European Journal of Cancer 120 (2019) 114–121

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

Superior skin cancer classification by the combination of human and artificial intelligence

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

[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 Kiel, Kiel, Germany*
[f] *Department of Dermatology, University Hospital Munich (LMU), Munich, Germany*
[g] *Department of Dermatology, University Hospital Regensburg, Regensburg, Germany*
[h] *Department of Dermatology, University Hospital Essen, Essen, Germany*
[i] *Department of Dermatology, University Hospital Würzburg, Würzburg, Germany*
[j] *Division of Biostatistics, German Cancer Research Center, Heidelberg, Germany*
[k] *Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA*

Received 28 June 2019; accepted 18 July 2019
Available online 10 September 2019

KEYWORDS
Artificial intelligence;
Deep learning;
Skin cancer;
Melanoma

Abstract ***Background:*** In recent studies, convolutional neural networks (CNNs) outperformed dermatologists in distinguishing dermoscopic images of melanoma and nevi. In these studies, dermatologists and artificial intelligence were considered as opponents. However, the combination of classifiers frequently yields superior results, both in machine learning and among humans. In this study, we investigated the potential benefit of combining human and artificial intelligence for skin cancer classification.

Methods: Using 11,444 dermoscopic images, which were divided into five diagnostic categories, novel deep learning techniques were used to train a single CNN. Then, both 112 dermatologists of 13 German university hospitals and the trained CNN independently classified a set of 300

* *Corresponding author*: National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 460, Heidelberg, 69120, Germany.
E-mail address: titus.brinker@dkfz.de (T.J. Brinker).
[1] These collaborators are listed in the acknowledgement section.

https://doi.org/10.1016/j.ejca.2019.07.019

biopsy-verified skin lesions into those five classes. Taking into account the certainty of the decisions, the two independently determined diagnoses were combined to a new classifier with the help of a gradient boosting method. The primary end-point of the study was the correct classification of the images into five designated categories, whereas the secondary end-point was the correct classification of lesions as either benign or malignant (binary classification).

Findings: Regarding the multiclass task, the combination of man and machine achieved an accuracy of 82.95%. This was 1.36% higher than the best of the two individual classifiers (81.59% achieved by the CNN). Owing to the class imbalance in the binary problem, sensitivity, but not accuracy, was examined and demonstrated to be superior (89%) to the best individual classifier (CNN with 86.1%). The specificity in the combined classifier decreased from 89.2% to 84%. However, at an equal sensitivity of 89%, the CNN achieved a specificity of only 81.5%

Interpretation: Our findings indicate that the combination of human and artificial intelligence achieves superior results over the independent results of both of these systems.

1. Background

With the advancement of artificial intelligence in medical diagnostics, many reader studies have been carried out to determine whether man or machine is the better diagnostician [1–4]. In these past studies, in particular in the field of skin cancer detection, man and machine were always regarded as competitors. However, this setting does not reflect clinical reality, where team-based diagnoses are considered to be more accurate than individual diagnoses [5–9]. This is practiced in the day-to-day routine, for example, by discussing difficult diagnostic questions in team meetings among colleagues and in grand rounds or by requesting second opinions. The same concept is used in machine learning, where so-called ensemble learners also use a finite set of different classifiers to obtain superior results to a single classifier [10]. However, to our knowledge, there is no data on the combination of human and artificial intelligence in the field of medical computer vision to date. Specifically, in the field of skin cancer detection, all past studies looked at the man against machine approach [1–4,11–19].

In this study, both physicians and a convolutional neural network (CNN) classified dermoscopic images of skin lesions into five diagnostic categories covering more than 90% of relevant lesions faced in skin cancer screening settings. Both systems decided independently of each other, and the decisions were then merged to form an overall classifier which would take the degree of certainty of both systems into account. If the errors of the two systems were only weakly correlated, a higher diagnostic accuracy was likely to be achievable.

