European Journal of Cancer 156 (2021) 202–216

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts

Sarah Haggenmüller [a,b], Roman C. Maron [a,b], Achim Hekler [a,b], Jochen S. Utikal [c,d], Catarina Barata [e], Raymond L. Barnhill [f], Helmut Beltraminelli [g], Carola Berking [h], Brigid Betz-Stablein [i], Andreas Blum [j], Stephan A. Braun [k,l], Richard Carr [m], Marc Combalia [n], Maria-Teresa Fernandez-Figueras [o], Gerardo Ferrara [p], Sylvie Fraitag [q], Lars E. French [r,ax], Frank F. Gellrich [s], Kamran Ghoreschi [t], Matthias Goebeler [u], Pascale Guitera [v,w], Holger A. Haenssle [x], Sebastian Haferkamp [y], Lucie Heinzerling [r], Markus V. Heppt [h], Franz J. Hilke [t], Sarah Hobelsberger [s], Dieter Krahl [z], Heinz Kutzner [aa], Aimilios Lallas [ab], Konstantinos Liopyris [ac], Mar Llamas-Velasco [ad], Josep Malvehy [n], Friedegund Meier [s], Cornelia S.L. Müller [ae], Alexander A. Navarini [af], Cristián Navarrete-Dechent [ag], Antonio Perasole [ah], Gabriela Poch [t], Sebastian Podlipnik [n], Luis Requena [ai], Veronica M. Rotemberg [aj], Andrea Saggini [aa], Omar P. Sangueza [ak], Carlos Santonja [al], Dirk Schadendorf [b,am], Bastian Schilling [u], Max Schlaak [t], Justin G. Schlager [r], Mildred Sergon [s], Wiebke Sondermann [am], H. Peter Soyer [i], Hans Starz [an], Wilhelm Stolz [ao], Esmeralda Vale [ap], Wolfgang Weyers [aq], Alexander Zink [ar], Eva Krieghoff-Henning [a,b], Jakob N. Kather [as], Christof von Kalle [at], Daniel B. Lipka [b,au], Stefan Fröhling [b,au], Axel Hauschild [av], Harald Kittler [aw], Titus J. Brinker [a,b,*]

[a] *Digital Biomarkers for Oncology Group, National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Heidelberg, Germany*
[b] *German Cancer Consortium (DKTK), Heidelberg, Germany*
[c] *Department of Dermatology, Heidelberg University, Mannheim, Germany*
[d] *Skin Cancer Unit, German Cancer Research Center (DKFZ), Heidelberg, Germany*

* *Corresponding author*: Digital Biomarkers for Oncology Group, National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120, Heidelberg, Germany.
E-mail address: titus.brinker@dkfz.de (T.J. Brinker).

https://doi.org/10.1016/j.ejca.2021.06.049

[e] Institute for Systems and Robotics (ISR/IST), Instituto Superior Técnico, University of Lisbon, Portugal
[f] Departments of Pathology and Translational Research, Institut Curie, Paris, France
[g] Department of Dermatology, Inselspital Bern University Hospital, University of Bern, Bern, Switzerland
[h] Department of Dermatology, University Hospital Erlangen, Erlangen, Germany
[i] The University of Queensland Diamantina Institute, The University of Queensland, Dermatology Research Centre, Brisbane, Australia
[j] Public, Private and Teaching Practice of Dermatology, Konstanz, Germany
[k] Department of Dermatology, Medical Faculty, Heinrich-Heine-University, Düsseldorf, Germany
[l] Department of Dermatology, University Hospital Münster, Germany
[m] Department of Pathology, Warwick Hospital, Warwick, UK
[n] Department of Dermatology, Hospital Clinic of Barcelona, IDIBAPS, University of Barcelona, Ciber de Enfermedades Raras ISCIII, Barcelona, Spain
[o] Hospital Universitari General de Catalunya, Grupo Quironsalud, Universitat Internacional de Catalunya, Sant Cugat Del Vallés, Barcelona, Spain
[p] Anatomic Pathology Unit, Macerata General Hospital, Macerata, Italy
[q] Department of Pathology, University Paris Descartes, Necker-Enfants Malades Hospital, Assistance Publique Hospitals of Paris, Paris, France
[r] Department of Dermatology and Allergy, University Hospital, LMU Munich, Munich, Germany
[s] Skin Cancer Center at the University Cancer Centre and National Center for Tumor Diseases Dresden, Department of Dermatology, University Hospital Carl Gustav Carus, Technische Universität Dresden, Germany
[t] Charité – Universitätsmedizin Berlin, Department of Dermatology, Venereology and Allergology, Berlin, Germany
[u] Department of Dermatology, Venereology and Allergology, University Hospital Würzburg, Würzburg, Germany
[v] Sydney Melanoma Diagnostic Centre, Royal Prince Alfred Hospital, Sydney, New South Wales, Australia
[w] Melanoma Institute Australia, And the University of Sydney, Australia
[x] Department of Dermatology, University Hospital Heidelberg, Heidelberg, Germany
[y] Department of Dermatology, University Hospital Regensburg, Regensburg, Germany
[z] Dres. Krahl Dermatopathology, Heidelberg, Germany
[aa] Dermatopathology Friedrichshafen, Friedrichshafen, Germany
[ab] First Department of Dermatology, School of Medicine, Faculty of Health Sciences, Aristotle University, Thessaloniki, Greece
[ac] Memorial Sloan Kettering Cancer Center, New York, NY, USA
[ad] Department of Dermatology, University Hospital La Princesa, Madrid, Spain
[ae] Institute for Histology, Cytology and Molecular Diagnostic, Trier, Germany
[af] Department of Dermatology, University Hospital of Basel, Switzerland
[ag] Department of Dermatology, Escuela de Medicina, Pontificia Universidad Católica de Chile, Santiago, Chile
[ah] Anatomic and Cytopathology, Az. ULSS 8 Berica, Regione Veneto, Ospedale San Bortolo, Vicenza, Italy
[ai] Dermatology Department, Hospital Fundación Jiménez Díaz, Madrid, Spain
[aj] Dermatology Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[ak] Dermatopathology, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA
[al] Pathology Department, Fundación Jiménez Díaz, Madrid, Spain
[am] Department of Dermatology, Venereology and Allergology, University Hospital Essen, University Duisburg-Essen, Essen, Germany
[an] Dermpath München, Munich, Germany
[ao] Department of Dermatology, Allergology and Environmental Medicine II, Hospital Thalkirchner Street, Munich, Germany
[ap] Department of Dermatology and Dermatopathology, Hospital da Luz, Lisbon, Portugal
[aq] Center for Dermatopathology, Freiburg, Germany
[ar] Department of Dermatology and Allergy, Faculty of Medicine, Technical University of Munich, 80802, Munich, Germany
[as] Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany
[at] Department of Clinical-Translational Sciences, Charité University Medicine and Berlin Institute of Health (BIH), Berlin, Germany
[au] Department of Translational Medical Oncology, National Center for Tumor Diseases (NCT) Heidelberg and German Cancer Research Center (DKFZ), Heidelberg, Germany
[av] Department of Dermatology, University Hospital of Schleswig-Holstein (UKSH), Campus Kiel, Kiel, Germany
[aw] ViDIR Group, Department of Dermatology, Medical University of Vienna, Vienna, Austria
[ax] Dr. Philip Frost, Department of Dermatology and Cutaneous Surgery, University of Miami Miller School of Medicine, Miami, FL, USA

Received 16 May 2021; received in revised form 18 June 2021; accepted 28 June 2021
Available online 8 September 2021

KEYWORDS
Skin cancer classification;

Abstract ***Background:*** Multiple studies have compared the performance of artificial intelligence (AI)–based models for automated skin cancer classification to human experts, thus setting the cornerstone for a successful translation of AI-based tools into clinicopathological practice.

