European Journal of Cancer 115 (2019) 79–83

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

Pathologist-level classification of histopathological melanoma images with deep neural networks

Achim Hekler [a], Jochen Sven Utikal [b,c], Alexander H. Enk [d], Carola Berking [e], Joachim Klode [f], Dirk Schadendorf [f], Philipp Jansen [f], Cindy Franklin [g], Tim Holland-Letz [h], Dieter Krahl [i], Christof von Kalle [a], Stefan Fröhling [a], Titus Josef 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] *Department of Dermatology, University Hospital Munich (LMU), Munich, Germany*
[f] *Department of Dermatology, University Hospital Essen, Essen, Germany*
[g] *Department of Dermatology, University Hospital Cologne, Cologne, Germany*
[h] *Department of Biostatistics, German Cancer Research Center (DKFZ), Heidelberg*
[i] *Private Laboratory of Dermatohistopathology, Mönchhofstraße 52, 69120 Heidelberg*

Received 17 March 2019; received in revised form 4 April 2019; accepted 10 April 2019
Available online 23 May 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 25–26% of discordance for classifying a benign nevus versus malignant melanoma. Deep learning was successfully implemented to enhance the precision of lung and breast cancer diagnoses. The aim of this study is to illustrate the potential of deep learning to assist human assessment for a histopathologic melanoma diagnosis.

Methods: Six hundred ninety-five lesions were classified by an expert histopathologist in accordance with current guidelines (350 nevi and 345 melanomas). Only the haematoxylin and eosin stained (H&E) slides of these lesions were digitalised using a slide scanner and then randomly cropped. Five hundred ninety-five of the resulting images were used for the training of a convolutional neural network (CNN). The additional 100 H&E image sections were used to test the results of the CNN in comparison with the original class labels.

Findings: The total discordance with the histopathologist was 18% for melanoma (95%

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

https://doi.org/10.1016/j.ejca.2019.04.021

confidence interval [CI]: 7.4–28.6%), 20% for nevi (95% CI: 8.9–31.1%) and 19% for the full set of images (95% CI: 11.3–26.7%).
Interpretation: Even in the worst case, the discordance of the CNN was about the same compared with the discordance between human pathologists as reported in the literature. Despite the vastly reduced amount of data, time necessary for diagnosis and cost compared with the pathologist, our CNN archived on-par performance. Conclusively, CNNs indicate to be a valuable tool 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 tissue specimen is cut into slices and is pretreated with different histopathological methods before it is observed on 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 whether further staining is necessary or whether the diagnoses of a nevus can safely be made. If the lesion is suspicious of melanoma, the histopathologist will order for 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 melanomas.

Lodha *et al.* [2] evaluated 392 diagnoses of thin melanoma or benign nevus. The histopathological slides were classified into four classes by two histopathologists: definite nevus, probable nevus, definite melanoma and probable melanoma. The pathologists fully agreed only in 54.5% of the cases. In 25% of the cases, a high level of disagreement (definite nevus and definite melanoma) was observed between both histopathologists.

Corona *et al.* [3] evaluated the discordance in the diagnoses of cutaneous melanoma vs. benign lesions: In 37 of 140 cases (26%), a disagreement in classifying the biopsy occurred.

Past research has revealed a lower variance in computer vision, and more specifically convolutional neural networks (CNNs), in clinical and dermoscopic melanoma diagnosis compared with human assessment [4–10]. CNNs were successfully applied to histopathological images in breast cancer, showing on-par performance with a group of 11 histopathologists [11] and in the field of non–small cell lung cancer, however, without the comparison with histopathologists [12].

A CNN autonomously learns statistical patterns in a data set which are relevant to a specific classification problem. This is done by 'showing' a large number of images to the CNN, processing them through the CNN and comparing the resulting output with the actual class of each image. The internal parameters are then adjusted over several training runs in such a way that the classification error is continuously reduced. If the training data set is representative, then, the CNN can also assign the previously unseen samples to the respective classes with high accuracy. The minimum number of samples/patients depends strongly on the specific classification task. For example, a study in 2018 presents a CNN for breast cancer diagnosis from pathology slides with a receiver operating characteristic (ROC) of 0.99. Thus, it was nearly perfect, and it was built from less than 270 slides [13]. The annual competition with the benchmark database ImageNet Large Scale Visual Recognition Competition (ILSVRC) contains approximately 1000 images per class [14].

This study is the first work that implemented deep learning into the histopathologic diagnosis of melanoma and compares it with the classification results of a board-certified histopathologist (set as the gold standard). 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. We aim to demonstrate that the discordance between the CNN and an expert pathologist is comparable with discordance rates between different pathologists as reported in the literature.

2. Methods

2.1. Study design

This comparative study was conducted from September 25, 2018 (design of study and submission to the ethics committee), to March 12, 2019 (completion 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 classified into two categories: nevi and melanomas. 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, reference number of the approval: 2018-630N-MA).