2. Methods

2.1. Considered problem and used image archive

We considered the image classification task of the most common lesions that should be distinguished in a skin cancer screening setting. About 90% of lesions found in the routine skin examination are covered by the following five skin lesion classes: (1) actinic keratosis, intraepithelial carcinoma/Bowen's disease, squamous cell carcinoma (class a), (2) basal cell carcinoma (class b), (3) benign keratosis including seborrhoeic keratosis, solar lentigo and lichen planus–like keratosis (class k), (4) melanocytic nevi (class n) and (5) melanoma (class m). This image classification task was carried out by three different classifiers: dermatologists from German university hospitals, a CNN and a combination of the two aforementioned classifiers.

All images were obtained from the International Skin Imaging Collaboration (ISIC) archive; most images came from the HAM10000 Dataset [20]. This archive contains dermoscopic images of heterogeneous populations that are publicly accessible, anonymous and taken by different camera systems and thus have a high external validity. Because some images from the HAM10000 Dataset show the same lesion from different magnifications and angles, the data set was cleaned so that only one image per lesion was used. This data set was supplemented with an additional 4291 images from the ISIC archive. Using these restrictions, this study used 11,444 images; 6390 of which had been biopsy verified.

In this study, the test set consisted only of biopsy-verified images from the HAM10000 Dataset. To prevent selection bias for the 300 test images (60 for each of the five disease classes) from the available biopsy-verified image set, we programmed a random generator in Python.

2.2. Implementation of the reader study

An electronic questionnaire with the 300 test images was sent to 13 skin cancer experts at 13 university hospitals in Germany (Aachen, Berlin, Bonn, Essen, Heidelberg, Kiel, Leipzig, Mannheim, Magdeburg, München,

Regensburg, Rostock, Würzburg) which in turn sent one electronic questionnaire to their employed dermatologists via their official university email accounts. The ethics committee of the University of Heidelberg waived ethical approval due to the anonymity of the survey and the dermatologic images.

Because a concentrated evaluation of 300 images at a time is unpractical, the test set was split into six questionnaires, with 50 images each, and randomly assigned to each clinic. In the following, the set of images for questionnaire $i \in \{1, \ldots 6\}$ is always referred to with Σ_i.

Each dermatologist viewed 50 images of biopsy-verified skin lesions and answered two relevant questions about each image for this study. First, they identified the type of lesion most likely shown in the image. There were five possible answers to this question: (1) melanoma (class m); (2) nevus (class n); (3) basal cell carcinoma (class b); (4) actinic keratosis, Bowen's disease or squamous cell carcinoma (class a) and (5) seborrhoeic keratosis, lentigo solaris or lichen ruber planus (class k).Then, for each image, the dermatologists rated their uncertainty regarding the diagnosis on a scale from 0 (= very uncertain) to 10 (= very certain), with a value of 5 corresponding to a maximum uncertainty. More formally, the dermatologists $l \in \{1, 2, \ldots, N\}$ selected a value $c \in \{0, 1, \ldots, 10\}$ for each image $j \in \{1, 2, \ldots, 50\}$ of one of the sets Σ_i to quantify the corresponding decision certainty. In the following, the value is referred to with $c_l(x_j^{(i)})$.

All parts and questions of the questionnaire were mandatory, and the participants received the correct answers to the differential diagnoses at the end of the survey. In anonymous surveys, the fact that some surveys are filled out carelessly or in a rush should be accounted for. To handle this issue, we predefined statistical criteria for outlier detection.

2.3. Training of the CNN

To maximise the training set, one CNN was trained for each test set Σ_i. The training and test set were made disjunctive by removing all test images belonging to a given questionnaire from the corresponding training set, leaving 11,394 images for training. Because the distribution of images across the classes was different (i.e., class imbalance), class n was down sampled by a factor of 2, whereas the others were up sampled by a factor of 3 (via data augmentation). Thus, the final training set consisted of 585 images of class a, 910 images of class b, 3101 images class k, 4219 images of class n and 3521 images of class m, giving a total of 12,336.

