Aus der Klinischen Kooperationseinheit Dermatoonkologie des Deutschen Krebsforschungszentrums (DKFZ) an der Klinik für Dermatologie, Venerologie und Allergologie der Medizinischen Fakultät Mannheim
(Leiter: Prof. Dr. med. Jochen Sven Utikal)

# The Skin Classification Project: How can an artificial intelligence-based algorithm for image-based melanoma detection be successfully implemented in clinical practice?

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum (Dr. sc. hum.)
der
Medizinischen Fakultät Mannheim
der Ruprecht-Karls-Universität
zu
Heidelberg

vorgelegt von
Sarah Haggenmüller

aus
Ulm, Deutschland
2023

Dekan: Prof. Dr. med. Sergij Goerdt
Referent: Prof. Dr. med. Jochen Sven Utikal

# FOREWORD

This publication-based doctoral thesis includes the following first authorships as main publications:

Publication 1: ***Haggenmüller****, Maron, Hekler et al. Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts, Original Research, published in European Journal of Cancer, IF 8.4*

Publication 2: ***Haggenmüller****, Maron, Hekler et al. Artificial intelligence for skin cancer diagnostics: patients and dermatologists require AI-systems with enhanced explainability and multiclass assessment, Original Article, recommended for revision as research letter, Journal of the American Academy of Dermatology, IF 13.8*

Publication 3: ***Haggenmüller****, Schmitt, Krieghoff-Henning et al. Comparison of federated learning for decentralized artificial intelligence in melanoma diagnostics, Original Investigation, accepted in JAMA Dermatology, shared first authorship, IF 10.9*

Summary of Haggenmüller's contributions to the main publications 1 to 3:

| | Publication 1 | Publication 2 | Publication 3 |
|---|---|---|---|
| Conception (%) | 90% | 100% | 50% |
| Literature review (%) | 100% | 100% | 80% |
| Ethics approval (%) | not applicable | not applicable | 100% |
| Animal testing application (%) | not applicable | not applicable | not applicable |
| Data collection (%) | 100% | 80% | 80% |
| Data analysis (%) | 100% | 100% | 60% |
| Interpretation of the results (%) | 100% | 100% | 70% |
| Writing of the manuscript (%) | 95% | 100% | 90% |
| Revision (%) | 80% | 80% | 80% |
| Figures/tables resulting from the doctoral thesis | all tables/figures | all tables/figures | Tab. 5/6, Fig. 3, Suppl. Fig. 10/11. |
| Data/figures/tables based on research findings of others | not applicable | not applicable | Fig. 4, Tab. 7, Suppl. Tab. 8, Suppl. Fig. 9 |

Furthermore, this publication-based doctoral thesis includes the following co-authorships as supplementary publications:

Supplementary publication 1: *Maron, **Haggenmüller**, von Kalle et al. Robustness of convolutional neural networks in recognition of pigmented skin lesions, Original Research, published in European Journal of Cancer, IF 8.4*

Supplementary publication 2: *Maron, Hekler, **Haggenmüller** et al. Model soups improve performance of dermoscopic skin cancer classifiers, Original Research, published in European Journal of Cancer, IF 8.4*

Supplementary publication 3: *Hekler, Maron, **Haggenmüller** et al. Using multiple real-world dermoscopic photographs of one lesion improves melanoma classification via deep learning, Original Article, recommended for revision as research letter, Journal of the American Academy of Dermatology, IF: 13.8*

TABLE OF CONTENTS

page

# ABBREVIATIONS

| | |
|---|---|
| MM | malignant melanoma |
| AI | artificial intelligence |
| CNN | convolutional neural network |
| ISBI | Symposium on Biomedical Imaging |
| cCNN | combined convolutional neural network |
| ACBC | adaptive choice-based conjoint analysis |
| CI | confidence intervals |
| FL | federated learning |
| AUROC | area under the receiver operating characteristic curve |
| IV | invasive melanoma |
| ECE | expected calibration error |
| BS | Brier score |
| NLL | negative log likelihood |
| mBCE | mean balanced corruption error |
| mFR | mean flip rate |
| MV-Artificial | multiview-artificial |
| MV-Real | multiview-real |
| MCC | maximum confidence change |
| KI | künstliche Intelligenz |

# 1 INTRODUCTION

## 1.1 Epidemiology of malignant melanoma

Malignant melanoma (MM) is the leading cause of death among skin cancer patients worldwide [1]. In 2020, approximately 325,000 new cases of MM were diagnosed, resulting in about 60,000 deaths [2]. This concerning trend is primarily attributed to the ever-increasing exposure to ultraviolet radiation, which is a well-known risk factor for the development of MM [3,4]. As a result, incidences are expected to rise dramatically in the future. Experts anticipate that by 2040, the number of new diagnoses of MM will reach approximately 510,000 cases per year, leading to an estimated 96,000 MM-related deaths annually [2].

While some cases of MM display aggressive behavior from an early stage, the probability of metastasis (i.e., the spread of cancer to other parts of the body) increases significantly with tumor thickness [5,6]. Detecting MM in its early stages substantially improves the survival chances of affected patients. Consequently, a rapid and accurate identification of MM carries unprecedented importance.

## 1.2 Diagnostic challenges of malignant melanoma

Early diagnosis, however, remains challenging due to frequent morphological overlap between MM and atypical nevi [7]. Clinical naked-eye examination allows the assessment of morphological features of a lesion with classification frameworks, such as the ABCDE rule [8], but is limited to the skin surface. To enhance diagnostic accuracy compared to naked-eye examination, dermatologists routinely employ dermoscopy, a technique that enables the visualization of deeper skin layers, revealing colors and

structures that are typically imperceptible to the unaided eye [9]. Despite these technical advancements, even experienced dermatologists rarely achieve sensitivity levels exceeding 80% [10]. In cases where a clinical suspicion of malignancy cannot be ruled out, a skin biopsy is routinely performed following dermoscopic examination. The subsequent histopathological examination of the biopsied lesion by a (dermato-)pathologist currently serves as the gold standard for diagnosing skin cancer. However, even this histopathological verification can yield inconclusive results, particularly in borderline cases or with thin MM (Breslow thickness <1mm), the latter accounting for the majority of cases detected during skin cancer screenings [11,12]. Notably, previous studies have shown a discordance between individual (dermato-)pathologists for classifying MM of up to 25% [13–15].

## 1.3 Consequences for patients and the healthcare system

The difficulty in differentiating MM from atypical benign lesions, combined with the potential discrepancy between individual physicians, can result in both overdiagnosis and underdiagnosis at various stages of the diagnostic process.

On one hand, numerous nevi are excised based on clinical suspicion, frequently leading to clinical overdiagnosis associated with unnecessary physical and psychological stress for the affected screening participants and avoidable costs for the healthcare system. For example, a study from 2013 revealed that approximately 1.8 percent of all screening participants in Germany were initially suspected of having MM. However, subsequent biopsies confirmed this diagnosis for only 0.1 percent of all individuals screened [11]. Considering that around 10 million people in Germany undergo annual

skin cancer screenings, this equates to approximately 170,000 avoidable biopsies every year.

Despite this common occurrence of overdiagnosis, certain cases of MM are still misdiagnosed as nevi or overlooked at clinical level, potentially leading to their discovery only at advanced stages [16]. Depending on the tumor thickness, this misdiagnosis can have critical consequences, markedly diminishing patients' chances of survival, and – in the worst-case scenarios – becoming the difference between life and death.

Addressing the issue of early MM detection while simultaneously minimizing the false positive rate requires the development of improved diagnostic systems. One particularly promising approach involves the application of artificial intelligence (AI) for improved MM detection.

## 1.4 Artificial intelligence-based skin cancer diagnostics

In this context, convolutional neural networks (CNNs), deep neural networks specifically designed for image-based classification, have shown promise for enhancing the diagnostic accuracy of MM detection [17]. CNNs are commonly trained via supervised learning, where they use labeled data, such as dermoscopic images with their corresponding diagnosis (i.e., ground truth), to perform end-to-end learning. This involves the network learning a direct relationship between the raw input data and the labels, enabling it to classify previously unseen skin lesion images.

In experimental settings, CNNs have already demonstrated comparable or even superior performance levels in comparison to experienced clinicians when using

clinical [18–22], dermoscopic [23–32] or histopathological whole-slide images [33–35] for various skin cancer classification tasks. These promising results indicate that AI-based approaches can potentially enhance the detection of MM. Thereby, offering the dual benefit of potentially enabling earlier interventions for otherwise overlooked cases of MM, while simultaneously reducing overdiagnosis and overtreatment.

However, the successful application of AI-assisted skin cancer diagnostics – as observed in experimental settings – has hardly been transferred to clinical practice so far [36]. In everyday clinical care, several additional challenges emerge. These challenges include the potential lack of acceptance by both patients and clinicians [37,38], as well as data privacy concerns, particularly when data are transferred to external institutions [39]. Moreover, state-of-the-art AI algorithms face issues related to algorithm robustness (i.e., the ability to consistently output the same diagnosis even if the target image is slightly changed, e.g., rotation) [40,41], uncertainty estimation (i.e., the ability to correctly estimate the uncertainty of a prediction) [42,43], as well as biological (e.g., various skin types) and/or technical (e.g., various acquisition systems) generalization (i.e., the ability to make accurate predictions on unseen images from different data distributions) [28,44]. The disparity between research conditions and clinical reality, renders it difficult to draw conclusions about the applicability of AI for skin cancer diagnostics outside the research environment.

## 1.5 Aims of the doctoral thesis

Against this background, the present doctoral thesis aims to conduct a feasibility study on the use of AI-systems for skin cancer diagnostics to investigate the research

question “How can an AI-based algorithm for image-based melanoma detection be successfully implemented in clinical practice?”.

The doctoral thesis can be divided into the following areas of investigation:

1. Compiling a review of state-of-the-art AI technology and comparative studies in the research field (see **publication 1**).
2. Investigating the patients’ and clinicians’ perspective on AI-systems for skin cancer diagnostics to facilitate a more demand-orientated AI development (see **publication 2**).
3. Exploring decentralized federated learning as a potentially more accessible and privacy-preserving alternative for the development of AI-systems for skin cancer diagnostics (see **publication 3**).
4. Examining various technical approaches for the development of potentially more robust diagnostic algorithms with improved generalization capabilities (see **supplementary publications 1 to 3**).

# 2 PUBLICATIONS

## 2.1 Publication 1: 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

[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

* Corresponding author: Digital Biomarkers for Oncology Group, 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).

The original publication is available at DOI: https://doi.org/10.1016/j.ejca.2021.06.049

### 2.1.1 Abstract

**Background:** Multiple studies have compared the performance of 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.

**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.

### 2.1.2 Introduction

Although 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 [45].

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 [9], even specialists rarely achieve sensitivity levels above 80% [10]. Beyond that, a significant variance depending on training and professional experience can be observed [30].

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% [13,15,30].

Against this backdrop, accurate distinction between benign and malignant skin lesions as well as the exact classification of skin cancer types through digital biomarkers is of great interest to reduce the number of missed MM as well as unnecessary excisions.

Digital biomarkers are data-driven indicators that provide information about the characteristics of a lesion and may predict health-related outcomes.

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 digital biomarkers [29].

This systematic review presents state of the art 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.

### 2.1.3 Material and methods

**Search strategy**

In 2017, *Esteva et al.* [27] 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 **Tabular Appendix 1**).

**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 Figure 1**). 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., [35,46–48]) 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., [49]). 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., [50,51]). Data were extracted from peer-reviewed articles exclusively. Data quality was assessed independently by two reviewers.

**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 data do not follow the training distribution and are often referred to as an external test set (e.g., from external clinics).

**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.

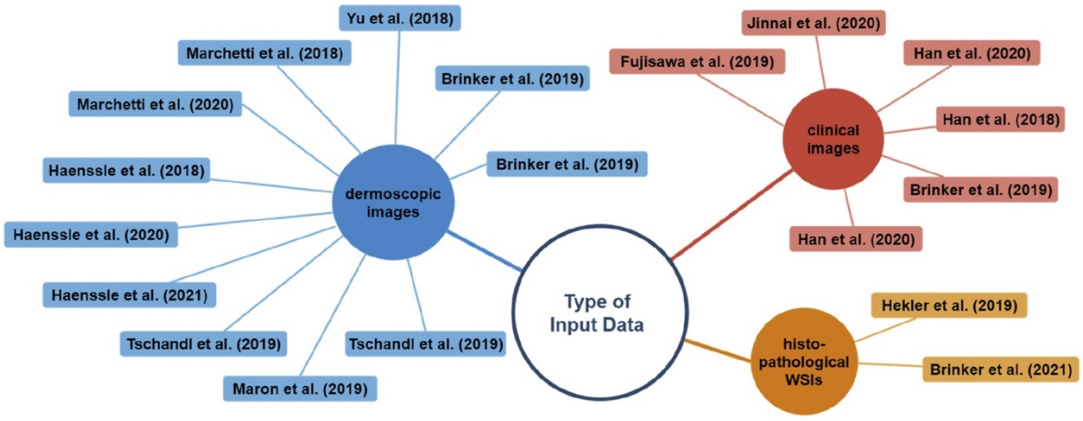
Sensitivity and specificity are not influenced by class imbalances and better reflect the performance for a specific class. However, both metrics require a dichotomous classification, where only one positive and one negative class are considered (e.g., MM vs. melanocytic nevus, benign vs. malignant or 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.

### 2.1.4 Results

A total of 19 comparative studies (since *Esteva et al.*'s [27] seminal article) were published that fulfilled the inclusion criteria. Most of the studies focused on dermoscopic images (n = 11) [25,26,28–32,52–55], followed by clinical image (n = 6) [18–22,24] and histopathological whole-slide image studies (n = 2) [33,34] (see **Figure 1**). In the following, the term histopathological whole-slide images refers to digitised hematoxylin-eosin-stained tissue sections processed with specialised slide scanners.

**Figure 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 [25,26,28–32,52–55], those based on clinical images [18–22,24] and those based on histopathological whole-slide images [33,34]. WSI: whole slide image




**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. **Tabular Appendix 2** contrasts the training and testing procedures of these approaches.

**Table 1. Overview reader studies based on dermoscopic images.**

| reader study | comparison with | scope of the reader study test set | meta-data y/n | origin of the reader study test set (Hold-out/OOD) | setting c/e | classification task | results |
|---|---|---|---|---|---|---|---|
| *Brinker et al.* [26] | 157 dermatologists<br>- 151 university hospital-based from 12 university hospitals in Germany:<br>- 88 junior clinicians<br>- 15 attendings<br>- 45 senior clinicians<br>- 3 chief clinicians<br>- 6 dermatologists in private practice | 100 images, randomly selected out of 20 735 images available at ISIC | n | ISIC image archive (Holdout) | e | binary: melanoma/ melanocytic nevi | CNN outperformed 136 out of 157 dermatologists |
| *Brinker et al.* [52] | 144 dermatologists from 9 university hospitals in Germany<br>- 92 junior clinicians<br>- 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.* [53] | - 2 general physicians<br>- 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),<br>Dongsan Hospital in the Keimyung University Health System, Daegu, Korea (Holdout) | e | binary: acral melanoma/ melanocytic nevi | comparable performance |
| *Marchetti et al.* [54] | 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.* [55] | 17 dermatologists<br>- 8 dermatologists from 4 countries<br>- 9 dermatologists in private practice 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.* [30] | 58 dermatologists from 17 countries, including 30 experts with more than 5 years of dermoscopic experience | selected 100 images with increased difficulty out of 300 images<br>I) dermoscopy only<br>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/ no action | significant superiority of the CNN |
| *Haenssle et al.* [31] | 96 dermatologists<br>- 17 beginners<br>- 29 skilled | 100 images with increased difficulty<br>I) dermoscopy only | y | Department of Dermatology, University of Heidelberg, Germany (OOD) | e | binary: (pre)malignant/ benign; excision or | comparable performance |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | - 40 experts<br>- 10 provided no information | II) in addition: clinical information and close-up images | | | | treatment/follow up or no action | |
| *Haenssle et al.* [32] | 64 dermatologists<br>- 9 beginners<br>- 20 skilled<br>- 30 experts<br>- 5 unknown | 100 images of face and scalp lesions in total<br>I) dermoscopy only<br>II) in addition: clinical information and close-up images | y | Department of Dermatology, University of Heidelberg, Germany (OOD)<br>Department of Dermatology Hospital Thalkirchner Street, Munich, Germany (OOD)<br>Department of Dermatology, Medical University Graz, Austria (OOD)<br>First Department of Dermatology, Aristotle University, Thessaloniki, Greece (OOD)<br>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.* [28] | 511 participants from 63 countries including 283 board-certified dermatologists, 118 dermatologists in private practice and 83 general practitioners | randomly selected 30 images per participant out of 1 511 images | n | HAM10000 data set, ISBI 2018 challenge, ISIC image archive (Holdout),<br>additional images from Turkey, New Zealand, Sweden and Argentina (OOD) | e | multiclass (7) | significant superiority of the CNN |
| *Maron et al.* [25] | 112 dermatologists<br>- 108 university hospital-based from 13 university hospitals in Germany:<br>- 67 junior clinicians<br>- 12 attendings<br>- 28 senior clinicians<br>- 1 chief physician<br>- 4 dermatologists in private practice | 6 subsets consisting of 50 images, 300 images in total | n | HAM10000 data set (Holdout) | e | binary: malignant/ benign;<br>multiclass (5) | significant superiority of the CNN |
| *Tschandl et al.* [29] | 95 human raters (medical personnel), including 62 board-certified dermatologists | randomly selected 50 images per participant out of 2 072 images | n | Primary skin cancer clinic in Queensland, Australia (Holdout),<br>Department of Dermatology of the Medical University of Vienna, Austria (OOD),<br>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, localization of the suspicious lesion)
y (yes)
n (no)
c (clinical setting)
e (experimental setting)

*Brinker et al.* [26] 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.* [52] 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 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.* [53] developed an algorithm focusing 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.* [54] 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.* [55] 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.* [30] 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 (excision 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.* [31] 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.* [32] 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.* [26,52], *Yu et al.* [53], *Marchetti et al.* [54,55] and *Haenssle et al.* [30–32], by carrying out multiclass classification tasks which covered more fine-grained diagnoses (see **Table 1**) [25,28,29]. **Tabular Appendix 2** outlines similarities and differences of these multiclass approaches with regard to individual training and testing procedures.

In 2019, *Tschandl et al.* [28] 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 **Tabular Appendix 5**), 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.* [25] proposed a similar reader study to *Tschandl et al.* [28] by developing a classifier to differentiate between MM and four other skin disease classes (see **Tabular Appendix 5**). 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.* [29] 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 **Tabular Appendix 5**) with comparable performance with that of 95 human raters (at participants' specificity of 51.3%: 80.5% vs. 77.6% sensitivity).

**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**). **Tabular Appendix 3** outlines the training and testing procedure of each individual approach.

**Table 2. Overview reader studies based on clinical images.**

| reader study | comparison with | scope of the reader study test set | meta-data y/n | origin of the reader study test set (Hold-out/OOD) | set-ting c/e | binary/ multiclass | results |
|---|---|---|---|---|---|---|---|
| *Fujisawa et al.* [18] | 22 dermatologists<br>- 9 dermatologic trainees<br>- 13 board-certified | randomly selected 140 images per participant out of 1 142 im-ages | n | University of Tsubuka Hospital, Japan (Holdout) | e | binary: malignant/benign, multiclass (14) | significant superiority of the CNN |
| *Jinnai et al.* [21] | 20 dermatologists<br>- 10 dermatologic trainees<br>- 10 board-certified | randomly selected 10 test samples of 200 images out of 1 114 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.* [20] | binary:<br>47 dermatologists<br>- 21 board-certified<br>- 26 dermatologists in private practice<br><br>multiclass:<br>4 dermatologists<br>- 2 board-certified<br>- 2 dermatology residents | randomly selected 240 images out of 2 201 images | n | SNU data set (OOD) | e | binary: malignant/benign, multiclass (134) | binary:<br>on par performance<br><br>multiclass:<br>comparable, but slightly worse perfor-mance of the CNN |
| *Han et al.* [22] | 16 dermatologist board members<br>- 6 clinicians (>10 years of experi-ence)<br>- 10 professors | randomly selected 480 im-ages<br>1) 260 images of 12 disorders out of 1276 images<br>2) 220 images of 10 disorders out of 1300 images | n | 1) Asan test set (Holdout)<br>2) Edinburgh data set (OOD) | e | multiclass<br>(12) | on par performance |
| *Brinker et al.* [24] | 145 dermatologists<br>- 142 university hospital-based:<br>- 88 junior clinicians<br>- 16 attendings<br>- 35 senior clinicians<br>- 3 chief clinicians<br>- 3 dermatologists in private practice | 100 images | n | MClass-Benchmark obtained from the MED-NODE database (OOD) | e | binary:<br>melanoma/ melanocytic nevi | on par performance |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *Han et al.* [19] | 1) 65 attending clinicians<br>2) 44 board-certified dermatologists | 1) 40 331 images from 10 426 cases of 43 disorders*<br>2) randomly selected 44 image batches of 30 patients out of 5 065 images | n | Department of Dermatology, Severance Hospital in Seoul, Korea (OOD) | 1) c<br>2) e | binary: malignant/benign, multiclass (32) | 1) significant superiority of the attending clinicians<br>2) binary: on par performance multiclass: significant superiority of the CNN |

*For multiclass classification 39 721 images from 10 315 cases of 32 disorders remained, after excluding cases belonging to too small and untrained classes.

metadata (additional information for readers beyond image input, e.g., age, gender, localization of the suspicious lesion)

y (yes)

n (no)

c (clinical setting)

e (experimental setting)

**Table 3. Overview reader studies based on histopathological whole-slide images.**

| reader study | comparison with | scope of the reader study test set | metadata y/n | origin of the reader study test set (Holdout/OOD) | setting c/e | classification task | results |
|---|---|---|---|---|---|---|---|
| *Hekler et al.* [33] | 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.* [34] | 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, 5-fold cross-testing) | e | binary: melanoma/ melanocytic nevi | on par performance |

metadata (additional information for readers beyond image input, e.g., age, gender, localization of the suspicious lesion)

y (yes)

n (no)

c (clinical setting)

e (experimental setting)

*Fujisawa et al.* [18] 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 **Tabular Appendix 5**). 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.* [21] proposed a similar reader study than *Fujisawa et al.* [18]. 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 **Tabular Appendix 5**). 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.* [20] 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 **Tabular Appendix 5**). 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.* [22] developed a model which focused exclusively on the multiclass discrimination of MM and 11 other skin diseases (see **Tabular Appendix 5**). 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.* [24] 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.* [19] 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 **Tabular Appendix 5**). The 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%).

