European Journal of Cancer 119 (2019) 57–65

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks

Roman C. Maron [a,1], Michael Weichenthal [b,1], Jochen S. Utikal [c,d], Achim Hekler [a], Carola Berking [e], Axel Hauschild [b], Alexander H. Enk [f], Sebastian Haferkamp [g], Joachim Klode [h], Dirk Schadendorf [h], Philipp Jansen [h], Tim Holland-Letz [i], Bastian Schilling [j], Christof von Kalle [a], Stefan Fröhling [a], Maria R. Gaiser [c,d], Daniela Hartmann [e], Anja Gesierich [j], Katharina C. Kähler [b], Ulrike Wehkamp [b], Ante Karoglan [k], Claudia Bär [k], Titus J. Brinker [a,f,*], Collabrators[2]

[a] *National Center for Tumor Diseases, German Cancer Research Center, Heidelberg, Germany*
[b] *Department of Dermatology, University Hospital Kiel, Kiel, Germany*
[c] *Department of Dermatology, Heidelberg University, Mannheim, Germany*
[d] *Skin Cancer Unit, German Cancer Research Center, Heidelberg, Germany*
[e] *Department of Dermatology, University Hospital Munich (LMU), Munich, Germany*
[f] *Department of Dermatology, University Hospital Heidelberg, Heidelberg, Germany*
[g] *Department of Dermatology, University Hospital Regensburg, Regensburg, Germany*
[h] *Department of Dermatology, University Hospital Essen, Essen, Germany*
[i] *Department of Biostatistics, German Cancer Research Center, Heidelberg, Germany*
[j] *Department of Dermatology, University Hospital Würzburg, Würzburg, Germany*
[k] *Department of Dermatology, University Hospital Magdeburg, Magdeburg, Germany*

Received 30 May 2019; received in revised form 19 June 2019; accepted 21 June 2019
Available online 14 August 2019

* *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 authors contributed equally to this work.
[2] These collaborators are listed in the acknowledgement section.

https://doi.org/10.1016/j.ejca.2019.06.013

KEYWORDS
Skin cancer;
Artificial intelligence;
Melanoma;
Skin cancer screening

Abstract ***Background:*** Recently, convolutional neural networks (CNNs) systematically outperformed dermatologists in distinguishing dermoscopic melanoma and nevi images. However, such a binary classification does not reflect the clinical reality of skin cancer screenings in which multiple diagnoses need to be taken into account.
Methods: Using 11,444 dermoscopic images, which covered dermatologic diagnoses comprising the majority of commonly pigmented skin lesions commonly faced in skin cancer screenings, a CNN was trained through novel deep learning techniques. A test set of 300 biopsy-verified images was used to compare the classifier's performance with that of 112 dermatologists from 13 German university hospitals. The primary end-point was the correct classification of the different lesions into benign and malignant. The secondary end-point was the correct classification of the images into one of the five diagnostic categories.
Findings: Sensitivity and specificity of dermatologists for the primary end-point were 74.4% (95% confidence interval [CI]: 67.0–81.8%) and 59.8% (95% CI: 49.8–69.8%), respectively. At equal sensitivity, the algorithm achieved a specificity of 91.3% (95% CI: 85.5–97.1%). For the secondary end-point, the mean sensitivity and specificity of the dermatologists were at 56.5% (95% CI: 42.8–70.2%) and 89.2% (95% CI: 85.0–93.3%), respectively. At equal sensitivity, the algorithm achieved a specificity of 98.8%. Two-sided McNemar tests revealed significance for the primary end-point ($p < 0.001$). For the secondary end-point, outperformance ($p < 0.001$) was achieved except for basal cell carcinoma (on-par performance).
Interpretation: Our findings show that automated classification of dermoscopic melanoma and nevi images is extendable to a multiclass classification problem, thus better reflecting clinical differential diagnoses, while still outperforming dermatologists at a significant level ($p < 0.001$).

1. Background

Skin cancer is the most common malignancy in white-skinned individuals, and melanoma accounts for most skin cancer–related deaths worldwide [1]. Early detection of melanoma improves survival, and thus, several institutions and entities of the European Union fund programs for skin cancer screening that lead to earlier diagnoses [2].

