European Journal of Cancer 111 (2019) 30–37

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark

Titus J. Brinker [a,b,*], Achim Hekler [a], Axel Hauschild [c], Carola Berking [d], Bastian Schilling [e], Alexander H. Enk [b], Sebastian Haferkamp [f], Ante Karoglan [g], Christof von Kalle [a], Michael Weichenthal [c], Elke Sattler [d], Dirk Schadendorf [h], Maria R. Gaiser [i,j], Joachim Klode [h,1], Jochen S. Utikal [i,j,1]

[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 Kiel, Kiel, Germany*
[d] *Department of Dermatology, University Hospital Munich (LMU), Munich, Germany*
[e] *Department of Dermatology, University Hospital Wuerzburg, Wuerzburg, Germany*
[f] *Department of Dermatology, University Hospital Regensburg, Regensburg, Germany*
[g] *Department of Dermatology, University Hospital Magdeburg, Magdeburg, Germany*
[h] *Department of Dermatology, University Hospital Essen, Essen, Germany*
[i] *Department of Dermatology, Heidelberg University, Mannheim, Germany*
[j] *Skin Cancer Unit, German Cancer Research Center (DKFZ), Heidelberg, Germany*

Received 21 October 2018; received in revised form 4 December 2018; accepted 5 December 2018
Available online 22 February 2019

KEYWORDS
Benchmark;
Artificial intelligence;
Deep learning;
Melanoma

Abstract ***Background:*** Several recent publications have demonstrated the use of convolutional neural networks to classify images of melanoma at par with board-certified dermatologists. However, the non-availability of a public human benchmark restricts the comparability of the performance of these algorithms and thereby the technical progress in this field.
Methods: An electronic questionnaire was sent to dermatologists at 12 German university hospitals. Each questionnaire comprised 100 dermoscopic and 100 clinical images (80 nevi images and 20 biopsy-verified melanoma images, each), all open-source. The questionnaire recorded factors such as the years of experience in dermatology, performed skin checks, age, sex and the

* *Corresponding author*: National Center for Tumor Diseases, German Cancer Research Center, Im Neuenheimer Feld 460, 69120 Heidelberg, Germany. Tel.: +496221 3219304; fax: +496221 566967
E-mail address: titus.brinker@nct-heidelberg.de (T.J. Brinker).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ejca.2018.12.016

rank within the university hospital or the status as resident physician. For each image, the dermatologists were asked to provide a management decision (treat/biopsy lesion or reassure the patient). Main outcome measures were sensitivity, specificity and the receiver operating characteristics (ROC).

Results: Total 157 dermatologists assessed all 100 dermoscopic images with an overall sensitivity of 74.1%, specificity of 60.0% and an ROC of 0.67 (range = 0.538–0.769); 145 dermatologists assessed all 100 clinical images with an overall sensitivity of 89.4%, specificity of 64.4% and an ROC of 0.769 (range = 0.613–0.9). Results between test-sets were significantly different ($P < 0.05$) confirming the need for a standardised benchmark.

Conclusions: We present the first public melanoma classification benchmark for both non-dermoscopic and dermoscopic images for comparing artificial intelligence algorithms with diagnostic performance of 145 or 157 dermatologists. Melanoma Classification Benchmark should be considered as a reference standard for white-skinned Western populations in the field of binary algorithmic melanoma classification.

1. Introduction

Melanoma accounts for the majority of skin cancer–related deaths worldwide [1]. Owing to rapid increase in prevalence over recent decades, several institutions have funded programs to improve measures for prevention and early detection/screening [2,3]. Despite special training and the use of dermoscopes, dermatologists rarely exceed a sensitivity of 80% [4].

In 2017, Esteva *et al.* was the first to report a deep-learning convolutional neural network (CNN) image classifier whose performance in determining the management of malignant lesions based on image analysis was comparable to that of 21 board-certified dermatologists [5]. The CNN deconstructed digital images of skin lesions and generated its own diagnostic criteria for melanoma detection during training.

Other subsequent landmark publications have claimed dermatologist-level skin cancer classification via CNNs [5–8]. However, these publications did not reveal the exact procedure or the images used for training. Moreover, the final test images used to measure performance of these algorithms were not made publicly available. Thus, the performance of these algorithms may only be evaluated by using the International Symposium on Biomedical Imaging (ISBI) challenge 2016 test-set as a benchmark, but this benchmark has never been fully compared with the performance of dermatologists for the 379 test images and, thus, provides limited information about the clinical value of an algorithm [9]. The status-quo restricts the comparison between algorithms and thereby the technical progress in this field [10].