Based on good previous experience with the ResNet50 architecture for skin lesion classification, we opted to use a ResNet50 model for this study. Kassani *et al.* [21] confirmed this experience with quantitative experiments. For more technical details about the training procedure, we refer the reader to Appendix 1.

2.4. Fusion of dermatologist's and CNN's decision

To fuse the two classification results of dermatologists and CNN into a more precise classifier, the CNN outputs and the dermatologists' responses had to be processed appropriately. For the considered problem, the output $o = f(x_j^{(i)}) \varepsilon [0, 1]^5$ of the CNN for a given image $x_j^{(i)} \varepsilon \Sigma_i$ was five-dimensional and was is given that $\| o \|_1 = 1$. This meant that the output values could be interpreted as a probability distribution over the given five classes. In addition to the decision for one class for each image, we asked the dermatologists how sure they felt about their decision. This specification of confidence was also used as an input for the fusion algorithm ($c_l(x_j^{(i)})$). For this, the given value was first divided by 10, so that the result was always in [0,1] and was interpreted as a probability value for the actually selected class. Because no further information about the certainty of the dermatologists had been collected, a uniform distribution of the remaining probability mass was assumed. As a result, the other four remaining classes were assigned the probability value $(1\text{-}c_l(x_j^{(i)}))/4$.

A naive approach would have been to fuse the two discrete distributions according to probability theory. This would, however, have required that the dermatologists' estimation of their certainty is as objective as the CNN's. Of course, this is generally not the case, and there is bias in their estimates. That is why we decided to use a different fusion method in this study. Instead of an analytical solution, the relationship between the true class label and the aforementioned probability distributions of the CNN and the dermatologists was learned by a machine learning algorithm, more specifically with XGBoost. XGBoost is an efficient tree algorithm with gradient boosting. Training was iterative by adding new trees that predicted the error of previous trees, which were then linked to the previous trees to produce a final classification. To obtain the largest possible overall test set, we tested six different XGBoost models for each of the six image sets Σ_i. For a given test image set Σ_i the union $\bigcup_j \Sigma_j$ for $j \varepsilon \{1, 2, \ldots 6\} \setminus \{i\}$ was used for training the model. Because only one questionnaire had been sent to the dermatologist, there were no assessments of the same dermatologists in training and test set. The hyperparameters of the XGBoost algorithm were tuned by means of randomised search. For this purpose, the function 'RandomizedSearchCV' available in scikit-learn was used, which uses cross-validation. The corresponding test set was therefore not used to optimise the algorithm's hyper parameter.

2.5. Statistical analysis

Clinical application of skin lesion classification has two potential goals: giving specific information and treatment options for a lesion and detecting early skin cancer with a reasonable sensitivity and specificity. The first

Table 1
Overview of achieved mean accuracies of physicians, CNN and fusion method.

Questionnaire (n specifies number of dermatologists)	Mean accuracy physicians	Mean accuracy CNN	Mean accuracy fusion method
1 (n = 17)	36.59%	78%	80.95%
2 (n = 14)	47%	82%	80.89%
3 (n = 16)	46%	88%	90.21%
4 (n = 14)	40.14%	84%	81.91%
5 (n = 21)	46.29%	80%	80.60%
6 (n = 30)	42.20%	80%	83.20%
Overall	**42.94%**	**81.59%**	**82.95%**

CNN, convolutional neural network.

task needs a precise diagnosis of a fine-granular class, so a multiclass problem, whereas the second demands a binary decision 'biopsy' versus 'no biopsy'.

Therefore, the results from the three classifiers (dermatologist, CNN and fusion of both) were evaluated from two standpoints:

The first assessed performance with respect to the differential diagnosis task. If the classes in the test set are totally balanced, as it was in our case, accuracy is an adequate and besides that easily understandable and interpretable measure to determine the classification quality in a multiclass problem.

