European Journal of Cancer 119 (2019) 11–17

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

Deep neural networks are superior to dermatologists in melanoma image classification

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

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

Received 27 April 2019; received in revised form 24 May 2019; accepted 28 May 2019
Available online 8 August 2019

KEYWORDS
Deep learning;
Melanoma;
Skin cancer;
Artificial intelligence

Abstract ***Background:*** Melanoma is the most dangerous type of skin cancer but is curable if detected early. Recent publications demonstrated that artificial intelligence is capable in classifying images of benign nevi and melanoma with dermatologist-level precision. However, a statistically significant improvement compared with dermatologist classification has not been reported to date.

Methods: For this comparative study, 4204 biopsy-proven images of melanoma and nevi (1:1) were used for the training of a convolutional neural network (CNN). New techniques of deep learning were integrated. For the experiment, an additional 804 biopsy-proven dermoscopic images of melanoma and nevi (1:1) were randomly presented to dermatologists of nine German university hospitals, who evaluated the quality of each image and stated their

* *Corresponding author*. National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 460, Heidelberg, 69120, Germany. Fax: +496221 3219304.
E-mail address: titus.brinker@dkfz.de (T.J. Brinker).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ejca.2019.05.023

recommended treatment (19,296 recommendations in total). Three McNemar's tests comparing the results of the CNN's test runs in terms of sensitivity, specificity and overall correctness were predefined as the main outcomes.

Findings: The respective sensitivity and specificity of lesion classification by the dermatologists were 67.2% (95% confidence interval [CI]: 62.6%–71.7%) and 62.2% (95% CI: 57.6%–66.9%). In comparison, the trained CNN achieved a higher sensitivity of 82.3% (95% CI: 78.3% –85.7%) and a higher specificity of 77.9% (95% CI: 73.8%–81.8%). The three McNemar's tests in 2 × 2 tables all reached a significance level of $p < 0.001$. This significance level was sustained for both subgroups.

Interpretation: For the first time, automated dermoscopic melanoma image classification was shown to be significantly superior to both junior and board-certified dermatologists ($p < 0.001$).

1. Background

Melanoma is accountable for the most skin cancer–related deaths, and early detection is the most relevant prognostic factor for survival [1]. In Western countries, melanoma is primarily detected via dermoscopy. However, the sensitivity of dermoscopic melanoma detection is mostly less than 80% for dermatologists in routine clinical settings [2]. Thus, new diagnostic tools that assist the dermatologists' diagnosis should be developed, evaluated and optimised.

Recent studies in digital skin diagnosis have used convolutional neural networks (CNNs) to classify images of melanoma and nevi with accuracies comparable with those achieved by dermatologists [3–5]. When training their algorithms, the prior studies used large numbers of images confirmed by consensus decisions. When using images confirmed in this manner, there is a high risk that the CNN will learn the decision-making process of dermatologists, including all possible misjudgements. In contrast, the purpose of this study was to demonstrate the first systematic outperformance of (board-certified) dermatologists by training our CNN with biopsy-verified images exclusively and using new techniques of enhanced deep learning.

2. Methods

2.1. Study design

This comparative study was conducted from 20th September 2018 (design of the study) to 20th February 2019 (completion of data analysis). The completion of the anonymous electronic questionnaires was undertaken from 5th December 2018 to 18th December 2018. The inclusion of dermatologists was conducted via randomly assigned links to the department directors of 9 university hospitals who would send two questionnaires to their employed dermatologists via the official university e-mail accounts. Ethical approval was waived by the Ethics Committee of the University of Heidelberg owing to the anonymity of the survey and the dermatologic images.

2.2. Training of the CNN

We used a pretrained [6] ResNet50 CNN [7]. To adapt the CNN for the classification of our test set, 4204 open-source and biopsy-proven images (1:1 = melanoma:-nevi) from the International Skin Imaging Collaboration image archive were used; this number of images is two times fewer than that used in prior studies [8]. For evaluation of the CNN, a test set of 804 biopsy-proven test images (melanoma:nevi = 1:1) was generated, which was separate from the training set.

