European Journal of Cancer 113 (2019) 47–54

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task

Titus J. Brinker [a,b,*], Achim Hekler [a], Alexander H. Enk [b], Joachim Klode [c], Axel Hauschild [d], Carola Berking [e], Bastian Schilling [f], Sebastian Haferkamp [g], Dirk Schadendorf [c], Tim Holland-Letz [h], Jochen S. Utikal [i,j,1], Christof von Kalle [a,1], Collaborators[2]

[a] *National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 460, 69120 Heidelberg, Germany*
[b] *Department of Dermatology, University Hospital Heidelberg, Heidelberg, Germany*
[c] *Department of Dermatology, University Hospital Essen, Essen, Germany*
[d] *Department of Dermatology, University Hospital Kiel, Kiel, Germany*
[e] *Department of Dermatology, University Hospital Munich (LMU), Munich, Germany*
[f] *Department of Dermatology, University Hospital Würzburg, Würzburg, Germany*
[g] *Department of Dermatology, University Hospital Regensburg, Regensburg, Germany*
[h] *Department of Biostatistics, German Cancer Research Center, Heidelberg, Germany*
[i] *Department of Dermatology, Heidelberg University, Mannheim, Germany*
[j] *Skin Cancer Unit, German Cancer Research Center (DKFZ), Heidelberg, Germany*

Received 15 February 2019; received in revised form 19 March 2019; accepted 2 April 2019
Available online 10 April 2019

KEYWORDS
Melanoma;
Skin cancer;
Artificial intelligence

Abstract ***Background:*** Recent studies have successfully demonstrated the use of deep-learning algorithms for dermatologist-level classification of suspicious lesions by the use of excessive proprietary image databases and limited numbers of dermatologists. For the first time, the performance of a deep-learning algorithm trained by open-source images exclusively is compared to a large number of dermatologists covering all levels within the clinical hierarchy.

Methods: We used methods from enhanced deep learning to train a convolutional neural network (CNN) with 12,378 open-source dermoscopic images. We used 100 images to compare the performance of the CNN to that of the 157 dermatologists from 12 university hospitals in Germany.

* *Corresponding author*: Dr. Titus J. Brinker, National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 460, 69120 Heidelberg, Germany. Tel.: +496221 3219304; fax: +496221 566967.
E-mail address: titus.brinker@dkfz.de (T.J. Brinker).

[1] These authors contributed equally to this work. [2] These collaborators are listed in the acknowledgement section.

https://doi.org/10.1016/j.ejca.2019.04.001

Outperformance of dermatologists by the deep neural network was measured in terms of sensitivity, specificity and receiver operating characteristics.

Findings: The mean sensitivity and specificity achieved by the dermatologists with dermoscopic images was 74.1% (range 40.0%–100%) and 60% (range 21.3%–91.3%), respectively. At a mean sensitivity of 74.1%, the CNN exhibited a mean specificity of 86.5% (range 70.8%–91.3%). At a mean specificity of 60%, a mean sensitivity of 87.5% (range 80%–95%) was achieved by our algorithm. Among the dermatologists, the chief physicians showed the highest mean specificity of 69.2% at a mean sensitivity of 73.3%. With the same high specificity of 69.2%, the CNN had a mean sensitivity of 84.5%.

Interpretation: A CNN trained by open-source images exclusively outperformed 136 of the 157 dermatologists and all the different levels of experience (from junior to chief physicians) in terms of average specificity and sensitivity.

1. Introduction

Skin cancer is the most common malignancy in fair-skinned populations, and melanoma accounts for the majority of skin cancer–related deaths worldwide [1]. Despite special training and the use of dermoscopes, dermatologists only rarely achieve clinical test sensitivities greater than 80% [2]. In 2017, Esteva *et al.* [3] were the first to report a deep-learning convolutional neural network (CNN) image classifier that performed as well as 21 board-certified dermatologists when identifying images with malignant lesions. The CNN deconstructed digital images of skin lesions and generated its own diagnostic criteria for melanoma detection during training. Several follow-up publications by other authors have demonstrated dermatologist-level skin cancer classification by using deep neural networks (CNN) [4–7]. However, these publications involved limited numbers of dermatologists and proprietary image databases and, thus, were neither fully reproducible nor allowed for a fine-grained comparison.