The success of skin cancer screening is highly dependent on the accuracy and diagnostic ability of dermatologists conducting the skin examination. However, dermatologists rarely achieve sensitivities exceeding 80% in skin cancer screening settings [3]. In 2017, Esteva *et al.* [4], were the first to report on a deep learning convolutional neural network (CNN) image classifier that performed. In addition, this CNN generated its own diagnostic criteria for melanoma detection during training. Several follow-up publications by other authors have demonstrated that CNNs may exceed a dermatologist's performance in distinguishing melanoma from nevi and suggested their regular use in clinical practice to assist physician diagnoses [5–9,17-19]. In addition, most recently deep learning indicated superiority over pathologists in distinguising melanoma from nevi in histopathological images [20,21]. However, a binary classification problem does not reflect the clinical reality of skin cancer screening in which entities of multiple diseases need to be taken into account.

As a result, different investigators started to conduct reader studies with multiple skin lesions that are more reflective of a skin cancer screening setting [10–13]. However, a systematic outperformance of dermatologists in such a multiclass study was not achieved to date.

In this article, we compared the sensitivity and specificity of 112 German dermatologists from 13 university hospitals with that of a single CNN in detecting the most common lesions that should be distinguished in a skin cancer screening setting: actinic keratosis, intraepithelial carcinoma/Bowen disease, squamous cell carcinoma, basal cell carcinoma, benign keratosis including seborrhoeic keratosis, solar lentigo and lichen planus–like keratosis, melanocytic nevi and melanoma. More than 90% of lesions. The majority of pigmented skin lesions found in routine skin examinations in skin cancer screening settings are covered by these lesions.

2. Methods

2.1. Study design

This comparative study was conducted from 15th January 2019 (design of study) to 25th May 2019 (completion of data analysis and manuscript approval by all authors). The completion of anonymous electronic questionnaires was undertaken from 23rd April 2019 to 7th May 2019. Dermatologists were included via randomly assigned links to the department directors of

13 university hospitals who were asked to send one questionnaire to their employed dermatologists via an official university email account. The Ethics Committee of Heidelberg University waived ethical approval owing to the anonymity of the survey and dermatologic images.

2.2. Data sets

2.2.1. Training and validation set

All images were obtained from the International Skin Imaging Collaboration (ISIC) archive; most images came from the HAM10000 Dataset [14]. This archive contains dermoscopic images of heterogeneous populations that are publicly accessible, anonymous and taken by different camera systems. Because some images from the HAM10000 Dataset show the same lesion from different magnifications and angles, duplicates were removed by lesion ID (provided by HAM10000 creators) so that only one image per lesion was used. This data set was supplemented with an additional 4291 images from the ISIC archive. A sufficiently large number of images per skin disease should be presented to dermatologists to attain a statistically reliable value for sensitivity and specificity. For this reason, we limited this study to the differential diagnoses of the most frequent skin diagnoses in the archive, split into five compound classes: (1) actinic keratosis (solar keratosis), intraepithelial carcinoma (Bowen disease), squamous cell carcinoma (akiec); (2) basal cell carcinoma (bcc); (3) benign keratosis, including seborrhoeic keratosis, solar lentigo and lichen planus–like keratosis (bkl); (4) melanocytic nevi (nv) and (5) melanoma (mel). Class division was based on the diagnostic categories set up by the HAM10000 Dataset creators. Using these restrictions, this study used 11,444 images, 6390 of which had been biopsy verified.

2.2.2. Test set

In this study, only biopsy-verified images from the HAM10000 Dataset were used for evaluating the algorithm. 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. The test set is available for downloading at (https://skinclass.de/TestSet.zip).

2.3. Participants and electronic questionnaire

To compare the CNN results with those of the dermatologists, an electronic questionnaire with the 300 test images was sent to 13 leading dermatologists at 13 university hospitals in Germany (Aachen, Berlin, Bonn, Essen, Heidelberg, Kiel, Leipzig, Mannheim, Magdeburg, München, Regensburg, Rostock and Würzburg). Because a concentrated evaluation of 300 images at a time is overtaxing, the test set was split into six questionnaires, with 50 images each, and randomly assigned to each clinic.