Digital biomarkers;
Convolutional neural network(s);
Artificial intelligence;
Machine learning;
Deep learning;
Dermatology;
Malignant melanoma

Objective: The objective of the study was to systematically analyse the current state of research on reader studies involving melanoma and to assess their potential clinical relevance by evaluating three main aspects: test set characteristics (holdout/out-of-distribution data set, composition), test setting (experimental/clinical, inclusion of metadata) and representativeness of participating clinicians.

Methods: PubMed, Medline and ScienceDirect were screened for peer-reviewed studies published between 2017 and 2021 and dealing with AI-based skin cancer classification involving melanoma. The search terms skin cancer classification, deep learning, convolutional neural network (CNN), melanoma (detection), digital biomarkers, histopathology and whole slide imaging were combined. Based on the search results, only studies that considered direct comparison of AI results with clinicians and had a diagnostic classification as their main objective were included.

Results: A total of 19 reader studies fulfilled the inclusion criteria. Of these, 11 CNN-based approaches addressed the classification of dermoscopic images; 6 concentrated on the classification of clinical images, whereas 2 dermatopathological studies utilised digitised histopathological whole slide images.

Conclusions: All 19 included studies demonstrated superior or at least equivalent performance of CNN-based classifiers compared with clinicians. However, almost all studies were conducted in highly artificial settings based exclusively on single images of the suspicious lesions. Moreover, test sets mainly consisted of holdout images and did not represent the full range of patient populations and melanoma subtypes encountered in clinical practice.

1. Introduction

Although malignant melanoma (MM) accounts for only 4% of skin cancers, it is responsible for about 75% of all skin cancer–associated deaths. Early detection and diagnosis are critical for survival chances of affected patients [1].

Early diagnosis, however, may be difficult, as MM and atypical melanocytic nevi frequently present with morphological overlap. Although dermoscopy improves diagnostic accuracy compared with naked eye examination [2], even specialists rarely achieve sensitivity levels above 80% [3]. Beyond that, a significant variance depending on training and professional experience can be observed [4].

In case of a suspected MM, skin biopsy is routinely performed to enable histopathological examination. Although histopathological analysis is currently considered the gold standard for skin cancer diagnosis, it is time-consuming, labour-intensive and can also be inconclusive in borderline cases. Previous studies revealed a discordance between individual pathologists for MM classification of up to 25% [5,6].

Against this backdrop, accurate distinction between benign and malignant skin lesions as well as the exact classification of skin cancer types through digital biomarkers (DBs) is of great interest to reduce the number of missed MM as well as unnecessary excisions. DBs are data-driven indicators that provide information about the characteristics of a lesion and may predict health-related outcomes.

Convolutional neural networks (CNNs) are deep neural networks with an architecture specifically designed for image analysis that are commonly trained via supervised learning. This means that CNNs use labelled data, for example dermoscopic images with their corresponding diagnosis/ground truth, to learn a relationship between the input data and the labels. Based on that, CNNs are able to apply learned operations to unknown images and classify them based on the extracted features. Because diagnosis in clinical dermatology and dermatopathology is largely based on the recognition of visual patterns, the use of CNNs could help to develop additional and/or improved clinically meaningful DBs [7].

This systematic review presents state of the art artificial intelligence (AI)-based automated skin cancer classification involving MM and comparing AI results with human experts. The included studies have been reviewed with particular reference to the clinical relevance of the reported results, thereby reflecting the actual impact and the forthcoming challenges expected with the implementation of AI-based classifiers into clinicopathological routine.

2. Material and methods

2.1. Search strategy

In 2017, Esteva et al. [8] first reported on a deep learning CNN-based image classifier that outperformed 21 board-certified dermatologists in the classification of clinical and dermoscopic images. We therefore screened PubMed, Medline and ScienceDirect for peer-reviewed studies published in English between 2017 and 2021

(search terms last accessed on 02/17/2021). The following search terms were combined: skin cancer classification, deep learning, convolutional neural network(s), melanoma (detection), digital biomarkers, histopathology and whole slide imaging (for a detailed overview of the comprehensive search strategy, see Supplementary Table 1).

2.2. Study selection

Search results were screened manually. Only publications that fulfilled the inclusion criteria listed in the following were selected (for a detailed overview of the systematic search procedure in accordance with PRISMA, see Supplementary Material 2 and 3). First, only studies that contained direct comparisons of AI classifiers with human experts were included, as these approaches better demonstrate the potential value of AI-based classifiers in clinicopathological practice. Non-comparative approaches (e.g. Refs. [9–12]) were excluded. Furthermore, only studies involving the diagnosis of MM were evaluated. As MM is the skin cancer subtype that is associated with the most skin cancer–related deaths, we discarded studies that completely excluded the diagnosis of MM (e.g. Ref. [13]). Finally, only studies that had a diagnostic classification as their main task were included. Studies concentrating on prognostic factors such as therapy response or long-term survival were explicitly not addressed (e.g. Refs. [14,15]). Data were extracted from peer-reviewed articles exclusively. Data quality was assessed independently by two reviewers.

2.3. Study analysis

The included studies were reviewed with particular reference to the potential clinical relevance of the reported results by assessing three main aspects: test set characteristics (holdout/out-of-distribution data set, composition), test setting (experimental/clinical, inclusion of metadata) and representativeness of the included clinicians.

Holdout data refer to data obtained from the same overall data set as the data used for training and validation of the algorithm. Thus, the test set follows the same probability distribution as the training set. Conversely, out-of-distribution (OOD) data do not follow the training distribution and are often referred to as an external test set (e.g. from external clinics).

2.4. Study performance metrics

In this systematic review, we focus on the performance metrics accuracy, sensitivity and specificity.

Accuracy is a meaningful metric if different classes within the test set are more or less evenly distributed and if the overall performance is of interest and not the performance for a specific class. Accuracy indicates the percentage of correctly classified skin lesions, that is the percent ratio between the total number of correctly classified lesions and the overall number of examined lesions.

Sensitivity and specificity are not influenced by class imbalances and better reflect the performance for a specific class. However, both metrics require a dichotomous classification, where only one positive and one negative class are considered (e.g. melanoma vs. melanocytic nevus, benign vs. malignant or one class vs. the rest in a multiclass classification setting). Sensitivity is calculated based on the actual positive cases; it is the percent ratio between cases that are correctly assigned as positive in comparison with the overall number of positive cases contained in the data set. By contrast, specificity is determined on the basis of the actual negative cases; it is the percent ratio between cases correctly allocated as negative and all negative cases of the data set.

3. Results

A total of 19 comparative studies (since Esteva et al.'s [8] seminal article) were published that fulfilled the inclusion criteria. Most of the studies focused on dermoscopic images (n = 11) [4,7,16–24], followed by clinical image (n = 6) [25–30] and histopathological whole slide image (WSI) studies (n = 2) [31,32] (see Fig. 1). In the following, the term histopathological WSI refers to digitised haematoxylin-eosin (H&E)–stained tissue sections processed with specialised slide scanners.

3.1. Automated skin cancer classification of dermoscopic images

Eleven studies based on the classification of dermoscopic images fulfilled the inclusion criteria (see Table 1). Out of these, eight publications were based on a binary classification system. Supplementary Table 4 contrasts the training and testing procedures of these approaches.