In this work, we trained a CNN with enhanced techniques to classify images of suspect lesions as melanoma or atypical nevi by the use of open-source images exclusively. The classification results of the CNN were compared with the efforts of 157 dermatologists from 12 German university hospitals of all levels of training including a small subsample of resident physicians.

2. Methods

2.1. Data sets

To develop the algorithm, dermoscopic images from melanomas and atypical nevi were obtained from the International Skin Imaging Collaboration (ISIC) image archive [8]. This image archive contained a total of 2169 melanomas and 18,566 atypical nevi as of 17 October 2018. The diagnoses of all melanomas were verified via histopathological evaluation of biopsies. The diagnosis of nevi was made either by histopathological examination (∼24%), expert consensus (∼54%) or by another diagnosis method, such as a series of images that showed no temporal changes (∼22%). All images were anonymous and open source.

To compare the performance of the digital automated diagnosis method with that of dermatologists, a test set with a total of 100 images of melanomas and atypical nevi was created. Using only 100 images allowed for the participation of a large number of dermatologists in the test, in light of the time required to review all the images. To avoid bias in the creation of the test set, we implemented a random generator, which selected 80 test images from all the atypical nevi and 20 test images from all the available melanomas in the ISIC archive. The chosen ratio of the classes was based on the test and training set for the International Symposium on Biomedical Imaging 2016 challenge [8]. While this proportion does not reflect the frequency of diagnosis in clinical practice, statistical quality of the test is enhanced when a sufficient number of melanomas are in the test set.

The training and validation images were also selected using a random generator from the set of available images in the ISIC archive, excluding the already selected test images. The ratio of training and validation data was set as 1:10, and the ratio of the two classes was kept at 1:4. This led to a training set consisting of 1888 melanomas and 10,490 atypical nevi, a validation set including 210 melanomas and 1049 atypical nevi, and a test set containing 20 melanomas and 80 atypical nevi. The test, training and validation sets were disjoint.

2.2. Development of the algorithm

From a mathematical perspective, deep neural networks can be interpreted as functions with millions of freely configurable parameters, called weights. These weights are adjusted for a given image classification task in such a way that the intensities of the pixels in an input image are mapped to a probability of class label. Because of the huge number of free parameters, training these functions requires a large number of images for which the class is already known. For each image, the output of the function is calculated, compared with the given class label and then the weights are slightly modified to reduce the error. This process is repeated many times for each image in the training set, and the function ‘learns’ how to precisely predict the class labels, given only the pixel intensities of each image. By using training data that adequately represent the possible input space, the result is a function that exhibits large generality when predicting the class labels for unknown images. In this work, we used CNNs which are characterised by a specific architecture. In regular neural networks, every weight, except for that of the first layer, is affected by the dependencies of all pixels. In contrast, CNNs first aggregate local adjacent pixels to recognise local features and then combine them into global features. This constraint on local connections results in faster training and lower model complexity.

In this work, a ResNet50 CNN model was used for the classification of melanomas and atypical nevi. The network parameters were initialised using the weights from the same network architecture trained to classify images in the ImageNet data set [9]. Details on the enhanced training procedures can be found in Appendix 1.

2.3. Evaluation of the CNN

The trained CNN outputs a continuous number between 0 and 1 for each input image, which can be interpreted as the probability that a melanoma was present in the input image. For a binary decision task, it is necessary to specify an operating value, that if exceeded, causes the input image to be classified as melanoma. This parameter selection allows the trade-off between sensitivity and specificity to be adjusted. Two operating values for the algorithm were selected; the first operating value approximated the mean specificity of 69.2% achieved by chief physicians on the test set, while the second operating value corresponded to a sensitivity of 76.7% for detecting melanomas, which is a necessary prerequisite for the application of the algorithm as a screening tool. This high sensitivity was achieved, on average, by resident physicians on the test set of 100 dermoscopic images. To evaluate the algorithm, the receiver operating curve (ROC) was plotted by varying the operating value between 0 and 1 and calculating the corresponding sensitivity and specificity.