The second approach measured how well benign lesions were distinguished from malignant ones. To convert the multiclass output from the dermatologists and the CNN into a binary output, each output class was mapped to either benign or malignant according to the nature of the class (e.g., class a is mapped to malignant). Then, the characteristic quantities of specificity and sensitivity were determined.

The accuracy of the multiclass problem as well as the sensitivity, specificity and overall classification of the binary problem were compared statistically by using separate (two-sided) McNemar-tests in the form of 2×2 tables. Statistical significance was a value of ($p < 0.05$).

For both standpoints, the confusion matrices were calculated for all three classifiers, which provided a good insight into the number of correctly classified and misclassified images including the corresponding distribution. Because the XGBoost contains stochastic components in the classification, not only one run was used to present the results but the result was averaged over 20 runs, respectively. Thereby, cherry picking is prevented.

3. Results

A total of 117 dermatologists from 13 German clinics participated in this study. Five were removed for not fulfilling the preset statistical specifications, so that 112 dermatologists remained in the study. Table 1 summarises the results of the multiclass classification problem by listing the mean achieved accuracies of the three classifiers. In Fig. 1, the boxplots of the achieved accuracies by physicians and by the new method are shown.

In Table 2 the results for the binary decision 'biopsy/treatment' versus 'no biopsy' are presented.

For each of the three classifiers the average sensitivity and specificity is listed. To show additionally the distribution of the classifications, the corresponding confusion matrices are depicted in Fig. 2 for both problems - the multiclass problem and the binary problem.

To better understand the lower specificity of the combined classifier for the binary problem, Fig. 3

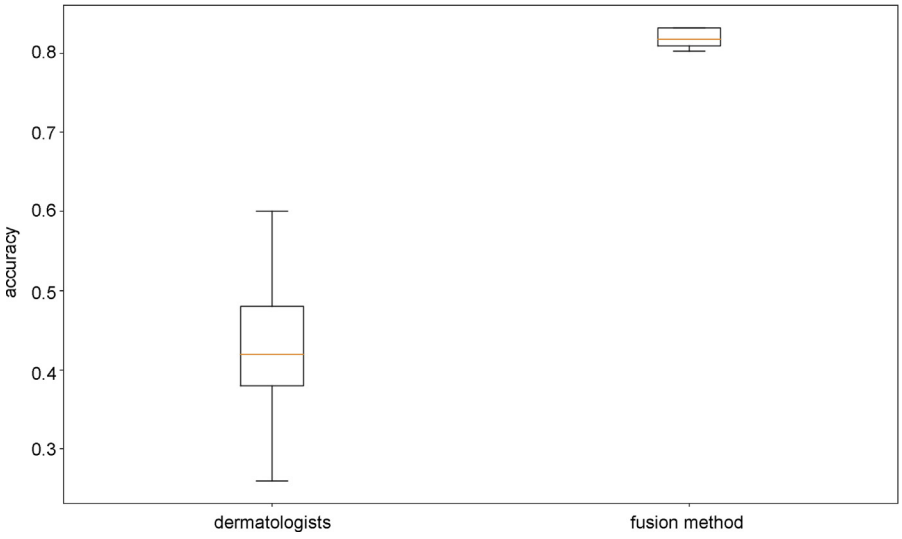

Fig. 1. Boxplot of the achieved accuracies by dermatologists and the new method.

Table 2
Comparison of the achieved sensitivities and specificities over the complete test set.

Used classifier	Mean sensitivity	Mean specificity
Physicians	66% (95% CI: 59.1%–72.9%)	62% (95% CI: 53.3%–70.7%)
CNN	86.1% (95% CI: 81.1%–91.2%)	89.2% (95% CI: 83.6%–94.7%)
Fusion method	**89% (95% CI: 84.4%–93.6%)**	**84% (95% CI: 77.4%–90.6%)**

CNN, convolutional neural network.

additionally shows the ROC curve of the CNN. CNN, convolutional neural network; ROC, receiver operating characteristic.