The first part of the questionnaire recorded metadata about the participating physician, including age, gender, years of dermatologic practice, years of dermoscopic experience, estimated number of performed skin cancer screenings in the last year and position within the medical hierarchy. In the second part of the questionnaire, the dermatologist viewed 50 images of biopsy-verified skin lesions. The participant answered four questions about each image. First, the participant evaluated the quality of the shown image (excellent, good, sufficient, poor, other image problems and no image visible). Second, the participant estimated whether the lesion image shown was benign (nevus, seborrhoeic keratosis, solar lentigo and lichen ruber planus) or malignant (melanoma, basal cell carcinoma, actinic keratosis, Bowen disease and squamous cell carcinoma). The participant then quantified the certainty of the decision on a scale of 0 (certain benign) to 10 (certain malignant). A value of 5 corresponded to a maximum uncertainty regarding this decision. The dermatologist then identified the type of lesion most likely shown in the image. There were five possible answers to this question: (1) melanoma (mel); (2) nevus (nv); (3) basal cell carcinoma (bcc); (4) actinic keratosis, Bowen disease or squamous cell carcinoma (akiec) and (5) seborrhoeic keratosis, lentigo solaris or lichen ruber planus (bkl). Finally, for each picture, 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. 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.

2.4. Training of the CNN model

To maximise the training set, one CNN was trained for each questionnaire. 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), nv was downsampled by a factor of 2, whereas the others were upsampled by a factor of 3 (via data augmentation). Thus, the final training set consisted of 585 akiec, 910 bcc, 3101 bkl, 4219 nv and 3521 mel images, giving a total of 12,336 images.

Based on good previous experience with the ResNet50 architecture for skin lesion classification, we opted to use a ResNet50 model for this study. Kassani and Kassani [15] confirmed this experience with quantitative experiments. Appendix 1 outlines details on the training procedure.

2.5. Survey evaluation

In anonymous surveys, the fact that some surveys are filled out carelessly or in a rush should be accounted for. To handle this issue, individual answers to each question are combined using methods tailored to the nature of the answers. For image quality (categorical), answers were converted into a grade (integer, range: 1–5, 1 = excellent, 5 = other image problems/no image visible), and the arithmetic mean was then taken. To get a combined estimate of whether the lesion was considered benign or malignant (numeric, range: 0–10), the median was taken and converted to benign if < 5 or malignant if ≥ 5. For the differential diagnosis question (categorical), the majority decision was used to assign a class to a single image by ranking the individual answers against each other and taking the most frequent class. To get a combined certainty estimate for the differential diagnoses, the median was taken. These values were considered the dermatologists' confidence levels in their decisions. For the CNN, confidence values were the output probability of the network for each class.

Of the 117 survey responses, 5 were removed for not fulfilling preset statistical specifications. A subset of 75 images was taken out of the 300 images because they scored above a certain quality threshold, and dermatologists experienced a performance drop when classifying this subset. The threshold was set at 2.5, excluding images that, on average, were marked insufficient or inadequate.

2.6. Performance and statistical analysis

The results from the dermatologists and the classifier were evaluated from two standpoints:

The first 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. akiec would be mapped to malignant). An additional approach was used for dermatologists to convert their estimate regarding benignancy and malignancy (survey question 2) into benign and malignant, as described in the previous section. From here on, the former will be referred to as DD (for differential diagnoses), with the latter referred to as Est. (for the estimation of benignancy/malignancy).

The second assessed performance with respect to the differential diagnosis task. Because sensitivity and specificity are metrics used in binary classification, multiclass classification results are binarised, with the classification performance of a single class versus the remaining classes calculated (one-vs-all approach).

Sensitivity, specificity and the overall rates of correct classification were compared statistically for the multiclass classification tasks by using 15 separate (two-sided) McNemar tests in the form of 2×2 tables. For the multiclass classification task, each class was considered individually in a one-vs-all approach, which resulted in five 2×2 tables for each metric. In a similar manner, another three tables were used for the binary classification task. Statistical significance was a value of $p < 0.05$ and was corrected after Bonferroni correction for the multiclass classification task to counteract the multiple comparison problems.

3. Results

A total of 112 dermatologists from 13 German-based clinics participated in this study. Thirty-six (32.1%) participants were men, and 76 (67.9%) were women. The median years of dermatologic practice was 4 years, with 47 (42.0%) dermatologists possessing less than 3 years of experience with dermoscopic examinations, 37 (33.0%) between three and 10 years, and 28 (25.0%) with more than 10 years. Table 1 shows the participants' distribution within the medical hierarchy. Each image was seen and diagnosed by a minimum of 14 and a maximum of 30 dermatologists (median: 17.5), with confidence levels ranging from 4 to 10 (median: 6.5). Fig. 1 compares the dermatologists' and the model's average and class-specific performance. Fig. 2 shows six mean receiver operating characteristic curves over a total of 10 runs (blue line) in comparison with the dermatologists' overall performance (blue circle) and with their performance broken down into the various employment types (coloured symbols). The plot in the upper left corner shows the average performance of the classifier and the dermatologists over all five classes (see also Table 1), whereas the other plots show the results for the one-vs-all approach, in which each class is considered individually against the rest (e.g. melanoma vs non-melanoma).

