Letter

Research letter

Diagnostic performance of artificial intelligence for histologic melanoma recognition compared to 18 international expert pathologists

To the Editor: Currently, pathologic melanoma classification is based on the—inevitably somewhat subjective—integration of several histologic features.[1] Thus, discordance between pathologists classifying the same lesions can be substantial, and objective assistance tools are needed. The classification of dermoscopic skin lesion images based on convolutional neural networks (CNNs) works well.[2] On a histologic level, our pilot studies provided a proof regarding the principle of CNN-based melanoma recognition using tiny sections of hematoxylin-eosin–stained digitized slides.[3,4]

We compared the ability of CNNs with that of 18 international expert pathologists from eight different countries to discriminate melanomas and nevi in a less artificial setting using hematoxylin-eosin–stained whole-slide images. Ensembles of 3 individual CNNs were trained and tested using single hematoxylin-eosin–stained whole-slide images of 50 individual melanomas and 50 nevi labeled by a panel of 2 experienced dermatopathologists according to the standard practice to provide the "ground truth" (Supplementary Figs 1 and 2 available via Mendeley at https://data.mendeley.com/datasets/j87c9jshxy/1, Supplementary Table I available via Mendeley at https://data.mendeley.com/datasets/j87c9jshxy/1). The same 100 digitized slides were diagnosed using a web-based survey by 18 international dermatopathologists, each with at least 5 years of experience.

With respect to the ground truth, the 18 individual pathologists achieved a mean sensitivity, specificity, and accuracy of 88.88% (SD = 6.66%), 91.77% (SD = 3.99%), and 90.33% (SD = 4.52%), respectively. Ensemble CNNs trained using slides with or without annotation of the tumor region as a region of interest performed at par with the experts (Fig 1) in terms of mean sensitivity, specificity, and accuracy (unannotated: 88% [SD = 0.0%], 88% [SD = 1.15%], and 88% [SD = 0.58%], respectively, and area under the curve [AUC] 0.95; annotated: 94% [SD = 0.0%], 90% [SD = 2.31%], and 92% [SD = 1.15%], respectively), with an AUC of 0.97. Majority of the votes of the expert panel yielded the best accuracy (98%). A statistical analysis of the performance differences is shown in Supplementary Table II (available via Mendeley at https://data.mendeley.com/datasets/j87c9jshxy/1).

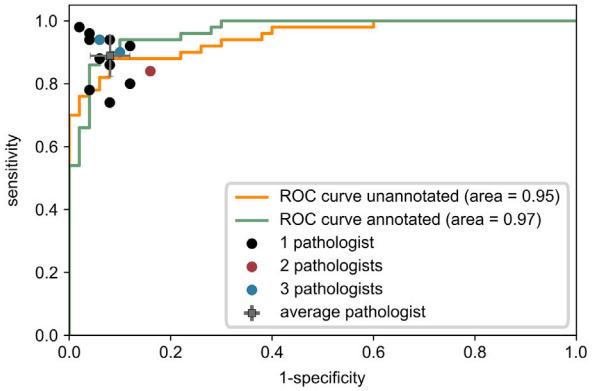

Fig 1. ROC curves of the ensemble CNNs trained on unannotated and preannotated WSI. ROC AUC shows the ratio of true-positive rate to false-positive rate of the CNN-based classification at all classification thresholds. The ROC curve for CNN trained on unannotated slides is shown in *orange* and the one for CNN trained on annotated slides in *green*. The *dots* represent the performance of 18 expert pathologists. The *blue* and *red dots* represent 3 and 2 pathologists, respectively, with the same results. Unique results of the pathologists are represented as a *black dot*. The *gray square* represents the average performance of all the pathologists, with the *error bars* denoting the standard deviation. The true-positive rate (sensitivity) is plotted on the y-axis and the false-positive rate (1-specificity) on the x-axis. *AUC*, Area under the curve; *CNN*, convolutional neural networks; *ROC*, receiver-operating characteristic; *WSI*, whole-slide images.