Brinker et al. [16] fine-tuned an algorithm for the binary discrimination between MM and melanocytic nevus. To compare the classifier performance with results obtained by human experts, 157 dermatologists indicated their corresponding management decision (biopsy or further treatment vs. reassurance of the patient) for 100 test images. This is how the authors compiled the most comprehensive binary dermoscopic reader study to date. Overall, the CNN outperformed 136 of 157 dermatologists across different levels of experience in terms of average specificity and sensitivity.

Subsequently, Brinker et al. [17] carried out a follow-up study comparing the diagnostic performance of the CNN with 144 dermatologists. In that study, only images with a histology-proven groundtruth (i.e. images of lesions suspicious for MM) were taken into consideration, thus presumably increasing the overall difficulty

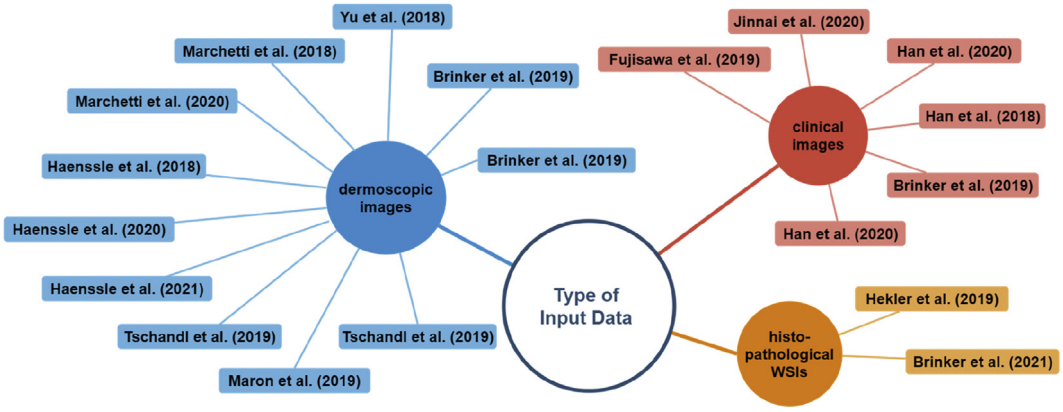

Fig. 1. **Categorisation of the included studies based on the type of input data.** Based on the input data, the included studies are grouped into three categories: those based on dermoscopic images [4,7,16–24], those based on clinical images [25–30] and those based on histopathological WSIs [31,32]. WSI, whole slide image.

of the test set. Nonetheless, for the first time, CNN-based MM classification was significantly superior to junior and board-certified dermatologists (82.3% vs. 68.9%/63.2% sensitivity and 77.9% vs. 58.0%/65.2% specificity, $p < 0.001$).

Yu et al. [19] developed an algorithm focussing on a binary classification (MM vs. melanocytic nevus) of lesions of the acral skin. The authors compared their CNN with the results achieved by two experienced dermatologists as well as with two non-trained general physicians. The CNN achieved mean sensitivity, specificity and accuracy levels that were comparable with those of the experienced dermatologists (92.6%, 71.8% and 81.9% vs. 96.6%, 67.0% and 81.4%), thus illustrating the potential of CNN-based automated melanoma detection for special subtypes such as acral MM on the hands and feet.

Marchetti et al. [20] published the first dermoscopic comparative study that used an ensemble approach to combine the classifier predictions of 25 participating teams of the International Symposium on Biomedical Imaging (ISBI) 2016 challenge. By investigating five different fusion approaches, the authors demonstrated that the top fusion approach was able to outperform eight experienced dermatologists. This was significant for both the binary classification of malignancy (at dermatologists' sensitivity of 82%: 76% vs. 59% specificity, $p = 0.02$) and for the consideration of management decisions (at dermatologists' sensitivity of 89%: 64% vs. 47% specificity, $p = 0.02$). In 2020, Marchetti et al. [21] proposed a similar reader study in which the best performing algorithm of the ISBI 2017 challenge significantly outperformed eight dermatologists and nine dermatology residents ($p < 0.001$).

Haenssle et al. [4] were the first to give additional clinical information to the clinicians within the reader study. The authors proposed a binary classification approach for automated MM classification and compared the diagnostic accuracy of the CNN with the results obtained by 58 dermatologists. The study was divided into two levels. In level I, participants reviewed the test set online and indicated their corresponding diagnosis (MM vs. melanocyctic nevus) as well as management decision (exicision or short term follow-up vs. no action) based solely on one dermoscopic image. In level II, the same dermatologists diagnosed the identical test set, but with additional clinical information and close-up images. Although additional information improved the diagnostic accuracy of the dermatologists, the CNN still significantly outperformed the average of the participants (at dermatologists' sensitivity of 88.9%: 82.5% vs. 75.7% specificity, $p < 0.01$).

In 2020, Haenssle et al. [18] replicated their previous reader study by comparing an updated version of their CNN with the results achieved by 96 dermatologists. In that study, they included a broader spectrum of disease classes ($n = 10$) which had to be classified into (pre)malignant and benign lesions. When fixing the specificity of the CNN at the dermatologists' mean specificity for their management decision in level II (80.4%), the sensitivity of the CNN was almost equal to that of human raters (95.0% vs. 94.1%).

Moreover, Haenssle et al. [24] proposed a reader study that focused exclusively on suspicious lesions of the face and scalp. In level II of that study, the CNN significantly outperformed 64 human experts in terms of management decision (at dermatologists' specificity of 69.4%: 96.2% vs. 84.2% sensitivity, $p < 0.001$). This difference resulted in an average of 6.2 more malignant lesions missed by dermatologists compared with the CNN (CNN: 2/52, dermatologists' mean: 8.2/52), thus outlining that the potential of CNN-based automated skin cancer classification can also be extended to special anatomic sites such as the face and scalp.

Three dermoscopic approaches expanded on the binary perspective (e.g. MM vs. melanocytic nevus, benign vs. malignant) presented by Brinker et al. [16,17], Yu et al. [19], Marchetti et al. [20,21] and Haenssle et al.

Table 1
Overview reader studies based on dermoscopic images.

Reader study	Comparison with	Scope of the reader study test set	Metadata	Origin of the reader study test set	Setting	Classification task	Results
Brinker et al. [16]	157 dermatologists - 151 university hospital–based from 12 university hospitals in Germany: - 88 junior clinicians - 15 attendings - 45 senior clinicians - 3 chief clinicians - 6 dermatologists in private practice	100 images, randomly selected out of 20735 images available at ISIC	n	ISIC image archive (holdout)	e	Binary: melanoma/ melanocytic nevi	CNN outperformed 136 out of 157 dermatologists
Brinker et al. [17]	144 dermatologists from 9 university hospitals in Germany - 92 junior clinicians - 52 board-certified dermatologists	6 subsets consisting of 134 images each, 804 images in total	n	ISIC image archive (holdout)	e	Binary: melanoma/ melanocytic nevi	Significant superiority of the CNN
Yu et al. [19]	4 participants: - 2 general physicians - 2 experienced dermatologists	2 subsets consisting of 362 images each, 724 images total	n	Severance Hospital in the Yonsei University Health System, Seoul, Korea (holdout) Dongsan Hospital in the Keimyung University Health System, Daegu, Korea (holdout)	e	Binary: acral melanoma/ melanocytic nevi	Comparable performance
Marchetti et al. [20]	8 experienced dermatologists from 4 different countries	Randomly selected 100 images out of 379 images	n	ISBI 2016 challenge, ISIC image archive (holdout)	e	Binary: malignant/ benign; biopsy/ observation or reassurance	Significant superiority of the CNN ensemble
Marchetti et al. [21]	17 dermatologists - 8 dermatologists from 4 countries - 9 dermatology residents from the United States	Randomly selected 150 images out of 600 images	n	ISIC image archive (holdout)	e	Binary: melanoma/ non-melanoma; biopsy/ observation	Significant superiority of the CNN
Haenssle et al. [4]	58 dermatologists from 17 countries - 17 beginners - 11 skilled - 30 experts	Selected 100 images with increased difficulty out of 300 images (I) dermoscopy only (II) in addition: clinical information and close-up images	y	Department of Dermatology, University of Heidelberg, Germany (OOD)	e	Binary: melanoma/ melanocytic nevi; excision or short-term follow-up/ no action	Significant superiority of the CNN
Haenssle et al. [18]	96 dermatologists - 17 beginners - 29 skilled - 40 experts	100 images with increased difficulty (I) dermoscopy only (II) in addition: clinical information and close-up images	y	Department of Dermatology, University of Heidelberg, Germany (OOD)	e	Binary: (pre) malignant/ benign; excision or treatment/ follow-up or no action	Comparable performance

