European Journal of Cancer 118 (2019) 91–96

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images

Achim Hekler [a], Jochen S. Utikal [b,c], Alexander H. Enk [d], Wiebke Solass [e], Max Schmitt [a], Joachim Klode [f], Dirk Schadendorf [f], Wiebke Sondermann [f], Cindy Franklin [g], Felix Bestvater [h], Michael J. Flaig [i], Dieter Krahl [j], Christof von Kalle [a], Stefan Fröhling [a], Titus J. Brinker [a,d,*]

[a] *National Center for Tumor Diseases, German Cancer Research Center, Heidelberg, Germany*
[b] *Department of Dermatology, Heidelberg University, Mannheim, Germany*
[c] *Skin Cancer Unit, German Cancer Research Center, Heidelberg, Germany*
[d] *Department of Dermatology, University Hospital Heidelberg, Heidelberg, Germany*
[e] *Institute of Pathology and Neuropathology, Eberhard-Karls-University Tuebingen and National Center for Pleura and Peritoneum, University of Tuebingen, Germany*
[f] *Department of Dermatology, University Hospital Essen, Essen, Germany*
[g] *Department of Dermatology, University Hospital Cologne, Cologne, Germany*
[h] *Core Facility Unit Light Microscopy, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany*
[i] *Department of Dermatology, University Hospital Munich (LMU), Munich, Germany*
[j] *Private Laboratory of Dermatohistopathology, Mönchhofstraße 52, 69120 Heidelberg*

Received 21 May 2019; received in revised form 18 June 2019; accepted 21 June 2019
Available online 18 July 2019

KEYWORDS
Melanoma;
Pathology;
Histopathology;
Deep learning;
Artificial intelligence

Abstract ***Background:*** The diagnosis of most cancers is made by a board-certified pathologist based on a tissue biopsy under the microscope. Recent research reveals a high discordance between individual pathologists. For melanoma, the literature reports on 25–26% of discordance for classifying a benign nevus versus malignant melanoma. A recent study indicated the potential of deep learning to lower these discordances. However, the performance of deep learning in classifying histopathologic melanoma images was never compared directly to human experts. The aim of this study is to perform such a first direct comparison.

Methods: A total of 695 lesions were classified by an expert histopathologist in accordance with current guidelines (350 nevi/345 melanoma). Only the haematoxylin & eosin (H&E) slides

* *Corresponding author*: National Center for Tumor Diseases, German Cancer Research Center, Im Neuenheimer Feld 460, Heidelberg, 69120, Germany.
E-mail address: titus.brinker@dkfz.de (T.J. Brinker).

https://doi.org/10.1016/j.ejca.2019.06.012

of these lesions were digitalised via a slide scanner and then randomly cropped. A total of 595 of the resulting images were used to train a convolutional neural network (CNN). The additional 100 H&E image sections were used to test the results of the CNN in comparison to 11 histopathologists. Three combined McNemar tests comparing the results of the CNNs test runs in terms of sensitivity, specificity and accuracy were predefined to test for significance ($p < 0.05$).

Findings: The CNN achieved a mean sensitivity/specificity/accuracy of 76%/60%/68% over 11 test runs. In comparison, the 11 pathologists achieved a mean sensitivity/specificity/accuracy of 51.8%/66.5%/59.2%. Thus, the CNN was significantly ($p = 0.016$) superior in classifying the cropped images.

Interpretation: With limited image information available, a CNN was able to outperform 11 histopathologists in the classification of histopathological melanoma images and thus shows promise to assist human melanoma diagnoses.

1. Background

Melanoma is accountable for most skin cancer−related deaths worldwide [1]. Just as most other cancers, it is primarily diagnosed via tissue biopsy. The standard procedure involves cutting the tissue biopsy into slices and subsequently treating it with different histopathological methods before it is observed as a slide under the microscope by a board-certified histopathologist. For the first assessment, H&E staining is standardly used to prepare the slide. The pathologist decides based on the H&E staining if further staining procedures are necessary or if the diagnosis of a nevus can safely be made. If the lesion is suspicious for melanoma, the histopathologist will order additional immunohistochemistry to confirm.

However, the gold standard in melanoma diagnosis via human-assessed biopsy alone is challenged by recent studies that revealed a diagnostic discordance between expert histopathologists in distinguishing between benign nevi and malignant melanoma of about 25%−26% [2,3].