3. Performance measurement of dermatologists

3.1. Electronic questionnaire

The test set, which consisted of 100 dermoscopic images, was examined by 175 dermatologists from 12 university hospitals in Germany [10]. Only physicians with clinical practice in dermatology participated in this study. Anonymous validation of the test set was performed using an electronic questionnaire. The first part recorded the practitioner’s age, gender, years of dermatologic practice/experience, estimated number of skin checks performed and position within the medical hierarchy. This was followed by the 100 dermoscopic images, with 80 of them being benign nevi and 20 biopsy-verified melanomas. For each image, the participants were asked for a management decision, to either recommend biopsy/further treatment or simply reassure the patient.

3.2. Outlier detection

Data quality is an important issue when using anonymous questionnaires, especially under conditions of obligatory participation. Careless and meaningless responses have to be identified and removed from the data set. In this work, we performed a two-step data cleaning process. To prevent bias in the selection of data entries, statistical methods were applied first. In the second validation step, we looked for contradictions in the respondent metadata. For example, no established physician could have zero years of professional experience.

For statistical outlier detection, we used the local outlier factor (LOF) method [11]. The management decision for each distinct image can be modelled as a categorical binary variable. Therefore, the space of all possible management decisions consisted of 100 dimensions, one for each test image, and each dimension was a discrete-valued variable with two possible values. The LOF algorithm is an unsupervised method that determines the local density deviation of a distinct point with respect to its neighbours. The factor is close to 1.0 if a point is located in a subspace where many other points can be found. In our case, this meant that there were very similar answers from dermatologists who differed only slightly from each other. For respondents who showed large deviations in their answers, the value was significantly larger, indicating the outliers. In this work, we considered the 30 nearest neighbours of each response, but the detected outliers were not dependent on the exact parameter selection.

3.3. Statistical analysis

As sensitivity and specificity of CNN depend on the chosen cut-off, these values could not be compared individually between methods. Instead, the ‘Youden index’ (YI = sensitivity + specificity-1) was compared, evaluated primarily at the cut-off of sensitivity 74.1%. Differences were tested for significance with a two-sided two-sample binomial test using the normal distribution approximation. The significance level was set as $\alpha = 0.05$.

3.4. Ethical approval

The ethical committee of the University of Heidelberg waived the need for ethical approval because all the dermatologists voluntarily participating in the reader study were anonymous and the training of an artificial intelligence algorithm was conducted with open-source images.

4. Results

4.1. Performance of dermatologists

Of 175 dermatologist-created data sets, 18 outliers were detected by the LOF method, which represented 10.3% of all entries. This value agrees in the order of magnitude with previous studies in the literature. Maniaci *et al.* found that about 3–9% of respondents to a questionnaire did not answer the questions carefully at all [12]. For validation of the chosen outlier detection method, we checked the provided metadata for contradictions. For five entries, the supplied information was considered very doubtful. All these suspicious entries had been detected by the LOF method as outliers, so we considered the outlier detection to be suitable. Finally, all 18 outliers were removed from the data set, and the valid answers of 157 dermatologists remained. In this set, 56 (35.7%) were male and 101 (64.3%) were female. The median of years of experience is 4 years, and the distribution for the participants is shown in Fig. 1.

Of the participants, 56.1% were junior physicians (dermatologic residents) and 43.9% were board certified. In addition to the 151 (96.2%) physicians practicing in hospitals, there were also six (3.8%) dermatologic resident physicians working in a private office. The performances of the dermatologists, expressed as various features, are summarised in Table 1.