(*continued on next page*)

Table 1 (*continued*)

Reader study	Comparison with	Scope of the reader study test set	Metadata	Origin of the reader study test set	Setting	Classification task	Results
	- 10 no information provided						
Haenssle et al. [24]	64 dermatologists - 9 beginners - 20 skilled - 30 experts - 5 no information provided	100 images of face and scalp lesions in total (I) dermoscopy only (II) in addition: clinical information and close-up images	y	Department of Dermatology, University of Heidelberg, Germany (OOD), Department of Dermatology Hospital Thalkirchner Street, Munich, Germany (OOD), Department of Dermatology, Medical University Graz, Austria (OOD), First Department of Dermatology, Aristotle University, Thessaloniki, Greece (OOD), Dermatology Office Based Clinic of Dermatology, Konstanz, Germany (OOD)	e	Binary: malignant/ benign; excision or treatment/ follow-up or no action	Significant superiority of the CNN
Tschandl et al. [22]	511 participants from 63 countries - 283 board-certified dermatologists - 118 dermatology residents - 83 general practitioners - 27 no information provided	Randomly selected 30 images per participant out of 1511 images	n	HAM10000 data set, ISBI 2018 challenge, ISIC image archive (holdout), additional images from Turkey, New Zealand, Sweden and Argentina (OOD)	e	Multiclass (7)	Significant superiority of the CNN
Maron et al. [23]	112 dermatologists - 108 university hospital–based from 13 university hospitals in Germany: - 67 junior clinicians - 12 attendings - 28 senior clinicians - 1 chief physician - 4 dermatologists in private practice	6 subsets consisting of 50 images, 300 images in total	n	HAM10000 data set (holdout)	e	Binary: malignant/ benign; multiclass (5)	Significant superiority of the CNN
Tschandl et al. [7]	95 participants from 29 countries: - 62 board-certified dermatologists - 12 dermatology residents - 17 general practitioners - 4 others	Randomly selected 50 images per participant out of 2072 images	n	Primary skin cancer clinic in Queensland, Australia (holdout), Department of Dermatology of the Medical University of Vienna, Austria (OOD), additional images from dermatologists from Sweden, Italy, Austria, France, Turkey, Germany (OOD)	e	Multiclass (8)	Comparable performance

Metadata (additional information for readers beyond image input, e.g. age, gender, localisation of the suspicious lesion).
y, yes.
n, no.
c, clinical setting.
e, experimental setting.
CNN, convolutional neural network.
OOD, out of distribution.
ISBI, International Symposium on Biomedical Imaging.

[4,18,24], by carrying out multiclass classification tasks which covered more fine-grained diagnoses (see Table 1) [7,22,23]. Supplementary Table 4 outlines similarities and differences of these multiclass approaches with regard to individual training and testing procedures.

In 2019, Tschandl et al. [22] compared the results obtained by 139 algorithms in the ISBI 2018 challenge with those obtained by 511 human readers, including 283 board-certified dermatologists, 118 dermatology residents and 83 general practitioners. This comparative approach constitutes the most comprehensive multiclass reader study to date. Regarding the discrimination between MM and six other skin diseases (for a more detailed specification of the classes, see Supplementary Table 7), the algorithms achieved an average of 19.9 correct diagnoses out of 30 with participants achieving an average of 17.9 correct diagnoses ($p < 0.0001$).

Maron et al. [23] proposed a similar reader study to Tschandl et al. [22] by developing a classifier to differentiate between MM and four other skin disease classes (see Supplementary Table 7). In that study, the CNN significantly outperformed 112 dermatologists from different levels of experience in the correct classification of images into five diagnostic categories (at dermatologists' sensitivity of 56.5%: 98.8% vs. 89.2% specificity, $p < 0.001$).

Tschandl et al. [7] were the first to propose a reader study integrating two different image types. They combined a CNN trained with dermoscopic images and a CNN trained on clinical close-up images into a combined CNN (cCNN). Focussing on amelanotic skin lesions, the authors showed that the cCNN was able to differentiate between MM and seven other skin diseases (see Supplementary Table 7) with comparable performance with that of 95 human raters (at participants' specificity of 51.3%: 80.5% vs. 77.6% sensitivity).

3.2. Automated skin cancer classification of clinical images

A total of six CNN-based classification approaches using clinical images fulfilled the inclusion criteria of this systematic review (see Table 2). Supplementary Table 5 outlines the training and testing procedure of each individual approach.

Fujisawa et al. [25] developed an algorithm for the binary discrimination between malignant and benign lesions, while simultaneously enabling a more fine-grained multiclass classification into MM and 13 other skin diseases (see Supplementary Table 7). The authors compared the classifier results with those of 13 board-certified dermatologists as well as nine dermatology trainees. The CNN achieved accuracy levels that significantly outperformed both groups with regard to binary (92.4% vs. 85.3%/74.4%, $p < 0.0001$) and multiclass classification (74.5% vs. 59.7%/41.7%, $p < 0.0001$).

Jinnai et al. [26] proposed a similar reader study than Fujisawa et al. [25]. The authors developed an algorithm for the distinction between malignant and benign skin lesions as well as for the precise classification into MM and five other disease classes (see Supplementary Table 7). In comparison with 10 dermatologists and 10 dermatology trainees, the used CNN significantly outperformed the participants in terms of accuracy for the binary (91.5% vs. 86.6%/85.3%, $p < 0.01$) and the multiclass approach (86.2% vs. 79.5%/75.1%, $p < 0.001$).

Han et al. [27] also addressed the binary discrimination between malignant and benign lesions and multiclass classification. The developed multiclass model enabled a differentiation into MM and 133 other skin diseases, therefore incorporating the broadest spectrum of diagnoses to date (see Supplementary Table 7). For the binary discrimination, the classifier performance was comparable with the results obtained by 47 medical professionals. Regarding the precise classification into the 134 disease categories, the CNN performed slightly worse in terms of accuracy (44.8% vs. 49.9%) than two board-certified dermatologists and two dermatology residents.

Unlike the previous approaches, Han et al. [28] developed a model which focused exclusively on the multiclass discrimination of MM and 11 other skin diseases (see Supplementary Table 7). Not only did their model output the diagnosis with the highest probability for a given image but also give a differential diagnosis once a defined threshold for any of the 12 considered disease classes was overcome. Based upon that, an experimental but more realistic comparison between the classifier performance and the diagnostic results of 16 dermatologist board members was possible. The algorithm achieved an accuracy of 57.3% and 55.7% on a holdout and out-of-distribution test set, respectively, which was comparable with the accuracy obtained by the dermatologist board members.