**Automated skin cancer classification of histopathological whole-slide images**

Whole-slide image scanners have enabled the efficient digitisation of haematoxylin-eosin-stained tissue sections, thereby setting the cornerstone for the development of AI-based digital skin cancer biomarkers for histopathology (e.g., [35,47]). Besides the proposed clinical and dermoscopic studies, two comparative approaches using histopathological whole-slide images met the inclusion criteria of this systematic review (see **Table 3**) [33,34]. **Tabular Appendix 4** summarises the training and testing procedures of both approaches.

*Hekler et al.* [33] were the first to compare the performance of a CNN developed for the classification of cropped image sections of whole-slide images 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.* [34] compared the ability of a CNN ensemble to differentiate MM from benign melanocytic nevi with that of 18 international expert pathologists using the

entire whole-slide images 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%).

### 2.1.5 Discussion

**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 clinically relevant differential diagnosis. The included studies covered three main image types (dermoscopic, clinical and histopathological whole-slide images). 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.

**Test set characteristics**

While a large proportion of clinical reader studies based their comparison on out-of-distribution test sets [19,20,22,24] (see **Table 2**), the vast majority of dermoscopic and histopathological approaches (8 out of 13, see **Table 1**, **Table 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 whole-slide images, 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 [28] 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 out-of-distribution 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 out-of-distribution images (e.g., from external clinics) should be considered the gold standard for future research [44,56].

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*. [44], 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 [28,29]. 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.* [53], *Haenssle et al.* [32] and *Tschandl et al.* [29] proved the potential of CNN-based classification for special anatomic sites such as the face and scalp [32] or acral MM on the hands and feet [53], as well as rare subtypes such as amelanotic MM [29]. 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.

**Test setting**

One possible limitation of almost all proposed publications (18 out of 19, see **Table 1**, **Table 2**, **Table 3**) is the experimental test setting of the conducted reader studies. The decision-making basis of 14 of the 19 (see **Table 1**, **Table 2**, **Table 3**) included reader studies was limited to a single image of the suspicious skin lesion. *Haenssle et al.* [30–32] 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 tele-dermatologists 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 [30–32] and other [19,29] 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., [57,58]).

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 **Table 1**, **Table 2**, **Table 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 [19]. 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.

**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., [24–26]). 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.

**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) [59,60] and small changes in image input (e.g., scaling or rotation) [61], 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 whole-

slide images may be the use as an assistance system, calling for a complementary instead of a comparative perspective (e.g., [62,63]).

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., [50,51]) to leverage the full potential of novel digital biomarkers.

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.

### 2.1.6 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 digital biomarkers. 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 out-of-distribution 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.

### 2.1.7 Acknowledgement

**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.

**Author contribution statement**

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.

**Declaration of interest statement**

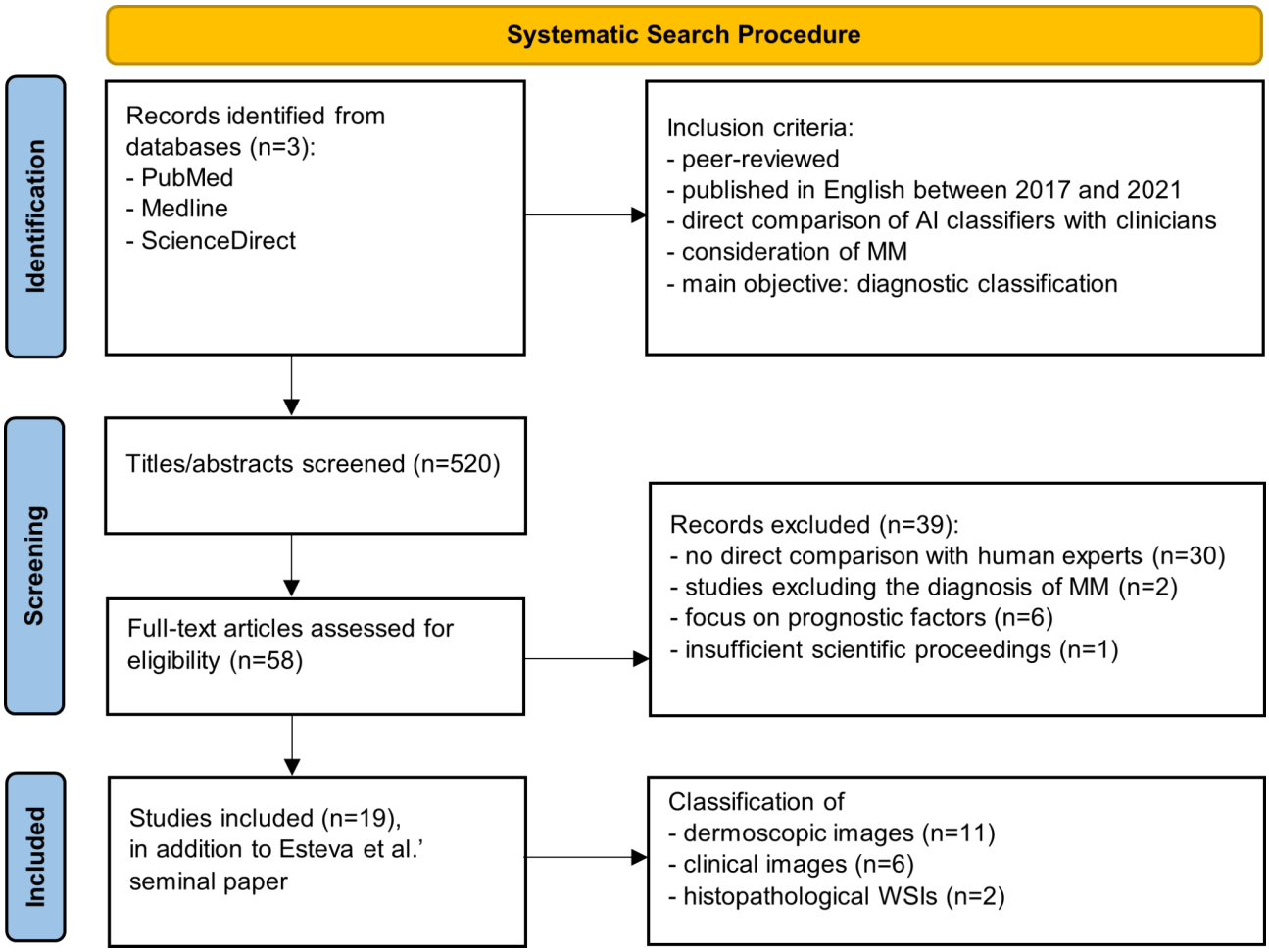
The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: J.S.U. is on the advisory board or has received honoraria and travel support from Amgen, Bristol Myers Squibb, GSK, LEO Pharma, Merck Sharp and Dohme, Novartis, Pierre 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 teledermatological 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 teledermatology 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.

### 2.1.8 Supplementary materials

**Supplementary Figure 1. Flow chart of the systematic search procedure in accordance with PRISMA.** MM: malignant melanoma



## 2.2 Publication 2: Artificial intelligence for skin cancer diagnostics: patients and dermatologists require AI-systems with enhanced explainability and multiclass assessment

Sarah Haggenmüller[a], MSc; Roman C. Maron[a], MSc; Achim Hekler[a], MSc; Eva Krieghoff-Henning, PhD[a]; Jochen S. Utikal[b,c,d], MD; Maria Gaiser[b,c,d], MD; Verena Müller[b,c,d], MD; Sascha Fabian[e], PhD; Friedegund Meier[f], MD; Sarah Hobelsberger[f], MD; Frank F. Gellrich[f], MD; Mildred Sergon[f], MD; Axel Hauschild[g], MD; Michael Weichenthal[g], MD; Lars E. French[h,i], MD; Lucie Heinzerling[h,k], MD; Justin G. Schlager[h], MD; Kamran Ghoreschi[j], MD; Max Schlaak[j], MD; Franz J. Hilke[j], PhD; Gabriela Poch[j], MD; Sören Korsing[j] ,MD; Carola Berking[k], MD; Markus V. Heppt[k], MD; Michael Erdmann[k], MD; Sebastian Haferkamp[l], MD; Konstantin Drexler[l], MD; Dirk Schadendorf[m], MD; Wiebke Sondermann[m], MD; Matthias Goebeler[n], MD; Bastian Schilling[n], MD; Jakob N. Kather[o], MD; Stefan Fröhling[p], MD; Katharina Kaminski[q]; Astrid Doppler[q]; Tabea Bucher[a], MSc; Titus J. Brinker[a,*], MD

[a] Digital Biomarkers for Oncology Group, National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Heidelberg, Germany

[b] Department of Dermatology, Venereology and Allergology, University Medical Center Mannheim, Ruprecht-Karls University of Heidelberg, Mannheim, Germany

[c] Skin Cancer Unit, German Cancer Research Center (DKFZ), Heidelberg, Germany

[d] DKFZ Hector Cancer Institute at the University Medical Center Mannheim, Mannheim, Germany

[e] Department of Economics, University of Applied Science Neu-Ulm, Neu-Ulm, Germany

[f] Department of Dermatology, Faculty of Medicine and University Hospital Carl Gustav Carus, Technische Universität Dresden, Dresden, Germany and Skin Cancer Center at the University Cancer Centre Dresden and National Center for Tumor Diseases, Dresden, Germany

[g] Department of Dermatology, University Hospital (UKSH), Kiel, Germany

[h] Department of Dermatology and Allergy, University Hospital, LMU Munich, Munich, Germany

[i] Dr. Phillip Frost Department of Dermatology and Cutaneous Surgery, University of Miami, Miller School of Medicine, Miami, FL, USA

[j] Department of Dermatology, Venereology and Allergology, Charité – Universitätsmedizin Berlin, Corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[k] Department of Dermatology, University Hospital Erlangen, Comprehensive Cancer Center Erlangen – European Metropolitan Region Nürnberg, CCC Alliance WERA, Erlangen, Germany

[l] Department of Dermatology, University Hospital Regensburg, Regensburg, Germany

[m] Department of Dermatology, University Hospital Essen, Essen and German Cancer Consortium, partner site Essen and National Center for Tumor Diseases (NCT), NCT-West, Campus Essen and University Alliance Ruhr, Research Center One Health, University Duisburg-Essen, Essen, Germany

[n] Department of Dermatology, Venereology and Allergology, University Hospital Würzburg and National Center for Tumor Diseases (NCT) WERA Würzburg, Germany

[o] Else Kroener Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, Technical University Dresden, Dresden, Germany

[p] Department of Translational Medical Oncology, National Center for Tumor Diseases (NCT) Heidelberg and German Cancer Research Center (DKFZ), Heidelberg, Germany

q Melanom Info Deutschland – MID e.V., Essen, Germany

* Corresponding author: Digital Biomarkers for Oncology Group, 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).

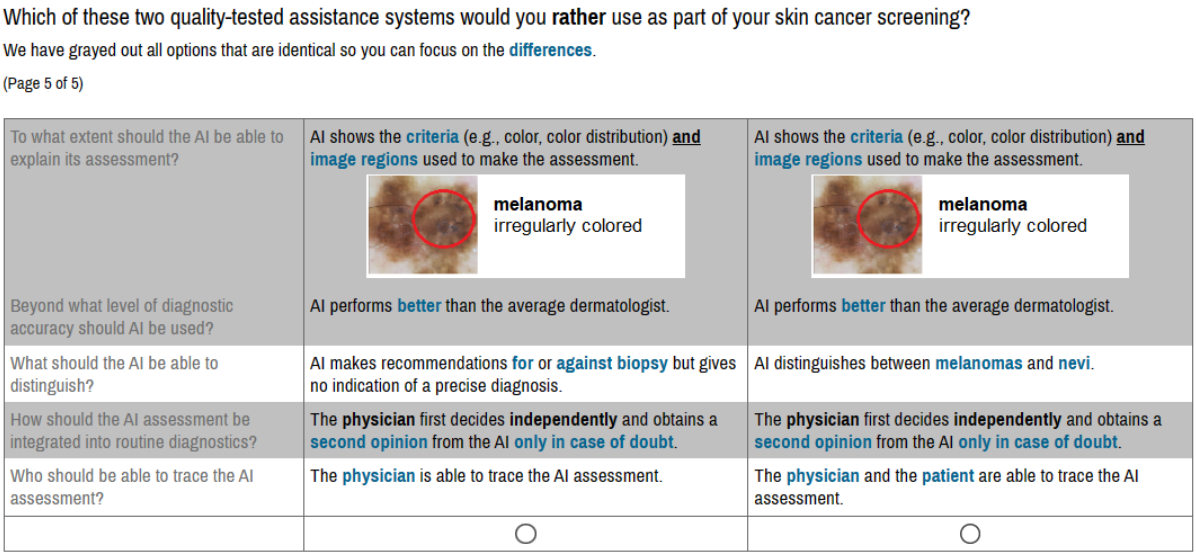
### 2.2.1 Research letter

AI has shown promise for improving diagnostics of skin cancer by matching or surpassing experienced clinicians [27]. However, successful clinical application of AI in skin cancer diagnostics depends on acceptance by patients and dermatologists.

In this prospective multicentric survey study, we therefore investigate the criteria required for patients and dermatologists to accept AI-systems and assess their importance on patients' and dermatologists' decision-making when considering the use of such systems. To this end, we perform an adaptive choice-based conjoint analysis (ACBC) and analyze it using hierarchical Bayes estimation [64]. By employing an ACBC, we investigate multiple influencing AI-features simultaneously (see **Table 4**)

whilst accounting for possible trade-offs from the patients' and dermatologists' perspective (see **Figure 2**). For details on questionnaire development, survey structure, participant recruitment, and statistical analysis, see **Supplementary Methods** in the **Supplementary Materials**).

**Figure 2. Example choice tournament of the present ACBC study design.** The survey was conducted in German, and this example choice tournament was translated into English for this illustration.




The data of 293 included respondents (178 patients and 115 dermatologists) showed a positive general attitude toward AI-systems (see **Supplementary Results** in the **Supplementary Materials** for participant characteristics and further details). However, AI-systems were considered unacceptable by 41.6% of patients (95% confidence interval (CI): 34.3-49.2%) and 47.8% of dermatologists (95% CI: 38.4-57.3%) if neither the dermatologist nor the patient could trace the assessment, and AI-systems were systematically ruled out by 36.5% of patients (95% CI: 29.4-44.1%) and 35.7% of dermatologists (95% CI: 26.9-45.1%) if they did not provide explanations on a case-by-case basis. Diagnostic accuracy and explainability were the most important AI-features in patients' and dermatologists' decision-making with an average importance of 20.6%

(95% CI: 19.3-22.0%) and 26.6% (95% CI: 26.0-27.3%) for patients, and 33.2% (31.1-35.2%) and 20.4% (19.4-21.4%) for dermatologists, respectively.

Participants preferred an increased explainability with display of both decision criteria and relevant image regions. Patients prioritized an AI assessment that is traceable for patients and clinicians, and dermatologists preferred a multiclass differentiation among various disorders (see **Supplementary Results** in the **Supplementary Materials** for further details). Specifically, the differentiation between MM and nevi, which has been the primary focus of AI research in dermatology [17], is considered insufficient. Consequently, there is a need for prospective studies to evaluate AI performance in multiclass assessments to provide a more accurate representation of clinically relevant differential diagnoses.

Current AI research is mainly performance-oriented (e.g., ISIC challenges [55]). However, patients and dermatologists require AI-systems that explain the rationale behind their decision-making and are at least somewhat traceable for both patients and dermatologists. This growing demand for explainable AI poses a key challenge for future research since state-of-the-art technology does not fully explain the reasoning behind its decisions due to the AI black box phenomenon [65].

Altogether, patients and dermatologists prioritized AI-systems with increased explainability (i.e., display of criteria and image regions) and traceability (i.e., understandable by patients and dermatologists) as well as the ability for multiclass decision-making. Therefore, future AI research must go beyond pure performance advancements and adhere to the criteria outlined above for a potentially more successful clinical adoption.

**Table 4. Overview of the AI-features and corresponding options within the ACBC design.** Five AI-features and corresponding options were included in the ACBC analysis based on insights from a literature review and semistructured interviews. The decision task feature was included only for the subgroup of dermatologists, and the input data feature was included only for the subgroup of patients.

| AI-feature | Options |
|---|---|
| **Integration**<br>How should the AI assessment be integrated into routine diagnostics? | • The physician first decides independently and then always obtains a second opinion from the AI.<br>• The physician first decides independently and obtains a second opinion from the AI only in case of doubt.<br>• The AI assessment is always obtained first, and the physician makes his or her decision based on it. |
| **Explainability**<br>To what extent should the AI be able to explain its assessment? | • AI shows the criteria (e.g., color, color distribution) and image regions used to make the assessment.<br>• AI cannot display the image regions, but it displays which criteria (e.g., color, color distribution) were used to make the assessment.<br>• AI cannot display any criteria, but it shows which image regions were used to make the assessment.<br>• AI does not have to explain its assessment on a case-by-case basis. However, it could be shown during the clinical trial that the AI pays attention to biologically relevant structures.<br>• AI does not have to explain its assessment on a case-by-case basis. It could not be shown during the clinical trial that the AI pays attention to biologically relevant structures. |
| **Traceability**<br>Who should be able to trace the AI assessment? | • The physician and the patient are able to trace the AI assessment.<br>• The physician is able to trace the AI assessment.<br>• Neither the physician nor the patient is able to trace the AI assessment. |
| **Diagnostic accuracy**<br>Beyond what level of diagnostic accuracy should AI be used? | • AI performs worse than the average dermatologist.<br>• AI performs equally well as the average dermatologist.<br>• AI performs better than the average dermatologist. |
| **Decision task**<br>**(only asked for dermatologists)**<br>What should the AI be able to distinguish? | • AI distinguishes between benign and malignant skin lesions but gives no indication of a precise diagnosis.<br>• AI makes recommendations for or against biopsy but gives no indication of a precise diagnosis.<br>• AI distinguishes between melanomas and nevi.<br>• AI distinguishes among melanomas, nevi and one category for other skin lesions.<br>• AI distinguishes between melanomas and non-melanomas.<br>• AI distinguishes among melanomas, one category for other types of skin cancer and one for benign skin lesions. |
| **Input data**<br>**(only asked for patients)**<br>What data should the AI use for its assessment? | • AI makes a diagnosis based on skin images exclusively.<br>• AI makes a diagnosis based on skin images and additional information about the skin lesion (e.g., diameter).<br>• AI makes a diagnosis based on skin images and additional information about the patient (e.g., age).<br>• AI makes a diagnosis based on skin images, additional information on the patient and the skin lesion. |

### 2.2.2 Acknowledgement

**Role of the funding source**

This study was funded by the Federal Ministry of Health, Berlin, Germany (grant: Skin Classification Project 2 (SCP2); grant holder: Titus J. Brinker, German Cancer Research Center, Heidelberg, Germany), the Ministry of Social Affairs, Health and Integration of the Federal State Baden-Württemberg, Germany (grant: AI-Translation-Initiative (KTI); grant holder: Titus J. Brinker, German Cancer Research Center, Heidelberg, Germany) and the Studienstiftung des deutschen Volkes, Bonn, Germany (scholarship holder: Sarah Haggenmüller). The sponsors 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.

**Author contribution statement**

Haggenmüller Conceptualization, Methodology, Software, Investigation, Formal analysis, Data Curation, Writing - original draft, Visualization. Maron Validation, Writing (Review & Editing). Hekler Validation, Writing (Review & Editing), Project administration, Supervision. Krieghoff-Henning Validation, Writing (Review & Editing). Utikal Resources, Writing (Review & Editing). Gaiser Resources, Writing (Review & Editing). Müller Resources, Writing (Review & Editing). Fabian Validation, Writing (Review & Editing). Meier Resources, Writing (Review & Editing). Hobelsberger Resources, Writing (Review & Editing). Gellrich Resources, Writing (Review & Editing). Sergon Resources, Writing (Review & Editing). Hauschild Resources, Writing (Review & Editing). Weichenthal Resources, Writing (Review & Editing). French Resources, Writing (Review & Editing). Heinzerling Resources, Writing (Review & Editing). Schlager Resources, Writing (Review & Editing). Ghoreschi Resources, Writing (Review & Editing).

Schlaak Resources, Writing (Review & Editing). Hilke Resources, Writing (Review & Editing). Poch Resources, Writing (Review & Editing). Korsing Resources, Writing (Review & Editing). Berking Resources, Writing (Review & Editing). Heppt Resources, Writing (Review & Editing). Erdmann Resources, Writing (Review & Editing). Haferkamp Resources, Writing (Review & Editing). Drexler Resources, Writing (Review & Editing). Schadendorf Resources, Writing (Review & Editing). Sondermann Resources, Writing (Review & Editing). Goebeler Resources, Writing (Review & Editing). Schilling Resources, Writing (Review & Editing). Kather Writing (Review & Editing). Fröhling Writing (Review & Editing). Kaminski Resources, Writing (Review & Editing). Doppler Resources, Writing (Review & Editing). Bucher Writing (Review & Editing). Brinker Resources, Project administration, Supervision, Writing (Review & Editing), Funding acquisition. This research is part of the doctoral thesis of Haggenmüller.

