management decision (excision, short-term follow-up, no action needed).

Statistical evaluation
Primary outcome measures were the CNN’s sensitivity, specificity and the diagnostic odds ratio (DOR). The CNN’s performance was compared with dermatologists by modelling an ‘average dermatologist’ (constructed from 1/3 ‘beginner’, 1/3 ‘skilled’ and 1/3 ‘expert’). Moreover, the CNN’s results were compared with individual groups of dermatologists ranked by their self-reported experience in dermoscopy. To account for correlations within the assessments of the images, we fitted logit binomial linear models allowing for repeated measurements within each dermatologist. Whenever all dermatologists uniformly rated an individual image of the test set, we added an opposing case with weight 0.001 to make the estimates robust. Compressions between groups of dermatologist and CNN were corrected for multiple testing by the method of Tukey–Kramer. Results were considered statistically significant at the $P < 0.05$ level. We did not estimate overall accuracy, positive or negative predictive values, because these measures (in contrast to DOR) are dependent on the melanoma prevalence within the tested sample. Clearly, the tested sample is not representative of the true melanoma prevalence in a larger population.

We also assessed the diagnostic performance resulting from an assumed collaboration of dermatologists and the CNN: scenario 1) CNN used to verify (or overrule) a dermatologist’s diagnosis of malignancy and scenario 2) CNN used to verify (or overrule) a dermatologist’s diagnosis of benignity. Strategy (i) aims to reduce false-positive diagnoses and strategy; (ii) to reduce false-negative diagnoses.

Moreover, we calculated the CNN’s area under the curve of receiver operating characteristics (ROC) for the diagnostic classification of lesions. Further endpoints included the assessment of the dermatologists’ diagnostic performance in their management decisions. For management decisions the option of a ‘short-term follow-up’ was positively accounted for both sensitivity and specificity calculations. Descriptive statistics as frequency, mean, range and standard deviation were used. The analyses were carried out using SPSS Version 24 (IBM, SPSS; Chicago, IL, USA) or SAS 9.4 (SAS Institute Inc., Cary, NC, USA) (Appendix S1).

Results

Diagnostic performance of dermatologists and CNN
Three out of 11 participating dermatologists (27.3%) indicated being a ‘beginner’ in dermoscopy (<2 years of experience), while 5 (45.5%) declared to be ‘skilled’ (2–5 years of experience) and 3 (27.3%) to be an ‘expert’ (>5 years of experience).

For an average dermatologist we estimated a sensitivity of 90.6% (95% CI [84.1–94.7]) and specificity of 71.0% (95% CI [62.6–78.1]). Both measures were lower than those attained by the CNN: sensitivity 97.1% (95% CI [82.7–99.6]; $P = 0.092$) and specificity 78.8% (95% CI [62.8–89.1]; $P = 0.256$); however,
significance was missed (Fig. 2). When using DOR as a single indicator of diagnostic performance that is independent of prevalence, the CNN attained a DOR of 34 (95% CI [4.8–239]), which was larger than that of an average dermatologist (DOR = 24; 95% CI [11.6–48.4], \( P = 0.1114 \); Tables 2 and 3).

The observed sensitivities in dermatologists with different levels of experience ranged from 86.7% (95% CI [77.7–92.4]) in experts to 90.9% (95% CI [82.4–95.5]) in beginners, but differences were not significant and should be regarded as due to random. However, the observed specificities increased with experience: from 55.1% (95% CI [45.7–64.2]) in beginners, to 74.2% (95% CI [64.4–82.0]) in skilled dermatologists, and 80.6% (95% CI [70.2–88.0]) in experts. The specificity in beginners was significantly lower than in skilled dermatologists or experts (all \( P < 0.008 \)). When calculating DOR, skilled dermatologists performed best [DOR = 40; (95% CI [16.0–102])], followed by experts [DOR = 27; (95% CI [11.6–62.5])] and beginners [DOR = 12; (95% CI [5.3–28.5])]. The CNN’s DOR of 34 was intermediate to skilled dermatologists (DOR = 40, \( P = 0.3117 \)) and experts (DOR = 27, \( P = 0.1738 \); Tables 2 and 3). There was a moderate but significant rank correlation between the CNN’s melanoma probability scores and the number of benign dermatologist diagnoses (Kendall tau = 0.59, \( P < 0.001 \); Fig. 3).