In contrast to existing works on melanoma classification, we used additional state-of-the-art training techniques:

1. Differential learning rates, rather than the same learning rate for all layers [9].
2. Reduction of the learning rate based on a cosine function [17].
3. Stochastic gradient descent with restart, to avoid local minima [17].

For more technical details, please see Appendix 1.

2.3. Comparison with dermatologists

The test set images were sent to dermatologists from nine German university hospitals via six randomly assigned electronic questionnaires, each containing 134 images of different skin lesions. The dermatologists were informed about the composition of the images (1:1 = melanoma/nevi) and were asked to both check the quality of the images and decide to either biopsy/treat the lesion or reassure the patient. No incentives were offered for participation; however, the

dermatologists were encouraged to test their knowledge. In total, the six questionnaires were completed 144 times (19,296 images were evaluated); 52 questionnaires were filled out by board-certified dermatologists (evaluation of 6968 images), and 92, junior dermatologists (evaluation of 12,328 images); each dermatologist was provided a maximum of two questionnaires. Each of the 804 individual images was evaluated by an average of 21.3 dermatologists (median = 21; standard deviation = 4.8; range = 4–31). Only images with at least 'sufficient' image quality, as rated by the participating dermatologists, were included in this study. In this study, we consider the image quality to be sufficient if the corresponding image is rated as 'excellent', 'good', or 'sufficient'. We excluded 11.1% of the answers owing to poor image quality as determined by the participating dermatologists. To ensure fair comparisons between the results determined by dermatologists and those determined by the CNN, we conducted 144 runs of the CNN; each test set of each CNN run corresponded exactly to the images rated as 'sufficient' by the dermatologists. A sample was regarded as a melanoma diagnosis by the CNN if the average melanoma probability from all runs was $\geq$50%. Equally, a sample was regarded as a melanoma diagnosis by the dermatologists if the majority selected that classification. Fig. 1 shows example images of skin lesions assigned to different classes by the dermatologists and the CNN.

2.4. Statistical analysis

To quantitatively evaluate the quality of the CNN classification and the performance of the dermatologists, images with known class labels were used to compare the class label assigned by the classifier with the actual class (as determined by biopsy).

Sensitivity and specificity were calculated separately for the summary decisions of the CNN and the dermatologists; exact binomial 95% confidence intervals (CIs) were calculated for sensitivity and specificity. Both sensitivity and specificity are statistical values that depend on the same configurable parameter, namely, the cut-off value. In a binary classification problem, this scalar value determines from which output value of the CNN the input is assigned to which class. The default value is normally 0.5, but can be adjusted to the respective question. If this parameter is lowered, the sensitivity increases and the specificity decreases and vice versa. The receiver operating characteristic (ROC) curve visualises this relationship. Sensitivity, specificity and overall rates of correct classifications (primary end-point) were compared statistically using three separate two-sided McNemar's tests in 2 × 2 tables. For the comparison of overall correctness, a joint 2 × 2 table was generated, which included all samples (melanoma and nevi) and showed the numbers of samples where none, one or both methods produced a correct diagnosis. The significance level was set at alpha = 5%. Sample size considerations can be found in Appendix.

All analyses were programmed via a Jupyter Notebook in Python.

3. Results

The confusion matrices of melanoma and nevi classifications in the test set are shown in Fig. 2.

The sensitivity and specificity for classification by the dermatologists were 67.2% (95% CI: 62.6–71.1%) and 62.2% (95% CI: 57.6–66.9%), respectively. The board-certified dermatologists achieved a sensitivity and specificity of 63.2% (95% CI: 58.7–68.1%) and 65.2% (95% CI: 60.5–69.8%), respectively. In contrast, the classification results of the junior physicians showed a higher sensitivity of 68.9% (95% CI: 64.4–73.4%), whereby the specificity with 58% (95% CI: 53.1–62.8%) is lower. The trained CNN achieved a sensitivity of 82.3% (95% CI: 78.3–85.7%) and specificity of 77.9% (95% CI: 73.8–81.8%).