In this work, we created the first publicly available Melanoma Classification Benchmark (MClass) for both dermoscopic and clinical images of melanocytic skin lesions accompanied by an open-source test-set. MClass enables researchers to compare their artificial intelligence algorithms for the classification of melanocytic images with that performed by dermatologists. The algorithm was validated with the help of 302 data sets (data set = responses by one dermatologist to one of the electronic questionnaire with 100 images of skin lesions) created by dermatologists from 12 German university hospitals, and eight data sets created by resident physicians from Germany (157 dermatologists completed the dermoscopic survey and 145 dermatologists completed the clinical survey). In addition, our work provides insights into the diagnostic performance of dermatologists for melanoma by illustrating the impact of major variables of interest (i.e. hierarchical position, residents versus university hospital physicians, sex and age).

2. Material and methods

2.1. Recruitment and data collection

The collaborative MClass benchmark project was introduced at the National German Skin Cancer Conference conducted in September 2018 at Stuttgart, Germany. Twelve leading dermatologists from 12 university hospitals in Germany (Berlin, Bonn, Erlangen, Essen, Hamburg, Heidelberg, Kiel, Magdeburg, Mannheim, Munich, Regensburg and Würzburg) agreed to participate. They encouraged their colleagues via their university email accounts to participate in the anonymous validation of the benchmark and to 'test their skills' pertaining to melanoma diagnosis via two online links to two separate questionnaires comprising 100 dermoscopic test images and 100 clinical test images, respectively. The ratio of nevocytic nevi (NZN)/melanoma images in the test-sets was not disclosed. At the end of the survey, participants learned about their diagnostic accuracy. Data were collected between 17th September 2018 and 1st October 2018.

2.2. Electronic questionnaire

Prior to data collection, both electronic questionnaires were developed by consensus between the authors. The first part of both questionnaires was identical and recorded age, sex, years of dermatologic practice/experience, estimated number of skin checks performed and position within the medical hierarchy. This was followed by 100 dermoscopic (link 1) or clinical (link 2) images of 80 benign nevi and 20 biopsy-verified melanomas, each. For each image, the participant was asked to make a management decision: (a) biopsy/further treatment or (b) reassure the patient. The same question was asked in the study by Esteva *et al.* [5]. A response for all images was mandatory, and participants were not allowed to skip any question. Dermatologists were able to use digital zoom and had to use desktop screens to answer the questionnaires. All originally used image files are available at www.skinclass.de/mclass.

2.3. Eligibility criteria

Only physicians with clinical training in dermatology were eligible. Every dermatologist was only allowed to participate once.

2.4. Used images

All images used were open-source and anonymous. We programmed a randomiser in Python for random selection of 100 images with an allocation of 80% NZN and 20% melanoma. The 80:20 ratio is based on the ISBI 2016 challenge test and training set hosted by the International Skin Imaging Collaboration (ISIC) [9]. In accordance, all dermoscopic images were sourced from the ISIC archive [9]. All melanomas were verified by histopathology and the nevi were either biopsy-verified (n = 29) or verified by single image expert consensus (n = 51) (test-set available for downloading [Multimedia Appendix 1]) [11]. The clinical images were obtained from the MED-NODE database, and only the melanomas were biopsy-verified; the nevi were declared as benign via expert consensus [12] (test-set available for downloaded here [Multimedia Appendix 2]). All images were publicly available together with an excel sheet enlisting the reader results per dermatologist per image for each dermoscopic and non-dermoscopic images in addition to the information of the underlying ground truth of each image and how it was determined under this link: www.skinclass.de/mclass.

2.5. Analysis

2.5.1. Data validation

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 dataset. 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. As a statistical outlier detection method, we applied the Local Outlier Factor (LOF) method [15]. The space of all possible management decisions consists of 100 dimensions, one for each test image, and each dimension is a binary variable. The LOF algorithm is an unsupervised method that determines the local density deviation of a distinct point with respect to its neighbors. 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 means that there are very similar answers from dermatologists who differ only slightly from each other. For respondents who show large deviations in their answers compared to other dermatologists, the value is significantly larger, indicating the outliers. In this work, we consider the 30 nearest neighbors of each response, but the detected outliers are not so sensitive to the exact parameter selection. As a result, 18 dermatologists were excluded from the dermoscopic survey group and 17 from the clinical survey group because of the following predefined exclusion criteria: age inconsistent with academic position ($N = 5$); identical response for all 100 images ($N = 2$) and double entry (participation by the same physician twice; $N = 2$).