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 specimen had the same range of vertical diameter (0.2 mm for in situ melanoma and up to about 3.0 mm for mostly stage 4 melanoma) and was 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. Training of the CNN

The whole slides were digitalised using 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; Fig. 1). The only criterion for the selection of the cropped area of the whole-slide image was that the epithelium 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 (and additional immunohistochemistry slides in case of melanoma diagnosis).

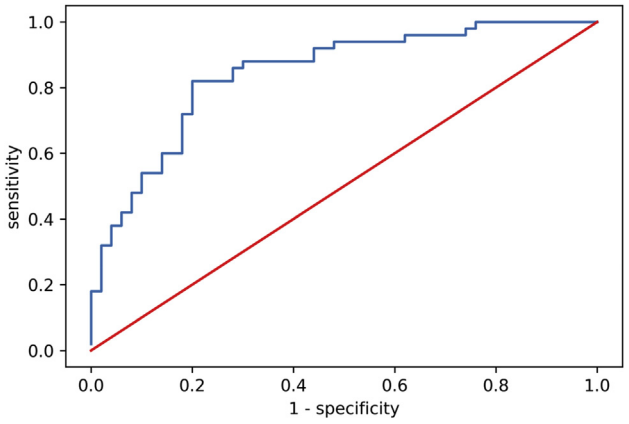

Fig. 2. Average receiver operating characteristic (ROC) curve of the CNN. CNN, convolutional neural network.

We used a pretrained [15] ResNet50 CNN [16]. 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 melanomas). 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. We chose random crop images of the

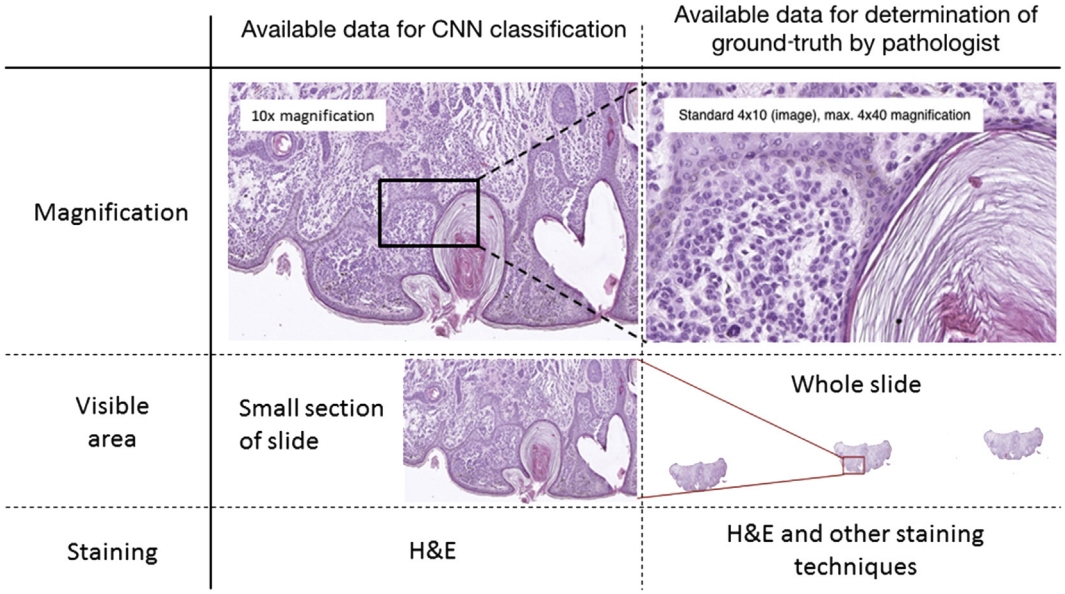

Fig. 1. Comparison of the data available for the pathologist vs. CNN to classify a biopsy. CNN, convolutional neural network; H&E, haematoxylin and eosin stain.

epithelium for both training and testing to reduce the time necessary for training and testing.

2.4. Comparison with the board-certified histopathologist

Misclassification rates between the average result of the CNN and the reference pathologist were calculated separately for melanomas and unspecific nevi, as well as for the combined set of all images. Ninety-five percent confidence intervals (CIs) were determined for all rates using the normal distribution approximation.

3. Results

The mean ROC curve over all 10 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 10 runs.

Misclassification rates of the trained CNN (total discordance with the histopathologist) were 18% for melanomas (95% CI: 7.4–28.6%), 20% for nevi (95% CI: 8.9–31.1%) and 19% for the full set of images (95% CI: 11.3–26.7%).