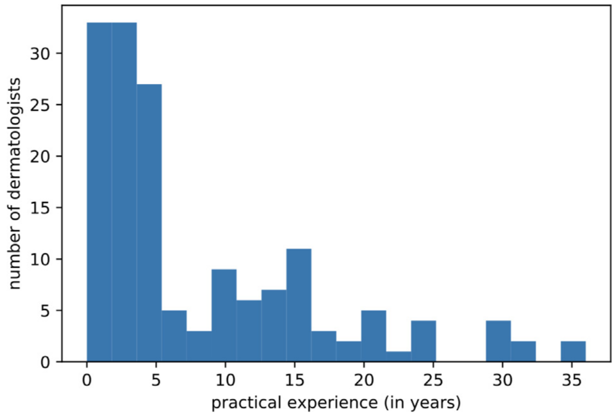

Fig. 1. Distribution of years of experience for participating dermatologists.

Table 1
Diagnostic performance of the dermatologists for the test set of 100 dermoscopic images.

Sample	Sensitivity	Specificity	ROC
All participants (n=157)	74.1%	60.0%	0.671
University hospital (n = 151)	74.0%	59.8%	0.669
Resident physicians (n = 6)	76.7%	65.8%	0.713
Position in hospital hierarchy			
Junior physicians (n = 88)	74.8%	58.2%	0.665
Attendings (n = 15)	72.7%	60.0%	0.664
Senior physicians (n = 45)	73.0%	62.3%	0.677
Chief physicians (n = 3)	73.3%	69.2%	0.713

4.2. Statistical analysis and performance comparison

The mean receiver operating characteristic (ROC) curve over all 10 runs is shown in Fig. 2 (blue line) in comparison to the 157 dermatologists (red dots).

The mean sensitivity and specificity of the dermatologists was 74.1% (range 40.0%–100%) and 60% (range 21.3%–91.3%), respectively (YI = 0.34). At a mean sensitivity of 74.1%, the CNN had a mean specificity of 86.5% (range 70.8%–91.3%, YI = 0.61). Compared with the dermatologists, this is a relevant but not significant difference (p = 0.31). For a mean specificity of 60%, a mean sensitivity of 87.5% (range 80%–95%, YI = 0.48) was achieved by our algorithm.

The average performance of the physicians from all different levels of hierarchy within dermatology (from junior physicians to chief physicians) is shown in Fig. 3. An outperformance of all of these subgroups in terms of average results was achieved by our algorithm.

The two operating values of the algorithm, the sensitivity and specificity, were calculated with respect to the class labels documented in the ISIC archive. Using the first operating value at high specificity, approximating the high mean specificity of chief physicians for the test set, the algorithm's mean sensitivity was 84.5%. This value outperformed the chief physicians' corresponding mean sensitivity of 73.3%.

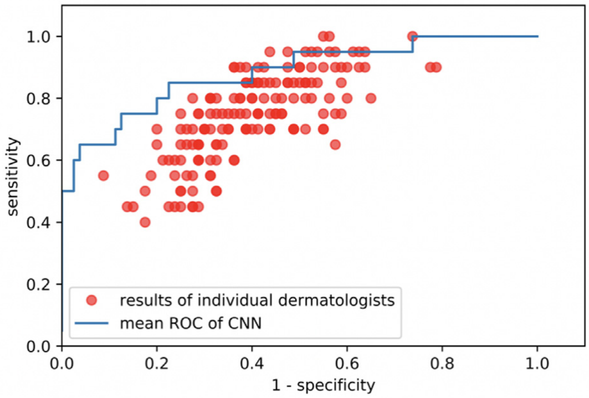

Fig. 2. The mean receiver operating characteristic (ROC) curve over all 10 runs. CNN, convolutional neural network.

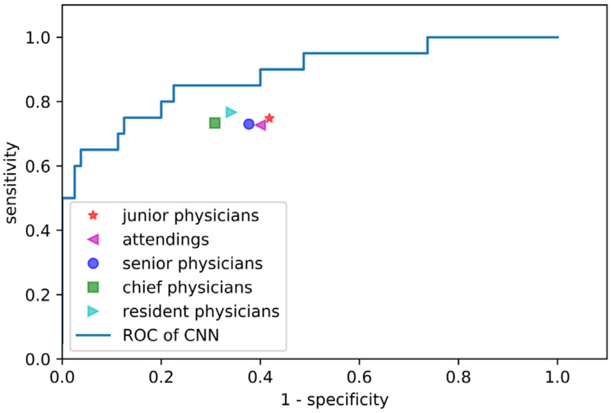

Fig. 3. The average performance of the physicians from all different levels of hierarchy within dermatology (from junior physicians to chief physicians). CNN, convolutional neural network.