2.5.2. Data analysis

The survey data were extracted in .csv format and imported into a Jupyter Notebook. The programming language Python was used for calculating sensitivity, specificity and receiver operating characteristics (ROC). On sub-group analysis, between-group differences were assessed using two-sided Chi-squared test programmed in the Jupyter Notebook. For dichotomous predictions, ROC is considered equivalent to the average of sensitivity and specificity.

3. Results

3.1. Total sample

Of the 337 dermatologist-created data sets, 35 were excluded during data validation. Thus, 302 data sets (89.6%) comprising 145 clinical and 157 dermoscopic data sets were included; 210 (64%) participants were female, and 118 (36%) were male. Median age was 30–34 years; 60% participants were junior physicians in their dermatologic residency; 320 dermatologists were from the 12 participating university hospitals and eight were resident physicians in private practice who formerly worked at one of these hospitals. Because a single invitation was sent for

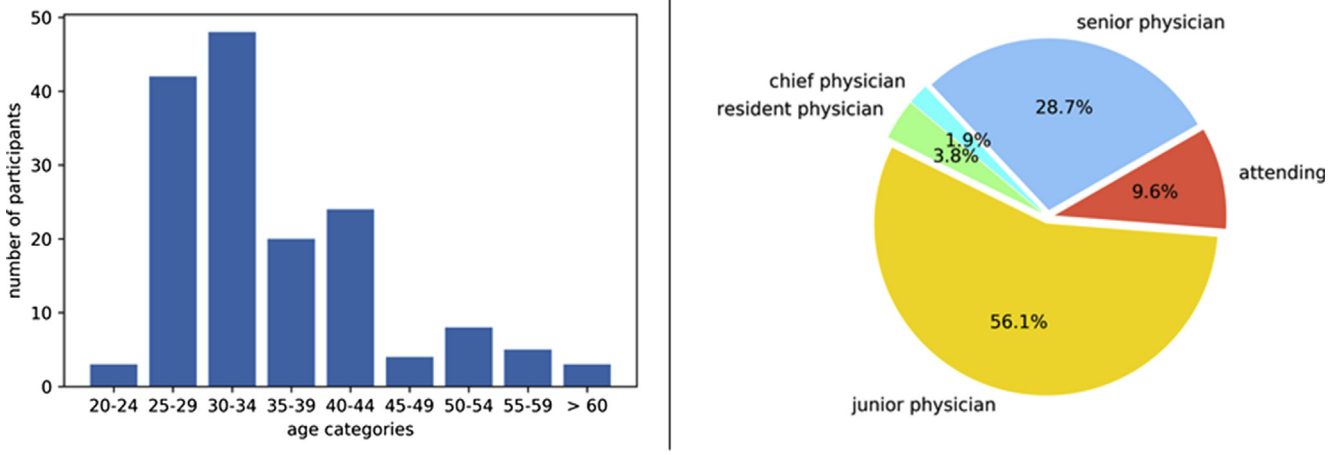

Fig. 1. Sample characteristics for the dermoscopic data set: age distribution (left); distribution of positions in the medical hierarchy (right).

this survey, at least 157 German dermatologists were involved in creating this benchmark.

3.1.1. Dermoscopic melanoma classification benchmark

3.1.1.1. Sample characteristics. Out of 175 dermatologists, 157 (56 [35.7%] males; 101 [64.3%] females) provided valid answers. Median age range was 30–34 years (Fig. 1; left); 56.1% were junior physicians (dermatologic residency), and 43.9% were board-certified (Fig. 1; right). Total 12 dermatologic university hospital departments in Germany provided 163 (95.9%) of these dermatologists, and seven (4.1%) were dermatologic residents involved with these departments.

The benchmark parameters for dermoscopic melanoma classification benchmark (MClass-D) as per various subgroups are summarised in Table 1. An overview of the results for the dermoscopic test-set is presented in Fig. 2, and Fig. 3 provides an overview of the easiest and hardest to diagnose lesions. None of the differences between the subgroups were statistically significant ($P > 0.05$). However, there was considerable variability in performance in the samples (mean ROC [range = 0.54–0.77]; best 25% > 0.732; best 50% > 0.709; best 75% > 0.691).