Past research revealed a lower variance of computer vision and more specifically deep learning in clinical and dermoscopic melanoma diagnosis compared to human assessment [4−14]. Deep learning was successfully applied to histopathological images in breast cancer, showing on par performance with a group of 11 histopathologists [15] and in the field of non-small cell lung cancer, however without the comparison to histopathologists [16].

While a recent study assessed the discordance between deep learning and human experts in classifying melanoma and compared it with the discordance between pathologists in the literature [17], a direct comparison was not conducted to date. However, the performance of a classifier is largely dependent on the given test set, and thus, for a fair comparison, the test images for both computer and humans should be the same [18].

This study performed the first head-to-head comparison of deep learning with the classification results for randomly cropped images with 11 practicing histopathologists. The original diagnosis made by a histopathologist with 20 years of practical experience and with whole slides as well as immunohistostaining available was set as the ground truth. The aim of this study is to illustrate the potential of deep learning not to replace but to supplement human assessment for a definite melanoma diagnosis, especially when limited information is at hand.

2. Methods

2.1. Study design

This comparative study was conducted from 29th September 2018 (design of study and submission to the ethics committee) to 24th March 2019 (finish of data analysis and manuscript approval by all authors). The anonymised slides were obtained from the largest regional dermatohistopathologic institute that follows the international guidelines for histopathologic diagnosis of melanoma (Dr. Dieter Krahl, Mönchhofstraße 52, 69120 Heidelberg) and were labelled into two categories: nevi and melanoma. The class labels were confirmed by the responsible board-certified histopathologist with more than 20 years of experience. Ethics approval was obtained from the ethics committee (Faculty of Mannheim of the University of Heidelberg, 68131 Mannheim, Germany).

2.2. Characteristics of the used specimen

The slides of the 350 nevi obtained from the institute of Dr. Krahl were obtained from biopsies of the past year and consisted of epidermal/junctional nevi and compound/epidermocorial nevi (1:1). Papillomatous nevi were not included because they mostly do not receive additional immunohistostaining. The vertical diameter of the specimen is in between 0.2 mm (epidermal/junctional nevi) and 3.0 mm (mostly congenital compound nevi).

The obtained melanoma specimens had the same range of vertical diameter (0.2 mm for in situ melanoma and up to 3.0 mm for mostly stage 4 melanoma) and were equivalent to the melanomas detected in the institute for the past year. The nevi were also diagnosed in the past year but were randomly picked from a large sample.

2.3. Participants

All participants were recruited by the study team via e-mail or phone call. The inclusion criteria were successful graduation from medical school and active clinical participation in histopathologic melanoma diagnosis. A total of 14 pathologists were invited to 'test their knowledge in comparison to an algorithm', but only 11 participated. The electronic questionnaire did not ask for the identity of the invited participants so that a direct comparison between the participants or the institutions they work at was made impossible.

2.4. Training of the convolutional neural network

The whole slides were digitalised by a NanoZoomer S360 digital slide scanner from the company Hamamatsu (Japan). Subsequently, image sections (0.06% of the whole slide on average) with a 10-fold magnification were randomly cropped (one crop per slide/patient). The only criterion for the selection of the cropped area of the whole slide image was that the epidermis was visible in it. The individual sections of the slide were then assigned to the respective class which the histopathologist had assigned to the whole slide.

We used a pretrained [19] ResNet50 convolutional neural network (CNN) [20]. To adapt the CNN for the classification of our test set, 595 cropped image sections of 595 histopathologic slides from 595 individual patients were used for transfer learning (300 nevi and 295 melanoma). For evaluation of the CNN, a test set of 100 additional randomly cropped test images (melanoma:nevi = 1:1) was generated, which was separate from the training set. For more technical details on the training procedure, please see Appendix 1.

2.5. Comparison with board-certified histopathologist

To quantitatively evaluate the quality of the CNN classification and the performance of the histopathologists, 100 images with known class labels were used to compare the class label assigned by the classifier with the actual class (as determined by the histopathologist with all methods and information at hand (Fig. 1)).

Sensitivity and specificity were calculated separately for the summary decisions of the CNN and the pathologists. Sensitivity, specificity and overall rates of correct classifications were compared statistically using three separate (two-sided) McNemar 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. All analyses were programmed via a Jupyter notebook in Python.