**Conflicts of interest statement**

Utikal is on the advisory board or has received honoraria and travel support from Amgen, Bristol Myers Squibb, GSK, Immunocore, LeoPharma, Merck Sharp and Dohme, Novartis, Pierre Fabre, Roche, Sanofi outside the submitted work. Meier has received travel support or/and speaker's fees or/and advisor's honoraria by Novartis, Roche, BMS, MSD and Pierre Fabre and research funding from Novartis and Roche. Hobelsberger reports clinical trial support from Almirall and speaker’s honoraria from Almirall, UCB and AbbVie and has received travel support from the following companies: UCB, Janssen Cilag, Almirall, Novartis, Lilly, LEO Pharma and AbbVie outside the submitted work. Gellrich has received travel support or/and speaker’s fees or/and advisor’s honoraria by Sun Pharma, Sanofi, Merck. Schlaak has received consultant or speaker fees or travel grants from BMS, MSD, Roche, Kyowa Kirin, Novartis, Sanofi Genzyme, Pierre Fabre, Sun Pharma, Immunocore. Erdmann declares honoraria and

travel support from Bristol-Meyers Squibb, Immunocore, and Novartis outside the submitted work. Haferkamp reports advisory roles for or has received honoraria from Pierre Fabre Pharmaceuticals, Novartis, Roche, BMS, Amgen and MSD outside the submitted work. Drexler has received honoraria from Pierre Fabre Pharmaceuticals and Novartis outside the submitted work. Hauschild reports clinical trial support, speaker's honoraria, or consultancy fees from the following companies: Agenus, Amgen, BMS, Dermagnostix, Highlight Therapeutics, Immunocore, Incyte, IO Biotech, Merck-Pfizer, MSD, NercaCare, Novartis, Philogen, Pierre Fabre, Regeneron, Roche, Sanofi-Genzyme, Seagen, Sun Pharma and Xenthera, outside the submitted work. French is on the advisory board or has received consulting/speaker honoraria from for Galderma, Janssen, Leo Pharma, Eli Lilly, Almirall, Union Therapeutics, Regeneron, Novartis, Amgen, Abbvie, UCB, Biotest, and InflaRx. Sondermann reports grants, speaker's honoraria, or consultancy fees from medi GmbH Bayreuth, Abbvie, Almirall, Amgen, Bristol-Myers Squibb, Celgene, GSK, Janssen, LEO Pharma, Lilly, MSD, Novartis, Pfizer, Roche, Sanofi Genzyme and UCB outside the submitted work. Schilling reports advisory roles for or has received honoraria from Pierre Fabre Pharmaceuticals, Incyte, Novartis, Roche, BMS and MSD, research funding from BMS, Pierre Fabre Pharmaceuticals and MSD, and travel support from Novartis, Roche, BMS, Pierre Fabre Pharmaceuticals and Amgen, outside the submitted work. Goebeler 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. Kather reports consulting services for Owkin, France, Panakeia, UK, and DoMore Diagnostics, Norway and has received honoraria for lectures by MSD, Eisai, and Fresenius. Brinker reports owning a company that develops mobile apps (Smart Health Heidelberg GmbH, Handschuhsheimer Landstr. 9/1, 69120 Heidelberg). 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.

### 2.2.3 Supplementary materials

This paragraph contains the following sections:

1. **Supplementary methods** where we provide a detailed description of the questionnaire development, the survey structure, the participant recruitment, and the statistical analysis.
2. **Supplementary results** where we describe the participant characteristics, further details on the patients' and dermatologists' perspective, and limitations of our study.

**Supplementary methods**

**Questionnaire development**

We first conducted a literature review on existing research, followed by nine semistructured expert interviews with patients (n=3), board-certified dermatologists (n=3), and AI professionals in healthcare (n=3) to elicit in-depth information. Then, the questionnaire for our survey was drafted and tested by the interviewed dermatologists and AI professionals, as well as individuals without a professional background in AI (n=19) to ensure comprehensibility and consistency.

**Survey structure**

The first part of the survey assessed participants' general outlook on AI in skin cancer diagnostics. Then, an ACBC was integrated to obtain a comprehensive understanding of the criteria required by patients and dermatologists to accept AI-based assistance systems. Finally, participants' demographics and prior experience with AI research were assessed.

For the ACBC, five AI-features and corresponding options (see **Table 4**) were defined based on the literature review and semistructured interviews. To prevent unrealistic combinations, specific prohibitions were implemented (see **Supplementary Table 1**). First, in the build-your-own section, participants created their customized assistance system to familiarize themselves with the five AI-features and corresponding options. Next, in the screening section, AI-systems with specific feature combinations were presented, and participants were asked whether they would consider using these AI-systems (see **Supplementary Figure 2**). During this section, the software analyzed participants' answers and suggested must-have or unacceptable criteria (see **Supplementary Figure 3**). Finally, in the choice tournament section, participants were shown a set of differently configured AI-systems based on their previous responses, from which they had to choose their preferred version (see **Figure 2**).

**Supplementary Table 1. Prespecified prohibited combinations within the ACBC design.** To ensure that no unrealistic combinations were presented within the ACBC part of the survey study, certain combinations were defined that were not permitted to be combined.

| **Prohibited Combinations** |
|---|
| AI shows the criteria (e.g., color, color distribution) and image regions used to make the assessment. AND Neither the physician nor the patient is able to trace the AI assessment. |
| AI cannot display the image regions, but it displays which criteria (e.g., color, color distribution) were used to make the assessment. AND Neither the physician nor the patient is able to trace the AI assessment. |
| AI cannot display any criteria, but it shows which image regions were used to make the assessment. AND Neither the physician nor the patient is able to trace the AI assessment. |
| AI does not have to explain its assessment on a case-by-case basis. However, it could be proven during the clinical trial that the AI pays attention to biologically relevant structures. AND The physician and the patient are able to trace the AI assessment. |
| AI does not have to explain its assessment on a case-by-case basis. However, it could be proven during the clinical trial that the AI pays attention to biologically relevant structures. AND The physician is able to trace the AI assessment. |
| AI does not have to explain its assessment on a case-by-case basis. It could not be proven during the clinical trial that the AI pays attention to biologically relevant structures. AND The physician and the patient are able to trace the AI assessment. |
| AI does not have to explain its assessment on a case-by-case basis. It could not be proven during the clinical trial that the AI pays attention to biologically relevant structures. AND The physician is able to trace the AI assessment. |

**Supplementary Figure 2. Example screening section of the present ACBC study design.** The survey was conducted in German, and this example screening section was translated into English for this illustration.

Please assume that all displayed assistance systems have been quality tested, and indicate for each specific system whether it is a **possibility** for you or **not**.

(Page 1 of 8)

| | | |
|---|---|---|
| What should the AI be able to distinguish? | AI distinguishes between **melanomas** and **nevi**. | AI distinguishes among **melanomas**, one category for **other types of skin cancer** and one for **benign skin lesions**. |
| How should the AI assessment be integrated into routine diagnostics? | The **physician** first decides **independently** and obtains a **second opinion** from the AI **only in case of doubt**. | The **AI assessment** is always obtained **first**, and the **physician** makes his or her decision **based on** it. |
| Who should be able to trace the AI assessment? | The **physician** is able to trace the AI assessment. | The **physician** and the **patient** are able to trace the AI assessment. |
| Beyond what level of diagnostic accuracy should AI be used? | AI performs **better** than the average dermatologist. | AI performs **worse** than the average dermatologist. |
| To what extent should the AI be able to explain its assessment? | AI cannot display the image regions, but it displays which **criteria** (e.g., color, color distribution) were used to make the assessmet.  **melanoma** irregularly colored | AI cannot display the image regions, but it displays which **criteria** (e.g., color, color distribution) were used to make the assessmet.  **melanoma** irregularly colored |
| | ◯ A possibility<br>◯ Not a possibility | ◯ A possibility<br>◯ Not a possibility |

**Supplementary Figure 3. Example must-have screener of the present ACBC study design.** The survey was conducted in German, and this example must-have screener was translated into English for this illustration.

**We do not want to jump to conclusions, however, we noticed that you selected AI-based assistance systems with the options listed below.**

**Would one of these options be an absolute must-have criterion for you to use an AI system for your skin cancer screening?**

- ◯ AI shows the **criteria** (e.g., color, color distribution) **and** **image regions** used to make the assessment.
- ◯ AI distinguishes between **benign** and **malignant** skin lesions but gives no indication of a precise diagnosis.
- ◯ The **physician** first decides **independently** and then **always** obtains a **second opinion** from the AI.
- ◯ The **physician** and the **patient** are able to trace the AI assessment.

- ◯ None of the options shown is an **absolute must-have** criterion.

## Participant recruitment and data collection

We conducted an anonymous online survey in German using Sawtooth SSI Web Lighthouse Studio 9.14.0 between May 06, 2022, and January 24, 2023. Individuals were prospectively enrolled at eight German university clinics and one private dermatology practice. Additionally, dermatologists were invited via institutional email accounts, and the survey was sent to melanoma support groups. Only patients who reported

suspicion of skin cancer within the last decade and dermatologists who participated in skin cancer screening were eligible for this study. All participants agreed to the analysis and publication of the anonymous data.

### Data analysis and statistics

To evaluate participants' general attitudes toward AI-systems in skin cancer screening, descriptive analysis was conducted. Ninety-five percent CIs were calculated, and two-sided chi-square tests were applied. A significance level of $p<0.05$ was set for all analyses. Statistical analysis was performed using SPSS, version 29.0.0.0 (IBM Corporation).

ACBC data were analyzed using hierarchical Bayes estimation and results were expressed in terms of counts, importance values, and utilities. Count analysis examined how often certain options were defined as unacceptable or must-have criteria. To determine the relevance of individual AI-features in decision-making, average importances were calculated by converting the utility ranges of each AI-feature (i.e., the difference between the perceived utility of the options that were regarded as most useful and least useful) to percentages using the following equation: feature importance (%) = (utility range of the specific feature/sum of the utility ranges of all features) × 100. Part-worth estimation was conducted to identify the options that were preferred most for each AI-feature. ACBC analysis was performed using Sawtooth SSI Web Lighthouse Studio 9.14.0.

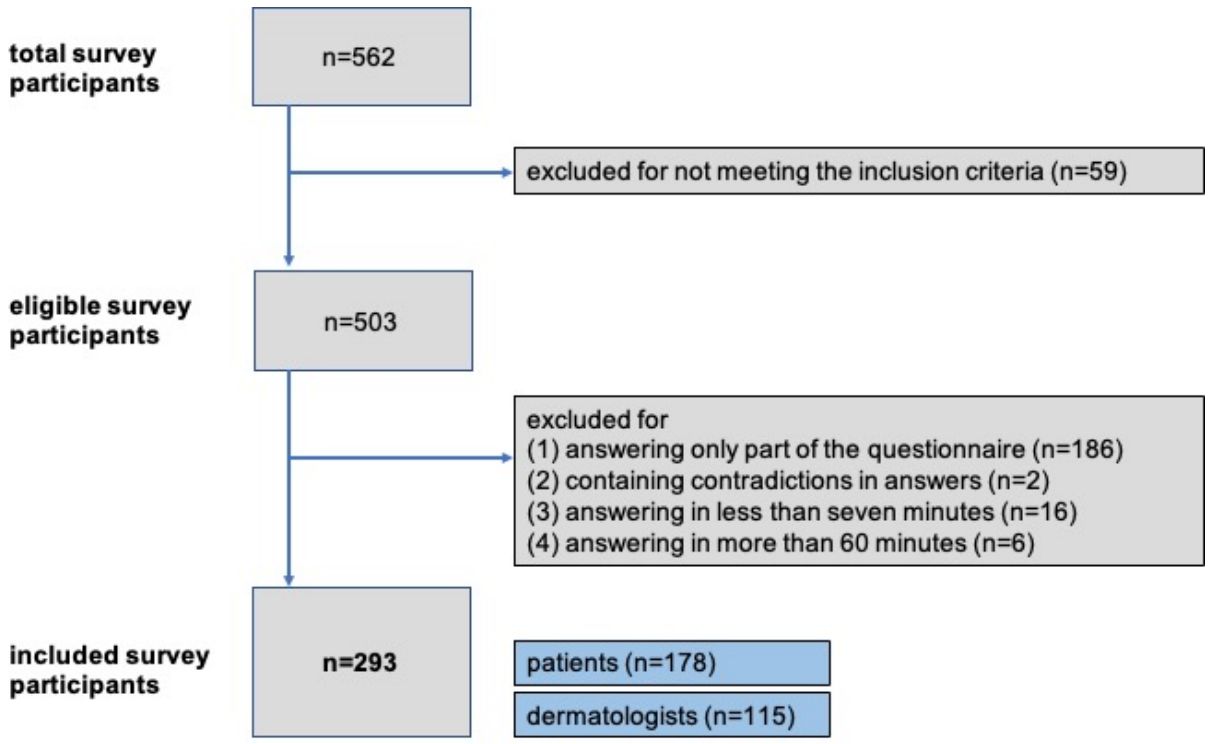
## Supplementary results

### Participant characteristics

After quality control (see **Supplementary Figure 4**), a validated dataset (N=293) of responses from 178 patients and 115 dermatologists remained (see **Supplementary**

**Table 2** for baseline characteristics). The included patients and dermatologists were predominantly female (62.9% and 62.6%, respectively) and a substantial number of participants indicated prior experience with AI research (28.1% and 50.4%, respectively). The majority of patients had received a histopathologically confirmed skin cancer diagnosis (69.1%) with MM being predominant (98/133; 79.7%). The patients' and dermatologists' median age was 53 years (range: 18-94 years) and 37 years (range: 25-60 years), respectively, and the dermatologists' median clinical experience was 7 years (range: 1-35 years).

**Supplementary Figure 4. CONSORT flow diagram of the quality control process.** From the 562 survey participants, we first disqualified participants who did not meet the inclusion criteria (n=59). Next, we excluded individuals who answered only part of the questionnaire (n=186) and those whose answers contained contradictions (e.g., inconsistencies between the physician's age and experience; n=2). In addition, using outlier detection (two times the standard deviation), we excluded participants who answered in less than seven minutes (n=16) or more than 60 minutes (n=6). Ultimately, a validated data set (N=293) consisting of responses from 178 patients and 115 dermatologists remained for analysis.



**Supplementary Table 2**. **Baseline characteristics of the included patients and dermatologists.**

| | | **Patients** | **Dermatologists** |
|---|---|---|---|
| Sociodemographic characteristics | | n=178 | n=115 |
| **Approximate Age (years)** | | Values, n (%) | Values, n (%) |
| | ≤ 30 | 9 (5.1%) | 24 (20.9%) |
| | 31 to 40 | 20 (11.2%) | 45 (39.1%) |
| | 41 to 50 | 35 (19.7%) | 28 (24.3%) |
| | 51 to 60 | 62 (34.8%) | 15 (13.0%) |
| | > 60 | 44 (24.7%) | 0 (0.0%) |
| | Unknown | 8 (4.5%) | 3 (2.6%) |
| **Gender** | | | |
| | Female | 112 (62.9%) | 72 (62.6%) |
| | Male | 65 (36.5%) | 42 (36.5%) |
| | Unknown | 1 (0.6%) | 1 (0.9%) |
| **Residence** | | | |
| | Big city (>100,000 inhabitants) | 68 (38.2%) | 98 (85.2%) |
| | Small town (10,000-100,000 inhabitants) | 55 (30.9%) | 13 (11.3%) |
| | Rural area (<10,000 inhabitants) | 54 (30.3%) | 3 (2.6%) |
| | Unknown | 1 (0.6%) | 1 (0.9%) |
| **Participation in AI Research** | | | |
| | Yes | 50 (28.1%) | 58 (50.4%) |
| | No | 103 (57.9%) | 51 (44.3%) |
| | Unknown | 25 (14.0%) | 6 (5.2%) |
| **Clinical Experience (years)** | | | |
| | <5 years | | 36 (31.3%) |
| | 5 to 14 years | | 42 (36.5%) |
| | 15 to 24 years | | 23 (20.0%) |
| | >24 years | | 14 (12.2%) |
| **Clinical Workplace** | | | |
| | University hospital | | 95 (82.6%) |
| | Private practice | | 10 (8.7%) |

| | | | |
|---|---|---|---|
| | Other | | 9 (7.8%) |
| | Unknown | | 1 (0.9%) |
| **Prevention Behavior (Skin Cancer Screening)** | | | |
| | More than once a year | 118 (66.3%) | |
| | Once a year | 45 (25.3%) | |
| | Less than once a year | 15 (8.4%) | |
| **(Previous) Diagnosis** | | multiple response* n=178, m=192 | |
| | No skin cancer | 49 (27.5%) | |
| | Skin cancer | 123 (69.1%) | |
| | Diagnosis pending | 13 (7.3%) | |
| | Unknown | 7 (3.9%) | |
| **Type of Skin Cancer** | | multiple response* n=123, m=133 | |
| | Melanoma | 98 (79.7%) | |
| | Other types of skin cancer (BCC, SCC, ...) | 31 (25.2%) | |
| | Unknown | 4 (3.2%) | |

*Respondents were allowed to check more than one answer option for this survey question (n=total number of respondents, m=total number of answers).

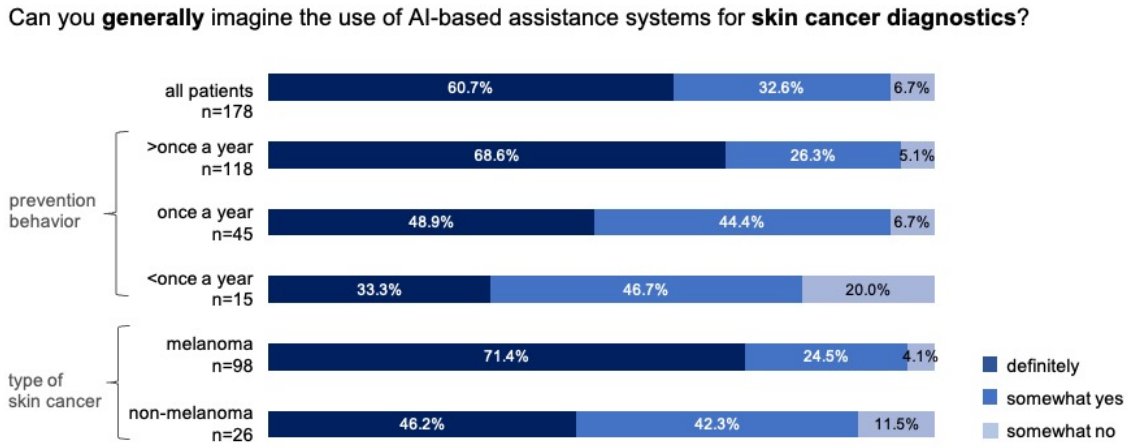
### Patients' perspectives

### Patients showed positive general attitudes toward AI-systems for skin cancer diagnostics

To investigate preferences among patients, we first aimed to determine their general attitudes toward AI-systems for skin cancer diagnostics. In this context, patients showed positive general attitudes toward AI-based assistance systems (108/178 could definitely imagine AI usage; 60.7%; 95% CI: 53.1-67.90%), albeit with significant differences by prevention behavior and skin cancer type (see **Supplementary Figure 5**). Among the included patients (n=178), 108 (60.7%; 95% CI: 53.1-67.90%) could definitely imagine and 58 could somewhat (32.6%; 95% CI: 25.8-40.0%) imagine AI usage for skin cancer diagnostics. Another 12 patients expressed hesitation ("somewhat no",

6.7%; 95% CI: 3.5-11.5%). Interestingly, none of the participants explicitly ruled out AI usage altogether.

**Supplementary Figure 5. Bar chart depicting patients' attitudes toward the use of AI-based assistance systems for skin cancer diagnostics.** Participants' general outlook on using AI-based assistance systems for skin cancer diagnostics was identified by asking, "Can you generally imagine the use of AI-based assistance systems for skin cancer diagnostics?" with response options of "definitely," "somewhat yes," "somewhat no," and "definitely not".




Subgroup analysis revealed significant differences based on patients' prevention behavior (p=.010). While 68.6% of the patients who reported undergoing skin cancer screening more than once a year (81/118; 95% CI: 59.5-76.9%) said they would definitely use AI-based assistance systems, only 48.9% of the patients who reported undergoing skin cancer screening once a year (22/45; 95% CI: 33.7-64.2%) and 33.3% of the patients who reported undergoing such screening less than once a year (5/15; 95% CI: 11.8-61.6%) said they would do so. Additionally, among the subgroup of confirmed cancer patients (123/178), the type of skin cancer had a significant effect on attitudes toward the use of AI-based assistance systems (p=.042), with patients diagnosed with MM being more open to the use.

**Patients rejected non-traceable and non-explainable AI-systems**

Next, we investigated must-have and unacceptable criteria. Seventy-four patients (41.6%; 95% CI: 34.3-49.2%) stated that they would not rely on an AI-system if neither the dermatologist nor the patient could trace how the AI made its decisions. In line with this, the majority of patients systematically ruled out AI-systems that did not provide some sort of case-by-case explanations. This was true regardless of whether the clinical trial could prove that the AI accounted for biologically relevant structures (40/178 patients, 22.5%; 95% CI: 16.6-29.3%) or not (65/178 patients; 36.5%; 95% CI: 29.4-44.1%). Consequently, patients rejected non-traceable and non-explainable AI-based assistance systems.

**Patients prioritized explainability in decision-making**

We also used hierarchical Bayes estimation to investigate the relative importance of different AI-features. For patients, the features explainability (i.e., the extent to which the AI can explain its assessment; 26.6%; 95% CI: 26.0-27.3%) and diagnostic accuracy (20.6%; 95% CI: 19.3-22.0%), on average, had the greatest relative importance (see **Supplementary Table 3**), closely followed by traceability (i.e., who should be able to trace the AI assessment; 19.8%; 95% CI: 18.9-20.6%) and AI integration (i.e., how the AI assessment is integrated into decision-making; 18.7%; 95% CI: 17.5-19.9%). Only the input data (14.3%; 95% CI: 13.4-15.2%), on average, played a relatively minor role (see **Supplementary Table 3**).