Table 1
Benchmark parameters for MClass-D.

Subset of dermatologists	Sensitivity (%)	Specificity (%)	ROC area
All participants (N = 157)	74.11	60.02	0.671
University hospital (N = 151)	74.01	59.79	0.669
Private practice (resident) (N = 6)	76.67	65.83	0.713
Practical experience (pe)			
pe ≤ 2 years (N = 46)	75.98	56.47	0.662
2 years < pe ≤ 4 years (N = 37)	73.78	59.09	0.664
4 years < pe ≤ 12 years (N = 32)	73.28	62.54	0.679
pe > 12 years (N = 42)	72.98	62.83	0.679
Position in university hospital			
Junior physician (N = 88)	74.77	58.15	0.665
Attending (N = 15)	72.67	60	0.663
Senior physician (N = 45)	73	62.31	0.677
Chief physician (N = 3)	73.33	69.17	0.713
Sex			
Female (N = 101)	77.33	57.34	0.673
Male (N = 56)	68.3	64.87	0.666
Age			
20–29 (N = 45)	74.89	57	0.659
30–34 (N = 48)	74.17	60.44	0.673
35–44 (N = 44)	75.45	61.7	0.686
>44 (N = 220)	69.25	62.13	0.657
Number of skin screenings (noS) (N = 156, one missing value)	73.94	60.13	0.670
noS ≤ 150 (N = 36)	78.47	53.13	0.658
150 < noS ≤ 500 (N = 40)	71.62	63.81	0.677
500 < noS ≤ 1000 (N = 39)	72.69	61.99	0.673
noS > 1000 (N = 41)	73.41	60.91	0.672

ROC, receiver operating characteristics; MClass-D, dermoscopic melanoma classification benchmark.

3.1.2. Non-dermoscopic melanoma classification benchmark

3.1.2.1. Sample characteristics. Out of the 162 dermatologists who participated in clinical image survey, 145 (50 males [34.5%]; 95 females [65.5%]) were included after data validation (89.5%). Median age (30–34 years) and the occupational profile of participants are summarised

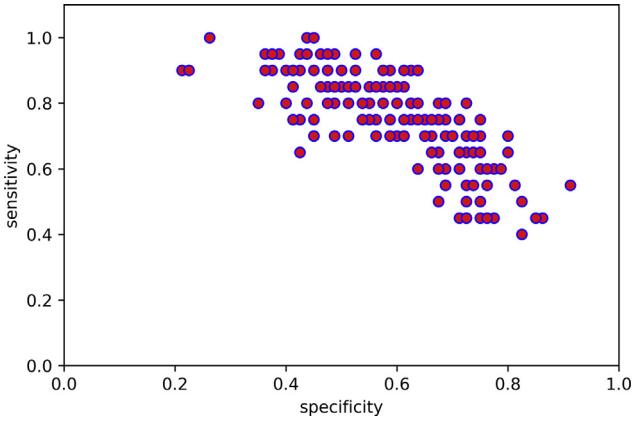

Fig. 2. Overview of results for the dermoscopic test-set: each dot represents the performance of an individual dermatologist.