Here, the output values of the CNN were averaged over all 10 test runs, and the specificity with the corresponding sensitivity was calculated for each cutoff value between 0 and 1. At a sensitivity of 89%, the CNN only achieved a specificity of 81.5%.

The results of the individual McNemar tests showed that the increase in sensitivity by the combined classifier in the binary problem was statistically significantly superior to the CNN alone ($p < 0.05$). All other tests did not reach statistical significance.

4. Discussion

In machine learning, the technique of combining multiple classifiers to form an ensemble is commonly used to improve performance. Recent reader studies compared man against machine and examined which of the two had the better diagnostic accuracy. In this study, we considered both classifiers not as counterparties but as two independent tests with their respective strengths and weaknesses, and we combined their results to a new ensemble classifier.

If the mean accuracies of the five-class problem are taken into account, it is noticeable that the CNN achieved a significantly higher accuracy than the dermatologists. However, when the decisions of dermatologists and the outputs of the CNN were combined using the XGBoost algorithm, the results of the CNN were exceeded even further by exactly 1.36%. While this result was not statistically significant, we did demonstrate that such a combination is feasible and has the potential to improve the already good performance of the CNN even further. Almost even more interesting is the result of the fusion in the binary decision problem. Owing to the combination of man and machine, the sensitivity increased from 86.1% (exclusively decided by CNN) to 89% by incorporating the human decision, and these results were significant ($p < 0.05$). At an equal sensitivity of 89%, the CNN achieved a specificity of only 81.5%.

In the opinion of the authors, the combination of human decisions with those of artificial intelligence to achieve a higher accuracy is not limited to skin cancer classficiation of static lesionsbut can also be applied to