**Supplementary Table 3**. Average relative importance of individual AI-features for patients' decision-making for or against the use of AI-based assistance systems according to hierarchical Bayes estimation of ACBC data.

| **Patients** n=178 | **Average Importance (in Percent)*** | **Standard Deviation** | **95% CIs** |
|---|---|---|---|
| Explainability | 26.6 | 4.7 | 26.0-27.3 |
| Diagnostic accuracy | 20.6 | 9.3 | 19.3-22.0 |
| Traceability | 19.8 | 5.9 | 18.9-20.6 |
| Integration | 18.7 | 8.3 | 17.5-19.9 |
| Input data | 14.3 | 6.3 | 13.4-15.2 |

*Relative importance is ratio-scaled; i.e., an AI-feature indicating an importance of ten percent is twice as important for the decision-making process as an AI-feature with an importance of five percent.

Note: Since ACBC can reflect the importance only in terms of the relative utility of the AI-features tested within the ACBC design; comparison between patients' and dermatologists' absolute importance values is not possible.

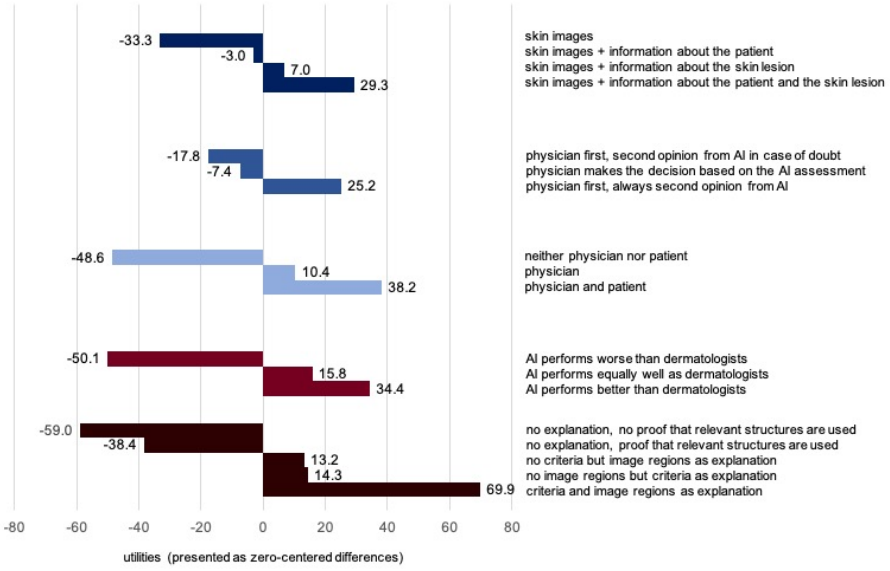
However, subgroup analysis showed that patients with prior AI research experience (103/178) attached significantly greater importance to the amount of input data used for the AI decision (prior experience 16.8%; 95% CI: 14.9-18.7; no experience 13.5%; 95% CI: 12.3-14.7%). Furthermore, AI explainability played a significantly greater role for this subgroup (prior experience 29.5%; 95% CI: 27.7-31.3%; no experience 26.3%; 95% CI: 25.5-27.1%), while the diagnostic accuracy was less important (prior experience 16.0%; 95% CI: 13.5-18.4%; no experience 21.8%; 95% CI: 20.2-23.4%).

### Patients' preferred options for individual AI-features

To determine patients' preferred options for each AI-feature, we performed part-worth estimates. We found that patients favored AI-systems where the explainability feature indicated not only the decision criteria, such as color distribution, but also the relevant image regions (see **Supplementary Figure 6**). Moreover, patients preferred AI-systems with decision-making that was traceable to them and their physician. Regarding

the way and extent to which AI was integrated into clinical routine, patients strongly preferred the unlimited use of AI for every suspicious skin lesion (i.e., not only in case of doubt), while simultaneously valuing independent decision-making between physicians and the assistance system. Additionally, patients favored AI-systems that incorporated patient- and lesion-specific metadata in their assessment.

**Supplementary Figure 6. Patients' perceived utilities of the individual options for each AI-feature under investigation.** Utilities are presented as zero-centered differences, a normalized approach where all utilities of one AI-feature are scaled to sum to zero. Negative utility does not imply that an option provides no or negative perceived utility for patients; rather, it indicates that an option is less preferred than an option indicating positive utility. Therefore, comparison of utilities is possible within the corresponding AI-feature but not across AI-features.



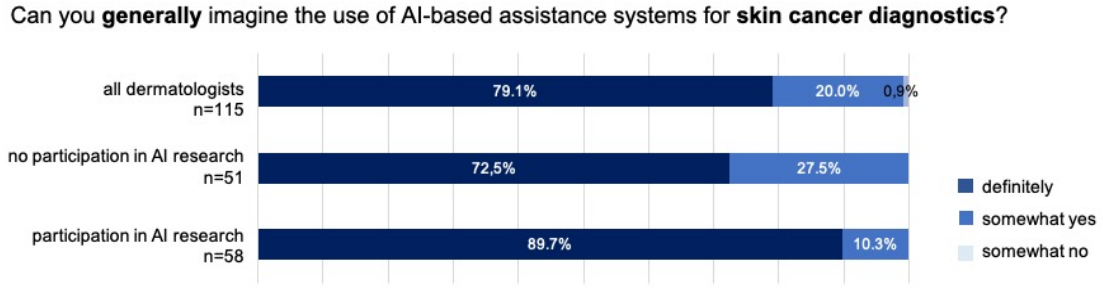

## Dermatologists' perspectives

### Dermatologists showed positive general attitudes toward AI-systems for skin cancer diagnostics

To investigate preferences regarding the use of AI among dermatologists, we first aimed to determine their general attitudes toward AI-systems for skin cancer screening. Overall, dermatologists showed positive general attitudes toward AI-based assistance systems for skin cancer diagnostics (91/115 could definitely imagine AI usages;

79.1%; 95% CI: 70.6-86.2%), albeit with significant differences by prior experience with AI research (see **Supplementary Figure 7).**

**Supplementary Figure 7. Bar chart depicting dermatologists' attitudes toward the use of AI-based assistance systems for skin cancer diagnostics.** Participants' general outlook on using AI-based assistance systems for skin cancer diagnostics was identified by asking, "Can you generally imagine the use of AI-based assistance systems for skin cancer diagnostics?" with response options of "definitely," "somewhat yes," "somewhat no," and "definitely not".




Among the included dermatologists (n=115), 91 participants (79.1%; 95% CI: 70.6-86.2%) could definitely imagine the use of AI-systems for skin cancer diagnostics. Another 23 participants (20.0%; 95% CI: 13.1-28.5%) indicated that they could somewhat imagine using such AI-systems, and one participant explicitly ruled it out.

In addition, subgroup analysis showed that previous participation in AI research had a positive significant effect on attitudes toward the use of AI-systems (p=.026).

**Dermatologists rejected non-traceable and non-explainable AI-systems**

In addition to examining dermatologists' general attitudes, we investigated their must-have and unacceptable criteria. Fifty-five dermatologists (47.8%; 95% CI: 38.4-57.3%) generally rejected AI-systems if neither dermatologist nor patient could understand how the decision was made. Additionally, a large proportion of dermatologists ruled out AI-systems that failed to explain their decision on a case-by-case basis regardless

of whether the clinical trial showed that the AI considers relevant biological aspects (judged unacceptable by 18/115 dermatologists, 15.7%; 95% CI: 9.6-23.6%) or not (41/115 dermatologists, 35.7%; 95% CI: 26.9-45.1%). Consequently, similar to patients, dermatologists rejected AI-based assistance systems that lacked traceability and explainability.

**Dermatologist prioritized diagnostic accuracy in decision-making**

We used hierarchical Bayes estimation to investigate the relative importance of different AI-features for dermatologists. The features of diagnostic accuracy (33.2%; 95% CI: 19.4-21.4%) and explainability (i.e., the extent to which the AI can explain its assessment; 20.4%; 95% CI: 19.4-21.4%), on average, had the greatest relative importance in decision-making. AI integration (i.e., how the AI assessment is integrated into the decision-making; 16.7%; 95% CI: 15.2-18.2%), the decision task (i.e., what the AI is able to distinguish; 15.5%; 95% CI: 14.4-16.5%) and traceability (i.e., who should be able to trace the AI assessment; 14.3%; 95% CI: 13.2-15.4%), on average, played a somewhat comparable yet subordinate role (see **Supplementary Table 4**).

**Supplementary Table 4**. Average relative importance of individual AI-features for dermatologists' decision-making for or against the use of AI-based assistance systems according to hierarchical Bayes estimation of ACBC data.

| **Dermatologists** n=115 | **Average Importance (in Percent)*** | **Standard Deviation** | **95% CIs** |
|---|---|---|---|
| Diagnostic accuracy | 33.2 | 11.1 | 31.1-35.2 |
| Explainability | 20.4 | 5.5 | 19.4-21.4 |
| Integration | 16.7 | 8.4 | 15.2-18.2 |
| Decision task | 15.5 | 5.9 | 14.4-16.5 |
| Traceability | 14.3 | 5.9 | 13.2-15.4 |

*Relative importance is ratio-scaled; i.e., an AI-feature indicating an importance of ten percent is twice as important for the decision-making process as an AI-feature with an importance of five percent.

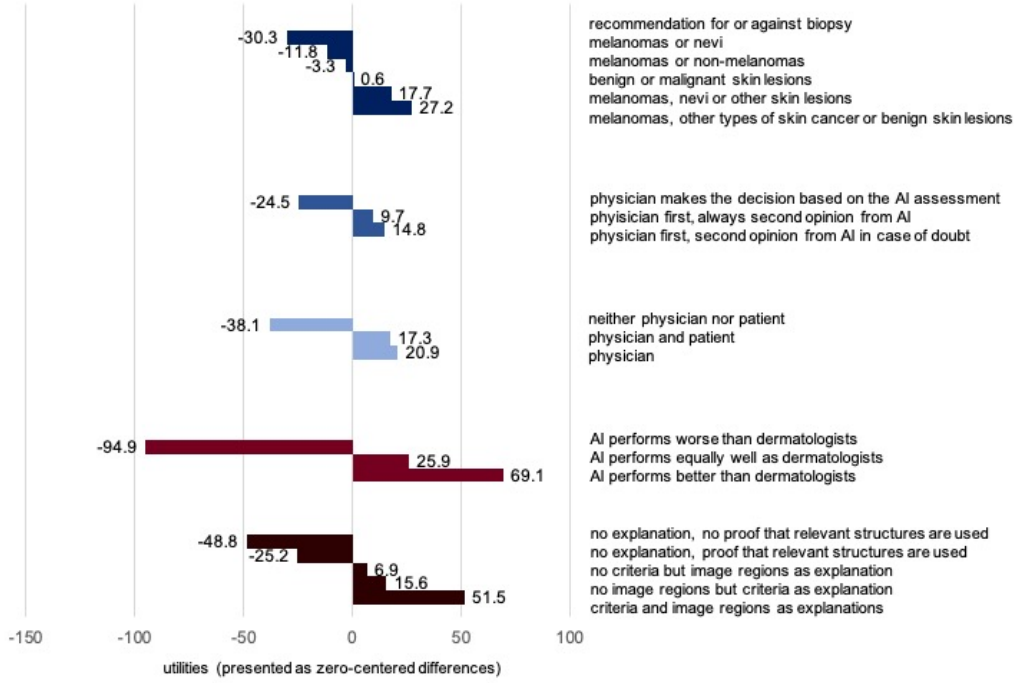
**Note:** Since ACBC can reflect the importance only in terms of the relative utility of the AI-features tested within the ACBC design; comparison between patients' and dermatologists' absolute importance values is not possible.

Subgroup analysis showed that explainability had greater importance in decision-making for women than for men (women 22.2%; 95% CI: 21.0-23.3%, men 18.4%; 95% CI: 16.5-20.2%). Furthermore, dermatologists with previous AI research experience considered the integration of AI into routine diagnostics as significantly more relevant than those with no prior involvement (prior experience 19.2%; 95% CI: 16.8-21.6%; no experience 13.7%; 95% CI: 11.7-15.6%).

**Dermatologists' preferred options of individual AI-features**

To further examine the subsample of dermatologists, we performed part-worth estimates to determine the perceived utilities of the corresponding options for each AI-feature (**Supplementary Figure 8**). The dermatologists preferred AI-systems that not only showed their decision criteria, such as color distribution, but also indicated the relevant image regions. Furthermore, dermatologists prioritized limited integration of AI into their clinical routine, i.e., to obtain a second opinion only for uncertain cases. Moreover, dermatologists preferred support systems that discriminated among melanomas, other types of skin cancer and benign skin lesions, closely followed by AI-systems that differentiated among melanomas, nevi and other diagnoses. Additionally, dermatologists prioritized AI-systems that enabled only them to trace the decision process as opposed to the option where both patient and clinician could understand it.

**Supplementary Figure 8. Dermatologists' perceived utilities for each AI-feature under investigation.** Utilities are presented as zero-centered differences, a normalized approach where all utilities of one AI-feature are scaled to sum to zero. Negative utility does not imply that an option provides no or negative perceived utility for patients; rather, it indicates that the option is less preferred than an option with positive utility. Therefore, comparison of utilities is possible within the corresponding AI-feature but not across AI-features.




## Limitations

Our study mainly recruited patients and dermatologists from German university hospitals; in addition, a substantial number of respondents indicated prior experience with AI research, and most of the included patients had a personal history of melanoma. Therefore, our results may not be fully generalizable. To overcome this limitation, future studies may involve multiple private dermatology practices, intensify recruitment of respondents with no experience with AI research, and patients with other types of skin cancer. Moreover, relative importances in ACBC designs are influenced by the range of options available for each AI-feature and the total number of AI-features being evaluated, which can never be exhaustive (e.g., adding a new option within an AI-feature could alter the importance of all other features).

## 2.3 Publication 3: Comparison of federated learning for decentralized artificial intelligence in melanoma diagnostics

Sarah Haggenmüller[†a], MSc; Max Schmitt[†a], MSc; Eva Krieghoff-Henning[a], PhD; Achim Hekler[a], MSc; Roman C. Maron[a], MSc; Christoph Wies[a], MSc; Jochen S. Utikal[b,c,d], MD; Friedegund Meier[e], MD; Sarah Hobelsberger[e], MD; Frank F. Gellrich[e], MD; Mildred Sergon[e], MD; Axel Hauschild[f], MD; Lars E. French[g,h], MD; Lucie Heinzerling[g,j], MD; Justin G. Schlager[g], MD; Kamran Ghoreschi[i], MD; Max Schlaak[i], MD; Franz J. Hilke[i], PhD; Gabriela Poch[i], MD; Sören Korsing[i], MD; Carola Berking[j], MD; Markus V. Heppt[j], MD; Michael Erdmann[j], MD; Sebastian Haferkamp[k], MD; Konstantin Drexler[k], MD; Dirk Schadendorf[l], MD; Wiebke Sondermann[l], MD; Matthias Goebeler[m], MD; Bastian Schilling[m], MD; Jakob N. Kather[n], MD; Stefan Fröhling[o], MD; Titus J. Brinker[*a], MD

[†]These authors contributed equally to this work.

[a] Digital Biomarkers for Oncology Group, National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Heidelberg, Germany

[b] Department of Dermatology, Venereology and Allergology, University Medical Center Mannheim, Ruprecht-Karls University of Heidelberg, Mannheim, Germany

[c] Skin Cancer Unit, German Cancer Research Center (DKFZ), Heidelberg, Germany

[d] DKFZ Hector Cancer Institute at the University Medical Center Mannheim, Mannheim, Germany

[e] Skin Cancer Center at the University Cancer Center and National Center for Tumor Diseases Dresden, Department of Dermatology, University Hospital Carl Gustav Carus, Technische Universität Dresden, Germany

[f] Department of Dermatology, University Hospital (UKSH), Kiel, Germany

[g] Department of Dermatology and Allergy, University Hospital, LMU Munich, Munich, Germany

h Dr. Phillip Frost Department of Dermatology and Cutaneous Surgery, University of Miami, Miller School of Medicine, Miami, FL, USA

[i] Department of Dermatology, Venereology and Allergology, Charité – Universitätsmedizin Berlin, Corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[j] Department of Dermatology, University Hospital Erlangen, Comprehensive Cancer Center Erlangen – European Metropolitan Region Nürnberg, CCC Alliance WERA, Erlangen, Germany

[k] Department of Dermatology, University Hospital Regensburg, Regensburg, Germany

[l] Department of Dermatology, Venereology and Allergology, University Hospital Essen, Essen, Germany

[m] Department of Dermatology, Venereology and Allergology, University Hospital Würzburg and National Center for Tumor Diseases (NCT) WERA Würzburg, Germany

[n] Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany

[o] Department of Translational Medical Oncology, National Center for Tumor Diseases (NCT) Heidelberg and 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 460, Heidelberg, 69120, Germany. E-mail address: titus.brinker@dkfz.de (T.J. Brinker).

### 2.3.1 Abstract

**Importance:** The development of AI-based melanoma classifiers typically calls for large centralized datasets, requiring hospitals to give away their patient data, which raises serious privacy concerns. To address this concern, decentralized federated learning (FL) has been proposed, where classifier development is distributed across hospitals.

**Objective:** To investigate whether a more privacy-preserving FL can achieve comparable diagnostic performance to a classical centralized (i.e., single-model) and ensemble learning approach for AI-based melanoma diagnostics.

**Design:** We developed a FL model for melanoma-nevus classification using histopathological whole-slide images prospectively acquired at six German university hospitals between April 2021 and February 2023, and benchmarked it using both a holdout and external test dataset.

**Setting:** A multicentric, single-arm study was conducted.

**Participants:** The study included 1025 whole-slide images of clinically melanoma-suspicious skin lesions from 923 patients, consisting of 388 histopathologically-confirmed invasive melanomas and 637 nevi.

**Exposures:** All whole-slide images were retrospectively analyzed by an AI-based classifier without influencing routine clinical care.

**Main Outcome and Measure(s):** The area under the receiver operating characteristic curve (AUROC) served as the primary endpoint for evaluating the diagnostic performance. Secondary endpoints included balanced accuracy, sensitivity and specificity.

**Results:** The federated approach (0.8579; 95% CI: 0.7693-0.9299) performed significantly worse than the classical centralized approach (0.9024; 95% CI: 0.8379-0.9565) in terms of AUROC on a holdout test dataset (pairwise Wilcoxon signed-rank, $P<.001$) but performed significantly better (0.9126; 95% CI: 0.8810-0.9412) than the classical centralized approach (0.9045; 95% CI: 0.8701-0.9331) on an external test dataset (pairwise Wilcoxon signed-rank, $P<.001$). Notably, the federated approach performed significantly worse than the ensemble approach on both the holdout (0.8867; 95% CI: 0.8103-0.9481) and external test dataset (0.9227; 95% CI: 0.8941-0.9479).

**Conclusions and Relevance:** The findings suggest that FL is a viable approach for the binary classification of invasive melanoma and nevi on a real-world distributed dataset. FL can improve privacy protection in AI-based melanoma diagnostics while simultaneously promoting collaboration across institutions and countries. Moreover, it may have the potential to be extended to other image classification tasks in digital cancer histopathology and beyond.

### 2.3.2 Introduction

CNNs – deep neural networks most commonly applied to image classification – have shown promise in improving diagnostic accuracy for various diseases [66–68],

including MM [17,20,27,31]. MM is the leading cause of death among skin cancer worldwide [1]. Early-stage detection increases the survival chances of affected patients significantly but is challenging due to frequent morphological overlap between MM and atypical nevi [13,14]. In experimental settings, CNNs have achieved performance on par or even superior to that of human experts for both dermatological [28,30,53,55] and histopathological [33,34] classification tasks. These results suggest that AI has the potential to revolutionize the diagnosis of melanoma in offering more accurate detection.

Nonetheless, AI models are highly data dependent, meaning that their performance correlates with the size and diverseness of the training set. The more diverse data an AI model is trained on, the more likely it is to perform well [69–71]. Therefore, to develop AI algorithms, patient data are typically transferred to one site for training and testing and stored in a centralized way (known as classical centralized learning). However, in the medical field, ensuring patient data confidentiality is of utmost importance; consequently, sharing patient data is heavily regulated. Thus, the transfer of patient data to an external facility to generate the envisaged algorithms can raise serious privacy concerns. Alternatively, institutions can use their own data and computing power to develop separate AI algorithms, whose decisions are subsequently merged into one (known as ensemble learning). However, clinical settings often face computational resource constraints, making it challenging to run complex ensemble models in real-time. These framework conditions pose difficulties for collaboration and data collection, particularly in multicenter studies or international research collaborations.

To address these challenges, new approaches, such as FL [72,73], have been developed to enable the decentralized training of AI algorithms using data kept at their origin,

while requiring less computational power on site. FL involves each institution training its own model with its own data, while communication and aggregation are executed by a central coordinator.

Previous studies have examined the use of FL in diagnosing melanoma [74,75] and other medical applications [76–79]. While *Bdair et al.* [74] and *Agbley et al.* [75] have demonstrated the promise of FL for classifying retrospective melanoma data, none has evaluated FL leveraging prospective collected real-world distributed melanoma data nor externally validated the performance of the proposed classifiers. These gaps in existing literature highlight the need for further research to explore the effectiveness of FL for melanoma diagnostics when leveraging prospective data and to assess the generalizability of the respective classifiers.

Therefore, we developed a model using a decentralized FL approach for the binary classification of invasive melanoma (IM) and nevi based on histopathological whole-slide images, and directly compared it retrospectively with the classical centralized and ensemble learning on both a holdout and external test dataset, using prospectively collected real-world distributed data from six German university hospitals.

Our findings highlight that FL represents a reliable alternative, particularly when leveraging external data. By providing a more accessible and privacy-preserving alternative that empowers institutions to contribute to the development of AI models, even with relatively small datasets or strict data protection rules, FL holds the potential to reshape AI-based melanoma diagnostics and may extend to other classification tasks in digital cancer histopathology and beyond.