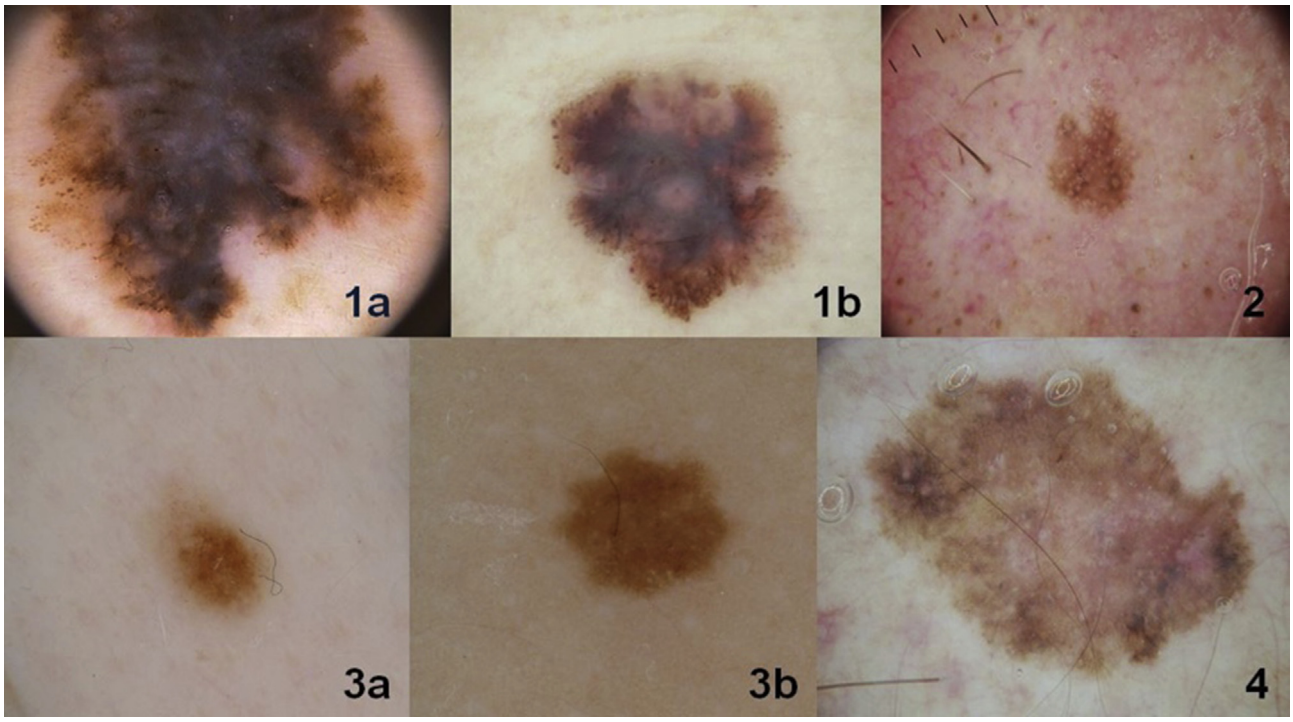

Fig. 3. Best/worst classification results. Upper row = melanoma: images 1a and 1b were associated with highest sensitivity (all 157 dermatologists opted for biopsy); for image 2, biopsy was recommended by 30 dermatologists (127 dermatologists opted to 'reassure the patient'). Lower row: benign nevi (biopsy-verified): images 3a (156 opted to "reassure patient"; one dermatologist recommended biopsy) and 3b (157 dermatologists opted to "reassure the patient") were associated with the highest specificity; for image 4, biopsy was recommended by 156 of the 157 dermatologists.

in Fig. 4. The benchmark parameters for non-dermoscopic melanoma classification benchmark (MClass-ND) as per various subgroups are summarised in Table 2. An overview of the results for the non-dermoscopic test-set is presented in Fig. 5, and Fig. 6 provides an overview of the easiest and hardest to diagnose lesions.

None of the differences between the subgroups were statistically significant ($P > 0.05$).

However, there was substantial variability in the performance (mean ROC [range = 0.615–0.9]; best 25% > 0.766; best 50% > 0.771; best 75% > 0.764).

3.1.2.2. Comparison of MClass-D and MClass-ND. MClass-ND based on expert opinion showed significantly better sensitivity and specificity than MClass-D ($P < 0.05$).

4. Discussion

In this work, we present the first public MClass for both dermoscopic (MClass-D) and non-dermoscopic (MClass-ND) images (based on 157 and 145 dermatologists, respectively) for evaluating artificial intelligence algorithms. Our results have high external validity owing to the largest number of dermatologists surveyed till date. Moreover, our results and the test-sets are available in the public domain. Previous landmark publications by Esteva *et al.*, Marchetti *et al.* and Hänßle *et al.* involved 21, 8 and 58 dermatologists, respectively, for evaluating their algorithms; moreover, the latter two studies only compared dermoscopic images [5,7,8]. More importantly, many groups could not compare their algorithms with clinical performance owing to the lack of availability of image sets to measure performance. Our work is of seminal importance as MClass-D and MClass-ND represent an open access standardised clinical benchmark to assess the performance of artificial intelligence (AI) algorithms against that of dermatologists with different sex and age and different levels of training.