3. Results

3.1. Participating pathologists

Eleven practicing pathologists were recruited who were all currently active in clinical practice and had different levels of experience. Two were practicing in an office-based pathology institute, and the remaining nine pathologists were practicing in university hospitals (Essen, Friedrichshafen, Mannheim, Munich and Tuebingen, all in Germany). Three were still junior physicians with less than 3 years of practical experience (but actively diagnosing diseases under the microscope), and the remaining 8 pathologists were board-certified and had more than 4 years of practical experience.

The mean receiver operating characteristic (ROC) curve over all 11 runs is shown in Fig. 2 (blue line).

It was determined by calculating the average-predicted class probability for each test image over all of the 11 runs. Therefore, the true positive rate (sensitivity) was plotted on the y-axis and the false positive rate (1-specificity) on the x-axis. Only two of the 11 pathologists performed on par with the CNN.

The 11 pathologists achieved a sensitivity/specificity/accuracy of 51.8% (SD = 9.8%)/66.5% (SD = 11.4%)/59.2% (SD = 5.3%). While specificity and overall accuracy increased for the group of pathologists with more than six years of practical experience (58.5% vs. 71.1% and 58.8% vs. 59.4%), the pathologists with less experience showed a higher sensitivity (59.0% vs. 48.0%).

For comparison, the CNN achieved a mean sensitivity/specificity/accuracy of 76% (SD = 7%)/60% (SD = 14.2%)/68% (SD = 8.3%) over the 11 test runs.

The CNN significantly outperforms our sample of pathologists (McNemar p = 0.016).

4. Discussion

For the first time, a deep learning algorithm was directly compared to practicing pathologists in diagnosing melanoma. Our CNN achieved systematic outperformance of 11 pathologists in the classification of melanoma and nevi (<0.05).

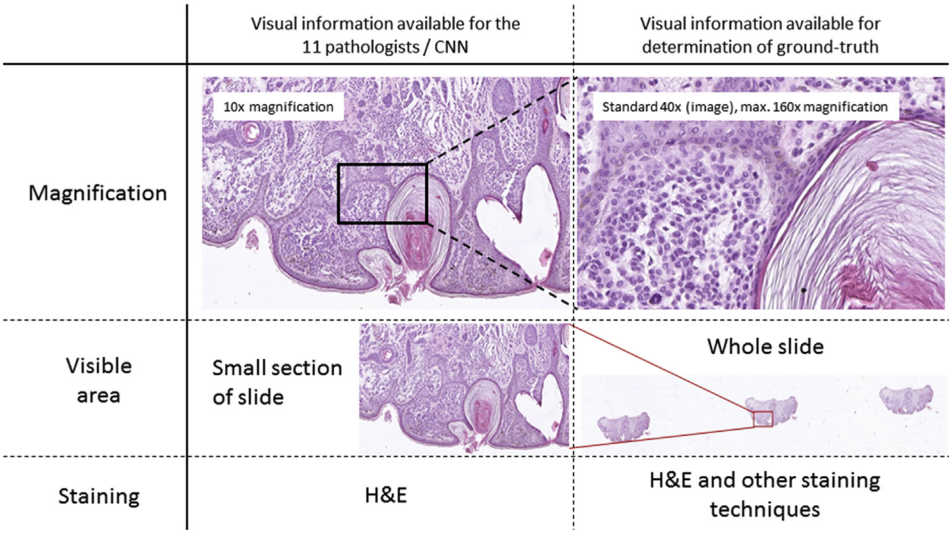
Visual information available for the 11 pathologists / CNN
Visual information available for determination of ground-truth
Magnification
10x magnification
Standard 40x (image), max. 160x magnification
Visible area
Small section of slide
Whole slide
Staining
H&E
H&E and other staining techniques

Fig. 1. Comparison of the data available for pathologists/CNN vs. determination of ground truth to classify a biopsy. CNN, convolutional neural network; H&E, haematoxylin & eosin.