Table 1
Dermatologists' diagnostic performance showing sensitivities and specificities for the binary (right) and multiclass classification task (left).

Sample	Differential diagnoses (5-class average)		Benign vs malignant (Est.)	
	Sensitivity	Specificity	Sensitivity	Specificity
All participants (n = 112)	56.5%	89.2%	74.4%	59.8%
Resident physicians (n = 4)	73.6%	93.4%	84.1%	27.6%
Chief physicians (n = 1)[a]				
Senior physicians (n = 28)	61.1%	90.3%	68.4%	72.8%
Attendings (n = 12)	50.4%	87.7%	69.2%	56.5%
Junior physicians (n = 67)	54.8%	88.7%	75.9%	54.3%

Est., estimation of benignancy/malignancy.
Metrics for the multiclass classification task were calculated by taking the arithmetic mean of sensitivity and specificity over all five classes.
[a] Sample size too small to make a statement.

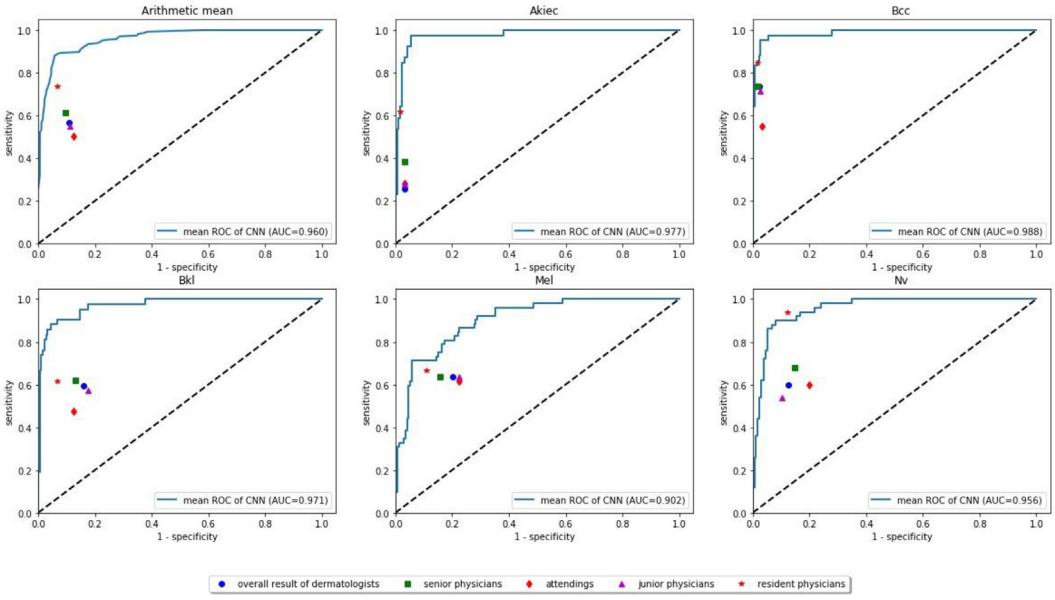

Fig. 1. **Comparison of the dermatologists' and the model's average and class-specific performance.** Each plot shows the classifier's mean ROC curve versus the majority voting of all dermatologists combined and grouped according to their position in the clinical hierarchy. Chief physicians' answers were omitted owing to the insufficient sample size. CNN, convolutional neural network; ROC, receiver operating characteristic.

Dermatologists achieved an overall mean sensitivity and specificity of 56.5% (95% confidence interval [CI]: 42.8–70.2%) and 89.2% (95% CI: 85.0–93.3%), respectively. At equal sensitivity, the average specificity of the classifier lies at 98.8%, with an AUC of 0.960. Table 2 shows a summary of both metrics broken down into five one-vs-all binary classifications, with the comparison of the overall dermatologists' score to the CNN's score for each class at equal sensitivity. The lowest overall sensitivity achieved by the dermatologists was 25.6% for akiec, a class in which the classifier demonstrated strong performance (area under the curve (AUC) of 0.977). Both dermatologists and the classifier performed best for the bcc class. The lowest score achieved for the classifier was for the mel class, with a specificity of 94.2% (AUC of 0.902).