4.1. Interpretation of results

In clinical practice, dermoscopy improves the sensitivity of naked-eye examination [4]. However, in our study, dermatologists performed significantly worse for dermoscopic images than for clinical images of different skin lesions ($P < 0.05$); this indicates that the performance is largely dependent on the images of nevi and melanoma selected for the test-set. Similar effect was observed by Esteva *et al.*; the ROC for dermoscopic images in their study was worse than that for non-dermoscopic images [5]. Another similarity with the work of Esteva *et al.* is the use of mostly biopsy-verified nevi for the dermoscopic set (obtained from the ISIC archive), which are difficult to distinguish from benign lesions (and therefore were sent to biopsy).

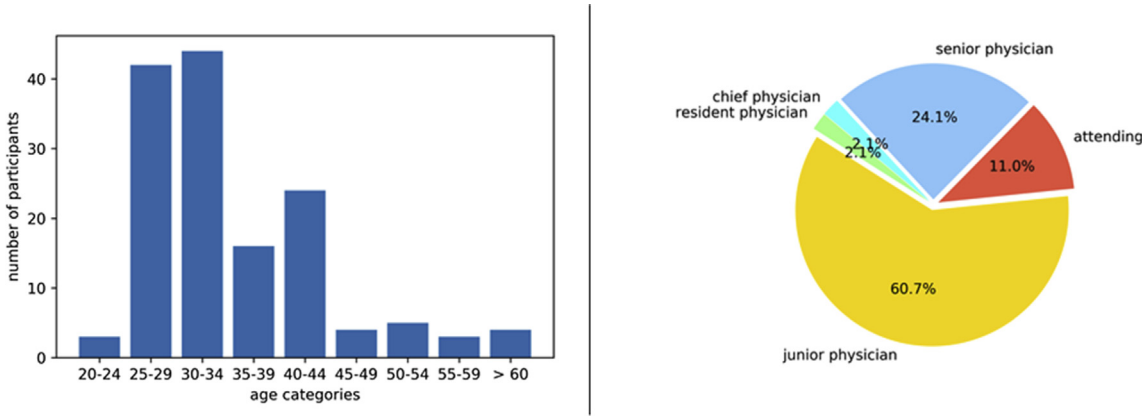

Fig. 4. Sample characteristics for the non-dermoscopic data set: age distribution (left); distribution of positions in the medical hierarchy (right).

However, the outcomes (both sensitivity and specificity) achieved with both test-sets are comparable to those of previous studies [13,14].

4.2. Generalisability

The performance of dermatologists may be different in other countries because of different education programs and different habits regarding use of dermoscopy.

Table 2
Benchmark parameters for MClass-ND.

Subset of dermatologists	Sensitivity (%)	Specificity (%)	ROC area
All participants (N = 145)	89.40	64.37	0.769
University hospital (N = 142)	89.44	64.18	0.768
Private practice (resident) (N = 3)	86.67	73.33	0.800
Practical experience (pe)			
pe ≤ 2 years (N = 42)	89.40	63.57	0.765
2 years < pe ≤ 4 years (N = 36)	87.92	64.86	0.764
4 years < pe ≤ 12 years (N = 31)	91.13	64.03	0.776
pe > 12 years (N = 36)	89.31	65.1	0.772
Position in university hospital			
Junior physician (N = 97)	87.68	64.45	0.761
Attending (N = 16)	92.81	57.66	0.752
Senior physician (N = 39)	88.71	65.8	0.773
Chief physician (N = 3)	91.67	58.75	0.752
Sex			
Female (N = 101)	88.71	63.44	0.761
Male (N = 57)	87.63	65.68	0.767
Age			
20–29 (N = 48)	87.5	64.43	0.76
30–34 (N = 50)	87.8	65.475	0.766
35–44 (N = 42)	90.12	62.83	0.765
>44 (N = 18)	86.15	63.85	0.75
Number of skin screenings (noS) (157, 1 missing value)	88.38	64.2	0.76
noS ≤ 150 (N = 35)	87.71	63.25	0.755
150 < noS ≤ 500 (N = 47)	85.86	67.96	0.769
500 < noS ≤ 2000 (N = 45)	88.97	65.22	0.771
noS > 2000 (N = 30)	89.67	62.83	0.763

ROC, receiver operating characteristics; MClass-ND, non-dermoscopic melanoma classification benchmark.