Only half (47%) of the diagnoses were unanimous. Overall, the discordance was 13.45%. Nine lesions were divergently classified by a third or more pathologists. Two were classified divergently based on majority (Fig 2, Supplementary Table I), of which 1 was originally classified as severely dysplastic acral nevus and the other as incompletely excised desmoplastic acral Spitz nevus with the melanocytic acral

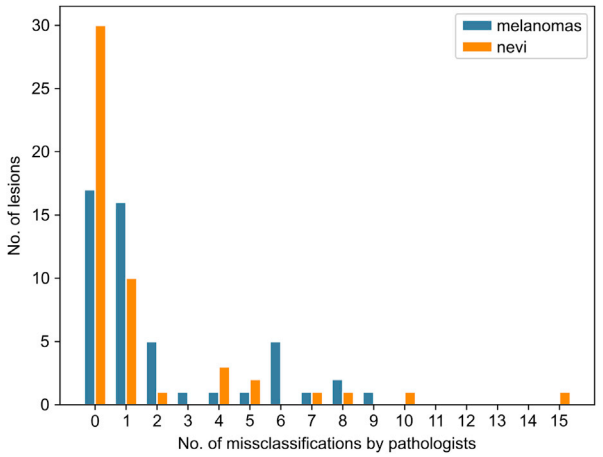

Fig 2. Distribution of the number of divergent classifications by the expert pathologists. To illustrate how many lesions were classified divergently and the frequency of this occurrence, the number of lesions was plotted on the y-axis and the number of divergent classifications per lesion by expert pathologists on the x-axis. Melanomas are shown in *blue* and nevi in *orange*.

nevus with intraepidermal ascents of cells phenomenon, both of which the ensemble CNNs classified as melanoma.

The diagnoses by ensemble CNNs trained using unannotated and annotated whole-slide images differed from the ground truth in 12 and 8, respectively, often pathologically unequivocal, cases. An ensemble CNN, trained and tested using an independent set of slides with the same methodology to confirm its reproducibility, achieved a mean sensitivity, specificity, and accuracy of 98% (SD = 0), 88% (SD = 0), and 93% (SD = 0), respectively, with an AUC of 0.97.

Thus, a high-accuracy classifier can be generated for a specific test environment with few images. Although such classifiers may not yield similar performances on slides from another institution,[5] the practical application of environment-specific assistance tools may be more realistic than an attempt to achieve broad generalization across all environments. Such systems might be the most beneficial for less experienced pathologists. For experienced pathologists, the systems could provide a triage.[5] Further studies are required to investigate CNN-based classifiers in a real-life setting.

We thank Dr Kenneth S. Resnick for his participation in the survey.

Titus J. Brinker, MD,[a] Max Schmitt, MSc,[a] Eva I. Krieghoff-Henning, PhD,[a] Raymond Barnhill, MD,[b] Helmut Beltraminelli, MD,[c] Stephan A. Braun, MD,[d,e] Richard Carr, MD,[f] Maria-Teresa Fernandez-Figueras, MD,[g] Gerardo Ferrara, MD,[b] Sylvie Fraitag, MD,[i] Raffaele Gianotti, MD,[j] Mar Llamas-Velasco, MD,[k] Cornelia S. L. Müller, MD,[l] Antonio Perasole, MD,[m] Luis Requena, MD,[n,o] Omar P. Sangueza, MD,[p] Carlos Santonja, MD,[n] Hans Starz, MD,[q] Esmeralda Vale, MD,[r] Wolfgang Weyers, MD,[s] Achim Hekler, Dipl-Inform,[a] Jakob N. Kather, MD,[t,u] Stefan Fröhling, MD,[u] Dieter Krahl, MD,[v] Tim Holland-Letz, PhD,[w] Jochen S. Utikal, MD,[x,y] Andrea Saggini, MD,[z] and Heinz Kutzner, MD[z]