Brinker et al. [30] were the first to investigate whether an algorithm benefits from training on high-resolution dermoscopic images even for clinical classification tasks. The authors trained an algorithm with dermoscopic images only and compared the classifier performance with the results of 145 dermatologists in a binary classification task on clinical images (MM vs. atypical melanocytic nevi). At dermatologists' sensitivity of 68.2%, the CNN achieved a slightly higher, but comparable, specificity (68.2% vs. 64.4%). For the first time, dermatologist-level image classification was achieved on a clinical image classification task without a specific training on clinical images.

Han et al. [29] established a direct comparison between the performance of a CNN-based classifier and the results obtained by dermatologists for the binary classification into malignant and benign lesions, as well as the automated discrimination between MM and 31 other skin diseases (see Supplementary Table 7). The

Table 2
Overview reader studies based on clinical images.

Reader study	Comparison with	Scope of the reader study test set	Metadata	Origin of the reader study test set	Setting	Binary/multiclass	Results
Fujisawa et al. [25]	22 dermatologists - 9 dermatologic trainees - 13 board-certified	Randomly selected 140 images per participant out of 1142 images	n	University of Tsubuka Hospital, Japan (holdout)	e	Binary: malignant/ benign, multiclass (14)	Significant superiority of the CNN
Jinnai et al. [26]	20 dermatologists - 10 dermatologic trainees - 10 board-certified	Randomly selected 10 test samples of 200 images out of 1114 images	n	Dermatologic Oncology in the National Cancer Center, Tokyo (holdout)	e	Binary: malignant/ benign, multiclass (6)	Significant superiority of the CNN
Han et al. [27]	Binary: 47 dermatologists - 21 board-certified - 26 dermatology residents Multiclass: 4 dermatologists - 2 board-certified - 2 dermatology residents	Randomly selected 240 images out of 2201 images	n	SNU data set (OOD)	e	Binary: malignant/ benign, multiclass (134)	Binary: on par performance Multiclass: comparable, but slightly worse performance of the CNN
Han et al. [28]	16 dermatologist board members - 6 clinicians (>10 years of experience) - 10 professors	Randomly selected 480 images 1) 260 images of 12 disorders out of 1276 images 2) 220 images of 10 disorders out of 1300 images	n	1) Asan test set (holdout) 2) Edinburgh data set (OOD)	e	Multiclass (12)	On par performance
Brinker et al. [30]	145 dermatologists - 142 university hospital–based: - 88 junior clinicians - 16 attendings - 35 senior clinicians - 3 chief clinicians - 3 dermatologists in private practice	100 images	n	MClass benchmark obtained from the MED-NODE database (OOD)	e	Binary: melanoma/ melanocytic nevi	On par performance

Table 2 (*continued*)

Reader study	Comparison with	Scope of the reader study test set	Metadata	Origin of the reader study test set	Setting	Binary/multiclass	Results
Han et al. [29]	1) 65 attending clinicians 2) 44 board-certified dermatologists	1) 40331 images from 10426 cases of 43 disorders[a] 2) Randomly selected 44 image batches of 30 patients out of 5065 images from 1320 cases	n	Department of Dermatology, Severance Hospital in Seoul, Korea (OOD)	1) c 2) e	Binary: malignant/ benign, multiclass (32)	1) Significant superiority of the attending clinicians 2) Binary: on par performance Multiclass: significant superiority of the CNN

Metadata (additional information for readers beyond image input, e.g. age, gender, localisation of the suspicious lesion).
y, yes.
n, no.
c, clinical setting.
e, experimental setting.
CNN, convolutional neural network.
OOD, out of distribution.
[a] For multiclass classification, 39721 images from 10315 cases of 32 disorders remained, after excluding cases belonging to too small and untrained classes.

Table 3
Overview reader studies based on histopathological WSIs.

Reader study	Comparison with	Scope of the reader study test set	Metadata	Origin of the reader study test set	Setting	Classification task	Results
Hekler et al. [31]	11 pathologists	100 cropped digitised H&E slides	n	Dermatohistopathologic Institute Dr. D. Krahl, Heidelberg, Germany (holdout)	e	Binary: melanoma/ melanocytic nevi	Significant superiority of the CNN
Brinker et al. [32]	18 pathologists from 8 different countries, each with at least 5 years of experience	100 digitised H&E slides	n	Routine files of 2 expert board-certified dermatopathologists from Friedrichshafen, Germany (holdout)	e	Binary: melanoma/ melanocytic nevi	On par performance

Metadata (additional information for readers beyond image input, e.g. age, gender, localisation of the suspicious lesion).
y, yes.
n, no.
c, clinical setting.
e, experimental setting.
CNN, convolutional neural network.
OOD, out of distribution.
WSI, whole slide image.
H&E, haematoxylin-eosin.

authors were the first to provide a clinical image reader study in a clinical setting by incorporating 65 attending clinicians that recorded their diagnoses during thorough examinations in clinical practice. The CNN was significantly outperformed by the attending participants regarding the binary (62.7% vs. 70.2% sensitivity and 90.0% vs. 95.6% specificity, $p < 0.0001$) and the multiclass classification task (42.6% vs. 65.4% accuracy). However, when conducting the reader study with 44 board-certified dermatologists that reviewed multiple images of the affected lesions in an experimental setting, the CNN achieved comparable results for the binary discrimination of images (66.9% vs. 65.8% sensitivity and 87.4% vs. 85.7% specificity) and significantly superior accuracy for the multiclass classification into 32 skin disorders (49.5% vs. 37.7%).

3.3. Automated skin cancer classification of histopathological WSI

WSI scanners have enabled the efficient digitisation of H&E-stained tissue sections, thereby setting the cornerstone for the development of AI-based digital skin cancer biomarkers for histopathology (e.g. Refs. [10,11]). Besides the proposed clinical and dermoscopic studies, two comparative approaches using histopathological WSIs met the inclusion criteria of this systematic review (see Table 3) [31,32]. Supplementary Table 6 summarises the training and testing procedures of both approaches.

Hekler et al. [31] were the first to compare the performance of a CNN developed for the classification of cropped image sections of WSIs with the results obtained by 11 pathologists. The CNN significantly outperformed the participants in terms of mean sensitivity, specificity and accuracy (76.0%, 60.0% and 68.0% vs. 51.8%, 66.5% and 59.2%, $p = 0.016$).

Brinker et al. [32] compared the ability of a CNN ensemble to differentiate MM from benign melanocytic nevi with that of 18 international expert pathologists using the entire WSIs instead of cropped image sections. Even when the tumour region was not annotated before training, the CNN ensemble achieved comparable results with that of the participants in terms of mean sensitivity, specificity and accuracy (88.0%, 88.0% and 88.0% vs. 88.9%, 91.8% and 90.3%).

4. Discussion

4.1. Principal findings

All 19 included reader studies demonstrated an at least equivalent classification performance of CNNs and clinicians. This was true not only for binary classification tasks but also for multiclass classification tasks, which reflect better the clinical relevant differential diagnosis. The included studies covered three main image types (dermoscopic, clinical and histopathological WSIs). Because the study designs were very heterogeneous and a direct comparison among them was mostly not possible, our discussion is mainly focused on their potential clinical relevance.