4.2.1. Limitations

4.2.1.1. Image only as input. Clinical encounter with the actual patient provides more information than that provided by an image. Hänßle *et al.* demonstrated that additional clinical information slightly improves the sensitivity (from 86.6% to 88.9%) and specificity (from 71.3% to 75.7%) of dermatologists [8]. However, currently tested algorithms only accept an image input; thus, for a current benchmark, the input data are restricted to an image for direct comparison of an image classification task with dermatologists.

4.2.1.2. Anonymity. The anonymity of the electronic questionnaire was mandatory to protect privacy. However, anonymity carries the risk of abuse. By involving physicians exclusively via their institutional email addresses and by predefining data validation strategies, this risk was minimised, and a high successful plausibility rate was achieved (157 of 175 participants for MClass-D and 145 of 162 participants for MClass-ND).

4.2.1.3. Allocation of images. The 1:5 ratio (Melanoma/Nevi) per image is equal to one from the ISBI test-set [9].

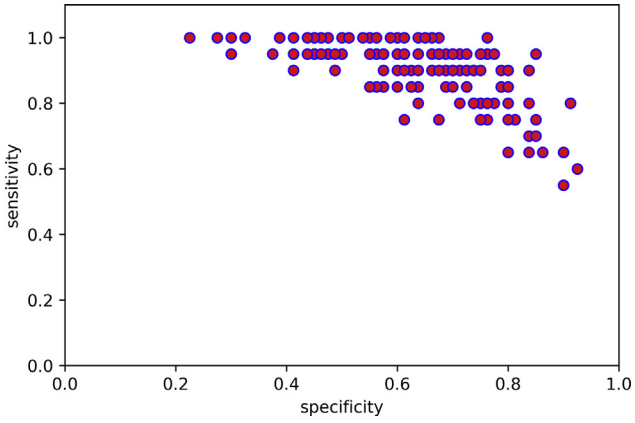

Fig. 5. Overview of the results for the non-dermoscopic test-set: each dot represents the performance of an individual dermatologist.