### 2.3.3 Methods

**Ethics statement and reporting standards**

Ethics approval was obtained from the ethics committee at the technical university of Dresden, the Friedrich-Alexander University Erlangen-Nuremberg, the LMU Munich, the university of Regensburg and the university hospital Wuerzburg. Patients provided informed written consent. This work was performed in accordance with the Declaration of Helsinki. The Standards for Reporting of Diagnostic Accuracy (STARD 2015) were followed for the reporting of this study (see **Tabular Appendix 6**) [80].

**Patient cohorts and slide acquisition**

Hematoxylin-eosin-stained reference slides of skin lesions were prospectively acquired at six German university hospitals (Berlin, Dresden, Erlangen, Munich, Regensburg, Wuerzburg) between April 2021 and February 2023. Study participants had to be at least 18 years old and were required to have clinically melanoma-suspicious skin lesions. Lesions were not allowed to be pre-biopsied nor located under the finger-/toenails. Diagnostic labels were histopathologically-confirmed by at least one reference dermatopathologist at the corresponding hospital as part of routine clinical practice. In collision cases involving multiple tumors, the label of the larger tumor region was assigned. Only histopathologically-confirmed IM and nevi were eligible for this study.

**Whole-slide image preprocessing**

A Leica Aperio AT2 DX was used to digitize the hematoxylin-eosin-stained reference slides of all enrolled patients at 40x magnification, producing whole-slide images with a resolution of 0.25 µm/px to generate patches for training and testing. After manually annotating the area of the epidermis (MS), the region of interest was tessellated into downscaled square patches. Each patch had a uniform edge length of 224 px,

corresponding to 103.04 µm. Whole-slide image annotation and tessellation were performed using QuPath 0.2.3 [81,82]. Additionally, blur detection was implemented with custom code written in Python 3.7.0. A patch was classified as blurry if it had a Laplacian below a manually set threshold of 510 and subsequently discarded.

**Model development**

ResNet18 pretrained on ImageNet was used to train one model with FL, one with centralized learning and one with ensemble learning. A small architecture was used to limit training and inference time and streamline the experimental procedures. The tree-structured Parzen estimator [83] was used to choose the hyperparameters to maximize the AUROC at lesion level for a validation set. For each approach, the learning rate, number of training epochs, amount of data used in one epoch per whole-slide image and, for FL specifically, the frequency of weight exchange were tuned for an equal number of optimization steps using the Python library Optuna [82]. During this process, 30% of the training data served as validation set and the training followed Leslie Smith's 'one cycle policy', which involves training the model with a gradually increasing learning rate for the first half of the training cycle, followed by a gradual decrease in the learning rate for the second half [84]. During inference, the confidence value of every patch of a whole-slide image was interpreted as the probability for classification as IM or nevus. The average of these probabilities was the final probability for each whole-slide image.

For the FL approach, each hospital's model was trained for a certain time interval with the same hyperparameters. The time interval was based on a synchronization factor which was tuned during training and was proportional to the size of the dataset of the respective hospital. After each interval, model weights were collected and merged into

a new model using a weighted average. The assigned weights were proportional to the amount of data available during training. Subsequently, the new model was (re)distributed to every hospital to continue training. Since communication between the participants in this approach was not the focus, this process was only simulated on one computational unit.

For the centralized approaches, the model $H_{full}$ represents the model that was trained using data from hospitals 1 to 5. The remaining five models (models $H_1$, $H_2$, $H_3$, $H_4$, and $H_{5)}$ were trained by excluding the data of hospitals 1, 2, 3, 4, or 5, respectively.

For the ensemble approach, five classifiers were trained separately using only one of the five training sets from hospitals 1 to 5 with individual hyperparameters. For inference, each model computed a probability for a given input. All five probabilities were subsequently averaged to calculate the final prediction.

Training and inference were implemented in Python 3.7.0 using PyTorch 1.13.0 [85] and fastai 2.7.10 [86].

**Statistical analysis**

Two-sided chi-square tests were employed to identify significant differences between the training and test datasets. The AUROC served as the primary endpoint for evaluating the performance of the developed models. Secondary endpoints included balanced accuracy, sensitivity and specificity. The mean values of the corresponding metrics were calculated using 1,000 iterations of bootstrapping to reduce the impact of stochastic events. 95% CIs were calculated using the nonparametric percentile method [87]. For statistical comparisons of the AUROCs, pairwise two-sided Wilcoxon signed-rank tests were applied. A significance level of $P<.05$ was set for all analyses.

Significance levels were adjusted to .025 (m=2) or .01 (m=5) according to Bonferroni correction in case of multiple tests. Statistical analysis was performed in SPSS 29.0.0.0.

### 2.3.4 Results

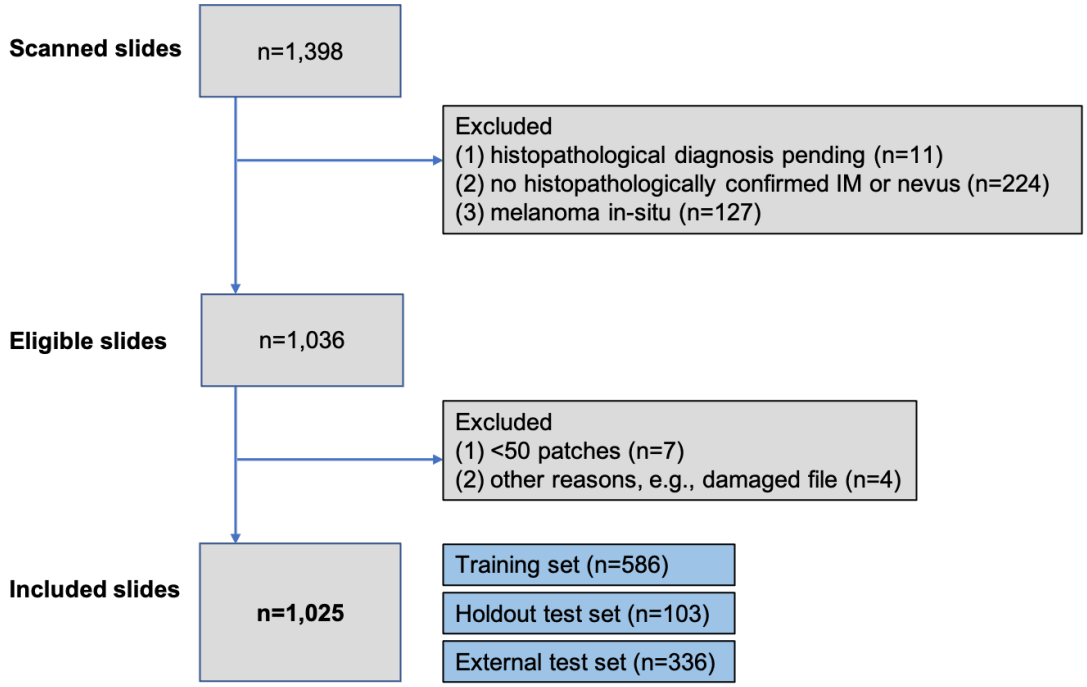
**Number of eligible slides and patients**

A total of 1025 slides from 923 patients, consisting of 388 IM and 637 nevi, were included in the analysis (see **Table 5**). A further 373 slides were excluded for not meeting the predefined inclusion criteria of this study (e.g., in-situ tumors, see **Figure 3)**. A total of 548,755 patches were derived from the eligible slides (296,141/252,614 – IM/nevus) for training and testing purposes (see **Supplementary Table 5)**.

**Table 5. Characteristics of the study sample.**

| Hospital | Slides (Patients) | Invasive melanomas | Nevi |
|---|---|---|---|
| Hospital 1 | 71 (62) | 19 | 52 |
| Hospital 2 | 97 (86) | 56 | 41 |
| Hospital 3 | 107 (103) | 59 | 48 |
| Hospital 4 | 178 (157) | 37 | 141 |
| Hospital 5 | 236 (215) | 75 | 161 |
| Hospital 6 | 336 (300) | 142 | 194 |
| **Total** | **1025 (923)** | **388** | **637** |

**Figure 3. Flowchart of the slide inclusion process.** Slides were excluded from the analysis if there was no histopathologically-confirmed label available (n=11) or if the lesion proved to be neither IM nor nevus (in situ tumors; n=127 or other diagnoses, e.g., basal cell carcinoma, squamous cell carcinoma; n=224). In addition, slides that exhibited <50 epidermal patches (n=7) or other technical issues (n=4) were removed. IM: invasive melanoma




## Patient characteristics and differences among datasets

The eligible cases in the training set (data from hospitals 1 to 5) and the holdout test dataset (data from hospitals 1 to 5) exhibited significant differences in lesion subtype and American Joint Committee on Cancer (AJCC) stage when compared to the external test dataset (data from hospital 6; $P$<.001). However, no significant differences were observed in lesion localization, age or Breslow thickness. The median age at diagnosis was 58 years (range 18 to 95) for the training set, 57 years (range 18 to 93) for the holdout and 61 years (range 18 to 95) for the external test dataset; the median Breslow thickness was 0.70 mm (range 0.10 to 34.00), 0.70 mm (range 0.20 to 14.40) and 0.80 mm (range 0.30 to 20.00), respectively. Thus, the training and holdout test

dataset were considered to be differently distributed than the external one. Patient characteristics of the study sample are presented in **Table 6**.

**Table 6. Patient characteristics of the study sample.** IM: invasive melanoma

| | Training set (hospitals 1 to 5) | | Holdout test dataset (hospitals 1 to 5) | | External test dataset (hospital 6) | |
|---|---|---|---|---|---|---|
| | IM n=209 | Nevus n=377 | IM n=37 | Nevus n=66 | IM n=142 | Nevus n=194 |
| Age at diagnosis (years) | | | | | | |
| <35 | 5 (2.4%) | 75 (19.9%) | 1 (2.7%) | 16 (24.2%) | 4 (2.8%) | 51 (26.3%) |
| 35-54 | 45 (21.5%) | 129 (34.2%) | 8 (21.6%) | 19 (28.8%) | 19 (13.4%) | 67 (34.5%) |
| 55-74 | 84 (40.2%) | 124 (32.9%) | 17 (45.9%) | 22 (33.3%) | 58 (40.8%) | 48 (24.7%) |
| >74 | 74 (35.4%) | 49 (13.0%) | 11 (29.7%) | 9 (13.6%) | 61 (43.0%) | 28 (14.4%) |
| Unknown | 1 (0.5%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) |
| Lesion localization | | | | | | |
| Palms/soles | 1 (0.5%) | 6 (1.6%) | 1 (2.7%) | 3 (4.5%) | 4 (2.8%) | 5 (2.6%) |
| Face/scalp/neck | 43 (20.6%) | 17 (4.5%) | 8 (21.6%) | 4 (6.1%) | 24 (16.9%) | 26 (13.4%) |
| Upper extremities | 37 (17.7%) | 38 (10.1%) | 5 (13.5%) | 9 (13.6%) | 18 (12.7%) | 13 (6.7%) |
| Lower extremities | 45 (21.5%) | 78 (20.7%) | 8 (21.6%) | 13 (19.7%) | 29 (20.4%) | 34 (17.5%) |
| Back | 54 (25.8%) | 134 (35.5%) | 8 (21.6%) | 18 (27.3%) | 43 (30.3%) | 59 (30.4%) |
| Abdomen | 13 (6.2%) | 48 (12.7%) | 3 (8.1%) | 9 (13.6%) | 9 (6.3%) | 29 (14.9%) |
| Chest | 12 (5.7%) | 37 (9.8%) | 2 (5.4%) | 8 (12.1%) | 10 (7.0%) | 16 (8.2%) |
| Buttock | 2 (1.0%) | 10 (2.7%) | 1 (2.7%) | 2 (3.0%) | 1 (0.7%) | 5 (2.6%) |
| Genitalia | 1 (0.5%) | 5 (1.3%) | 1 (2.7%) | 0 (0.0%) | 1 (0.7%) | 3 (1.5%) |
| Unknown | 1 (0.5%) | 4 (1.1%) | 0 (0.0%) | 0 (0.0%) | 3 (2.1%) | 4 (2.1%) |
| Lesion subtype | | | | | | |
| Superficial spreading melanoma | 142 (67.9%) | | 24 (64.9%) | | 35 (24.6%) | |
| Nodular melanoma | 25 (12.0%) | | 4 (10.8%) | | 20 (14.1%) | |
| Lentigo maligna melanoma | 29 (13.9%) | | 5 (13.5%) | | 9 (6.3%) | |
| Acral lentiginous melanoma | 8 (3.8%) | | 2 (5.4%) | | 6 (4.2%) | |
| Desmoplastic melanoma | 0 (0.0%) | | 0 (0.0%) | | 2 (1.4%) | |
| Spitzoid melanoma | 1 (0.5%) | | 1 (2.7%) | | 0 (0.0%) | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | Other types of IM/ combined forms of IM/ subtype unknown | 4 (1.9%) | | 1 (2.7%) | | 70 (49.3%) | |
| | Spitz nevus and variants | | 6 (1.6%) | | 0 (0.0%) | | 4 (2.1%) |
| | Dysplastic nevus/Clark nevus | | 155 (41.1%) | | 30 (45.5%) | | 110 (56.7%) |
| | Acral nevus | | 7 (1.9%) | | 4 (6.1%) | | 12 (6.2%) |
| | Recurrent nevus | | 1 (0.3%) | | 0 (0.0%) | | 1 (0.5%) |
| | Blue nevus | | 21 (5.6%) | | 3 (4.5%) | | 6 (3.1%) |
| | Other types of nevi/ combined forms of nevi/ subtype unknown | | 187 (49.6%) | | 29 (43.9%) | | 61 (31.4%) |
| AJCC stage[a] | | | | | | | |
| | IA | 87 (41.6%) | | 13 (35.1%) | | 70 (49.3%) | |
| | IB | 23 (11.0%) | | 8 (21.6%) | | 30 (21.1%) | |
| | IIA | 13 (6.2%) | | 3 (8.1%) | | 6 (4.2%) | |
| | IIB | 7 (3.3%) | | 0 (0.0%) | | 14 (9.9%) | |
| | IIC | 7 (3.3%) | | 3 (8.1%) | | 7 (4.9%) | |
| | IIIA | 4 (1.9%) | | 0 (0.0%) | | 3 (2.1%) | |
| | IIIB | 4 (1.9%) | | 0 (0.0%) | | 5 (3.5%) | |
| | IIIC | 12 (5.7%) | | 1 (2.7%) | | 6 (4.2%) | |
| | IV | 2 (1.0%) | | 1 (2.7%) | | 1 (0.7%) | |
| | Unknown | 50 (23.9%) | | 8 (21.6%) | | 0 (0.0%) | |
| Breslow thickness[b] | | | | | | | |
| | ≤ 1.00 mm (T1) | 126 (60.3%) | | 23 (62.1%) | | 89 (62.7%) | |
| | 1.01 to 2.00 mm (T2) | 25 (12.0%) | | 6 (16.2%) | | 16 (11.3%) | |
| | 2.01 to 4.00 mm (T3) | 27 (12.9%) | | 1 (2.7%) | | 19 (13.4%) | |
| | > 4.00 m (T4) | 23 (11.0%) | | 6 (16.2%) | | 17 (12.0%) | |
| | Unknown | 8 (3.8%) | | 1 (2.7%) | | 1 (0.7%) | |

[a]AJCC staging constitutes the gold standard for histopathological reporting of IM.

[b]Breslow thickness describes the extent of anatomic spread and serves as an important prognostic factor for IM.

**Comparison of federated learning with other approaches**

To compare the performance of FL, a total of 586 lesions (209 IM, 377 nevi) derived from five hospitals were used to train three distinct models (see **Supplementary Figure 9**): first, the federated approach, where a model was built through decentralized training of individual models that were merged at regular intervals [88]; second, the centralized approach ($H_{full}$), where a model was built using all available data on a centralized server [89]; and third, the ensemble approach, where a model was built for each participating hospital, and the results of all models were aggregated into one final prediction [90]. A randomly sampled holdout test dataset – from the same hospitals already involved in model training – consisting of 103 lesions (37 IM, 66 nevi) and an external test dataset – from another hospital not involved in model training – consisting of 336 lesions (142 IM, 194 nevi) were used to evaluate the performances of the approaches.

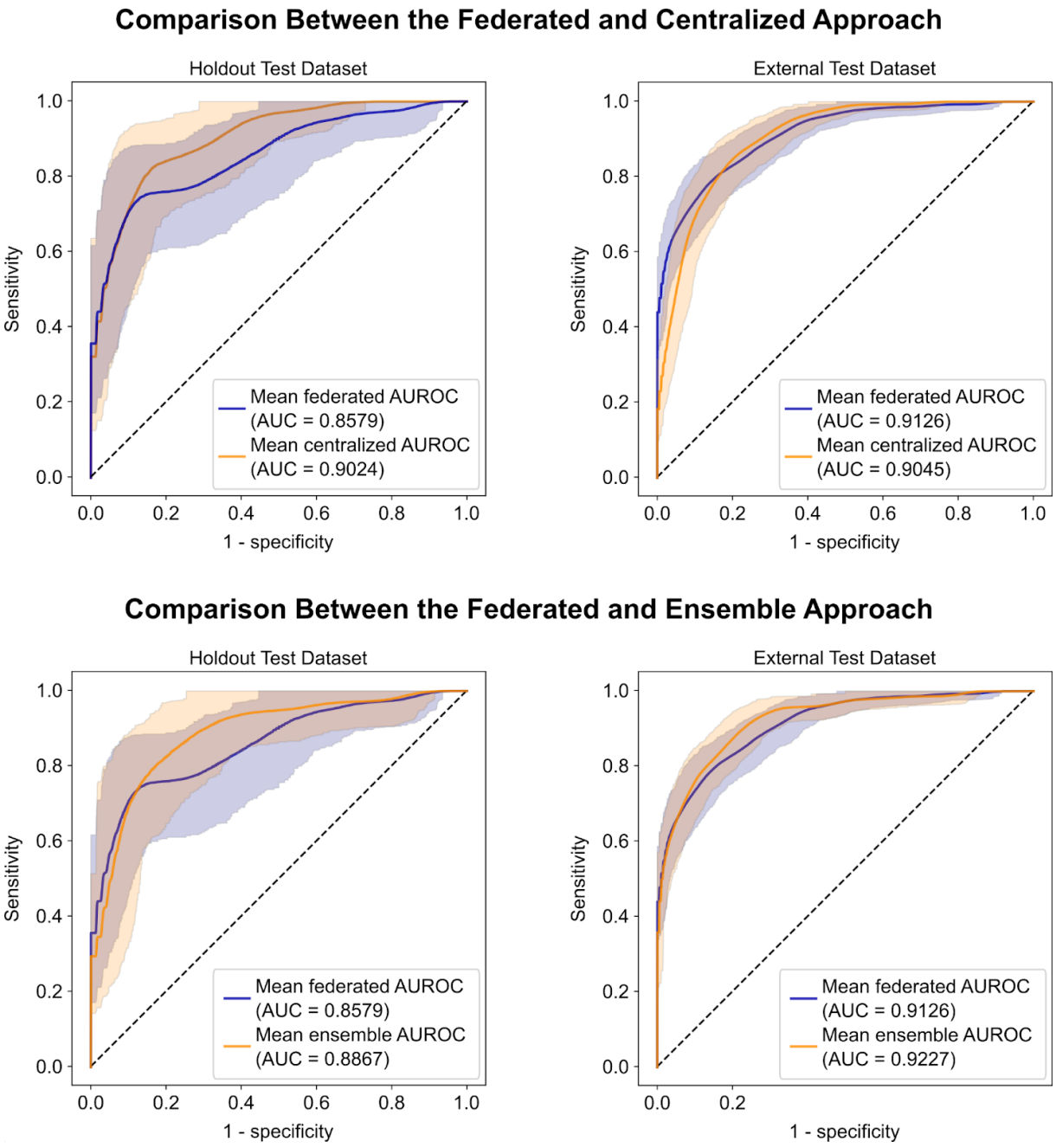
**Federated learning performs the worst on the holdout test dataset**

On the holdout test dataset, FL performed the worst (see **Table 7**), with a AUROC of 0.8579 (95% CI: 0.7693-0.9299, see **Figure 4**), followed by the ensemble approach with a mean AUROC of 0.8867 (95% CI: 0.8103-0.9481). The centralized approach (model $H_{full}$) performed best, with a mean AUROC of 0.9024 (95% CI: 0.8379-0.9565). The results indicate that on the holdout test dataset, the classical centralized model performed significantly better than the federated and ensemble approaches in terms of AUROC (pairwise Wilcoxon signed-rank, $P<.001$). For a detailed overview of the confusion matrices on the holdout test dataset, see **Supplementary Figure 10**.

**Federated learning outperforms classical centralized learning on the external test dataset**

On the external test dataset, a different ranking was observed (see **Table 7**). The centralized approach (model $H_{full}$) performed the worst, achieving a mean AUROC of 0.9045 (95% CI: 0.8701-0.9331), while FL demonstrated a mean AUROC of 0.9126 (95% CI: 0.8810-0.9412, see **Figure 4**). The ensemble approach performed best on the external test dataset, with a mean AUROC of 0.9227 (95% CI: 0.8941-0.9479). Altogether, on the external test dataset the federated approach yielded significantly better results than the centralized model in terms of AUROC (pairwise Wilcoxon signed-rank, $P<.001$). Notably, both the FL and centralized models performed significantly worse than the ensemble approach (pairwise Wilcoxon signed-rank, $P<.001$). For a detailed overview of the confusion matrices on the external test dataset, see **Supplementary Figure 11**.