4. Discussion

For the first time, a deep learning algorithm was implemented in histopathologic melanoma classification. Our algorithm was discordant for 19% of images on average (95% CI: 11.3–26.7%). Consequently, the discordance between the CNN and the histopathologist is on par with the discordance between expert histopathologists as described in the literature (25% [Lodha *et al.* [2]] or 26% [Corona *et al.* [3]]) in the classification of melanomas and nevi [3].

This finding is remarkable in light of the fact that the CNN had a much more limited amount of data available (0.06% of the scanned H&E slide; see Fig. 1). In addition, a histopathologist may order additional tests, examine the whole slide and use the magnification of the microscope (i.e. zoom in and zoom out). The results are reported as the average value of the performed 10 test runs. Thus, the measured results did not occur randomly but due to the training.

The promise of digital pathology is the potential to augment the pathologist's eye with information/intelligence that cannot be gleaned by human examination [17]. In this work, randomly cropped images from digital whole-slide images were used to compare machine learning with discordance rates of human pathologists. The on-par performance may be explained by the ability of artificial intelligence to mine 'subvisual' image features [18] 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 subvisual image features offer the opportunity for better quantitative modelling of disease appearance and hence possibly improved prediction of disease aggressiveness and patient outcome with a lower amount of input data [18].

4.1. Limitations

A major 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 only decide whether a lesion is more likely a nevus or a melanoma without even determining any subtypes of these. In addition, prospective studies implemented in the clinical setting are necessary to confirm a clinical impact of CNNs in assisting melanoma diagnoses.

It should also be noted that the obtained slides stem from a single dermatohistopathologic institute which adheres to guideline-recommended procedures for preparing the biopsy. Thus, it is unknown how the algorithm would perform on slides that stem from institutes not adhering to the standard operating procedures as defined by the guidelines or in different countries with different standard operating procedures.

Furthermore, the only criterion for inclusion in the training and test set is that part of the epithelium is visible. It can happen that an image section is randomly selected from a slide containing malignant tissue where no malignant structures are visible. Accordingly, the results of our classifier could be further improved in the future if the sections used for training and test data are examined/annotated by a pathologist review panel [16] to evaluate if malignant structures are visible in a section or by the use of whole digitalised slides for training and testing.

Choosing a single pathologist following the histopathologic guidelines as the ground truth could further be improved by including independent pathologist review panels which have demonstrated to reduce discordance and improve the precision of histopathologic melanoma diagnosis [19].

In addition, we did not stratify our sample for criteria such as vertical diameter of the lesion. However, the risk of a selection bias is reduced because the test set was pulled randomly from the whole sample of 695 slides. Moreover, the composition of the 695 slides reflects the clinical reality of the composition of melanoma cases and was not preselected (all diagnosed melanomas from the past year), and the composition of nevi reflects the clinical reality of nevi which receive immunostaining (random selection of nevi that received immunostaining in the past year). Since the test-set that was used for the calculation of the discordance between pathologists in the literature was different from ours and classifier performance is also dependent on test-sets, our results should be interpreted with caution and need further confirmation.

Recent deep learning publications in the field of breast and lung cancer [11,12] demonstrated the implementation of automatic annotation methods which has potential for both: to highlight areas of interest for improved processing by a subsequent algorithm and/or pathologist. Such a system is not available for melanoma but should be addressed in future research.

5. Conclusion

For the first time, deep learning was applied to classification of histopathological images of nevi and melanomas. The CNN showed pathologist-level discordance on a test set of 100 images of nevi and melanomas despite using considerably few data and less time to make a decision. Conclusively, CNNs indicate to be a valuable tool to assist human melanoma diagnoses. Additional studies are necessary to confirm this *post hoc* finding.

Conflict of interest statement

None declared.

Funding

This research received no funding.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejca.2019.04.021

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 DN. Discordance in the histopathologic diagnosis of difficult melanocytic neoplasms in the clinical setting. J Cutan Pathol 2008;35(4):349–52.
[3] Corona R, Mele A, Amini M, De Rosa G, Coppola G, Piccardi P, et al. Interobserver variability on the histopathologic diagnosis of cutaneous melanoma and other pigmented skin lesions. J Clin Oncol 1996;14(4):1218–23.
[4] 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 taskal. Eur J Canc 2019;vol. 111: 148–54.
[5] 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.
[6] 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 August 2018;29(8):1836–42.
[7] 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.
[8] 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.
[9] 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.
[10] 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.
[11] Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama 2017;318(22):2199–210.
[12] Yu K-H, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun 2016; vol. 7:12474.
[13] 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.
[14] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. Int J Comp Vis 2015;115(3):211–52.
[15] 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.
[16] 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, vol. 2016; 2016. p. 770–8.
[17] Acs B, Rimm D. Not just digital pathology, intelligent digital pathology. JAMA Oncol 2018;4(3):403–4.
[18] Madabushi A, Lee G. Image analysis and machine learning in digital pathology: challenges and opportunities. Med Image Anal 2016;33:170–5.
[19] 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.