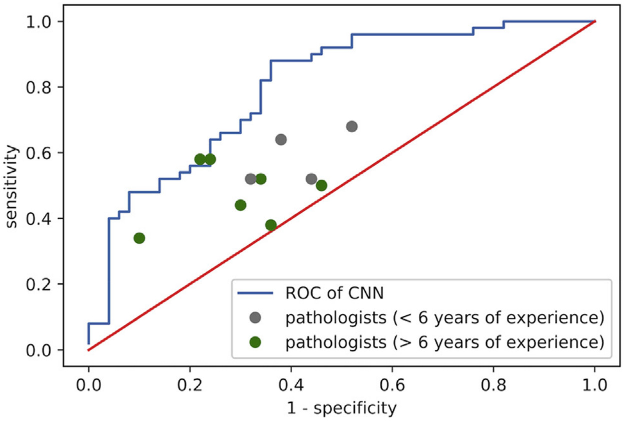

Fig. 2. Average receiver operating characteristic (ROC) curve of the CNN. CNN, convolutional neural network.

The promise of digital pathology is the potential to augment the pathologist's eye with information/intelligence that cannot be gleaned by human examination [21]. In this work, randomly cropped images from digital whole-slide images were used to compare machine learning to pathologists. The clear outperformance may be explained by the ability of artificial intelligence to mine 'sub-visual' image features [22] that may not be visually discernible by a pathologist. Consequently, computer vision is able to gather more information with diagnostic relevance from an image section than a pathologist. Thus, these sub-visual image features offer the opportunity for better quantitative modelling of disease appearance with a lower amount of input data [22].

The minimum number of samples/patients for the training of a CNN depends strongly on the specific classification task. For example, a study of 2018 presents a CNN for breast cancer diagnosis from pathology slides with an ROC of 0.99. The nearly perfect algorithm was built from only 270, indicating that the small number of input data used in this study is tolerable [23].

4.1. Limitations

Even though H&E slides are also evaluated in daily routine, in a normal setting, a pathologist is able to look at the whole slide instead of just a section and order additional immunostaining. We chose randomly cropped images of the epidermis for both training and testing to reduce the time necessary for training and testing and to have an acceptable benchmark (=classification by a board-certified pathologist as defined by the guidelines).

The pathologists filled out the electronic questionnaires in front of desktop screens. They were asked to analyse the cropped images and diagnose the image as melanoma or nevus. While the images had the original resolution as presented to the CNN, the desktop screens itself may have limited the resolution (=below retina resolution).

Another limitation of this study is the binary nature of the algorithm: A pathologist has to exclude a broad spectrum of differential diagnoses, while our algorithm can and will only decide whether a lesion is more likely a nevus or a melanoma. In addition, prospective studies implemented in the clinical setting are necessary to confirm a clinical impact of CNNs in assisting melanoma diagnoses.

Finally, it should be noted that the defined ground truth has to be interpreted with caution: While the procedure that led to the definition of the ground truth is the standard of care in histopathological melanoma diagnosis, around 25–26% of discordance is found between two pathologists who assess the same slides for melanoma vs. nevi [2,3]. Possible alternatives for future research to improve the ground truth include consent decisions of groups of histopathologists, genomic analyses and the integration of cancer registry data.

5. Conclusions

With limited image information available, a CNN was able to systematically outperform 11 histopathologists in the classification of histopathological melanoma images and thus shows great potential to assist human melanoma diagnoses. Prospective studies that use whole slides for testing are necessary to confirm this preliminary finding.

Funding

This research received no funding.

Acknowledgements

The authors would like to thank the participating pathologists who asked to remain anonymous except for the following colleagues: Essen: and Tobias Schimming; Friedrichshafen: Leo Schärer; Munich: Daniela Hartmann and Michael Flaig; Mannheim: Maria Gaiser.

Conflicts of interest statement

None declared.

Appendix 1

A ResNet50 conventional neural network (CNN) model that was pretrained on the large ImageNet data set, comprising approximately 1.28 million images with 1000 class labels, was used in this study. This model was trained using transfer learning for our classification task and data set. In CNNs, the dependencies of all pixels affect each weight, except that of the first layer. By contrast, CNNs first aggregate adjacent local pixels to recognise local characteristics; these are then combined into global features. This restriction on local connections leads to faster training and reduced complexity of the model. To develop the algorithm in this study, 595 histopathologic images of melanomas and nevi were used (295 melanomas and 300 nevi). The test set consists of additional 100 images (50 melanomas and 50 nevi) and is disjunct to the training set.