A second operating value for the algorithm was evaluated, based on the high sensitivity of resident physicians. Using this operating value, the algorithm had a sensitivity of 76% and a specificity of 81.7%, on average. Compared with the results of the resident physicians, who achieved a mean sensitivity of 67.7% and a mean specificity of 65.8% on the test set, the mean specificity of the CNN was better by 15.9 percentage points at approximately the same sensitivity.

Fig. 4 shows all lesions in which the majority of human raters and the majority of CNN-test runs disagreed:

5. Discussion

A CNN that was trained with open-source images exclusively was capable to outperform dermatologists of all hierarchical categories of experience (from junior to chief physicians) in dermoscopic melanoma image classification. Only seven of 157 dermatologists had better corresponding values for specificity and sensitivity than the CNN. Previous landmark publications comparing the performance of a CNN to dermatologists involved 8, 21 or 58 dermatologists [3–5]. This study exceeds these numbers significantly by including 157 dermatologists from 12 German university hospitals. This allows for a more fine-grained comparison with higher external validity which encompasses all hierarchical positions in the landscape of dermatologic experience and expertise. In addition, all the cited publications used proprietary images from large archives of dermatologic departments [3–5] and, thus, could not be reproduced publicly because the training and the test set images were not made publicly available. Because we only used open-source images and provide our test set as an appendix of this publication while disclosing the full training procedure of our algorithm, our experiment is entirely reproducible (Appendix 2).

A CNN for the diagnosis of melanocytic lesions offers many advantages, including consistent interpretation—because the CNN assigns a distinct class to each specific image every time—and more accurate diagnoses than human experts of all levels of training. Additionally, by setting the operational value, the trade-off between sensitivity and specificity can be adapted to the requirements of the specific clinical setting. For example, in a screening setting, high sensitivity is desired, so the operating value can be decreased accordingly. Fig. 4 illustrates the lesions on which the majority of dermatologists and the majority of CNN-test runs disagreed: CNNs and humans apply different techniques for identifying melanoma which could complement each other in order for more accurate diagnoses in the form of assistant systems.

When analysing the results of dermatologists based on their positions in the clinical hierarchy, it is noticeable that junior physicians showed high sensitivity but low specificity. They tend to overdiagnose lesions to miss as few melanomas as possible. With high-ranking hospital respondents who had more years of professional experience, the specificity increased substantially, while the sensitivity remained approximately the same.

In contrast with previous publications [3–5] that compared the performance of a CNN with that of dermatologists, our study reports the stochastic nature of the result. We believe that it is mandatory to describe the overall performance of an algorithm, because the training and evaluation procedure of a CNN includes stochastic components, such as the random splitting of training and validation images, stochastic gradient descent, and random initialisation of the parameters.

When comparing the results of different training runs, it is notable that the quality of the classification differed only slightly. In contrast, the performances of dermatologists showed large variance.

There are some limitations to this system. It remains an open question whether the design of the questionnaire had any influence on the performance of the dermatologists compared with clinical settings. Furthermore, clinical encounters with actual patients provide more information than that can be provided by images alone. Hänßle *et al.* showed that additional clinical data improve the sensitivity and specificity of dermatologists slightly [5]. Machine learning techniques can also include this information in their decisions. However, even with this slight improvement, the CNN would still outperform the dermatologists.