Fig. 3 shows the average ROC curve of the CNN.

Sensitivity, specificity and overall rates of correct classifications were compared statistically using three separate (two-sided) McNemar's tests in 2 × 2 tables. For the comparison of overall correctness (primary objective), a joint 2 × 2 table was generated, which included all samples (melanoma and nevi) and showed the numbers of samples where none, one or both methods produced a correct diagnosis. Two additional tables were created showing only melanoma and only nevi, respectively, and sensitivity and specificity were calculated and compared based on these tables (secondary objective). The significance level was set at alpha = 5%. The CNN (trained exclusively with open-source images) outperformed our sample of dermatologists ($p < 0.001$). If board-certified dermatologists and junior physicians were considered separately in the statistical test, the CNN also showed a significant outperformance for both cases (McNemar's $p < 0.001$). The secondary classification results regarding both sensitivity and specificity showed a p-value <0.001.

4. Discussion

For the first time, automated dermoscopic melanoma image classification was shown to be significantly superior to both junior and board-certified dermatologists ($p < 0.001$).

Past studies used non–biopsy-verified images for training and would often calculate the overall performance on the basis of the sensitivities and specificities of the individual dermatologists or not disclose the composition of images but train their CNN with this exact composition [8]. These points were addressed by

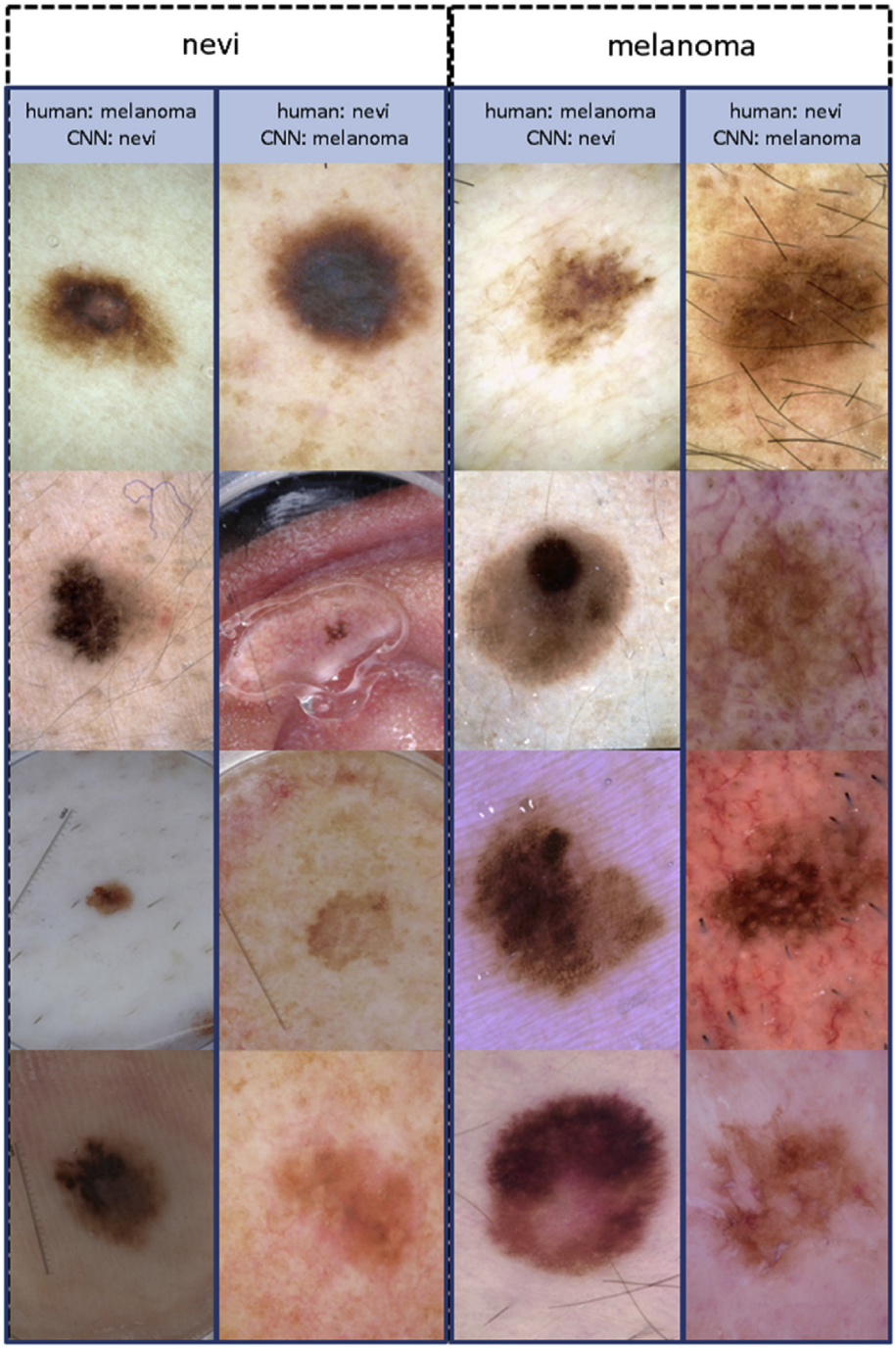

Fig. 1. Example images from the test set that were classified differently by the majority of dermatologists and the convolutional neural network (CNN).

our design; majority decisions on images and the rating of quality of all the test set images enable a high reliability on the answers and minimise the impact of redundancy [10]. The use of biopsy-verified images exclusively allowed systematic outperformance of decisions made by board-certified dermatologists. The composition of the images was disclosed to the dermatologists before answering the survey.

Moreover, the present study is the first conducting not only one dedicated training and test run but also the same number of test runs as performed by the dermatologists. To allow the reproducibility of our results and enable other groups to compare their algorithms, we provide the test set, the underlying ground truth per image and the majority answers per image for public use (Appendix 2).