Using the numeric estimate of benignancy/malignancy for the evaluation of the binary classification task (Est.), the dermatologists achieved an overall sensitivity of 74.4% (95% CI: 67.0–81.8%), at a specificity of 59.8% (95% CI: 49.8–69.8%). If mapping of differential diagnosis answers to benignancy/malignancy was used for evaluation (DD) instead, an overall sensitivity and specificity of 69.2% (95% CI: 61.3–77.0%) and 70.7% (95% CI: 61.3–80.0%) was achieved, respectively. At the highest sensitivity of 74.4%, the classifier's specificity was 91.3% (95% CI: 85.5–97.1%) (AUC of 0.928) (Fig. 2).

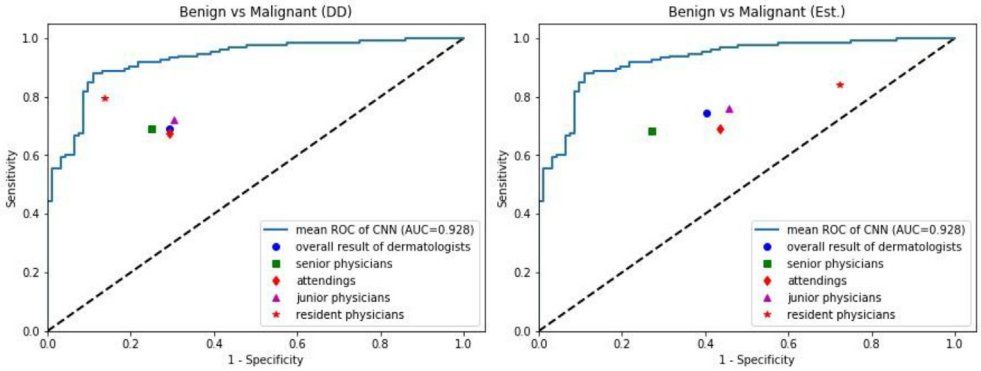

Fig. 2. **Comparison of dermatologists' and the model's performance for the binary classification task.** The model's ROC curve is identical across both plots. The dermatologists' performance varied depending on the method used to classify their answers. DD, differential diagnoses; Est., estimation of benignancy/malignancy; CNN, convolutional neural network; ROC, receiver operating characteristic.

Table 2
Comparison of dermatologists' overall specificity to the classifier's specificity at equal sensitivity for each class.

	Sensitivity	Specificity	
		Dermatologists	CNN
Akiec	25.6% (95% CI: 11.9–39.3%)	96.8% (95% CI: 94.2–99.3%)	99.5% (95% CI: 98.5–100%)
Bcc	73.8% (95% CI: 60.5–87.1%)	97.8% (95% CI: 95.7–99.9%)	99.5% (95% CI: 98.5–100%)
Bkl	59.5% (95% CI: 44.7–74.4%)	84.2% (95% CI: 78.9–89.4%)	99.5% (95% CI: 98.5–100%)
Mel	63.5% (95% CI: 50.4–76.5%)	79.8% (95% CI: 73.8–85.8%)	94.2% (95% CI: 90.7–97.7%)
Nv	60.0% (95% CI: 46.4–73.6%)	87.4% (95% CI: 82.5–92.3%)	97.1% (95% CI: 94.6–99.6%)

CI, confidence interval; CNN, convolutional neural network.

For the binary task, the CNN significantly outperformed the dermatologists regardless of the method used to classify the dermatologists' answers (McNemar $p < 0.001$). When considering the multiclass classification task, the classifier outperformed them (McNemar ≤ 0.001) for all classes except the bcc class. If the complete test set (i.e. 300 images) was used for analysis, outperformance was achieved even for the bcc class.

4. Discussion

In this study, we used images that were entirely available from open-source databases to construct a CNN that was trained and able to outperform dermatologists significantly from all hierarchical categories in the differential diagnosis of a five-class classification problem, except for the bcc class, where performance was on par. In addition, outperformance was also achieved for the clinically relevant decision of benign vs. malignant lesions.

In machine learning, the technique of combining multiple classifiers to form an ensemble is commonly used to improve performance [16]. In this study, the answers obtained from dermatologists were combined to form an ensemble of sorts, which served as a kind of outlier removal to provide a more robust answer.