**Figure 4. Mean AUROCs of the three investigated approaches.** Mean AUROCs on the holdout and external test dataset after 1000 iterations of bootstrapping including the corresponding 95% CIs (orange- and violet-colored areas) are illustrated for the FL and the centralized approach (model $H_{full}$) on the top and for the FL and the ensemble approach on the bottom. AUROC: area under the receiver operating characteristic curve




## Comparison of federated learning with a more realistic centralized approach

Furthermore, the classical centralized approach was subjected to retraining using several smaller datasets (models $H_1$, $H_2$, $H_3$, $H_4$ and $H_5$), for comparison with the original federated approach, which was trained with all available training data. This comparison

was conducted to investigate whether FL would achieve at least comparable results to centralized approaches when it had access to more data (ranging from 71 to 236 more cases). Thereby, we explored the feasibility of potential future clinical FL application scenarios where hospitals might be more willing to participate in the development and refinement of a classifier when no patient data needs to be transferred to an external institution.

After retraining, the centralized approach maintained its superiority on the holdout test dataset in terms of AUROC (see **Table 7**), regardless of which hospital was omitted for classifier training (models $H_1$, $H_2$, $H_3$, $H_4$ and $H_5$; pairwise Wilcoxon signed-rank, $P$<.001). However, on the external test dataset, the model developed with the FL approach held its performance advantage (see **Table 7**) over all five centralized models developed using smaller datasets (pairwise Wilcoxon signed-rank, $P$<.001). These results suggest that a surplus of training data does not necessarily result in superior classification performance for FL.

**Table 7. Performance metrics of the different classification approaches on the holdout and external test datasets (top).** Performance metrics of the original federated approach and all five retrained "leave-one-hospital-out" approaches on the holdout and external test datasets (bottom). AUROC: area under the receiver operating characteristic curve

| **Performance metrics of the different classification approaches** | | | | |
|---|---|---|---|---|
| **Holdout** | **AUROC (95% CIs)** | **Balanced accuracy (95% CIs)** | **Sensitivity (95% CIs)** | **Specificity (95% CIs)** |
| FL model | 0.8579 (0.7693-0.9299) | 76.76% (67.70-84.89%) | 59.54% (42.86-75.00%) | 93.99% (87.84-98.55%) |
| Ensemble model | 0.8867 (0.8103-0.9481) | 81.46% (73.10-88.94%) | 84.02% (70.59-94.59%) | 78.89% (68.57-88.06%) |
| Centralized model | 0.9024 (0.8379-0.9565) | 85.23% (77.30-92.31%) | 83.91% (70.97-94.59%) | 86.55% (77.46-93.94%) |
| **External** | **AUROC (95 % CIs)** | **Balanced accuracy (95 % CIs)** | **Sensitivity (95 % CIs)** | **Specificity (95 % CIs)** |
| FL model | 0.9126 (0.8810-0.9412) | 81.73% (77.36-85.77%) | 80.92% (74.21-86.90%) | 82.54% (77.07-87.92%) |

| | | | | |
|---|---|---|---|---|
| Ensemble model | 0.9227 (0.8941-0.9479) | 76.47% (72.69-80.48%) | 95.79% (92.19-98.65%) | 57.16% (50.51-63.96%) |
| Centralized model | 0.9045 (0.8701-0.9331) | 80.56% (76.71-84.38%) | 93.66% (89.21-97.22%) | 67.46% (60.87-74.05%) |
| **Performance metrics of the original federated approach and all five retrained "leave-one-hospital-out" approaches** | | | | |
| Holdout | AUROC (95% CIs) | Balanced accuracy (95% CIs) | Sensitivity (95% CIs) | Specificity (95% CIs) |
| FL model | 0.8579 (0.7693-0.9299) | 76.76% (67.70-84.89%) | 59.54% (42.86-75.00%) | 93.99% (87.84-98.55%) |
| model H1 | 0.9139 (0.8508-0.9648) | 79.30% (70.90-87.40%) | 67.59% (52.63-82.50%) | 91.02% (83.33-97.06%) |
| model H2 | 0.8874 (0.8041-0.9529) | 82.76% (74.68-90.05%) | 72.91% (57.89-86.67%) | 92.61% (86.15-98.41%) |
| model H3 | 0.8675 (0.7879-0.9337) | 74.15% (65.63-82.90%) | 54.23% (37.50-70.97%) | 94.06% (87.67-98.59%) |
| model H4 | 0.8851 (0.8099-0.9511) | 81.55% (73.26-89.44%) | 81.19% (68.29-93.55%) | 81.91% (72.06-90.77%) |
| model H5 | 0.8710 (0.7961-0.9401) | 84.10% (75.96-91.18%) | 89.24% (78.38-97.50%) | 78.95% (68.75-88.06%) |
| External | AUROC (95% CIs) | Balanced accuracy (95% CIs) | Sensitivity (95% CIs) | Specificity (95% CIs) |
| FL model | 0.9126 (0.8810-0.9412) | 81.73% (77.36-85.77%) | 80.92% (74.21-86.90%) | 82.54% (77.07-87.92%) |
| model H1 | 0.8868 (0.8517-0.9207) | 76.90% (72.60-80.99%) | 89.49% (84.09-94.24%) | 64.31% (57.43-70.77%) |
| model H2 | 0.8941 (0.8585-0.9252) | 79.69% (75.59-83.84%) | 89.43% (84.29-93.92%) | 69.95% (63.37-76.22%) |
| model H3 | 0.8831 (0.8465-0.9172) | 78.82% (74.30-82.76%) | 88.66% (82.99-93.48%) | 68.99% (62.43-75.13%) |
| model H4 | 0.8670 (0.8281-0.9020) | 76.29% (71.84-80.39%) | 86.61% (81.21-91.88%) | 65.97% (59.28-72.77%) |
| model H5 | 0.8296 (0.7837-0.8698) | 72.39% (67.77-76.60%) | 88.78% (83.45-93.63%) | 55.99% (49.46-62.78%) |

### 2.3.5 Discussion

In this study, we aimed to develop and externally validate a decentralized trained FL model for melanoma-nevus classification using histopathological whole-slide images. Additionally, we directly compared FL with classical centralized and ensemble learning that are commonly applied for melanoma classification tasks. In this context, FL achieved a mean AUROC of 0.8579 (95% CI: 0.7693-0.9299) on the holdout test dataset and 0.9126 (95% CI: 0.8810-0.9412) on the external test dataset, thus representing a reliable alternative.

The utilized datasets encompassed a comprehensive representation of the IM cases encountered in day-to-day clinical care due to the prospective and consecutive data collection from multiple centers. By avoiding selection bias that may have arisen in previous melanoma classification studies that applied FL but collected data retrospectively [74,75], we minimized the risk of over-/underestimating the performance of the compared classifiers. A strength of our study is the long-tailed distribution of localizations and IM subtypes (including rare subtypes such as spitzoid melanoma), and all possible AJCC stages and Breslow thickness categories [91]. Training the model on such a heterogeneous dataset that captures the complexity of real-world IM data enables the model to effectively recognize lesions of different types, severity levels, and depths, and allows the model to learn spatial patterns and specific characteristics associated with diverse body regions. This enhances its overall generalizability, ultimately leading to robust performance.

Overall, the classical centralized model ($H_{full}$) significantly outperformed FL on the holdout test dataset (i.e., tested on data from hospitals involved in model training) in terms of AUROC (0.9024 versus 0.8579), while FL performed significantly better (0.9126 versus 0.9045) on the external test dataset (i.e., on data from a hospital not involved in model training). The findings demonstrate that FL techniques may not be as well suited to solve in-distribution classification problems (i.e., same distribution as the training data), as indicated by the inferior performance on the holdout test dataset. On the other hand, they show that FL may provide additional advantages in terms of out-of-distribution generalizability, as indicated by the enhanced performance on the external test datasets (similar observations see [72,77]). The observed superior performance on the external test set could be due to the FL model not fully converging during training, possibly introducing a slight regularization effect. This phenomenon of non-

convergence is frequently encountered in FL due to the challenging task of training on data from different distributions [92].

While the observed differences between FL and the centralized approach may not be large in absolute terms, they are consistent over 1,000 iterations of bootstrapping (i.e., paired data comparisons), thereby demonstrating a sustained outperformance of the centralized approach. Despite the comparatively lower statistical power of the Wilcoxon signed-rank test, this marginal yet persistent performance improvement is clinically highly relevant, as any melanoma misclassifications can lead to fatal outcomes.

Despite these positive findings, the ensemble approach continued to outperform FL and the classical centralized approach in terms of AUROC (0.9227 versus 0.9126 and 0.9045, respectively). Nevertheless, an ensemble approach poses extensive challenges for the explainability of the results, since understanding multiple sets of model weights is more difficult than dealing with one set in the FL approach. This is particularly relevant given the legislative requirement that medical devices must be explainable to a certain extent [65] as well as its substantial influence on patients' and physicians' acceptance [93].

Although the whole-slide images were digitized using the same slide scanner (Leica Aperio AT2 DX), heterogeneity was ensured by different staining and cutting protocols of the participating hospitals. While the labels for this study were established based on the gold standard of care (i.e., histopathological verification), caution should be exercised in interpreting the results as previous studies observed a discordance between pathologists of up to 25% in classifying melanoma [13,14]. Future studies may involve the integration of independent pathologist panels or epigenetic analyses (e.g., methylation analyses) to further reduce interrater variability.

### 2.3.6 Conclusion

The results of this study demonstrate that FL can achieve a comparable performance to that of classical centralized or ensemble approaches, making it a reliable alternative for the classification of IM and nevi. Additionally, FL empowers institutions to contribute to the development of AI models, even with relatively small datasets or strict data protection rules, thereby fostering collaboration across institutions and countries. Moreover, FL may have the potential to be further extended to other image classification tasks in digital cancer histopathology and beyond. Future research could build on this work by assessing its effectiveness with different types of medical images (e.g., dermoscopic or hyperspectral images), evaluating its feasibility for diagnosing various types of cancer, and investigating its effectiveness using technically different (e.g., attention-based methods) AI models. In our ongoing research, we are exploring the scalability of FL for refined diagnostic tasks by incorporating in-situ tumors as a clinically highly-relevant but separate classification class.

### 2.3.7 Acknowledgement

**Role of the funding source**

This study was funded by the Federal Ministry of Health, Berlin, Germany (grants: Skin Classification Project 2 (SCP2) and Tumor Behavior Prediction Initiative (TPI); grant holder in both cases: Titus J. Brinker, German Cancer Research Center, Heidelberg, Germany) and the Studienstiftung des deutschen Volkes, Bonn, Germany (scholarship holder: Sarah Haggenmüller). The sponsors 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.

**Author contribution statement**

Haggenmüller Conceptualization, Methodology, Investigation, Formal analysis, Data Curation, Writing - original draft, Visualization, Ethics Approval. Schmitt Methodology, Software, Validation, Writing (original draft), Visualization. Krieghoff-Henning Conceptualization, Writing (Review & Editing), Project administration, Supervision. Hekler Validation, Writing (Review & Editing), Project administration, Supervision. Maron Validation, Writing (Review & Editing). Wies Validation, Writing (Review & Editing). Utikal Resources, Writing (Review & Editing). Meier Resources, Writing (Review & Editing). Hobelsberger Resources, Writing (Review & Editing). Gellrich Resources, Writing (Review & Editing). Sergon Resources, Writing (Review & Editing). Hauschild Writing (Review & Editing). French Resources, Writing (Review & Editing). Heinzerling Resources, Writing (Review & Editing). Schlager Resources, Writing (Review & Editing). Ghoreschi Resources, Writing (Review & Editing). Schlaak Resources, Writing (Review & Editing). Hilke Resources, Writing (Review & Editing). Poch Resources, Writing (Review & Editing). Korsing Resources, Writing (Review & Editing). Berking Resources, Writing (Review & Editing). Heppt Resources, Writing (Review & Editing). Erdmann Resources, Writing (Review & Editing). Haferkamp Resources, Writing (Review & Editing). Drexler Resources, Writing (Review & Editing). Schadendorf Resources, Writing (Review & Editing). Sondermann Resources, Writing (Review & Editing). Goebeler Resources, Writing (Review & Editing). Schilling Resources, Writing (Review & Editing). Kather Writing (Review & Editing). Fröhling Writing (Review & Editing). Brinker Conceptualization, Project administration, Supervision, Writing (Review & Editing), Funding acquisition. Brinker had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. This research is part of the doctoral thesis of Haggenmüller.

**Conflict of interest statement**

Utikal is on the advisory board or has received honoraria and travel support from Amgen, Bristol Myers Squibb, GSK, Immunocore, LeoPharma, Merck Sharp and Dohme, Novartis, Pierre Fabre, Roche, Sanofi outside the submitted work. Meier has received travel support or/and speaker's fees or/and advisor's honoraria by Novartis, Roche, BMS, MSD and Pierre Fabre and research funding from Novartis and Roche. Hobelsberger reports clinical trial support from Almirall and speaker's honoraria from Almirall, UCB and AbbVie and has received travel support from the following companies: UCB, Janssen Cilag, Almirall, Novartis, Lilly, LEO Pharma and AbbVie outside the submitted work. Gellrich has received travel support or/and speaker's fees or/and advisor's honoraria by Sun Pharma, Sanofi, Merck. Schlaak has received consultant or speaker fees or travel grants from BMS, MSD, Roche, Kyowa Kirin, Novartis, Sanofi Genzyme, Pierre Fabre, Sun Pharma, Immunocore. Erdmann declares honoraria and travel support from Bristol-Meyers Squibb, Immunocore, and Novartis outside the submitted work. Haferkamp reports advisory roles for or has received honoraria from Pierre Fabre Pharmaceuticals, Novartis, Roche, BMS, Amgen and MSD outside the submitted work. Drexler has received honoraria from Pierre Fabre Pharmaceuticals and Novartis outside the submitted work. Hauschild reports clinical trial support, speaker's honoraria, or consultancy fees from the following companies: Agenus, Amgen, BMS, Dermagnostix, Highlight Therapeutics, Immunocore, Incyte, IO Biotech, Merck-Pfizer, MSD, NercaCare, Novartis, Philogen, Pierre Fabre, Regeneron, Roche, Sanofi-Genzyme, Seagen, Sun Pharma and Xenthera, outside the submitted work. French is on the advisory board or has received consulting/speaker honoraria from for Galderma, Janssen, Leo Pharma, Eli Lilly, Almirall, Union Therapeutics, Regeneron, Novartis, Amgen, Abbvie, UCB, Biotest, and InflaRx. Sondermann reports grants, speaker's honoraria, or consultancy fees from medi GmbH Bayreuth, Abbvie, Almirall, Amgen,

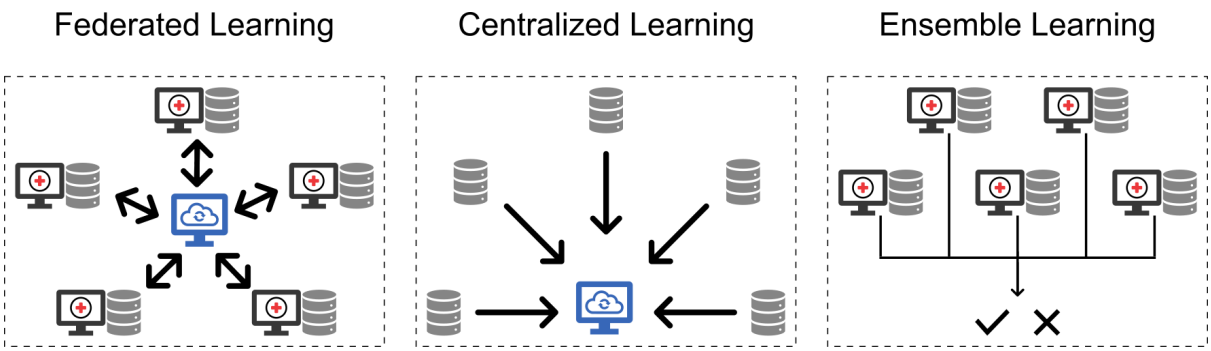
Bristol-Myers Squibb, Celgene, GSK, Janssen, LEO Pharma, Lilly, MSD, Novartis, Pfizer, Roche, Sanofi Genzyme and UCB outside the submitted work. Schilling reports advisory roles for or has received honoraria from Pierre Fabre Pharmaceuticals, Incyte, Novartis, Roche, BMS and MSD, research funding from BMS, Pierre Fabre Pharmaceuticals and MSD, and travel support from Novartis, Roche, BMS, Pierre Fabre Pharmaceuticals and Amgen, outside the submitted work. Goebeler 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. Kather reports consulting services for Owkin, France, Panakeia, UK, and DoMore Diagnostics, Norway and has received honoraria for lectures by MSD, Eisai, and Fresenius. Brinker reports owning a company that develops mobile apps (Smart Health Heidelberg GmbH, Handschuhsheimer Landstr. 9/1, 69120 Heidelberg). 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.

### 2.3.8 Supplementary materials

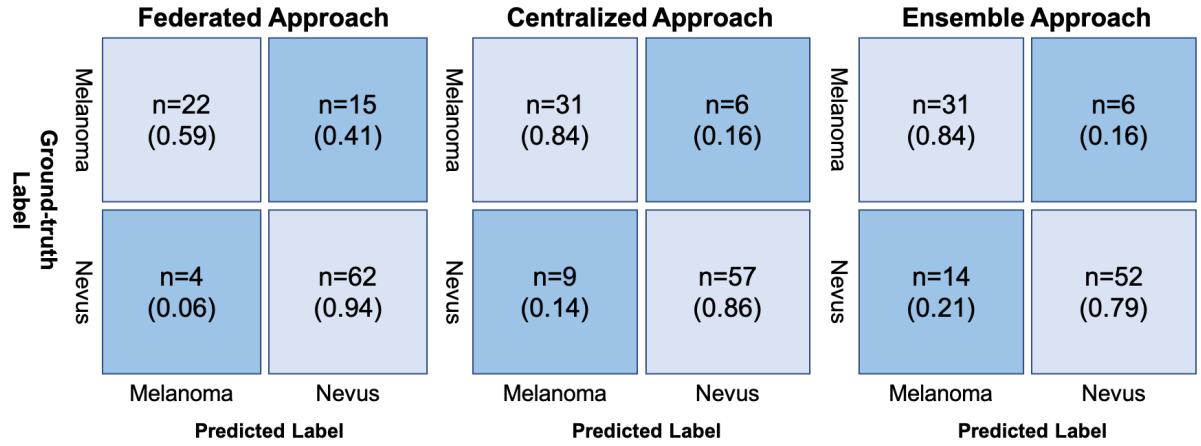
**Supplementary Table 5. Dataset characteristics at the patch level.**

| Hospital | Overall (#patches) | Melanoma (#patches) | Nevi (#patches) |
|---|---|---|---|
| Hospital 1 | 80706 | 30180 | 50526 |
| Hospital 2 | 32486 | 22910 | 9576 |
| Hospital 3 | 51919 | 35693 | 16226 |
| Hospital 4 | 85474 | 32226 | 53248 |
| Hospital 5 | 80655 | 33748 | 46907 |
| Hospital 6 | 217057 | 141384 | 75673 |
| **Total** | **548297** | **296141** | **252156** |

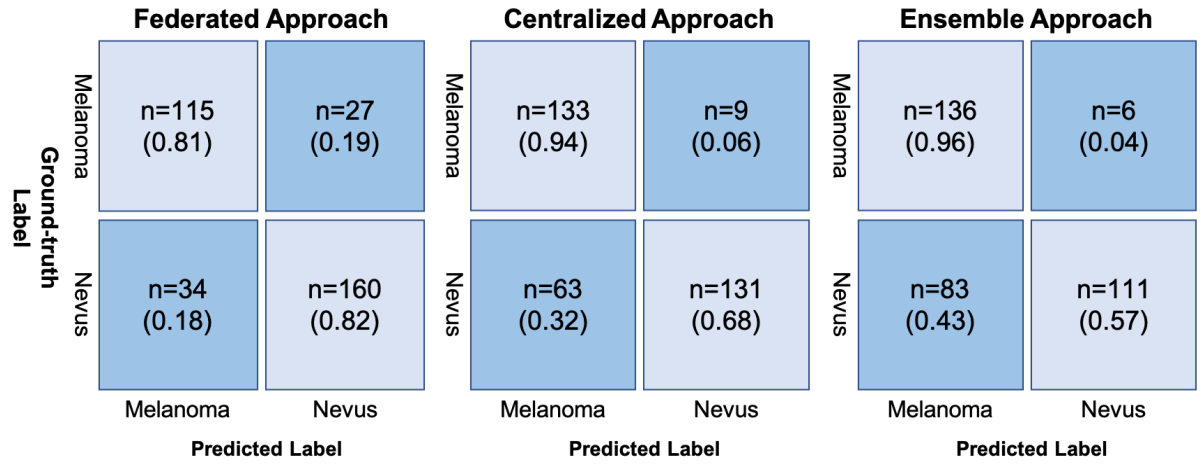
**Supplementary Figure 9. Workflow of the three implemented approaches.** On the left side, the federated approach is depicted, where every hospital (represented by the red +) has its own data (gray database) and computing power (monitors), but communication and aggregation are executed by a third party (blue monitor) that serves as a central coordinator. In the middle, the centralized approach is represented. In this case, the hospitals transfer their data to a third party, which uses it to train a centralized model. On the right side, the ensemble approach is depicted, where each hospital uses their own data and computing power to train a separate model. The decisions over all models are averaged to obtain a final prediction for a given image.




**Supplementary Figure 10. Confusion matrices of the three approaches on the holdout test dataset.** Distribution of correct and incorrect predictions on the holdout test dataset for the federated approach, the centralized approach $H_{ful}$ and the ensemble approach. The ground truth was determined by at least one reference dermatopathologist at the corresponding hospital as part of routine clinical practice.



**Supplementary Figure 11. Confusion matrices of the three approaches on the external test dataset.** Distribution of correct and incorrect predictions on the external test dataset for the federated approach, the centralized approach $H_{ful}$ and the ensemble approach. The ground truth was determined by at least one reference dermatopathologist at the corresponding hospital as part of routine clinical practice.