melanoma — all dermatologists

CNN	n	m
n	38	33
m	94	237

nevi — all dermatologists

CNN	n	m
n	207	106
m	43	46

melanoma — board-certified dermatologists

CNN	n	m
n	38	32
m	109	223

nevi — board-certified dermatologists

CNN	n	m
n	209	90
m	53	50

melanoma — junior physician

CNN	n	m
n	41	30
m	84	247

nevi — junior physician

CNN	n	m
n	194	120
m	39	49

Fig. 2. Confusion matrices on the test set for melanoma and nevi classification. The overall result is listed, as well as the results of the board-certified dermatologists and the junior physicians. On the left side, the classification results (n = nevi, m = melanoma) of dermatologists and the CNN for the 402 biopsy-proven melanoma are shown. If the entirety of the dermatologists in the survey is considered, 237 melanomas are classified correctly, 38 melanomas are misclassified as nevi by both classifiers. Ninety-four melanomas are detected by the CNN, with dermatologists misclassifying these skin lesions as nevi. Thirty-three melanomas are correctly classified by the majority of dermatologists, whereby these skin lesions are not detected as melanoma by the CNN. The other matrices should be read in the same way. CNN, convolutional neural network.

Past research conducted by the authors consists of (a) a melanoma classification benchmark for both clinical and dermatoscopic images together with the sensitivity and specificity of 157 German dermatologists [11], (b) a CNN trained on the basis of dermoscopic images but tested with the benchmark of clinical images [12] and (c) a (non-systematic) outperformance of 136 of 157 of the dermatologists for dermoscopic images [13]. In all preceding publications, non–biopsy-verified images were used for training. In addition, our sample size calculation revealed that the test set used in the previously published benchmark consisting of only 20 melanomas and 80 nevi is too small to demonstrate systematic outperformance with high external validity [11]. Therefore, the test set in this work consists of 402 melanomas and 402 nevi.

The diagnostic performance of dermatologists was lower than that typically found in past research owing to the fact that all images in our test set were biopsy verified (and therefore all suspicious of melanoma).