During the training, the weights were slightly modified to reduce loss; in particular, loss is described mathematically by a function that models the difference between class labels predicted for a given parameter setting and actual class labels. The learning rate includes a hyperparameter controlling these adjustments extent with respect to the loss function's gradient. We used different learning rates for each layer as opposed to existing approaches that apply an identical learning rate to all CNN layers. Specifically, slower learning rates were used for layers closer to input, while for layers closer to output, faster learning rates were applied. The rationale behind this enhanced technique, known as differential learning rates, is that earlier layers (closer to input) contain more general features (e.g. edges or gradients); therefore, their weights do not require extensive changes to the new classification task. Accordingly, the learning rates for the earlier layers were set at low values, resulting in moderate weight adjustment. By contrast, there are application-specific features in the last layers (closer to output). Consequently, higher learning rates were assigned to these layers, resulting in greater changes in the corresponding weights compared with the weights of the early layers. We divided the layers into three groups to realise this concept and applied a different learning rate for each group. The first six residual units had a learning rate of 0.009, the following eight residual blocks had a learning rate of 0.003 and a learning rate of 0.01 is assigned to the fully connected layers. Specific learning rates were selected based on practical experience with other image classification tasks.

As the model approaches the minimum, a common practice comprises stepwise reduction of the learning rate such that the optimisation settles close to the minimum, rather than passing beyond the minimum. In this article, we used a cosine annealing method, which decreased the learning rate based on a cosine function. A third enhanced training technique addressed the problem that the optimisation process can become fixed in a local minimum rather than a global minimum. To resolve this problem, the learning rate was rapidly increased at some specific time steps; thus, the optimisation process could escape a local minimum and reach the global minimum. This technique is known as a stochastic gradient descent with warm restarts and is described in detail in the study by Loshchilov et al. [24].

References

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

[2] Lodha S, Saggar S, Celebi JT. Silvers DNJJocp: discordance in the histopathologic diagnosis of difficult melanocytic neoplasms in the clinical settingvol. 35; 2008. p. 349–52. 4.

[3] Corona R, Mele A, Amini M, De Rosa G, Coppola G, Piccardi P, et al. Faraggiana TJJocO: interobserver variability on the histopathologic diagnosis of cutaneous melanoma and other pigmented skin lesionsvol. 14; 1996. p. 1218–23. 4.

[4] 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.

[5] Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, 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.

[6] Sondermann S, Utikal JS, Brinker TJ. Prediction of melanoma evolution in melanocytic nevi via artificial intelligence: a call for prospective data. Eur J Cancer 2019. forthcoming.

[7] Maron RC, Weichenthal M, Brinker TJ. Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. Eur J Cancer 2019. forthcoming.

[8] Brinker TJ, Hekler A, et al. Deep neural networks are superior to dermatologists in melanoma image classification. Eur J Cancer 2019. forthcoming.

[9] Hekler A, Utikal JS, et al. Superior skin cancer classification by the combination of human and artificial intelligence. Eur J Cancer 2019. forthcoming.

[10] Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, et al. Fröhling SJEJoC: a convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification taskvol. 111; 2019. p. 148–54.

[11] 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.

[12] 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 Aug 1;29(8):1836–42.

[13] 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.

[14] 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 Cancer 2019;113:47–54.

[15] Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G, et al. Balkenhol MJJ: diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancervol. 318; 2017. p. 2199–210. 22.

[16] Yu K-H, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL. Snyder MJNc: predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image featuresvol. 7; 2016. p. 12474.

[17] Hekler A, Utikal J, Enk A, Berking C, Klode J, Schadendorf D, et al. Pathologist-level classification of histopathological melanoma images with deep neural networks. Eur J Cancer 2019; 113.

[18] 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 Cancer 2019;111:30–7.

[19] 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.

[20] 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.

[21] Acs B, Rimm D. Not just digital pathology, intelligent digital pathology. JAMA Oncology 2018;4(3):403–4.

[22] Madabushi A, Lee G. Image analysis and machine learning in digital pathology: challenges and opportunities. Med Image Anal 2016;33:170–5.

[23] Liu Y, Kohlberger T, Norouzi M, Dahl G, Smith J, Mohtashamian A, et al. Artificial intelligence-based breast cancer nodal metastasis detection. Archives of pathology & laboratory medicine; 2018.

[24] Loshchilov I, Hutter F. Sgdr: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 2016.