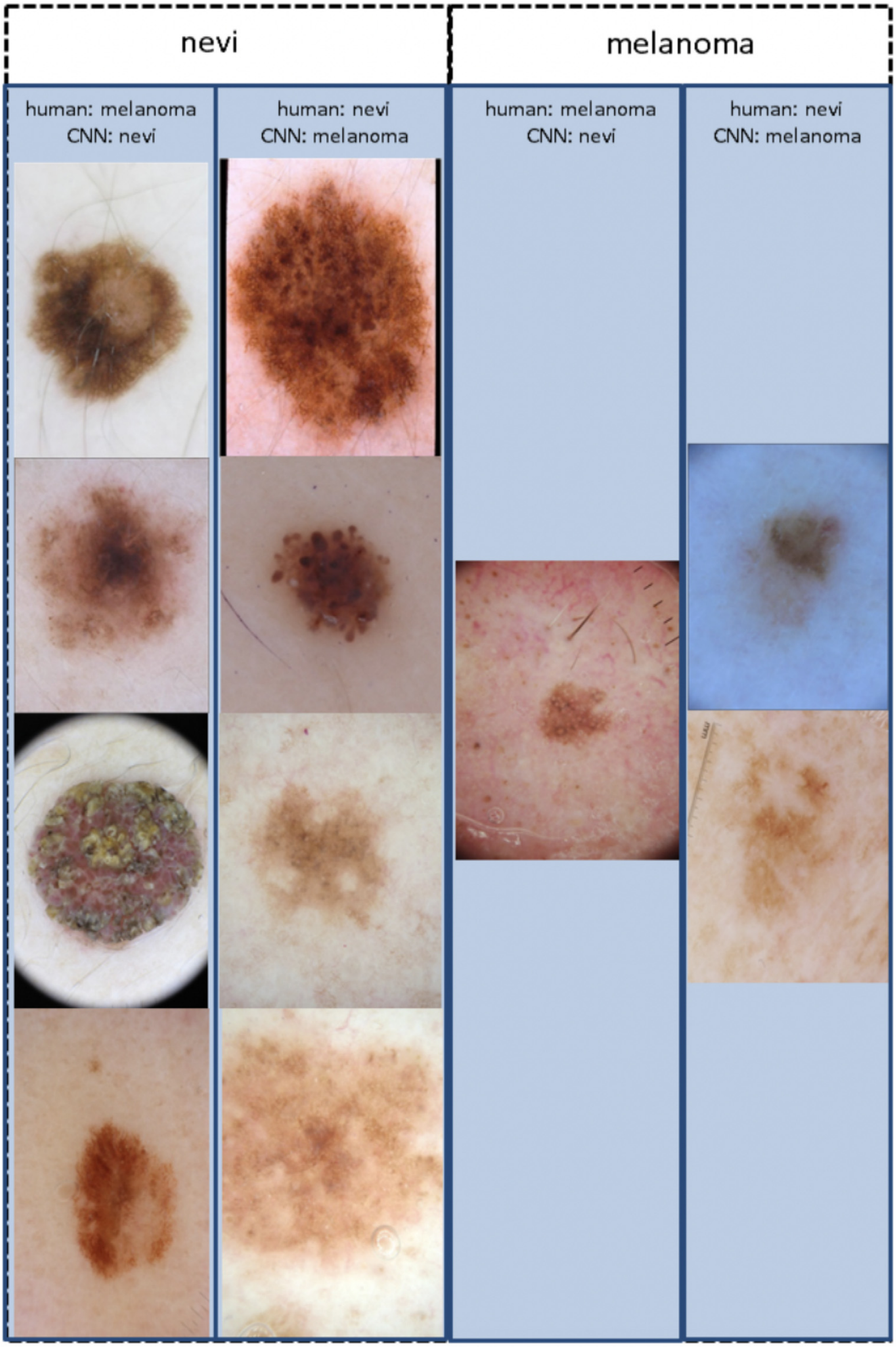

Fig. 4. Lesions on which the majority of human raters and the majority of CNN test runs disagreed. CNN, convolutional neural network.

6. Conclusion

A CNN that was trained with open-source images exclusively was capable to outperform dermatologists of all hierarchical categories of experience (from junior to chief physicians) in dermoscopic melanoma image classification. Our findings suggest that artificial intelligence algorithms may successfully assist dermatologists with melanoma detection in clinical practice which needs to be carefully evaluated in prospective trials.