Our findings add to a growing body of literature demonstrating that in modern CNN architectures, large numbers of images are not needed for training to achieve high accuracies for classification but rather the quality of the training data is important [14]. In computer vision, this is mostly attributable to enhanced data extraction features of modern CNNs. Recent research indicates that CNNs are also capable of enhancing the precision of histopathological melanoma diagnoses and may predict a nevis' oncologic transformation [18–21].

In our reader study, the sensitivity of junior physicians is higher than that of more experienced colleagues, whereas the specificity is substantially lower, which are both confirmatory to our recently published benchmark [11]. The authors hypothesise that the higher sensitivity is mainly due to the fact that in case of doubt, the junior

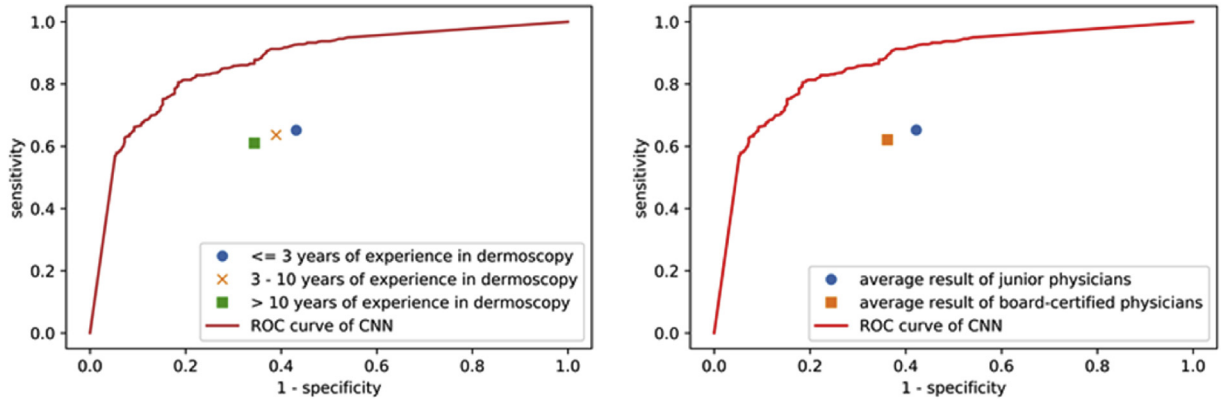

Fig. 3. Mean receiver operating characteristic (ROC) curve of the convolutional neural network (red): the dots represent the average group results of dermatologists differentiated by years of experience in dermoscopy (left panel) and the position in hierarchy (right panel). Physicians with less professional experience show a slightly higher sensitivity than their more experienced colleagues. In case of doubt, they therefore tend to classify a skin lesion as melanoma to avoid leaving a malignant case untreated. With increasing professional experience, nevi are classified more precisely, resulting in a higher specificity. CNN, convolutional neural network.

physicians classify lesions as malignant rather than benign to not miss any melanoma owing to less clinical experience and confidence.

4.1. Limitations

The main limitation of this study is the lack of clinical information on the images, which has however shown to improve the answers of dermatologists only slightly [4] and thus would be unlikely to preclude significance. In addition, it should be mentioned that the decision of the developed algorithm is binary and thus does not reflect clinical practice with many options to take into account for the differential diagnosis. As a consequence, the use of current binary melanoma classification algorithms should be regarded as an assisting tool for dermatologists that may improve accuracy but not as a replacement for independent diagnoses without a supervising dermatologist. In addition, the setting of the investigation does not reflect clinical practice in which one may ask questions to the patient and which allows careful palpation of the skin lesion as additional diagnostic information (i.e. how soft/hard a lesion feels). The clinical routine may not be reflected by a study conducted in front of the computer. However, the most important diagnostic aspect for clinicians in dermoscopy is visual.

Another limitation may be that we chose a ratio of 1:1 (melanoma:nevi) in the test set for statistical reasons, which does not reflect clinical practice. Accordingly, not disclosing the ratio to the participating dermatologists could have led to a systematic bias. However, we made sure that all participants were informed of this ratio before answering the survey (reading about the ratio was mandatory before starting to answer the questionnaire), and thus, the likelihood of this ratio to have an impact on the reader study is reduced.