## 2.4 Supplementary publication 1: Robustness of convolutional neural networks in recognition of pigmented skin lesions

Roman C. Maron[a], Sarah Haggenmüller[a], Christof von Kalle[b], Jochen S. Utikal[c,d], Friedegund Meier[e], Frank F. Gellrich[e], Axel Hauschild[f], Lars E. French[g,p], Max Schlaak[g], Kamran Ghoreschi[h], Heinz Kutzner[i], Markus V. Heppt[j], Sebastian Haferkamp[k], Wiebke Sondermann[l], Dirk Schadendorf[l], Bastian Schilling[m], Achim Hekler[a], Eva Krieghoff-Henning[a], Jakob N. Kather[n], Stefan Fröhling[o], Daniel B. Lipka[o], Titus J. Brinker[a,*]

[a] Digital Biomarkers for Oncology Group, National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Heidelberg, Germany

[b] Department of Clinical-Translational Sciences, Charité – University Medicine and Berlin Institute of Health (BIH), Berlin, Germany

[c] Department of Dermatology, Heidelberg University, Mannheim, Germany

[d] Skin Cancer Unit, German Cancer Research Center (DKFZ), Heidelberg, Germany

[e] 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

[f] Department of Dermatology, University Hospital (UKSH), Kiel, Germany

[g] Department of Dermatology and Allergy, University Hospital, LMU Munich, Munich, Germany

[h] Department of Dermatology, Venereology and Allergology, Charité – Universitätsmedizin Berlin, Berlin, Germany

[i] Dermatopathology Laboratory, Friedrichshafen, Germany

[j] Department of Dermatology, University Hospital Erlangen, Erlangen, Germany

[k] Department of Dermatology, University Hospital Regensburg, Regensburg, Germany

[l] Department of Dermatology, University Hospital Essen, Essen, Germany

[m] Department of Dermatology, University Hospital Würzburg, Würzburg, Germany

[n] Division of Translational Medical Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany

[o] National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Heidelberg, Germany

[p] Dr. Phillip Frost Department of Dermatology and Cutaneous Surgery, University of Miami, Miller School of Medicine, Miami, FL, USA

* Corresponding author: Digital Biomarkers for Oncology Group, 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).

The original publication is available at DOI: https://doi.org/10.1016/j.ejca.2020.11.020

### 2.4.1 Abstract

**Background:** A basic requirement for AI-based image analysis systems, which are to be integrated into clinical practice, is a high robustness. Minor changes in how those images are acquired, for example, during routine skin cancer screening, should not change the diagnosis of such assistance systems.

**Objective:** To quantify to what extent minor image perturbations affect the CNN-mediated skin lesion classification and to evaluate three possible solutions for this problem (additional data augmentation, test-time augmentation, anti-aliasing).

**Methods:** We trained three commonly used CNN architectures to differentiate between dermoscopic melanoma and nevus images. Subsequently, their performance and susceptibility to minor changes (‘brittleness’) was tested on two distinct test sets with multiple images per lesion. For the first set, image changes, such as rotations or zooms, were generated artificially. The second set contained natural changes that stemmed from multiple photographs taken of the same lesions.

**Results:** All architectures exhibited brittleness on the artificial and natural test set. The three reviewed methods were able to decrease brittleness to varying degrees while still maintaining performance. The observed improvement was greater for the artificial than for the natural test set, where enhancements were minor.

**Conclusions:** Minor image changes, relatively inconspicuous for humans, can have an effect on the robustness of CNNs differentiating skin lesions. By the methods tested

here, this effect can be reduced, but not fully eliminated. Thus, further research to sustain the performance of AI classifiers is needed to facilitate the translation of such systems into the clinic.

### 2.4.2 Introduction

AI-based image classification by CNNs has the potential to assist clinicians with diagnostic tasks that are based on the visual inspection of potentially malignant lesions. In experimental settings, CNNs have achieved performances in medical image classification tasks that were on par or even exceeded the results obtained by human experts [94–97]. In particular, CNNs have shown very promising results in macroscopic and microscopic skin lesion classification, both individually [22–28,33,35,52] and as assistance systems for dermatologists [20,62,63,98]. And while such systems are as of yet, mostly unable to predict malignant oncologic transformations due to a lack of prospective training data [99], they are already used in practice. In fact, CNN-based systems have begun to enter clinical dermatological practice as skin cancer screening tools, for example, as a market-approved computer-aided diagnostic system [30,99], which has demonstrated superior performance to more conventional computer-aided diagnostic systems [100].

While CNN-based image analysis has advantages over human observation with respect to objective and quantitative feature extraction, an obvious drawback is that in contrast to human experts, CNNs have difficulty distinguishing biologically significant features from insignificant features and artifacts. Depending on the data set that is used for CNN training, spurious and unwanted correlations within the training set can be picked up and hamper generalization [101–103]. Moreover, deceptively created input images specifically designed to fool a CNN (adversarial attacks) have been shown

to pose a real threat [104]. Both shortcomings also apply to CNNs in the field of dermatology [59,105–107].

Another observed shortcoming is the brittleness of modern CNNs in image analysis. Brittleness in this context refers to the phenomenon that small changes in the input image, such as scaling or rotation, can have a large effect on the classification of the CNN. It is therefore different to adversarial attacks, as image changes are not designed to deceive the CNN, but reflect fluctuations in image acquisition occurring in daily clinical routine. The resulting vulnerability of AI-based tools contradicts the assumption that CNNs are invariant to small transformations and is reported in the machine learning community [40,41,106–108]. As this lack of robustness and reliability may have a detrimental effect in a clinical setting, it needs to be overcome to facilitate the successful translation of AI-based diagnostic tools into routine clinical care.

In this study, we investigate the brittleness of three commonly used CNN architectures, which could serve as backends of CNN-based diagnostic systems, by testing them on images that have undergone transformations, which model variations that may occur when dermatologists photograph suspicious skin lesions. Moreover, we investigate three possible techniques (data augmentation, test time augmentation, anti-aliased networks) regarding their effectiveness in solving the problem of CNN brittleness.

### 2.4.3 Materials and methods

**Study design**

We trained three commonly used CNN architectures (ResNet50, DenseNet121, VGG16) to distinguish between dermoscopic nevus and MM images. To establish the models' susceptibility to image changes, each classifier was evaluated on a test set containing unmodified, original images and several additional sets containing

duplicated images that were digitally modified. Transformations were chosen to mimic events, which might occur in a clinical setting. Moreover, the magnitude of transformations was limited to an extent which would not render the lesion unrecognizable for a physician. Subsequently, a range of pre-existing methods which address AI brittleness were tested to assess if they are indeed effective in reducing brittleness without impairing performance.

As the test set transformations described above were artificial, the models and methods were additionally tested on an independent test set where at least two dermoscopic images with natural changes resulting from differences in real-life image acquisition were available for each lesion.

Ethics approval was waived by the ethics committee of the University of Heidelberg, as images were open source and anonymous.

**Data sets**

Dermoscopic images were obtained from the ISIC archive [109], the HAM10000 data set [110], the PH2 data set [111], the SKINL2 data set [112], the BCN20000 data set [112,113], and PROP, a proprietary data set. The training set was made up exclusively of ISIC, HAM10000, and BCN20000 images. The artificial test set consisted of a holdout component in ISIC and an external component in PH2 and SKINL2. Similarly, the natural test set consisted of a holdout component in BCN20000 and an external component in PROP. Exact details on training and test set composition are listed in the **Supplementary Materials**.

The artificial test set was duplicated 11 times. Each of the 11 duplicated sets was modified according to one previously defined transformation type and magnitude. Available types were a change in orientation, zoom, or brightness. In addition, the artificial test set was duplicated six more times, but this time combinations of transformations were applied and the magnitude was increased (see **Supplementary Materials**).

The additional natural test set contained at least two separately taken dermoscopic images per lesion. Thus, the changes between these images were not produced retrospectively using a computer. As this makes it impossible to define an original test set against which deviations should be measured, all possible image combinations were compiled and evaluated. Because different photographs of the same lesion often looked extremely different, e.g., because of an altered zoom by more than 50%, images for each lesion were manually sorted into similarly looking groups using the four-eyes principle.

**Classifier development**

All classifiers, regardless of architecture, were trained using the same training set and protocol. Furthermore, all architectures had the same set of fully connected layers on top of the individual feature extractor, which was made up of fastai's [86] default custom head. Online data augmentation was applied during training, where the type and magnitude of augmentations were adapted from the fastai library, which has sensible preset values. For exact details on the training procedure and used augmentations, see **Supplementary Materials**.

All work was carried out in Python 3.7.7 using fastai 1.0.61 in combination with torch 1.5.1 [85] and torchvision 0.6.1. Training was carried out on a single NVIDIA GeForce RTX 2080 Ti.

**Methods to reduce brittleness**

Three methods were tested for their effectiveness against brittleness. The first approach used a more extreme form of data augmentation during the training stage, where the magnitudes of the applied transformations were increased. The second approach used test-time augmentation during the inference stage. Instead of the model just rating one version of an input image, it rates a collection of slightly modified duplicates and averages the output. In our case, eight modified duplicates were rated, which were transformed using a flip coupled with a zoom into all four image corners. These transformations were set to be deterministic to allow reproducibility. The third approach replaced the original model architecture by an anti-aliased architecture, which reduces anti-aliasing effects in downsampling layers (strided convolutions, max-/average-pooling) [107]. This is achieved by upgrading all downsampling layers to include a low-pass filter. While originally intended to address shift-invariance, a general positive effect on model robustness was observed [107].

**Analysis**

To obtain robust performance estimates that encompass the stochastic nature of the training process, each training and evaluation run was repeated five times. Thus, all calculated metrics are averaged over five runs.

Classifier performance was captured using the AUROC. As the receiver operating curve shows the sensitivity and specificity of a dichotomous outcome for all possible

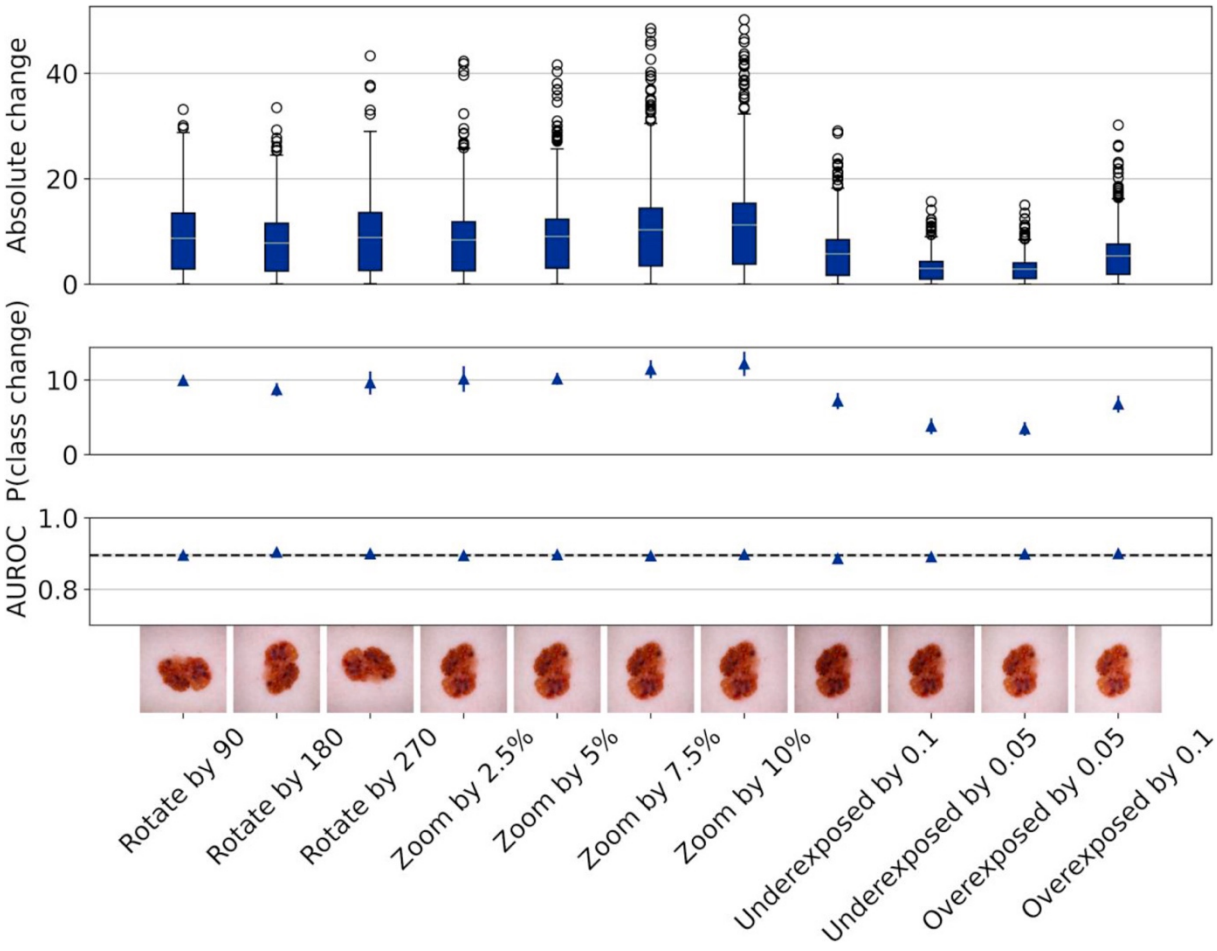
classification thresholds, the area under this curve provides a single summary measure, which captures a classifier's overall performance. The classifiers' susceptibility to change was measured using P(class change) and mean absolute change, two metrics adapted from *Azulay et al.* [40]. P(class change) represents the probability that the classifier changes its prediction from MM to nevus or vice versa, after the input image is transformed. This measure is independent of small confidence fluctuations, which do not have an impact on the classification, e.g. when a model changes its lesion diagnosis from 95% nevus to 85% nevus, this change is ignored by P(class change). Mean absolute change measures by how much on average the model's output probability changes after the input image is transformed. This metric allows us to verify if class changes are mainly a result of lesions being diagnosed divergently when the model was unsure to begin with. For a robust classifier, both metrics should be minimised.

### 2.4.4 Results

**Baseline performance and brittleness**

All baseline CNNs achieved an AUROC of approximately 0.9. This was comparable with the AUROCs obtained across the 11 artificially transformed test sets (see **Figure 5** and **Supplementary Materials**). For ResNet50, the mean absolute change varied from 2.9% ± 0.4% to 11.2% ± 1.2% and resulted in a P(class change) ranging from 3.5% ± 0.9% to 12.2% ± 1.6%. Variations in mean absolute change and P(class change) were slightly lower for DenseNet121, with VGG16 showing the lowest variation out of all three architectures (see **Supplementary Materials**).

**Figure 5. Individual performance and brittleness metrics for the baseline ResNet50 model across all artificially transformed test sets.** Top row shows the absolute change distribution over each artificially transformed test set. The grey line within the box plot indicates the mean absolute change. Middle row and bottom row show the mean P(class change) and AUROC, respectively, for each individually and artificially transformed test set. In addition, the AUROC for the unmodified test set is shown as a dashed line. Results for the other architectures were similar (see **Supplementary Materials**). AUROC: area under the receiver operating characteristic curve



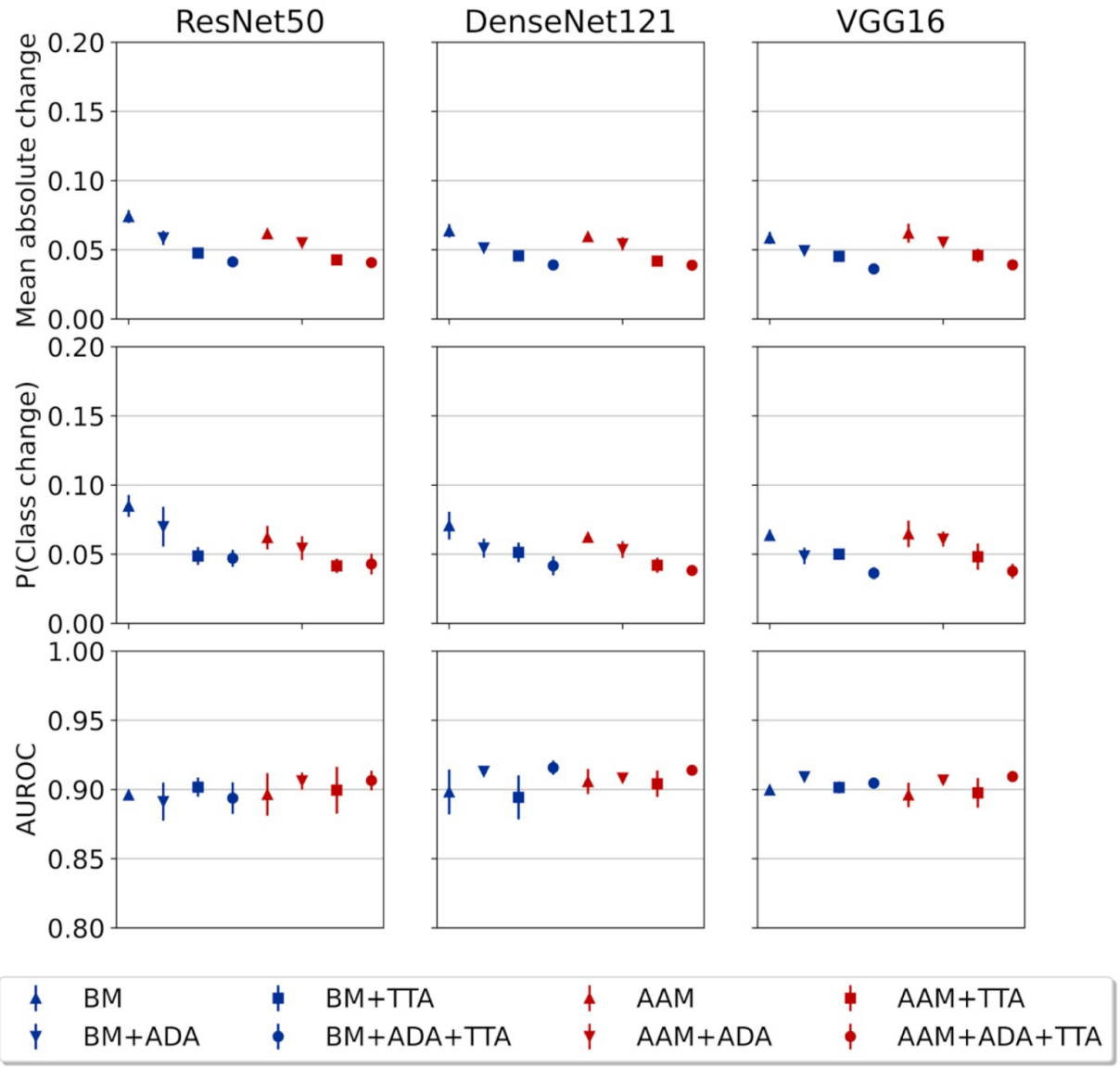

Averaging performance and robustness metrics across all twelve artificial test sets shows that both metrics were always better on the holdout than on the external test set regardless of used architecture (see **Supplementary Materials**). Moreover, there was a clear ranking between architectures with VGG16 having the best overall performance and brittleness scores, followed by DenseNet121 and ResNet50.

### Effectiveness of tested methods on artificial transformations

The three tested methods, which were additional data augmentation, test-time augmentation and anti-aliasing, were able to reduce overall brittleness when applied

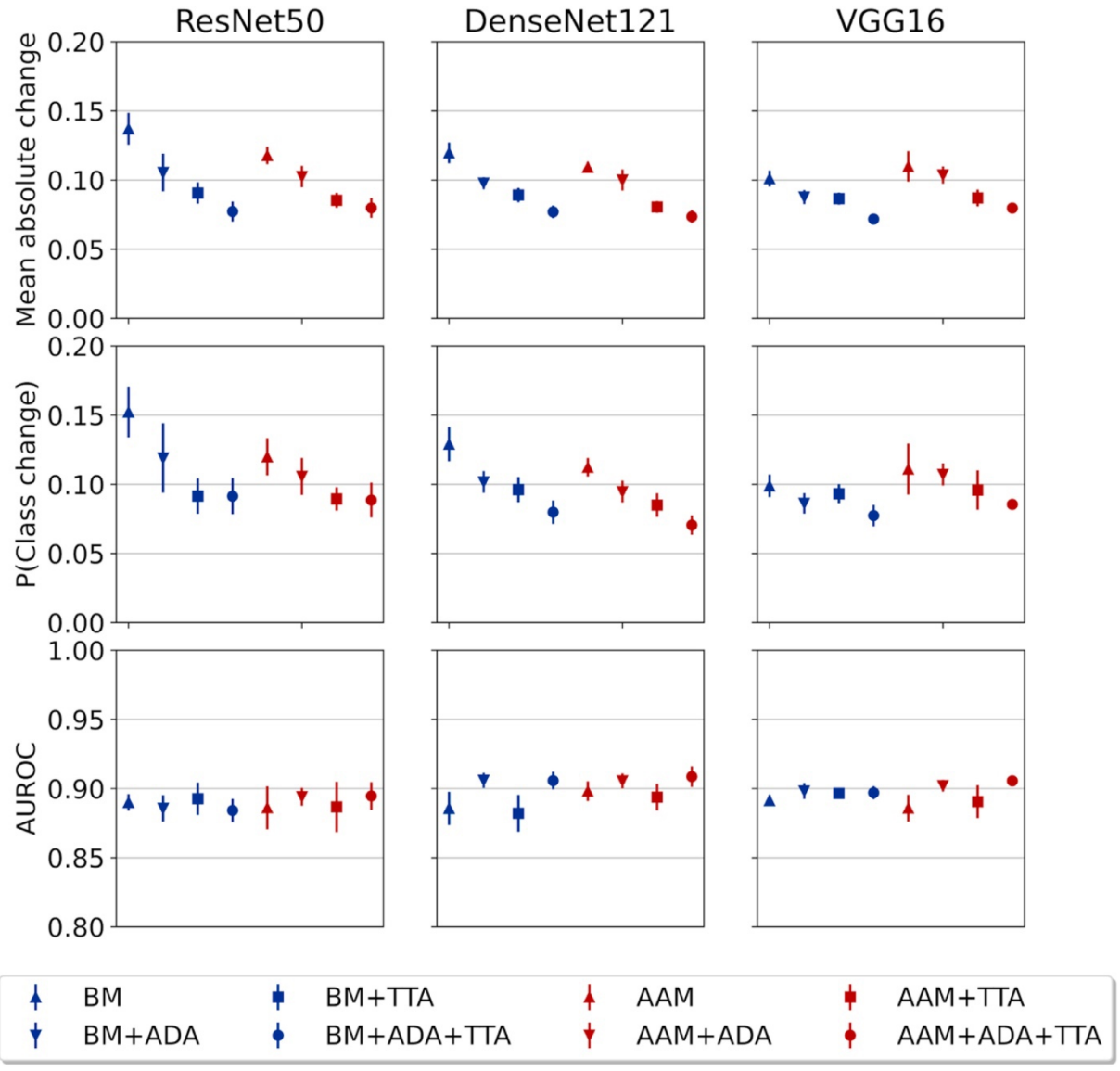
individually and especially when used in combination. This was true for all three architectures to a similar extent and did not result in performance deterioration (see **Figure 6**). Depending on the type of transformation that was applied to the test set, the used methods showed varying degrees of effectiveness. Generally, larger improvements were observed for rotations and zooms than for brightness (see **Supplementary Materials**).