Comparing the two methods used to obtain a binary classification result from the dermatologists (differential diagnosis vs. estimation of benignancy/malignancy—see Methods for details), the latter

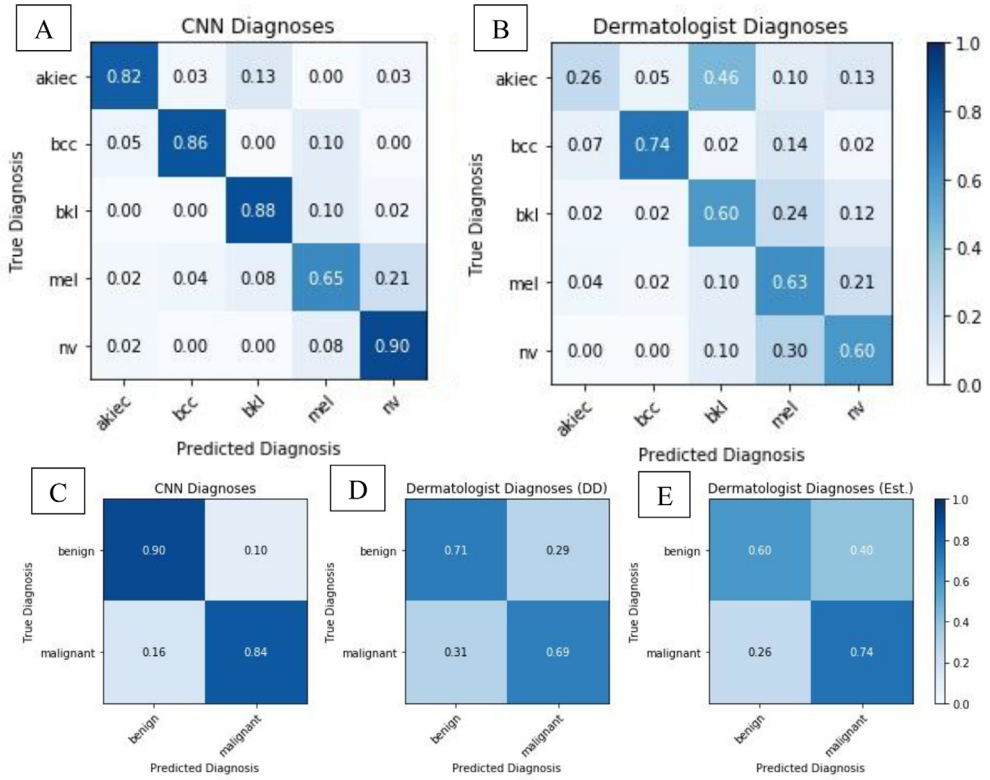

Fig. 3. **Distribution of correct and incorrect predictions by dermatologists and the classifier.** (A and B) The distribution for the five-class multiclassification task; (C–E) the distribution for the binary classification task. CNN, convolutional neural network; DD, differential diagnoses; Est., estimation of benignancy/malignancy.

showed higher sensitivity and, would be considered the technically correct method to use. Converting the answer from the differential diagnosis does not take into account that the questions' sole focus lies on the correct classification but not on the clinically important binary decision. A dermatologist swaying between nevi and melanoma may choose nevi; however, if the same dermatologist was asked for therapy management, they likely would opt for a biopsy, thus tending towards a malignant lesion. Regardless of the used approach, the classifier showed systematic outperformance.

For dermatologists, the binary task revealed a higher sensitivity at a lower specificity than the average overall performance for the multiclass classification task. This decrease in specificity for the binary task was to be expected because the general high specificity for the multiclass classification is attributed to the one-vs-all approach. Indeed, this approach introduces a high class imbalance, leading to a low frequency of the singled-out class (true positives and false negatives) and a high frequency of the remaining classes (true negatives and false positives). As specificity measures the true negative rate, the occurring imbalance naturally would favour the higher specificity values.

Regarding the differential diagnosis, the class with the lowest accuracy score for the classifier was a