Funding

No external funding was obtained.

Conflict of interest statement

The authors declare no competing interests with relevance to this study.

Acknowledgements

This work is part of the Skin Classification Project which is funded by the Federal Ministry of Health in Germany. The grant is held by Dr. Titus J. Brinker (principal investigator). The authors would like to thank and acknowledge the dermatologists who actively and voluntarily spend much time to participate in the reader study (=claimed to have filled out the anonymous questionnaire with 100 dermoscopic images); some participants did not ask to be mentioned despite their declared participation and the authors also thank these colleagues for their commitment. Berlin (Charité): Wiebke Ludwig-Peitsch; Bonn: Judith Sirokay; Erlangen: Lucie Heinzerling; Essen: Magarete Albrecht, Katharina Baratella, Lena Bischof, Eleftheria Chorti, Anna Dith, Christina Drusio, Nina Giese, Emmanouil Gratsias, Klaus Griewank, Sandra Hallasch, Zdenka Hanhart, Saskia Herz, Katja Hohaus, Philipp Jansen, Finja Jockenhöfer, Theodora Kanaki, Sarah Knispel, Katja Leonhard, Anna Martaki, Liliana Matei, Johanna Matull, Alexandra Olischewski, Maximilian Petri, Jan-Malte Placke, Simon Raub, Katrin Salva, Swantje Schlott, Elsa Sody, Nadine Steingrube, Ingo Stoffels, Selma Ugurel, Anne Zaremba. Hamburg: Christoffer Gebhardt, Nina Booken, Maria Christolouka; Heidelberg: Kristina Buder-Bakhaya, Therezia Bokor-Billmann, Alexander Enk, Patrick Gholam, Holger Hänßle, Martin Salzmann, Sarah Schäfer, Knut Schäkel, Timo Schank; Kiel: Ann-Sophie Bohne, Sophia Deffaa, Katharina Drerup, Friederike Egberts, Anna-Sophie Erkens, Benjamin Ewald, Sandra Falkvoll, Sascha Gerdes, Viola Harde, Axel Hauschild, Marion Jost, Katja Kosova, Laetitia Messinger, Malte Metzner, Kirsten Morrison, Rogina Motamedi, Anja Pinczker, Anne Rosenthal, Natalie Scheller, Thomas Schwarz, Dora Stölzl, Federieke Thielking, Elena Tomaschewski, Ulrike Wehkamp, Michael Weichenthal, Oliver Wiedow; Magdeburg: Claudia Maria Bär, Sophia Bender-Säbelkampf, Marc Horbrügger, Ante Karoglan, Luise Kraas Mannheim: Jörg Faulhaber, Cyrill Geraud, Ze Guo, Philipp Koch, Miriam Linke, Nolwenn Maurier, Verena Müller, Benjamin Thomas, Jochen Sven Utikal; Munich: Ali Saeed M. Alamri, Andrea Baczako, Carola Berking, Matthias Betke, Carolin Haas, Daniela Hartmann, Markus V. Heppt, Katharina Kilian, Sebastian Krammer, Natalie Lidia Lapczynski, Sebastian Mastnik, Suzan Nasifoglu, Cristel Ruini, Elke Sattler, Max Schlaak, Hans Wolff; Regensburg: Birgit Achatz, Astrid Bergbreiter, Konstantin Drexler, Monika Ettinger, Sebastian Haferkamp, Anna Halupczok, Marie Hegemann, Verena Dinauer, Maria Maagk, Marion Mickler, Biance Philipp, Anna Wilm, Constanze Wittmann and Würzburg: Anja Gesierich, Valerie Glutsch, Katrin Kahlert, Andreas Kerstan, Bastian Schilling and Philipp Schrüfer.