For a real-life clinical setting, one may assume collaboration between dermatologists and CNN. Therefore, we also investigated potential improvements in the diagnostic performance of dermatologists when using the CNN to verify or falsify an initial clinical diagnosis. If the CNN was presumably applied to verify an initial melanoma diagnosis (scenario 1), the specificity would increase for all groups of dermatologists. In this setting, the average dermatologist would achieve a specificity of 90.3% (95% CI

Figure 1 Representative lesions of the study test set including the saliency maps which reveal relevant pixels for the convolutional neural networks (CNN’s) prediction. Breslow thickness is indicated in millimetre (mm). CNN, convolutional neural network; SSM, superficial spreading melanoma.

Figure 2 Receiver operating characteristics (ROC) curve of the convolutional neural networks (CNN’s) diagnostic performance in relation to the dermatologists. ROC curve of the CNN in relation to the results of all individual dermatologists \((n = 11, \text{ red dots})\). Moreover, the average (±SD) sensitivity and specificity of all dermatologists (mean: green circle; ±SD: green error bars) is depicted. The point of operation of the CNN is indicated by a blue circle.
Recognition of combined naevi by a CNN

Table 2 Dermatologists’ and convolutional neural networks (CNN’s) diagnostic performance

| Diagnosis       | Sensitivity (%) | 95% CI  | specificity (%) | 95% CI  | DOR  | 95% CI |
|-----------------|----------------|--------|----------------|--------|------|--------|
| Beginners (n = 3) | 90.9           | 82.4-95.5 | 55.1           | 45.7-64.2 | 12.2 | 5.3-28.5 |
| Skilled (n = 5)  | 93.3           | 86.3-96.9 | 74.2           | 64.4-82.0 | 40.3 | 16.0-102.0 |
| Experts (n = 3)  | 86.7           | 77.7-92.4 | 80.6           | 70.2-88.0 | 26.9 | 11.6-62.5 |
| av. dermatologist† | 90.6           | 84.1-94.7 | 71.0           | 62.6-78.1 | 23.7 | 11.6-48.4 |
| CNN             | 97.1           | 82.7-99.6 | 78.8           | 62.8-89.1 | 33.9 | 4.8-239 |

Management

| Diagnosis       | Sensitivity (%) | 95% CI  | specificity (%) | 95% CI  | DOR  | 95% CI |
|-----------------|----------------|--------|----------------|--------|------|--------|
| Beginners (n = 3) | 100.0          | 99.8-100.0 | 17.9           | 10.1-29.6 | 869 | 108-6970 |
| Skilled (n = 5)  | 100.0          | 99.8-100.0 | 42.5           | 33.7-51.8 | 2690 | 356-20 351 |
| Experts (n = 3)  | 100.0          | 99.8-100.0 | 71.8           | 60.0-81.0 | 9350 | 1197-73 042 |
| av. dermatologist† | 100.0          | 99.8-100.0 | 47.6           | 36.6-58.8 | 2796 | 366-21 357 |

CNN in scenario 1:

| Diagnosis       | Sensitivity (%) | 95% CI  | specificity (%) | 95% CI  | DOR  | 95% CI |
|-----------------|----------------|--------|----------------|--------|------|--------|
| Beginners (n = 3) | 91.1           | 80.9-96.1 | 90.5           | 80.0-95.8 | 97.7 | 28.4-336.3 |
| Skilled (n = 5)  | 84.2           | 73.5-91.1 | 92.6           | 81.2-97.3 | 66.5 | 19.1-231.7 |
| Experts (n = 3)  | 89.8           | 79.3-95.3 | 87.0           | 74.5-93.9 | 58.9 | 18.2-190.4 |
| av. dermatologist† | 88.7           | 79.3-94.1 | 90.3           | 79.8-95.6 | 34.2 | 13.5-86.6 |