The images used in the test set were all biopsy verified; however, they were not verified by an independent pathology review panel, which would have improved the underlying ground truth [15].

The diagnostic performance of dermatologists was lower than that typically found in past research owing to the fact that all images in our test set were biopsy verified (and therefore all suspicious of melanoma) [16].

4.2. External validity of the algorithm

The classifier's performance was established on a test-set disjunct from the training and validation set. However, the test images originated from the same overall dataset which was used for training (ISIC), thus raising concern about the classifier's ability to generalise on a truly external test set (i.e. a set of images where a subset was not used for training/validation). A valid concern as factors intrinsic to the training dataset (e.g. type of dermatoscope, lighting or pre-processing) could be picked up during training and result in the network better classifying images sharing these intrinsic factors.

In a preliminary study, a binary-classification CNN (naevus vs melanoma), trained on ISIC images, showed good performance on an ISIC test set but performed worse on an external test set from the PH2 dermoscopic image database [22]. Using just 100 images from the external test set for fine-tuning the CNN (training of the last fully-connected layers), sufficed to completely restore performance.

This specific limitation needs further investigation evaluating points such as

1) whether this is a general phenomenon or occurred due to overtraining on ISIC,
2) is fine-tuning an option for every external set and
3) is there a transferable fine-tuning procedure?

5. Conclusions

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 clinical trials. Future research should test our results in a clinical setting with patients at hand. We suggest implementation after the clinical diagnosis is made by the dermatologist to avoid bias.

Funding

This work is part of the Skin Classification Project that is funded by the Federal Ministry of Health in Germany. The grant is held by Carola Berking, Dirk Schadendorf and Titus J. Brinker (principal investigator).

Conflict of interest statement

None declared.

Acknowledgements

The authors would like to thank all participating dermatologists.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejca.2019.05.023.

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] Vestergaard M, Macaskill P, Holt P, et al. 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;vol. 159(3):669–76.
[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] Haenssle HA, 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.
[5] 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.
[6] 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.
[7] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition: 2016; 2016. p. 770–8.
[8] 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 Intern Res 2018; 20(10):e11936.
[9] Howard J, Ruder S. Universal language model fine-tuning for text classification. arXiv preprint arXiv:180106146. 2018.
[10] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs 2016;316(22):2402–10.
[11] Brinker TJ, Hekler A, Hauschild A, Berking C, Schilling B, Enk AH, et al. Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark. Eur J Canc 2019;111:30–7.
[12] Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task 2019;111:148–54.
[13] Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur J Canc 2019;113:47–54.
[14] Liu Y, Kohlberger T, Norouzi M, Dahl G, Smith J, Mohtashamian A, et al. Artificial intelligence-based breast cancer nodal metastasis detection. Arch Pathol Lab Med 2018.
[15] Elmore JG, Barnhill RL, Elder DE, Longton GM, Pepe MS, Reisch LM, et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. BMJ 2017;357:j2813.
[16] 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(3):669–76.
[17] Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 2018;5:180161.
[18] Sondermann W, Utikal JS, Enk AH, Schadendorf D, Klode J, Hauschild A, et al. Prediction of melanoma evolution in melanocytic nevi via artificial intelligence: a call for prospective data. Eur J Cancer 2019;119:30–4.
[19] Maron RC, Weichenthal M, Utikal JS, Hekler A, Berking C, Hauschild A, et al. Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. Eur J Cancer 2019;119:57–65.
[20] Hekler A, Utikal JS, Enk AH, Solass W, Schmitt M, Klode J, et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. Eur J Cancer 2019; 118:91–6.
[21] Hekler A, Utikal JS, Enk AH, Hauschild A, Weichenthal M, Maron RC, et al. Superior skin cancer classification by the combination of human and artificial intelligence. Eur J Cancer 2019 [in press].
[22] Teresa Mendonça, Ferreira Pedro M, Marques Jorge S, Marçal André RS, Rozeira Jorge. PH2-A dermoscopic image database for research and benchmarking. In: 2013 35th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2013. p. 5437–40.