From the Digital Biomarkers for Oncology Group,[a] National Center for Tumor Diseases (NCT),[u] Division of Biostatistics,[w] and Skin Cancer Unit,[x] National Center for Tumor Diseases (NCT) German Cancer Research Center (DKFZ), Heidelberg, Germany; Departments of Pathology and Translational Research, Institut Curie, Paris, France[b]; Department of Dermatology, Inselspital Bern University Hospital, University of Bern, Bern, Switzerland[c]; Department of Dermatology, Medical Faculty, Heinrich-Heine-University, Düsseldorf, Germany[d]; Department of Dermatology, University of Münster, Münster, Germany[e]; Department of Pathology, Warwick Hospital, Warwick, United Kingdom[f]; University General Hospital of Catalonia, Grupo Quironsalud, International University of Catalonia, Sant Cugat del Vallés, Barcelona, Spain[g]; Anatomic Pathology Unit, Macerata General Hospital, Macerata, Italy[h]; Department of Pathology, University Paris Descartes, Necker-Enfants Malades Hospital, Assistance Publique Hospitals of Paris, Paris, France[i]; Dermatopathology Lab, Dermatological Clinic, University of Milan, Milan, Italy[j]; Department of Dermatology, University Hospital La Princesa, Madrid, Spain[k]; Department of Dermatology, Venereology and Allergology, Saarland University Medical Center, Homburg/Saar, Germany[l]; Anatomic and Cytopathology, Az. ULSS 8 Berica, Regione Veneto, Ospedale San Bortolo, Vicenza, Italy[m]; Dermatology Department, Fundación Jiménez Díaz, Autonomous University of Madrid, Madrid, Spain[n]; Anatomic Pathology Service, Fundación Jiménez Díaz, Madrid, Spain[o]; Dermatopathology, Wake Forest School of Medicine, Winston-Salem, North Carolina[p]; Dermpath München, Munich, Germany[q]; Service of Dermatology, Medical-Surgical Center Lisbon, Lisbon, Portugal[r]; Center for Dermatopathology, Freiburg, Germany[s]; Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany[t];

Dres. Krahl Dermatopathology, Heidelberg, Germany Skin Cancer Unit, German Cancer Research Center (DKFZ), Heidelberg, Germany[y]; and Department of Dermatology, Venereology and Allergology, University Medical Center Mannheim, Ruprecht-Karl University of Heidelberg, Mannheim, Germany[y]; and Dermatopathology Friedrichshafen, Friedrichshafen, Germany.[z]

Funding sources: This work was supported by the German Federal Ministry of Health, Berlin, Germany [grant: Skin-Classification-Project to TJB; German Cancer Research Center].

IRB approval status: Ethics approval was obtained from the ethics committee of the Medical Faculty of Mannheim of the University of Heidelberg, 68131 Mannheim, Germany, before the study was initiated.

Reprints not available from the authors.

Correspondence to: Titus J. Brinker, Digital Biomarkers for Oncology Group, German Cancer Research Center, im Neuenheimer Feld 460, Heidelberg, 69120, Germany

E-mail: titus.brinker@dkfz.de

Conflicts of interest

Dr Brinker would like to disclose that he owns a health technology company (Smart Health Heidelberg GmbH; https://smarthealth.de), which develops mobile apps, outside the submitted work. Dr Beltraminelli would like to disclose that he received honoraria for his role on the Takeda Pharma advisory board, outside the submitted work. The other authors have no conflicts of interest to declare.

REFERENCES

1. Adamson AS, Welch HG. Machine learning and the cancer-diagnosis problem—no gold standard. *N Engl J Med*. 2019;381(24):2285-2287.
2. Tschandl P, Codella N, Akay BN, 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;20(7):938-947.
3. Hekler A, Utikal JS, Enk AH, et al. Pathologist-level classification of histopathological melanoma images with deep neural networks. *Eur J Cancer*. 2019;115:79-83.
4. Hekler A, Utikal JS, Enk AH, et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *Eur J Cancer*. 2019;118:91-96.
5. Niazi MKK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. *Lancet Oncol*. 2019;20(5):e253-e261.

https://doi.org/10.1016/j.jaad.2021.02.009