CNN in scenario 2:

| Diagnosis       | Sensitivity (%) | 95% CI  | specificity (%) | 95% CI  | DOR  | 95% CI |
|-----------------|----------------|--------|----------------|--------|------|--------|
| Beginners (n = 3) | 100.0          | 99.8-100.0 | 64.4           | 51.4-75.6 | 6520 | 831-51 152 |
| Skilled (n = 5)  | 100.0          | 99.8-100.0 | 66.6           | 52.9-78.1 | 7194 | 907-57 049 |
| Experts (n = 3)  | 98.1           | 88.2-99.7 | 46.3           | 35.4-57.5 | 45  | 6.1-330 |
| av. Dermatologist† | 99.9           | 99.5-100.0 | 59.4           | 47.4-70.3 | 1052 | 231-4796 |

†Average dermatologist constructed from 1/3 ‘beginner’, 1/3 ‘skilled’ and 1/3 ‘experienced dermatologist’; ‡Expert, dermatologists indicated to have ≥5 years of experience; ‡Skilled, dermatologists indicated to have 2-5 years of experience; ‡Beginner, dermatologists indicated to have <2 years of experience. ‡Diagnostic performance resulting from an assumed collaboration of dermatologists and the CNN: clinical scenario 1) CNN used to verify (or overrule) a dermatologist’s melanoma diagnosis; and scenario 2) CNN used to verify a dermatologist’s benign naevus diagnosis.

Table 3 Significance of group comparisons

| Diagnosis       | Sensitivity | Specificity |
|-----------------|-------------|-------------|
|                 | CNN         | Beginner    | Skilled     | Expert |
|                 |             |             |             |        |
| CNN             | 1           | 1           | 1           |        |
| Beginner (n = 3) | 0.535       | 0.750       | 0.767       |        |
|                  | 0.090       | 1           | 0.008       | <0.001 |
| Skilled (n = 5)  | 0.843       | 1           | 0.053       |        |
|                  | 0.957       | 1           | 0.433       |        |
| Expert (n = 3)   | 0.456       | 1           | 0.947       |        |
|                  | 0.256       |             |             |        |

Management

| Diagnosis       | Sensitivity | Specificity |
|-----------------|-------------|-------------|
|                 | CNN         | Beginner    | Skilled     | Expert |
|                 |             |             |             |        |
| Beginner (n = 3) | 1           |             |             |        |
|                  |             |             |             | <0.001 |
| Skilled (n = 5)  | 1           |             |             |        |
|                  |             |             |             | <0.001 |
| Expert (n = 3)   | 1           |             |             |        |

CNN in scenario 1:

| Diagnosis       | Sensitivity | Specificity |
|-----------------|-------------|-------------|
|                 | CNN         | Beginner    | Skilled     | Expert |
|                 |             |             |             |        |
| Beginner (n = 3) | 1           | 0.848       | 0.337       |        |
|                  |             |             |             | 0.269  |
| Skilled (n = 5)  | 1           | 0.055       | 1           |        |
|                  |             |             |             | 0.757  |
| Expert (n = 3)   | 1           |             |             |        |

CNN in scenario 2:

| Diagnosis       | Sensitivity | Specificity |
|-----------------|-------------|-------------|
|                 | CNN         | Beginner    | Skilled     | Expert |
|                 |             |             |             |        |
| Beginner (n = 3) | 1           | 0.007       | 0.007       |        |
|                  |             |             |             | <0.001 |
| Skilled (n = 5)  | 1           |             |             |        |
|                  |             |             |             | 0.773  |
| Expert (n = 3)   | 1           |             |             |        |