4.1.1. Test set characteristics

While a large proportion of clinical reader studies based their comparison on OOD test sets [27–30] (see Table 2), the vast majority of dermoscopic and histopathological approaches (8 out of 13, see Tables 1 and 3) grounded their reader study on holdout images exclusively. While this may partially be due to the limited amount of publicly available data sets for histopathological WSIs, there are already several public dermoscopic data sets available. This makes the omission of external testing for dermoscopic studies questionable. The authors of a large international challenge which included many AI models competing against hundreds of clinicians [22] showed that the difference between human experts and the top three challenge algorithms was significantly lower for test images that came from a different source than the training images. This highlights that generalisability to OOD data is not guaranteed. To provide comparisons that account for the variance between image records from different sources, as in clinical reality, reader studies that allow classifiers to be evaluated on OOD images (e.g. from external clinics) should be considered the gold standard for future research [33,34].

To achieve more general statements about the performance of automated skin cancer classification in comparison with clinicians, it is important to use test data that are as representative of the world population as possible and at least include the relevant skin diseases that are commonly encountered in clinical practice. Navarrete-Dechent et al. [33], for example, showed that the sensitivity of a skin cancer algorithm was considerably lower when applied to a different patient population, thus limiting its generalisability. However, few studies have explicitly expanded their test data with skin lesions from different ethnicities to ensure diversity of skin types [7,22]. Regarding the 6 clinical reader studies, 3 of these studies recruited images from an Asian skin–type population exclusively. On the other hand, the images of the ISIC database (used as a test set for 6 out of 11 dermoscopic reader studies) mainly encompassed light-skinned skin lesions from patients in Europe, Australia and the United States, whereas Asian and dark-skinned populations were underrepresented. Yu et al. [19], Haenssle et al. [24] and Tschandl et al. [7] proved the potential of CNN-based classification for special anatomic sites such as the face and scalp [24] or

acral MM on the hands and feet [19], as well as rare subtypes such as amelanotic MM [7]. However, other special anatomic sites (e.g. genital area), rare subtypes (e.g. mucosal or desmoplastic MM) and the simultaneous incorporation of all relevant factors for a representative test set composition (i.e. diversity of skin types, skin diseases and anatomical sites) remain poorly investigated.

4.1.2. Test setting

One possible limitation of almost all proposed publications (18 out of 19, see Tables 1–3) is the experimental test setting of the conducted reader studies. The decision-making basis of 14 of the 19 (see Tables 1–3) included reader studies was limited to a single image of the suspicious skin lesion. Haenssle et al. [4,18,24] showed that dermatologists performed somewhat better, when provided with additional close-up images and patient information such as age, sex or lesion location. The authors highlighted the value of clinical data in addition to visual data. Clinicians assess patients with all their lesions, aiming to identify the 'ugly duckling' throughout physical examination. Even teledermatologists are trained to leverage information from multiple sources. The CNNs considered in this systematic review, however, have been trained to assign a label for images only, disregarding the clinical context. Therefore, comparative studies that are solely based on single images fall short of the clinical routine. Interestingly enough, in these [4,18,24] and other [7,29] studies in which multiple images were provided to human experts, the participants only attained at most equivalent results in comparison with CNN-based classification. Nevertheless, to enable a fair comparison, future reader studies should not only provide clinicians but also provide CNNs with additional close-up images and patient information (e.g. Refs. [35,36]).

One reason why participants with additional patient information did not outperform CNNs might be that the setting was still artificial. In most of the analysed studies (18 out of 19, see Tables 1–3), including those with additional clinical or image data, the recording of the participants' diagnoses took place through web-based rating applications or online questionnaires, thereby substantially differing from the decision-making process occurring in daily clinical practice. Only one study had its participants record their diagnosis during clinical examination of the patient [29]. Under these conditions, the CNN was significantly outperformed by the participating dermatologists, regardless of the classification task. This finding highlights that no conclusions about the added value of automated MM detection should be drawn solely based on experimental comparisons.

4.1.3. Representativeness of the included clinicians

A considerable number of publications already included clinicians with different levels of experience, ranging from dermatology trainees to board-certified dermatologists. However, from a statistical point of view, the number of incorporated clinicians from certain subgroups (e.g. level of experience) did not reach the necessary threshold of $n = 30$ to get reasonable mean averages (in accordance with the central limit theorem), hence raising concerns about adequate statistical representativeness. Moreover, only few studies included dermatologists in private practices (e.g. Refs. [16,23,30]). Given that dermatologists in private practices carry out skin cancer screenings for most of the population, we believe that they were not represented adequately in the assessed studies of this systematic review. Comparative studies with a larger number and variance of human experts would help in making the results more representative of the actual physician population that is encountered in clinical practice.

4.2. Limitations and outlook

This systematic review is limited to approaches that considered direct comparison between CNN-based skin cancer classification and clinicians. However, AI-based systems are susceptible to the influence of confounding factors (e.g. skin markings, skin hairs) [37,38] and small changes in image input (e.g. scaling or rotation) [39], therefore requiring a 'plausibility check' by human experts to avoid false diagnoses. Thus, one of the main practical uses of AI with dermoscopic, clinical and histopathological WSIs may be the use as an assistance system, calling for a complementary instead of a comparative perspective (e.g. Refs. [40,41]).

We explicitly addressed studies that had a diagnostic classification task as their main objective. This is, however, only one of many aspects that are important for improved personalised patient care. To further enhance precision medicine and therapy selection in addition to mere cancer identification using AI-based assistance systems, we should not only consider studies comparing computer-aided diagnosis but also expand on studies focussing on prognostic end-points such as therapy response or long-term survival (e.g. Refs. [14,15]) to leverage the full potential of novel DBs.

Finally, because positive studies outlining statistically significant results are more likely to be published than negative studies that did not reject the null hypotheses, we cannot exclude the risk of publication bias.

5. Conclusions

All 19 included reader studies—regardless of the classification task and the type of input data—showed

superior or at least equivalent performance of CNN-based classifiers in comparison with clinicians. This indicates the potential of CNN-based approaches to evolve into novel DBs. However, almost all studies were conducted in an experimental setting based exclusively on single images of the suspicious lesions. To increase clinical relevance of the results, future comparison studies should be conducted under less artificial conditions, with use of external OOD test sets reflecting the full range of ethnicities and melanoma subtypes occurring in clinical practice. Furthermore, there is a need for truly prospective studies comparing the clinicians' diagnoses after real-life face-to-face patient examinations with the results of AI-based classification models. Ideally, such studies would also measure the impact of the CNN classifications on the final management decisions of clinicians.