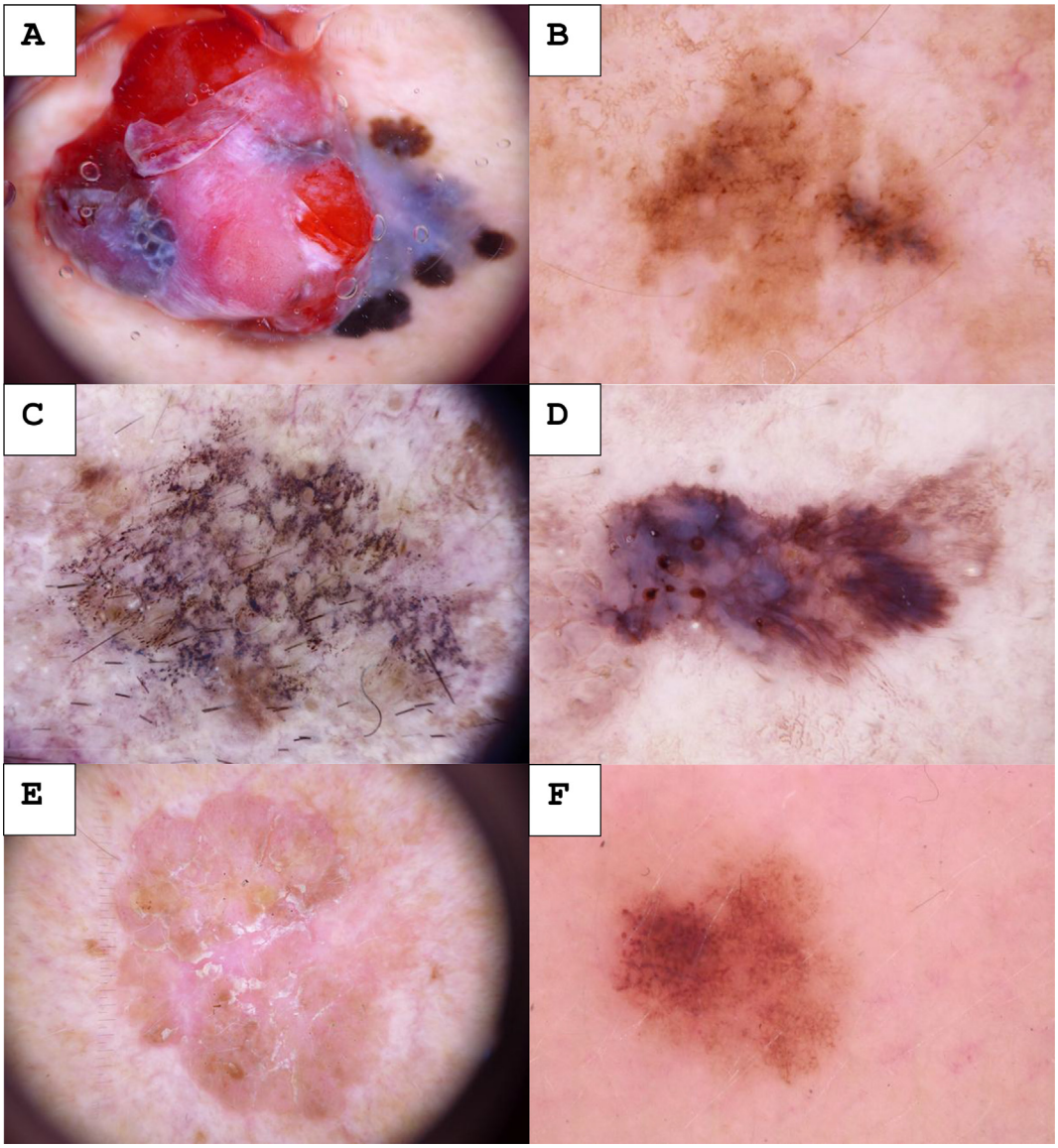

Fig. 4. **Differential diagnoses with confidence levels by the dermatologists and the classifier.** (A) mel lesion diagnosed as mel the dermatologists (10) and CNN (97.4%). (B) nv lesion diagnosed as mel by the dermatologists (8) and CNN (93.8%). (C) bkl lesion diagnosed as mel by dermatologists (5) and bkl by the CNN (96.0%). (D) bkl lesion diagnosed as mel by dermatologists (9) and bkl by the CNN (90.0%). (E) akiec lesion diagnosed as akiec by dermatologists (8) and bcc by the CNN (44.6%). (F) mel lesion diagnosed as mel by dermatologists (6) and nv by the CNN (96.6%). CNN, convolutional neural network.

melanoma (mel) with 65%, a class where dermatologists performed similarly with 63% (Fig. 3).

Both the dermatologists and the classifier frequently (21%) misclassified melanoma as nevi. On the other hand, misclassifying nevi as melanoma was infrequent for the classifier (8%), but high for the dermatologists (30%). Fig. 3 shows the frequencies of correct classifications and misclassifications for the dermatologists and CNN. These results suggest that the differential diagnosis of mel, bkl and nv is inherently difficult. Fig. 4 illustrates some of the lesions where dermatologists and classifiers agreed or disagreed, along with their confidence levels.