Titus J. Brinker, Jochen Sven Utikal and Achim Hekler had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Appendix 1

As described in the summary of our publication, the weights were slightly modified during training to reduce the loss. The loss is mathematically described by a function that models the difference between the class labels predicted by the function for a given parameter setting and actual class labels. The learning rate is a hyperparameter that controls how much these adjustments are made with respect to the gradient of the loss function. In contrast with existing approaches that apply the same learning rate to all layers of the convolutional neural network (CNN), we used different learning rates for each layer. In particular, slower learning rates were used for layers closer to the input, whereas faster learning rates were used for layers closer to the output. The intuition behind this enhanced technique, which is called differential learning rates, is that the earlier layers contain more general features, such as edges or gradients. Therefore, their weights do not need to be changed significantly for the new classification task. Thus, the learning rates for the earlier layers are set to low values, resulting in a moderate adjustment of the corresponding weights. In contrast, the later layers contain application-specific features. Consequently, these layers are assigned higher learning rates, which causes the corresponding weights to be modified more in relation to each other compared with the weights of the early layers. To realise this concept, we split the layers into three groups and applied a different learning rate for each group. The first six residual units had a learning rate of 0.009, the subsequent eight residual blocks had a value of 0.003 and the fully connected layers used 0.01. The selection of the specific learning rates was based on practical experience with other image classification tasks.

For each adjustment during training, the parameters normally approach a minimum in the loss function. As the model gets closer to the minimum, it is a common practice to decrease the learning rate stepwise so that the optimisation settles as close as possible to the minimum, instead of overshooting it. In this article, we used a cosine annealing method, which decreases the learning rate based on a cosine function.

The third enhanced training technique addressed the problem that the optimisation process can get stuck in a local, rather than a global, minimum. To overcome this problem, the learning rate was suddenly increased at some specific time steps, and thus the optimisation process may be able to escape a local minimum and reach the global minimum. This technique is called stochastic gradient descent with restart (SGDR), an idea shown to be highly effective by Loshchilov *et al.* [1].

To document the performance of the algorithm and the enhanced training techniques as accurately as possible, we retrained the CNN a total of 10 times, and each training run consisted of 13 epochs.

1. Loshchilov I, Hutter F. Stochastic gradient descent with warm restarts. 2016:2–8; https://arxiv.org/abs/1608.03983.

Appendix A. Supplementary data

Test set of 100 dermoscopic images (the use of the test-set requires the citation of this article). Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejca.2019.04.001.

References

[1] Schadendorf D, van Akkooi AC, Berking C, Griewank KG, Gutzmer R, Hauschild A, et al. Melanoma. Lancet 2018;392(10151):971–84.

[2] Carli P, Quercioli E, Sestini S, Stante M, Ricci L, Brunasso G, 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):981–4.

[3] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115.

[4] Marchetti MA, Codella NC, Dusza SW, Gutman DA, Helba B, Kalloo A, et al. Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. J Am Acad Dermatol 2018;78(2):270–7. e271.

[5] Haenssle H, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018;29(8):1836–42.

[6] Brinker Titus J, Hekler Achim, Enk Alexander H, Klode Joachim, Hauschild Axel, Berking Carola, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. Eur J Cancer 2019;111:148–54.

[7] Brinker TJ, Hekler A, Utikal JS, Grabe N, Schadendorf D, Klode J, et al. Skin cancer classification using convolutional neural networks: systematic review. J Med Internet Res 2018;20(10):e11936.

[8] Gutman D, Codella NC, Celebi E, Helba B, Marchetti M, Mishra N, et al. Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC). 2016. arXiv preprint arXiv:160501397.

[9] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. Int J Comput Vis 2015;115(3):211–52.

[10] Brinker Titus J, Hekler Achim, Hauschild Axel, Berking Carola, Schilling Bastian, Enk Alexander H, et al. Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark. Eur J Cancer 2019;111:30–7.

[11] Breunig MM, Kriegel H-P, Ng RT, Sander JLOF. Identifying density-based local outliers. In: ACM sigmod record: 2000. ACM; 2000. p. 93–104.

[12] Maniaci MR, Rogge R. Caring about carelessness: participant inattention and its effects on research. J Res Personal 2014;48:61–83.