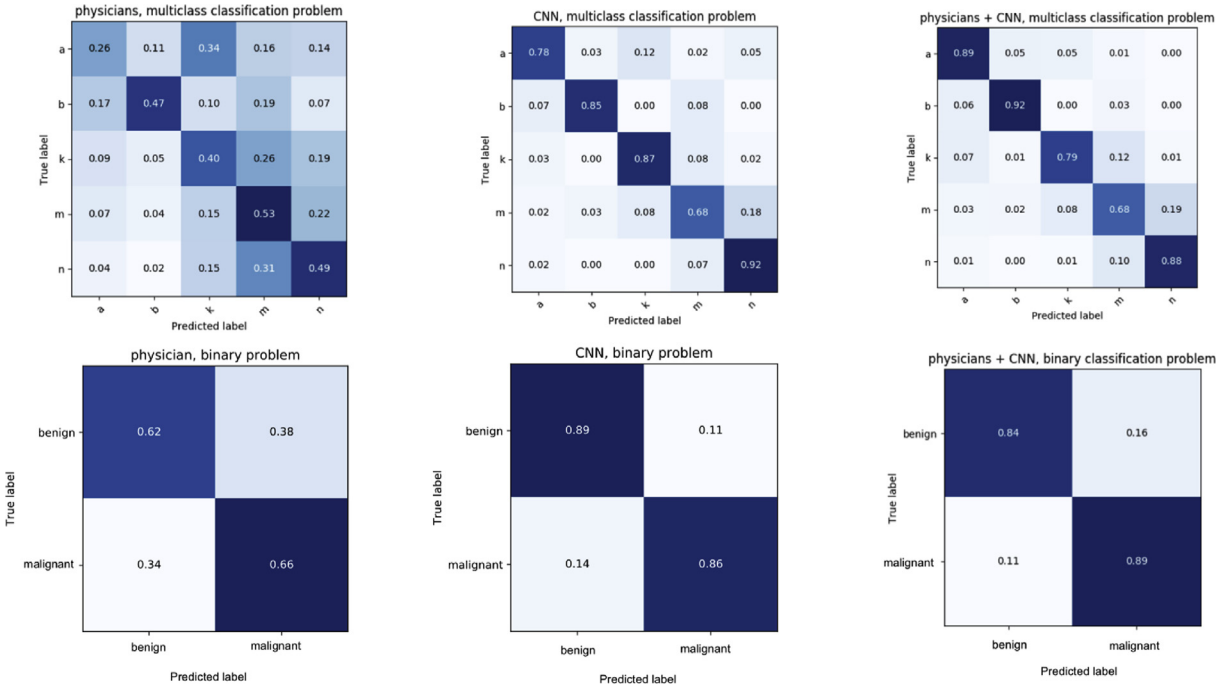

Fig. 2. **Confusion matrices of the three classifiers.** The upper row shows the classification results of the multiclass problem; the lower row shows the results of the binary classification problem.

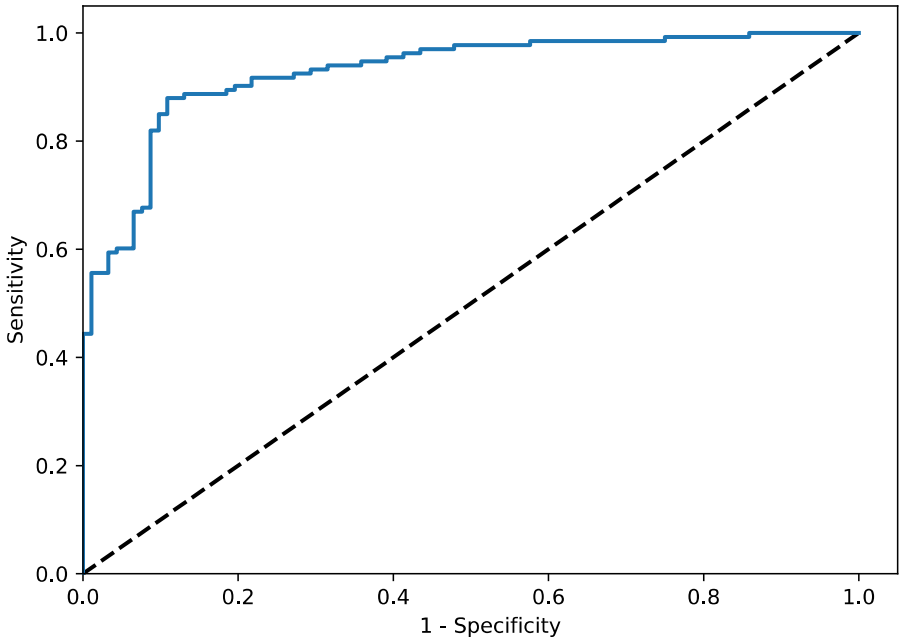

Fig. 3. Averaged ROC curve over the 10 test runs of the CNN. CNN, convolutional neural network.

the prospective assessment of the oncologic transformation of nevi as recently outlined in this Journal [22] and to other medical questions.

4.1. Limitations

4.1.1. Data selection

As only biopsy-verified images were considered for the survey, a certain bias is introduced, as these lesions are difficult to diagnose which provides an explanation for the relatively low accuracy of the participating dermatologists.

4.1.2. Anonymity

To comply with the privacy policy, the survey provided to the dermatologists was conducted anonymously. However, anonymity carries the risk of abuse and carelessly provided answers. By involving physicians exclusively through their institutional email address, this risk was minimised, and a high plausibility rate was achieved.

4.1.3. Dermatology decision based exclusively on one image

During clinical examination of the patient, more information is available to the physician for the diagnosis than just the visual impression of the examined skin area.

For example, a palpation examination can be performed, or the affected skin area can be related to other skin lesions of the patient. Other clinical data, such as age or family history, also contribute to decision making. Hänßle *et al.* showed integrating this additional clinical information can slightly improve dermatologists' sensitivity and specificity [2]. In principle, this information may also be taken into account by machine learning methods, leading to better classification quality in the future.