Author contribution section

Sarah Haggenmüller: Conceptualization, Methodology, Investigation, Formal analysis, Validation, Writing – original draft, Visualization. **Roman C. Maron:** Validation, Writing – original draft, Visualization. **Achim Hekler:** Validation, Writing – Review & Editing, Project administration. **Jochen S. Utikal:** Resources, Writing – Review & Editing, Visualization. **Catarina Barata:** Resources, Writing – Review & Editing, Visualization. **Raymond L. Barnhill:** Resources, Writing – Review & Editing, Visualization. **Helmut Beltraminelli:** Resources, Writing – Review & Editing, Visualization. **Carola Berking:** Resources, Writing – Review & Editing, Visualization. **Brigid Betz-Stablein:** Resources, Writing – Review & Editing, Visualization. **Andreas Blum:** Resources, Writing – Review & Editing, Visualization. **Stephan A. Braun:** Resources, Writing – Review & Editing, Visualization. **Richard Carr:** Resources, Writing – Review & Editing, Visualization. **Marc Combalia:** Resources, Writing – Review & Editing, Visualization. **Maria-Teresa Fernandez-Figueras:** Resources, Writing – Review & Editing, Visualization. **Gerardo Ferrara:** Resources, Writing – Review & Editing, Visualization. **Sylvie Fraitag:** Resources, Writing – Review & Editing, Visualization. **Lars E. French:** Resources, Writing – Review & Editing, Visualization. **Frank F. Gellrich:** Resources, Writing – Review & Editing, Visualization. **Kamran Ghoreschi:** Resources, Writing – Review & Editing, Visualization. **Matthias Goebeler:** Resources, Writing – Review & Editing, Visualization. **Pascale Guitera:** Resources, Writing – Review & Editing, Visualization. **Holger A. Haenssle:** Resources, Writing – Review & Editing, Visualization. **Sebastian Haferkamp:** Resources, Writing – Review & Editing, Visualization. **Lucie Heinzerling:** Resources, Writing – Review & Editing, Visualization. **Markus V. Heppt:** Resources, Writing – Review & Editing, Visualization. **Franz J. Hilke:** Resources, Writing – Review & Editing, Visualization. **Sarah Hobelsberger:** Resources, Writing – Review & Editing, Visualization. **Dieter Krahl:** Resources, Writing – Review & Editing, Visualization. **Heinz Kutzner:** Resources, Writing – Review & Editing, Visualization. **Aimilios Lallas:** Resources, Writing – Review & Editing, Visualization. **Konstantinos Liopyris:** Resources, Writing – Review & Editing, Visualization. **Mar Llamas-Velasco:** Resources, Writing – Review & Editing, Visualization. **Josep Malvehy:** Resources, Writing – Review & Editing, Visualization. **Friedegund Meier:** Resources, Writing – Review & Editing, Visualization. **Cornelia S. L. Müller:** Resources, Writing – Review & Editing, Visualization. **Alexander A. Navarini:** Resources, Writing – Review & Editing, Visualization. **Cristián Navarrete-Dechent:** Resources, Writing – Review & Editing, Visualization. **Antonio Perasole:** Resources, Writing – Review & Editing, Visualization. **Gabriela Poch:** Resources, Writing – Review & Editing, Visualization. **Sebastian Podlipnik:** Resources, Writing – Review & Editing, Visualization. **Luis Requena:** Resources, Writing – Review & Editing, Visualization. **Veronica M. Rotemberg:** Resources, Writing – Review & Editing, Visualization. **Andrea Saggini:** Resources, Writing – Review & Editing, Visualization. **Omar P. Sangueza:** Resources, Writing – Review & Editing, Visualization. **Carlos Santonja:** Resources, Writing – Review & Editing, Visualization. **Dirk Schadendorf:** Resources, Writing – Review & Editing, Visualization. **Bastian Schilling:** Resources, Writing – Review & Editing, Visualization. **Max Schlaak:** Resources, Writing – Review & Editing, Visualization. **Justin G. Schlager:** Resources, Writing – Review & Editing, Visualization. **Mildred Sergon:** Resources, Writing – Review & Editing, Visualization. **Wiebke Sondermann:** Resources, Writing – Review & Editing, Visualization. **H. Peter Soyer:** Resources, Writing – Review & Editing, Visualization. **Hans Starz:** Resources, Writing – Review & Editing, Visualization. **Wilhelm Stolz:** Resources, Writing – Review & Editing, Visualization. **Esmeralda Vale:** Resources, Writing – Review & Editing, Visualization. **Wolfgang Weyers:** Resources, Writing – Review & Editing, Visualization. **Alexander Zink:** Resources, Writing – Review & Editing, Visualization. **Eva Krieghoff-Henning:** Writing – Review & Editing, Visualization, Project administration. **Jakob N. Kather:** Resources, Writing – Review & Editing, Visualization. **Christof von Kalle:** Resources, Writing – Review & Editing, Visualization. **Daniel B. Lipka:** Resources, Writing – Review & Editing, Visualization. **Stefan Fröhling:** Resources, Writing – Review & Editing, Visualization. **Axel Hauschild:** Resources, Writing – Review & Editing, Visualization. **Harald Kittler:** Resources, Writing – Review & Editing, Visualization. **Titus J. Brinker:** Conceptualization, Writing – Review

& Editing, Validation, Supervision, Project administration, Funding acquisition.

Role of the funding source

This study was funded by the Federal Ministry of Health, Berlin, Germany (grant: Skin Classification Project 2; grant holder: T.J.B., German Cancer Research Center, Heidelberg, Germany). The sponsor had no role in the design and conduct of the study; collection, management, analysis and interpretation of the data; preparation, review or approval of the manuscript and decision to submit the manuscript for publication. This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA008748. This research is part of the doctoral thesis of Haggenmüller S.

Conflict of interest statement

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: J.S.U. is on the advisory board or has received honoraria and travel support from Amgen, Bristol Myers Squibb, GSK, LEO Pharma, Merck Sharp and Dohme, Novartis, Pierre Fabre and Roche, outside the submitted work. M.G. has received speaker's honoraria and/or has served as a consultant and/or member of advisory boards for Almirall, Argenx, Biotest, Eli Lilly, Janssen Cilag, LEO Pharma, Novartis and UCB, outside the submitted work. H.A.H. worked as a consultant or received honoraria and travel support from Heine Optotechnik GmbH, JenLab GmbH, FotoFinder Systems GmbH, Magnosco GmbH, SciBase AB, Beiersdorf AG, Almirall Hermal GmbH and Galderma Laboratorium GmbH. V.M.R. is on the advisory board or has received honoraria or ownership in Inhabit Brands, Inc. unrelated to this work. Sondermann W. reports grants from medi GmbH Bayreuth, personal fees from Janssen, grants and personal fees from Novartis, personal fees from Lilly, personal fees from UCB, personal fees from Almirall, personal fees from LEO Pharma and personal fees from Sanofi Genzyme, outside the submitted work. H.P.S. is a shareholder of MoleMap NZ Limited and e-derm consult GmbH and undertakes regular tele-dermatological reporting for both companies. H.P.S. is a medical consultant for Canfield Scientific, Inc., MoleMap Australia Pty Ltd and Revenio Research Oy and a medical advisor for First Derm. M.L-V. has received speaker's honoraria and/or received grants and/or participated in clinical trials of AbbVie, Almirall, Amgen, Celgene, Eli Lilly, Janssen Cilag, LEO Pharma, Novartis and UCB, outside the submitted work. A.Z. has been an advisor and/or received speaker's honoraria and/or received grants and/or participated in clinical trials of AbbVie, Almirall, Amgen, Beiersdorf Dermo Medical, Bencard Allergy, Celgene, Eli Lilly, Janssen Cilag, LEO Pharma, Novartis, Sanofi-Aventis and UCB Pharma, outside the submitted work. Kittler H. received speaker's honoraria from FotoFinder Systems GmbH and received non-financial support from Heine Optotechnik GmbH, Derma Medical and 3Gen. T.J.B. reports owning a company that develops mobile apps, including the tele-dermatology services AppDoc (https://online-hautarzt.de) and Intimarzt (https://Intimarzt.de); Smart Health Heidelberg GmbH, Handschuhsheimer Landstr. 9/1, 69120 Heidelberg, https://smarthealth.de. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejca.2021.06.049.

References

[1] Mahbod A, Schaefer C, Ellinger I, Ecker R, Pitiot A, Wang C. Fusing fine-tuned deep features for skin lesion classification. Comput Med Imaging Graph 2019;71:19–29. https://doi.org/10.1016/j.compmedimag.2018.10.007.