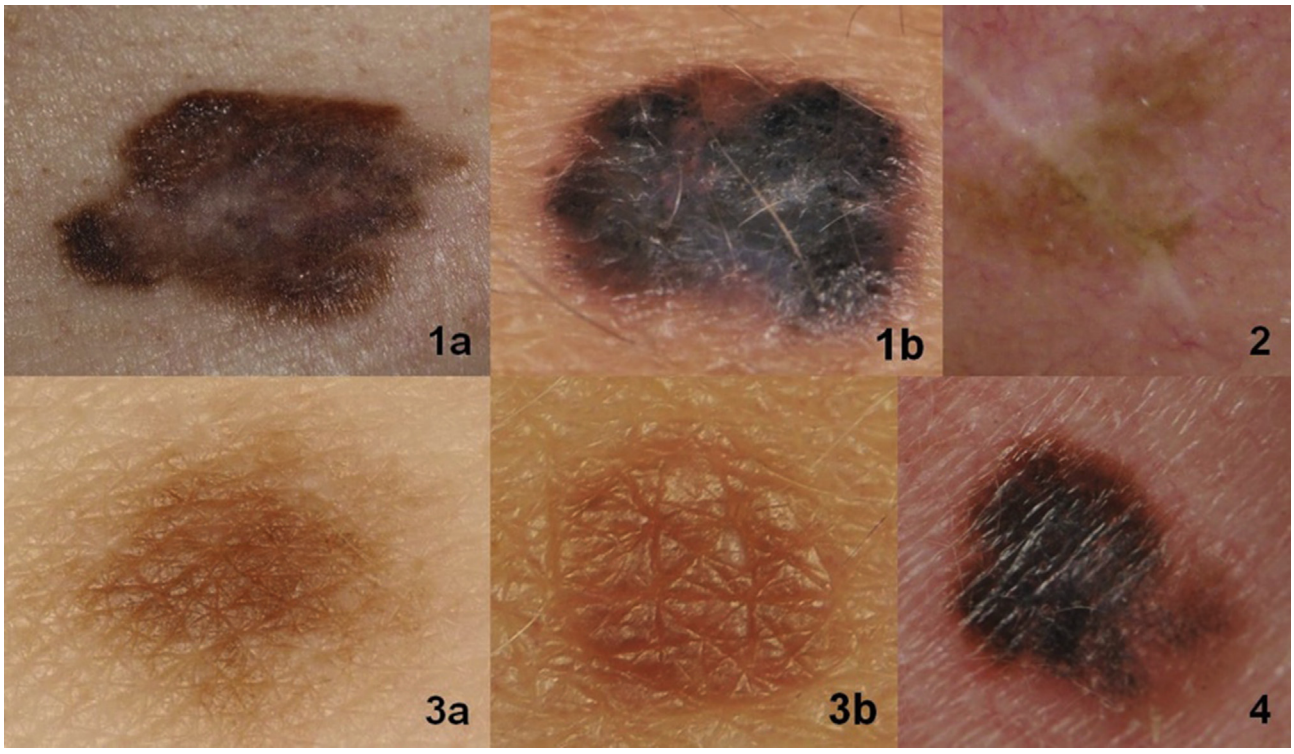

Fig. 6. Best/worst classification results. Upper row (melanoma): 1a and 1b were associated with the highest sensitivity (all dermatologists opted for biopsy); for image 2, biopsy was recommended by 45 cases (100 dermatologists opted to 'reassure the patient'). Lower row (benign nevi): 3a and 3b were associated with the highest specificity (100% opted to 'reassure the patient'); 4 had the lowest specificity (three dermatologists opted for reassurance of patient and 142 recommended biopsy).

In clinical practice, a 1:50 ratio would be more realistic. However, use of this ratio would have necessitated 10 times more test images per data set to achieve the same number of classified melanomas (total 2000 test images), which would have drastically reduced the number of dermatologists willing to participate.

4.2.1.4. Generalisability. MClass may be used as a benchmark for binary decisions on images of melanocytic lesions to distinguish nevi from melanoma trained on and for classification of images from white-skinned Western populations. Age, sun exposure and other factors of the original lesions could not be controlled in our benchmark but might cause slight differences in performance of algorithms. However, most past publications were tested for white-skinned Western populations [5,7,8]. In addition, other factors such as age and sun exposure were not controlled for in past publications in this field [5–8].

5. Conclusions

We present the first public melanoma classification benchmark for both non-dermoscopic (MClass-ND) and dermoscopic (MClass-D) images for comparison of artificial intelligence algorithms with diagnostic performance of 145 and 157 dermatologists, respectively. Future publications should consider MClass as a reference standard for classification of melanocytic images of white-skinned Western populations for binary classification tasks.

Acknowledgements

The authors would like to thank and acknowledge the dermatologists who actively and voluntarily spend much time to participate in the reader study; some participants asked to remain anonymous, and they also thank these colleagues for their commitment. **Berlin**: 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, Dr. 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; **Würzburg:** Anja Gesierich, Valerie Glutsch, Katrin Kahlert, Andreas Kerstan, Bastian Schilling and Philipp Schrüfer.

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejca.2018.12.016.

Conflict of interest statement

None declared.

References

[1] Schadendorf D, van Akkooi ACJ, Berking C, et al. Melanoma. Lancet 2018;392(10151):971–84.

[2] Gordon LG, Rowell D. Health system costs of skin cancer and cost-effectiveness of skin cancer prevention and screening: a systematic review. Eur J Cancer Prev 2015;24(2):141–9.

[3] Brinker TJ, Klode J, Esser S, Schadendorf D. Facial-aging app availability in waiting rooms as a potential opportunity for skin cancer prevention. JAMA dermatology 2018;154(9):1085–6.

[4] Vestergaard M, Macaskill P, Holt PE, Menzies SW. 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.

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

[6] Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermatol 2018;138(7):1529–38.

[7] Marchetti MA, Codella NCF, Dusza SW, 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.

[8] Haenssle H, Fink C, Schneiderbauer R, 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.

[9] Gutman D, Codella NCF, Celebi E, 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:1605.01397.

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

[11] Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset: a large collection of multi-source dermatoscopic images of common pigmented skin lesions. 2018. arXiv preprint arXiv:1803.10417.

[12] Giotis I, Molders N, Land S, Biehl M, Jonkman MF, Petkov N. MED-NODE: a computer-assisted melanoma diagnosis system using non-dermoscopic images. Expert Syst Appl 2015;42(19):6578–85.

[13] Carli P, Quercioli E, Sestini S, 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.

[14] Dolianitis C, Kelly J, Wolfe R, Simpson P. Comparative performance of 4 dermoscopic algorithms by nonexperts for the diagnosis of melanocytic lesions. Arch Dermatol 2005;141(8):1008–14.

[15] Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: identifying density-based local outliers. In: ACM sigmod record; 2000, May.