†Average dermatologist constructed from 1/3 ‘beginner’, 1/3 ‘skilled’ and 1/3 ‘experienced dermatologist’; ‡Expert, dermatologists indicated to have ≥5 years of experience; ‡Skilled, dermatologists indicated to have 2-5 years of experience; ‡Beginner, dermatologists indicated to have <2 years of experience. ‡Diagnostic performance resulting from an assumed collaboration of dermatologists and the CNN: clinical scenario 1) CNN used to verify (or overrule) a dermatologist’s melanoma diagnosis; and scenario 2) CNN used to verify a dermatologist’s benign naevus diagnosis. CNN, convolutional neural networks.
For instance, the specificity of the average dermatologist would be accompanied by a non-ignorable loss of specificity. In contrast, in scenario 2, i.e. using the CNN to verify an initial diagnosis of a benign combined naevus, the sensitivity would increase to 99.9% (95% CI [99.5–100.0]) for all three groups of dermatologists. However, this improvement would be accompanied by a non-ignorable loss of specificity. For instance, the specificity of the average dermatologist would drop from 71.0% to 59.4% (95% CI [47.4–70.3]; Tables 2 and 3).

Saliency maps
To show, which pixels in the image are most important for the CNN’s diagnostic classifications we created saliency maps of the dermoscopic images. By vanilla gradient descent backpropagation, it was shown that the lesions’ pixels are responsible for CNN’s prediction.12

Management decisions of dermatologists
Besides the dermatologists’ binary diagnostic classifications (benign/malignant) we also assessed their management decisions. Dermatologists were offered ‘excision’, ‘short-term follow-up’ or ‘send away/no action needed’ as management decisions in order to reflect options at hand in daily clinical routine. For melanomas, ‘excision’ and ‘short-term follow-up’ were counted as true positives, while in combined naevi ‘send away/no action needed’ and ‘short-term follow-up’ were counted as true negatives. In this setting, all dermatologists performed at a sensitivity of 100.0% (95% CI [99.8–100.0]), however at the cost of a much-decreased specificity (47.6%, 95% CI [36.6–58.8]). As observed for dermatologists’ diagnostic classifications, their experience also had a significant impact on their specificity in management decisions. We observed an increase in specificity from 17.9% (95% CI [10.1–29.6]) in beginners to 42.5% (95% CI [33.7–51.8]) in skilled dermatologists, and 71.8% (95% CI [60.0–81.0]) in experts, all pairwise comparisons being significant (all P ≪ 0.001).

Discussion
Recent advances in the application of deep learning CNNs have fuelled hopes for a new and game-changing strategy in early melanoma detection.4,5,7 However, it is crucial to identify limitations before any large-scale application. We hypothesized that combined naevi that represent one important type of benign ‘melanoma simulators’ might frequently trigger a melanoma diagnosis by the CNN leading to an increase of the false-positive rate. In a clinical routine setting, this scenario could potentially result in increased excision rates of benign naevi and unnecessary concerns of patients receiving a false CNN-based melanoma diagnosis.

For the dichotomous classification of the 72 dermoscopic images (36 combined naevi and 36 melanomas), dermatologists showed a sensitivity, specificity and DOR of 90.6%, 71.0% and 24%, respectively. This diagnostic performance was superior compared with earlier studies showing a dermatologists’ sensitivity for melanoma detection of mostly less than 80%.13,14 In order to create a setting closer to daily clinical practice dermatologists were additionally offered the options ‘excision’, ‘short-term follow-up’ or ‘no action needed’ as management decisions. In this setting, an increased sensitivity of 100% was achieved and indicated that irrespective of the experience level no melanoma was missed by any of the participating dermatologists. In parallel, we observed a considerable decrease in specificity that was mostly caused by beginners extensively using the ‘excision’ option. Not unexpectedly, it appears that beginners attempted to offset their heightened uncertainty by choosing to ‘be on the safe side’ by selecting more often the option ‘excision’ than the skilled or expert group. On the other hand, the less invasive option of a ‘short-term follow-up’ was chosen more often by the experts.