4.1.4. Generalisability of the performance of our algorithm

Our test, training and validation set were disjunct. However, they stem from the same database. While this database is of heterogenous character (i.e. a high number of different cameras were used to create the images) it still may be assumed that the algorithms performance would be worse on an entirely external dataset of images. In a recent study, a binary-classification CNN (nevus vs. melanoma) was trained on the same set of images that we used (ISIC database). It showed good performance on an ISIC test set but performed worse on an external test set from the PH2 dermoscopic image database [28]. However, using just 100 images from the external test set for training of the last fully-connected layers, sufficed to completely restore performance. This indicates that deep learning algorithms can easily be calibrated to new dermatoscopes or other forms of preprocessing by the help of training with relatively few images with these new properties.

4.1.5. Complementary efforts to increase awareness

Skin cancer screenings may only be effective if individuals at risk are aware of their importance. We therefore want to underline recent complementary research efforts focusing on the increase of awareness by the use of the free photoaging mobile app "Sunface" developed by the authors. This selfie-app shows the user the results of UV-radiation on their own face and has

photoaged more than >500,000 faces to date. Our research data indicate that this app may successfully introduce behavioral change on a population-level [23–27].

5. Conclusions

To the best of our knowledge, this is the first study in the field of digital skin diagnostics that has combined the decisions of dermatologists and artificial intelligence. Although our results are not significant, they indicate superiority of the 'man with machine' approach on 300 test images obtained from a heterogenous data set with high external validity.

Funding

This research is funded by the Federal Ministry of Health in Germany (Skin Classification Project (skin-class.de); grant holder: Dr. Titus J. Brinker, MD, German Cancer Research Center (DKFZ), Heidelberg, Germany).

Acknowledgements

The authors would like to thank and acknowledge the dermatologists who actively and voluntarily spent much time to participate in the reader study; some participants asked to remain anonymous, and the authors also thank these colleagues for their commitment. Aachen: Laurenz Schmitt; Berlin (Charité): Wiebke K. Peitsch; Bonn: Friederike Hoffmann; Essen: Jürgen C. Becker, Christina Drusio, Philipp Jansen, Joachim Klode, Georg Lodde, Stefanie Sammet, Dirk Schadendorf, Wiebke Sondermann, Selma Ugurel, Jeannine Zader; Heidelberg: Alexander Enk, Martin Salzmann, Sarah Schäfer, Knut Schäkel, Julia Winkler, Priscilla Wölbing; Kiel: Hiba Asper, Ann-Sophie Bohne, Victoria Brown, Bianca Burba, Sophia Deffaa, Cecilia Dietrich, Matthias Dietrich, Katharina Antonia Drerup, Friederike Egberts, Anna-Sophie Erkens, Salim Greven, Viola Harde, Marion Jost, Merit Kaeding, Katharina Kosova, Stephan Lischner, Maria Maagk, Anna Laetitia Messinger, Malte Metzner, Rogina Motamedi, Ann-Christine Rosenthal, Ulrich Seidl, Jana Stemmermann, Kaspar Torz, Juliana Giraldo Velez; Leipzig: Jennifer Haiduk; Magdeburg: Mareike Alter, Claudia Bär, Paul Bergenthal, Anne Gerlach, Christian Holtorf, Ante Karoglan, Sophie Kindermann, Luise Kraas; Mannheim: Moritz Felcht, Maria R Gaiser, Claus-Detlev Klemke, Hjalmar Kurzen, Thomas Leibing, Verena Müller, Raphael R. Reinhard, Jochen Utikal, Franziska Winter; Munich: Carola Berking, Laurie Eicher, Daniela Hartmann, Markus Heppt, Katharina Kilian, Sebastian Krammer, Diana Lill, Anne-Charlotte Niesert, Eva Oppel, Elke Sattler, Sonja Senner, Jens Wallmichrath, Hans Wolff; Würzburg: Anja Gesierich, Tina Giner, Valerie Glutsch, Andreas Kerstan, Dagmar Presser, Philipp Schrüfer, Patrick Schummer, Ina Stolze, Judith Weber; Regensburg: Konstantin Drexler, Sebastian Haferkamp, Marion Mickler, Camila Toledo Stauner; Rostock: Alexander Thiem.