By exclusively using open-source images and limiting the training process to a single neural network, this study is entirely reproducible, even with limited available computing power. Furthermore, this limitation makes the approach realistic for adaptation to a real-world scenario. As such, commonly used ensemble approaches are able to achieve better results but are costly to train and run.

4.1. Limitations

4.1.1. Data selection

To ensure reproducibility, images for training and testing the classifier were obtained solely from open-source databases. Supplementing training data sets with additional images from proprietary sources is an efficient method to increase classifier performance, but significantly limits the ease of reproducibility. Because only biopsy-verified images were considered for the survey, a certain bias is introduced as these lesions are presumably difficult to diagnose.

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 and using the majority decision for further analysis, 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. Haenssle *et al.* [9] showed that integrating this additional clinical information can slightly improve dermatologists' sensitivity and specificity. 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

The classifier's performance was established on a test set disjunct from the training and validation set. However, the test images originated from the same overall dataset which was used for training (ISIC), thus raising concern about the classifier's ability to generalise on a truly external test set (i.e. a set of images where a subset was not used for training/validation (out-of-distribution sample)). A valid concern as factors intrinsic to the training dataset (e.g. type of dermatoscope, lighting or pre-processing) could be picked up during training and result in the network better classifying images sharing these intrinsic factors.

In a preliminary study, a binary-classification CNN (naevus vs melanoma), trained on ISIC images, showed good performance on an ISIC test set but performed worse on an external test set from the PH2 dermoscopic image database [Mendonça, Teresa, Pedro M. Ferreira, Jorge S. Marques, André R. S. Marçal and Jorge Rozeira. "PH2 - A dermoscopic image database for research and benchmarking." 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2013): 5437-5440]. 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.

5. Conclusion

Collectively, our findings show that the automated binary classification of dermoscopic melanoma and nevi images can be extended to a multiclass classification problem, thus better reflecting clinical differential diagnoses, while still outperforming dermatologists at a significant level ($p \leq 0.001$).

Conflict of interest statement

None declared.

Funding

This research is funded by the Federal Ministry of Health in Germany (Skin Classification Project; grant holder: Titus J. Brinker, MD).

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 they 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: 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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejca.2019.06.013.

References

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

[2] Crocetti E, Mallone S, Robsahm TE, Gavin A, Agius D, Ardanaz E, et al. Survival of patients with skin melanoma in Europe increases further: results of the EUROCARE-5 study. Eur J Canc 2015;51(15):2179–90.

[3] Vestergaard M, Macaskill P, Holt P, Menzies S. 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.

[4] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks 2017;542(7639):115.

[5] 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).

[6] 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 Canc 2019;113:47–54.

[7] 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 task 2019;111:148–54.

[8] 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 Canc 2019;111:30–7.

[9] Haenssle H, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018;29(8):1836–42.

[10] 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 Investig Dermatol July 2018;138(7):1529–38.

[11] Tschandl P, Kittler H, Argenziano G. A pretrained neural network shows similar diagnostic accuracy to medical students in categorizing dermatoscopic images after comparable training conditions. Br J Dermatol 2017;177(3):867–9.

[12] Yap J, Yolland W, Tschandl P. Multimodal skin lesion classification using deep learning. Exp Dermatol 2018;27(11):1261–7.

[13] Tschandl P, Rosendahl C, Akay BN, Argenziano G, Blum A, Braun RP, et al. Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks. JAMA Dermatol. 2019;155(1):58–65.

[14] Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. ScientifIc Data 2018;5:180161.

[15] Kassani SH, Kassani PH. A comparative study of deep learning architectures on melanoma detection. Tissue Cell 2019;58:76–83.

[16] Polikar R. Ensemble based systems in decision making. IEEE Circuits Syst Mag 2006;6(3):21–45.

[17] Sondermann W, Utikal JS, et al. Prediction of melanoma evolution in melanocytic nevi via artificial intelligence: a call for prospective data. Eur J Cancer 2019;119:30–4.

[18] Brinker TJ, et al. Deep neural networks are superior to dermatologists in melanoma image classification. Eur J Cancer 2019; 119:11–7.

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

[20] Hekler A, Utikal JS, Enk AH, Solass W, Schmitt M, Klode J, et al. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. Eur J Cancer 2019; 118:91–6.

[21] Hekler A, Utikal JS, Enk AH, Berking C, Klode J, Schadendorf D, et al. Pathologist-level classification of histopathological melanoma images with deep neural networks. Eur J Cancer 2019;115:79–83.