[2] Salerni G, Terán T, Puig S, Malvehy J, Zalaudek I, Argenziano G, et al. Meta-analysis of digital dermoscopy follow-up of melanocytic skin lesions: a study on behalf of the International Dermoscopy Society. J Eur Acad Dermatol Venereol 2013;27:805–14.

[3] Vestergaard ME, 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;159:669–76.

[4] Haenssle HA, 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;29:1836–42.

[5] Lodha S, Saggar S, Celebi JT, Silvers DN. Discordance in the histopathologic diagnosis of difficult melanocytic neoplasms in the clinical setting. J Cutan Pathol 2008;35:349–52.

[6] Corona R, Mele A, Amini M, De Rosa G, Coppola G, Piccardi P, et al. Interobserver variability on the histopathologic diagnosis of cutaneous melanoma and other pigmented skin lesions. J Clin Oncol 1996;14:1218–23.

[7] 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;155:58–65.

[8] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8.

[9] Nasr-Esfahani E, Samavi S, Karimi N, Soroushmehr SMR, Jafari MH, Ward K, et al. Melanoma detection by analysis of clinical images using convolutional neural network. In: 2016 38th annual international conference of the IEEE engineering in medicine and biology society. EMBC; 2016. p. 1373–6.

[10] De Logu F, Ugolini F, Maio V, Simi S, Cossu A, Massi D, et al. Recognition of cutaneous melanoma on digitized histopathological slides via artificial intelligence algorithm. Front Oncol 2020;10:1559.
[11] 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 Canc 2019;115:79–83.
[12] Brinker TJ, Hekler A, Enk AH, von Kalle C. Enhanced classifier training to improve precision of a convolutional neural network to identify images of skin lesions. PloS One 2019;14:e0218713.
[13] Hart SN, Flotte W, Norgan AP, Shah KK, Buchan ZR, Mounajjed T. Classification of melanocytic lesions in selected and whole-slide images via convolutional neural networks. J Path Inform 2019;10(5). https://doi.org/10.4103/jpi.jpi_32_18.
[14] Acs B, Ahmed FS, Gupta S, Wong PF, Gartrell RD, Sarin Pradhan J, et al. An open source automated tumor infiltrating lymphocyte algorithm for prognosis in melanoma. Nat Commun 2019;10:5440.
[15] Kulkarni PM, Robinson EJ, Sarin Pradhan J, Gartrell-Corrado RD, Rohr BR, Trager MH, et al. Deep learning based on standard H&E images of primary melanoma tumors identifies patients at risk for visceral recurrence and death. Clin Canc Res 2020;26:1126–34.
[16] 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.
[17] Brinker TJ, Hekler A, Enk AH, Berking C, Haferkamp S, Hauschild A, et al. Deep neural networks are superior to dermatologists in melanoma image classification. Eur J Canc 2019;119:11–7.
[18] Haenssle HA, Fink C, Toberer F, Winkler J, Stolz W, Deinlein T, et al. Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. Ann Oncol 2020;31:137–43.
[19] Yu C, Yang S, Kim W, Jung J, Chung K-Y, Lee SW, et al. Acral melanoma detection using a convolutional neural network for dermoscopy images. PloS One 2018;13:e0193321.
[20] Marchetti MA, Codella NCF, Dusza SW, Gutman DA, Helba B, Kalloo A, et al. Results of the 2016 international skin imaging collaboration international Symposium on biomedical imaging challenge: comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. J Am Acad Dermatol 2018;78:270–277.e1.
[21] Marchetti MA, Liopyris K, Dusza SW, Codella NCF, Gutman DA, Helba B, et al. Computer algorithms show potential for improving dermatologists' accuracy to diagnose cutaneous melanoma: results of the International Skin Imaging Collaboration 2017. J Am Acad Dermatol 2020;82:622–7.
[22] Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. Lancet Oncol 2019;20:938–47.
[23] Maron RC, Weichenthal M, Utikal JS, Hekler A, Berking C, Hauschild A, et al. Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. Eur J Canc 2019;119:57–65.
[24] Haenssle HA, Winkler JK, Fink C, Toberer F, Enk A, Stolz W, et al. Skin lesions of face and scalp - classification by a market-approved convolutional neural network in comparison with 64 dermatologists. Eur J Canc 2021;144:192–9.
[25] Fujisawa Y, Otomo Y, Ogata Y, Nakamura Y, Fujita R, Ishitsuka Y, et al. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. Br J Dermatol 2019;180:373–81.
[26] Jinnai S, Yamazaki N, Hirano Y, Sugawara Y, Ohe Y, Hamamoto R. The development of a skin cancer classification system for pigmented skin lesions using deep learning. Biomolecules 2020;10. https://doi.org/10.3390/biom10081123.
[27] Han SS, Park I, Eun Chang S, Lim W, Kim MS, Park GH, et al. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. J Invest Dermatol 2020;140:1753–61.
[28] 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;138:1529–38.
[29] Han SS, Moon IJ, Kim SH, Na J-I, Kim MS, Park GH, et al. Assessment of deep neural networks for the diagnosis of benign and malignant skin neoplasms in comparison with dermatologists: a retrospective validation study. PLoS Med 2020;17:e1003381.
[30] 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. Eur J Canc 2019;111:148–54.
[31] 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 Canc 2019;118:91–6.
[32] Brinker TJ, Schmitt M, Krieghoff-Henning EI, Barnhill R, Beltraminelli H, Braun SA, et al. Diagnostic performance of artificial intelligence for histologic melanoma recognition compared to 18 international expert pathologists. J Am Acad Dermatol 2021. https://doi.org/10.1016/j.jaad.2021.02.009.
[33] Navarrete-Dechent C, Dusza SW, Liopyris K, Marghoob AA, Halpern AC, Marchetti MA. Automated dermatological diagnosis: hype or reality? J Invest Dermatol 2018;138:2277–9.
[34] Navarrete-Dechent C, Liopyris K, Marchetti MA. Multiclass Artificial intelligence in dermatology: progress but still room for improvement. J Invest Dermatol 2020. https://doi.org/10.1016/j.jid.2020.06.040.
[35] Höhn J, Krieghoff-Henning E, Jutzi TB, von Kalle C, Utikal JS, Meier F, et al. Combining CNN-based histologic whole slide image analysis and patient data to improve skin cancer classification. Eur J Canc 2021;149:94–101.
[36] Li W, Zhuang J, Wang R, Zhang J, Zheng W-S. Fusing metadata and dermoscopy images for skin disease diagnosis. In: 2020 IEEE 17th international Symposium on biomedical imaging. ISBI); 2020. p. 1996–2000.
[37] Winkler JK, Fink C, Toberer F, Enk A, Deinlein T, Hofmann-Wellenhof R, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatol 2019. https://doi.org/10.1001/jamader-matol.2019.1735.
[38] Maron RC, Hekler A, Krieghoff-Henning E, Schmitt M, Schlager JG, Utikal JS, et al. Reducing the impact of confounding factors on skin cancer classification via image segmentation: technical model study. J Med Internet Res 2021;23:e21695.
[39] Maron RC, Haggenmüller S, von Kalle C, Utikal JS, Meier F, Gellrich FF, et al. Robustness of convolutional neural networks in recognition of pigmented skin lesions. Eur J Canc 2021;145:81–91.
[40] Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human-computer collaboration for skin cancer recognition. Nat Med 2020;26:1229–34.
[41] Hekler A, Utikal JS, Enk AH, Hauschild A, Weichenthal M, Maron RC, et al. Superior skin cancer classification by the combination of human and artificial intelligence. Eur J Canc 2019;120:114–21.