Conflict of interest statement

None declared.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejca.2019.07.019.

References

[1] Esteva A, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115.
[2] Haenssle HA, 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.
[3] Rajpurkar P, et al. Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. 2017. arXiv preprint arXiv:1711.05225.
[4] Brinker TJ, 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.
[5] Barnett ML, et al. Comparative accuracy of diagnosis by collective intelligence of multiple physicians vs individual physicians. JAMA Ntwork Oen 2019;2(3). e190096-e190096.
[6] Wolf M, et al. Collective intelligence meets medical decision-making: the collective outperforms the best radiologist. PLoS One 2015;10(8). e0134269.
[7] Kurvers RH, et al. Detection accuracy of collective intelligence assessments for skin cancer diagnosis. JAMA Dermatol 2015;151(12):1346–53.
[8] Kurvers RH, et al. Boosting medical diagnostics by pooling independent judgments. Proc Natl Acad Sci 2016;113(31):8777–82.
[9] Kämmer JE, et al. The potential of collective intelligence in emergency medicine: pooling medical students' independent decisions improves diagnostic performance. Med Decis Mak 2017;37(6):715–24.
[10] Murphy KP. Machine learning: a probabilistic perspective. MIT press; 2012.
[11] Maron R, et al. Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. Eur J Canc 2019;119:57–65.
[12] Brinker T, et al. Skin cancer classification using convolutional neural networks: systematic review. J Med Internet Res 2018;20(10):e11936.
[13] Hekler A, et al. Pathologist-level classification of histopathological melanoma images with deep neural networks. Eur J Canc 2019;115:79–83.
[14] Brinker TJ, et al. Deep neural networks are superior to dermatologists in melanoma image classification. Eur J Canc 2019;119:11–7.
[15] Hekler A, et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. Eur J Canc 2019;118:91–6.

[16] Brinker TJ, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. Eur J Canc 2019; 111:148–54.
[17] Brinker TJ, et al. Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark. Eur J Canc 2019;111:30–7.
[18] Tschandl P, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. JAMA Dermatol 2019;155(1):58–65.
[19] Tschandl P, 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 July 2019;20(7):891–2.
[20] 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.
[21] Kassani SH, Kassani PH. A comparative study of deep learning architectures on melanoma detection. Tissue Cell 2019;58:76–83.
[22] Sondermann Wiebke, et al. Prediction of melanoma evolution in melanocytic nevi via artificial intelligence: a call for prospective data. Eur J Cancer 2019;119:30–4.
[23] Brinker Titus Josef, et al. Photoaging mobile apps as a novel opportunity for melanoma prevention: pilot study. JMIR mHealth uHealth 2017;5.7:e101.
[24] Brinker Titus Josef, et al. Photoaging mobile apps in school-based melanoma prevention: pilot study. J Med Int Res 2017;19.9:e319.
[25] Brinker Titus J, et al. Facial-aging app availability in waiting rooms as a potential opportunity for skin cancer prevention. JAMA Dermatol 2018;154.9:1085–6.
[26] Brinker Titus Josef, et al. A skin cancer prevention facial-aging mobile app for secondary schools in Brazil: appearance-focused interventional study. JMIR mHealth uHealth 2018;6.3:e60.
[27] Brinker Titus Josef, et al. A skin cancer prevention photoageing intervention for secondary schools in Brazil delivered by medical students: protocol for a randomised controlled trial. BMJ Open 2018;8.3:e018299.
[28] Mendonça Teresa, 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.