Within this study we compared the diagnostic performance of dermatologists to a CNN with regulatory approval for the European market (MoleAnalyzer-Pro; FotoFinder Systems GmbH). Within the same set of images the CNN revealed an excellent diagnostic performance displayed by a sensitivity of 97.1%, specificity of 78.8% and DOR of 34. However, significance for a statistically meaningful difference between dermatologists and CNN was missed. Using an image set of morphologically matched combined naevi and melanoma the dermatologists’ and CNN’s diagnostic performance was quite convincing and comparable to data published previously by Haenssle et al.4 In this study, 58 dermatologists achieved a sensitivity and specificity of
86.6% and 71.3% while the investigated CNN attained a sensitivity and specificity of 95% and 80%, respectively.6

Interestingly, when assuming the use of the CNN as an assistance system to verify an initial diagnosis of a melanoma we observed an increased specificity of 90.3% at an almost unchanged sensitivity. In this scenario, the group of beginners particularly benefitted from the CNN application as shown by a considerable increase in DOR from 12 to 98.

This cross-sectional study shows several limitations. First, a larger group of dermatologists with various levels of experience would be desirable to allow for a broader generalization of results. Second, while our collection of combined naevi cases with dermoscopic images is the largest reported in the literature, it does not display the full range of potential combinations of naevus types. Particularly, our study included only one case with a Spitz naevus component. This is noteworthy, as one previous study investigating dermoscopic patterns in combined naevi showed that whenever a combined naevus displayed a Spitz naevus component malignancy was suspected.15 Third, in a real-life clinical setting dermatologists would be able to consider additional criteria into decision-making (comparison to other lesions within a patient, history of recent change in a lesion, palpation). Therefore, one might assume a better diagnostic performance of dermatologists in a less artificial setting with access to more information.

In summary, the tested CNN showed a highly accurate diagnostic performance. We provide a first proof-of-principle that an adequately trained CNN may be capable of a precise diagnostic classification even in so-called ‘melanoma simulators’ such as combined naevi. When applied as an assistance system to confirm/overrule a malignant dermatologists’ diagnosis the diagnostic performance of dermatologists could be improved; this particularly holds true for dermoscopically less experienced dermatologists.

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

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

References
1 Arnold M, Holterhaus C, Hollestein I, et al. Trends in incidence and predictions of cutaneous melanoma across Europe up to 2015. J Eur Acad Dermatol Venereol 2014; 28: 1170–1178.
2 Shellenberger R, Nabhani M, Kakaraparthi S. Melanoma screening: a plan for improving early detection. Ann Med 2016; 48: 142–148.
3 Vestergaard M, Macaskill P, Holt P, Menzies S. 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 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 with 58 dermatologists. Ann Oncol 2018; 29: 1836–1842.
5 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542: 115–118.
6 Tschandl P, Rosendahl C, Ayak BN, et al. Expert-level diagnosis of non-pigmented skin cancer by combined convolutional neural networks. JAMA Dermatol 2019; 155: 58–65.
7 Marchetti MA, Codelia 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 2017; 76: 270–277.e1.
8 Scolyer RA, Zhuang L, Palmer AA, Thompson JR, McCarthy SW. Combined naevus: a benign lesion frequently misdiagnosed both clinically and pathologically as melanoma. Pathology 2004; 36: 419–427.
9 Gartmann H, Müller HD. Über das gemeinsame Vorkommen von blauem Návus und Návuszellnávus in ein und derselben Geschwulst (“combined naevus”). Z Hautkr 1977; 52: 389–398.
10 Weedon D, Little JH, Spindle and epitheloid cell naevi in children and adults. A review of 211 cases of the Spitz naevus. Cancer 1977; 40: 217–225.
11 World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 2013; 310: 2191–2194.
12 Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps, 2013. URL https://arxiv.org/abs/1312.6034 (last accessed: 20 August 2019).
13 Dolianitis C, Kelly J, Wolfe R, Simpson P. Comparative performance of 4 dermoscopic algorithms by nonexperts for the diagnosis of melanocytic lesions. Arch Dermatol 2003; 141: 1008–1014.
14 Carl J, 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: 981–984.
15 de Giorgi V, Massi D, Salvini C, Trez E, Mannone F, Carl J. Dermoscopic features of combined melanocytic naevi. J Cutan Pathol 2004; 31: 600–604.

Supporting information
Additional Supporting Information may be found in the online version of this article:
Appendix S1. Methods.
Figure S1. The global Inception_v4 network architecture.