**Figure 6. Average performance and brittleness metrics across all artificially transformed test sets for the various method combinations using individual transformations.** The three proposed methods, ADA, TTA and AAM were tested individually and in combination. Metrics were established on all individually transformed test sets and averaged. AUROC: area under the receiver operating characteristic curve, BM: baseline model, ADA: additional data augmentation, TTA: test-time augmentation, AAM: anti-aliased model.



When combining the artificial transformations to act on an image together, brittleness increased even more and performance deteriorated slightly. However, all reviewed methods were still effective in reducing brittleness while upholding performance (see **Figure 7**).

**Figure 7. Average performance and brittleness metrics across all artificially transformed test sets for the various method combinations using combined transformations.** The three proposed methods, ADA, TTA and AAM were tested individually and in combination. Metrics were established and averaged over all transformed test sets, which were modified using a combination of individual transformations. AUROC: area under the receiver operating characteristic curve, BM: baseline model, ADA: additional data augmentation, TTA: test-time augmentation, AAM: anti-aliased model.

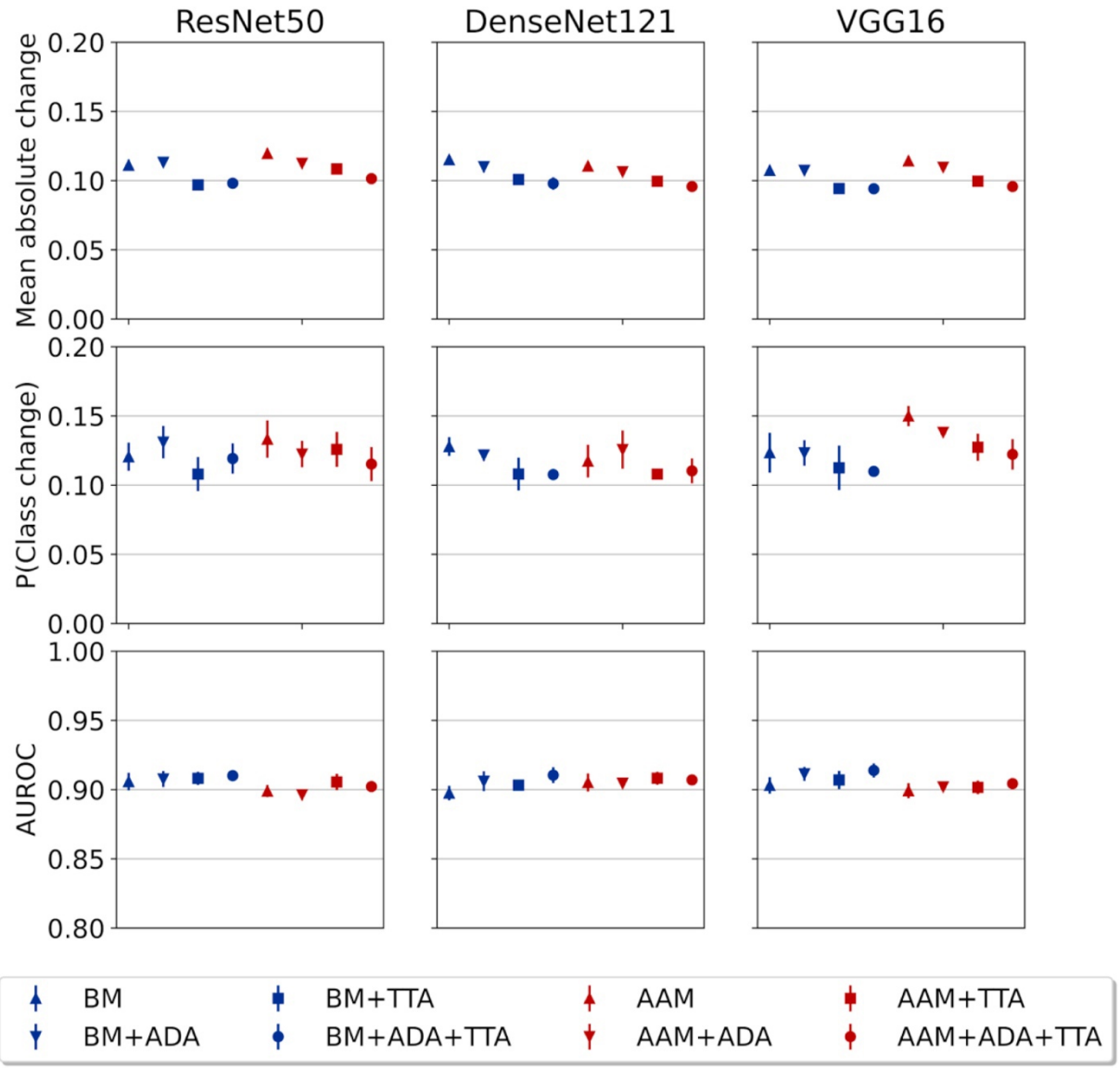


Regardless whether the artificial transformations were used individually or in combination, additional data augmentation and test-time augmentation always showed improvements for brittleness and were most effective when applied in combination. Anti-aliasing worked well for ResNet50 and DenseNet121; however, the anti-aliased VGG16 suffered an increase in brittleness.

**Effectiveness of tested methods on natural transformations**

Average performance and brittleness of all three baseline models on the natural test set was in-between that of the artificial test set with individual transformations and the artificial test set with combined transformations. However, effectiveness of the employed methods was far less pronounced on the natural test set than on either of the two artificial test sets (see **Figure 8**). Trends were less consistent and while one method showed improvements for a certain architecture, it did not do so for another. For example, ResNet50 experienced slightly worse brittleness with additional data augmentation while DenseNet121 did not. Regardless of architecture, test-time augmentation always improved both performance and brittleness.

**Figure 8. Performance and brittleness metrics across the natural transformed test set for the various method combinations.** The three proposed methods, ADA, TTA and AAM, were tested individually and in combination. AUROC: area under the receiver operating characteristic curve, BM: baseline model, ADA: additional data augmentation, TTA: test-time augmentation, AAM: anti-aliased model.




### 2.4.5 Discussion

**Practical implications**

This study demonstrated brittleness i.e., vulnerability of CNNs toward small input changes for three commonly used CNN architectures (ResNet50, DenseNet121, VGG16). Although this phenomenon has been reported throughout the machine learning community, its potential impact on AI-based assistance systems in the clinic has

not received proper attention [40,41,106–108]. We reviewed three different methods to reduce brittleness (additional data augmentation, test-time augmentation, anti-aliasing) and found them to be partially effective on artificial image transformations such as rotations, altered brightness or zooms, but less so on natural image transformations resulting from image acquisition differences.

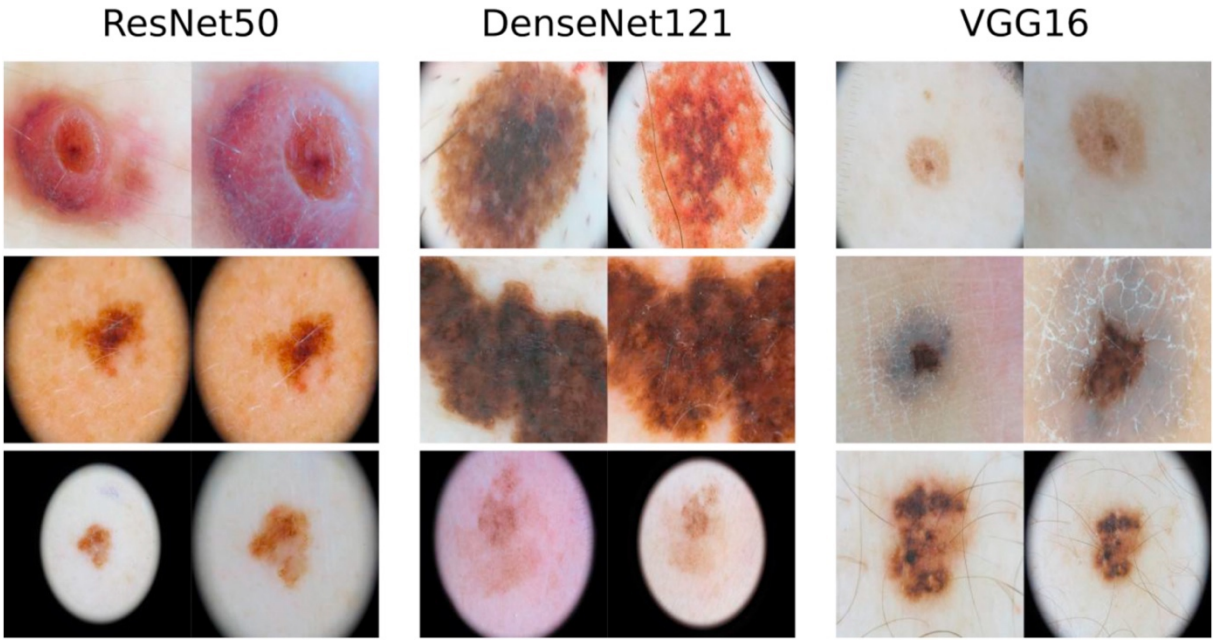
For our models, we chose architectures and training techniques that are commonly used throughout image classification tasks for skin cancer [22,62,114] and other cancer subtypes [95,115–118]. Thus, we believe our baseline models to be suitably representative of existing or future models, which could serve as the backbone of a diagnostic system.

While the change of diagnosis i.e., P(class change) is independent of monotonic confidence fluctuations and intuitive to grasp, we also consider the mean absolute change. In a clinical setting, it is unlikely that an assistance system, which solely presents a plain diagnosis such as MM, will be accepted by physicians or patients. Inclusion of the model's confidence level may increase trust in the system as it enables the physician to judge the weight he/she should attribute to the model's classification. Low-confidence decisions by the system would therefore be less likely to influence the physician's management decision to begin with. In such a setting, brittleness would be partially compensated as the observed confidence changes would often only alter the CNN's classification if its confidence was low to begin with. If, however, high-confidence classifications show these fluctuations, the range of confidences for similar images can be highly disconcerting to the physician.

The techniques we evaluated to reduce brittleness, namely additional data augmentation, test-time augmentation and anti-aliasing, substantially reduced this phenomenon in an artificial setting, but even when used in combination did not completely eliminate it. Depending on the architecture, some methods worked better than others; for example, anti-aliasing did not reduce brittleness for VGG16. When all three methods were used in combination, brittleness and performance always improved in comparison with the baseline model.

The observed improvement was much more limited on naturally transformed images. Even when combinations were applied, improvements were minor or non-existent. **Figure 9** shows a selection of natural image pairs where our models, regardless of the applied method, always came to a divergent diagnosis on an image pair of the same lesion, even though some of the paired images appear almost identical. Thus, it may hardly be possible for a physician to determine how to photograph a lesion ‘correctly’, which they intend to diagnose with a CNN-based lesion classification system. Such problems limit the applicability of the technology in the clinic and therefore have to be solved.

**Figure 9. Natural image pairs for selected lesions where models disagreed constantly.** Each lesion was photographed twice and rated by all possible combinations of proposed methods (i.e., BM, BM + ADA, BM + TTA, AAM, etc.). Regardless of the applied method, none of the selected image pairs received the same diagnosis. BM: baseline model; ADA: additional data augmentation, TTA: test-time augmentation, AAM: anti-aliased model.




Against this background, we would like to inform physicians to not consider CNN-based systems as error free and be aware of such limitations. We also want to encourage deep learning practitioners to actively minimise brittleness on a case-by-case basis in the same way performance is optimised. The reported improvements could be further enhanced through method-specific optimisations, alternative techniques for robustness [119,120] or an ensemble-approach, which showed even better improvements than model-specific techniques (see **Supplementary Materials**). Finally, future work should also investigate alternatives, which do not solely focus on the training/inference procedure or on architectural modifications but rather on other architectures such as Capsule Neural Networks [121] which could be better suited to handle small affine transformations.

**Limitations**

The artificial image changes were designed in such a way as to be relatively inconspicuous to a human observer. The inconspicuousness was determined using the four-eye principle and is therefore subjective. But even if images changes are not deemed as inconspicuous, such transformations are still likely to arise in a clinical setting and therefore any CNN-based system should be invariant against such changes.

The natural test set contained multiple photographs per lesion, where some looked extremely distinct, to the point where there was no overlap between images. Thus, suitably similar image pairs for each lesion were manually chosen using the four-eye principle. As this was largely subjective, the reported results for the natural test set could change depending on how the images are sorted.

### 2.4.6 Conclusions

Minor image changes, relatively inconspicuous for humans, can have an effect on the confidence and diagnosis of CNNs differentiating skin lesions. Using the methods tested here, this effect was reduced but not fully eliminated. Therefore, we would like to remind deep learning practitioners and physicians in dermatology but also in medicine in general, that brittleness needs to be explicitly targeted and overcome to facilitate translation from bench-to-bedside.

### 2.4.7 Acknowledgement

**Role of the funding source**

This study was funded by the Federal Ministry of Health, Berlin, Germany (grant: Skin Classification Project; grant holder: Titus J. Brinker, 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.

**Authorship contribution statement**

Roman C. Maron: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft. Sarah Haggenmüller: Conceptualization, Methodology, Writing - original draft, Visualization. Christof von Kalle: Resources, Writing - review & editing. Jochen S. Utikal: Resources, Writing - review & editing. Friedegund Meier: Resources, Writing - review & editing. Frank F. Gellrich: Resources, Writing - review & editing. Axel Hauschild: Resources, Writing - review & editing. Lars E. French: Resources, Writing - review & editing. Max Schlaak: Resources, Writing - review & editing. Kamran Ghoreschi: Resources, Writing - review & editing. Heinz Kutzner: Resources, Writing - review & editing. Markus V. Heppt: Resources, Writing - review & editing. Sebastian Haferkamp: Resources, Writing - review & editing. Wiebke Sondermann: Resources, Writing - review & editing. Dirk Schadendorf: Resources, Writing - review & editing. Bastian Schilling: Resources, Writing - review & editing. Achim Hekler: Conceptualization, Methodology, Writing - review & editing. Eva Krieghoff-Henning: Conceptualization, Writing - original draft. Jakob N. Kather: Resources, Writing - review & editing. Stefan Fröhling: Resources, Writing - review & editing. Daniel B. Lipka: Resources, Writing - review & editing. **Titus J. Brinker:** Conceptualization, Writing - review & editing, Supervision, Project administration, Funding acquisition.

**Conflict of interest statement**

Sebastian H. reports advisory roles for or has received honoraria from Pierre Fabre Pharmaceuticals, Novartis, Roche, BMS, Amgen and MSD outside the submitted work. Axel H. reports clinical trial support, speaker's honoraria, or consultancy fees from the

following companies: Amgen, BMS, Merck Serono, MSD, Novartis, Oncosec, Philogen, Pierre Fabre, Provectus, Regeneron, Roche, OncoSec, Sanofi-Genzyme, and Sun Pharma, outside, the submitted work. BS reports advisory roles for or has received honoraria from Pierre Fabre Pharmaceuticals, Incyte, Novartis, Roche, BMS and MSD, research funding from BMS, Pierre Fabre Pharmaceuticals and MSD, and travel support from Novartis, Roche, BMS, Pierre Fabre Pharmaceuticals and Amgen; outside the submitted work. JSU is on the advisory board or has received honoraria and travel support from Amgen, Bristol Myers Squibb, GSK, LeoPharma, Merck Sharp and Dohme, Novartis, Pierre Fabre, Roche, outside the submitted work. WS received travel expenses for attending meetings and/or (speaker) honoraria from Abbvie, Almirall, Bristol-Myers Squibb, Celgene, Janssen, LEO Pharma, Lilly, MSD, Novartis, Pfizer, Roche, Sanofi Genzyme and UCB outside the submitted work. FM has received travel support or/and speaker's fees or/and advisor's honoraria by Novartis, Roche, BMS, MSD and Pierre Fabre and research funding from Novartis and Roche. TJB reports owning a company that develops mobile apps (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.

### 2.4.8 Supplementary materials

Supplementary Materials are available at https://doi.org/10.1016/j.ejca.2020.11.020

## 2.5 Supplementary publication 2: Model soups improve performance of dermoscopic skin cancer classifiers

Roman C. Maron[a], Achim Hekler[a], Sarah Haggenmüller[a], Christof von Kalle[b], Jochen S. Utikal[c,d,p], Verena Müller[c,d,p], Maria Gaiser[c,d,p], Friedegund Meier[e], Sarah Hobelsberger[e], Frank F. Gellrich[e], Mildred Sergon[e], Axel Hauschild[f], Lars E. French[g,o], Lucie Heinzerling[g], Justin G. Schlager[g], Kamran Ghoreschi[h], Max Schlaak[h], Franz J. Hilke[h], Gabriela Poch[h], Sören Korsing[h], Carola Berking[i], Markus V. Heppt[i], Michael Erdmann[i], Sebastian Haferkamp[j], Dirk Schadendorf[k], Wiebke Sondermann[k], Matthias Goebeler[l], Bastian Schilling[l], Jakob N. Kather[m], Stefan Fröhling[n], Daniel B. Lipka[n], Eva Krieghoff-Henning[a], Titus J. Brinker[a,*]

[a] Digital Biomarkers for Oncology Group, National Center for Tumor Diseases (NCT), German Cancer Research Center (DKFZ), Heidelberg, Germany

[b] Department of Clinical-Translational Sciences, Charité – University Medicine and Berlin Institute of Health (BIH), Berlin, Germany

[c] Department of Dermatology, Venereology and Allergology, University Medical Center Mannheim, Ruprecht-Karl University of Heidelberg, Mannheim, Germany

[d] Skin Cancer Unit, German Cancer Research Center (DKFZ), Heidelberg, Germany

[e] Skin Cancer Center at the University Cancer Center and National Center for Tumor Diseases Dresden, Department of Dermatology, University Hospital Carl Gustav Carus, Technische Universität Dresden, Germany

[f] Department of Dermatology, University Hospital (UKSH), Kiel, Germany

[g] Department of Dermatology and Allergy, University Hospital, LMU Munich, Munich, Germany

[h] Department of Dermatology, Venereology and Allergology, Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[i] Department of Dermatology, University Hospital Erlangen, Comprehensive Cancer Center Erlangen – European Metropolitan Region Nürnberg, CCC Alliance WERA, Erlangen, Germany

[j] Department of Dermatology, University Hospital Regensburg, Regensburg, Germany

[k] Department of Dermatology, Venereology and Allergology, University Hospital Essen, Essen, Germany

[l] Department of Dermatology, Venereology and Allergology, University Hospital Würzburg, Würzburg, Germany

[m] Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany

[n] Department of Translational Medical Oncology, National Center for Tumor Diseases (NCT) Heidelberg and German Cancer Research Center (DKFZ), Heidelberg, Germany

[o] Dr. Phillip Frost Department of Dermatology and Cutaneous Surgery, University of Miami, Miller School of Medicine, Miami, FL, USA

[p] DKFZ Hector Cancer Institute at the University Medical Center Mannheim, Mannheim, Germany

* Corresponding author: Digital Biomarkers for Oncology Group, 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).

The original publication is available at DOI: https://doi.org/10.1016/j.ejca.2022.07.002

### 2.5.1 Abstract

**Background:** Image-based cancer classifiers suffer from a variety of problems which negatively affect their performance. For example, variation in image brightness or different cameras can already suffice to diminish performance. Ensemble solutions, where multiple model predictions are combined into one, can improve these problems. However, ensembles are computationally intensive and less transparent to practitioners than single model solutions. Constructing model soups, by averaging the weights of multiple models into a single model, could circumvent these limitations while still improving performance.

**Objective:** To investigate the performance of model soups for a dermoscopic MM-nevus skin cancer classification task with respect to (1) generalisation to images from other clinics, (2) robustness against small image changes and (3) calibration such that the confidences correspond closely to the actual predictive uncertainties.

**Methods:** We construct model soups by fine-tuning pre-trained models on seven

different image resolutions and subsequently averaging their weights. Performance is evaluated on a multi-source dataset including holdout and external components.

**Results:** We find that model soups improve generalisation and calibration on the external component while maintaining performance on the holdout component. For robustness, we observe performance improvements for pertubated test images, while the performance on corrupted test images remains on par.

**Conclusions:** Overall, souping for skin cancer classifiers has a positive effect on generalisation, robustness and calibration. It is easy for practitioners to implement and by combining multiple models into a single model, complexity is reduced. This could be an important factor in achieving clinical applicability, as less complexity generally means more transparency.

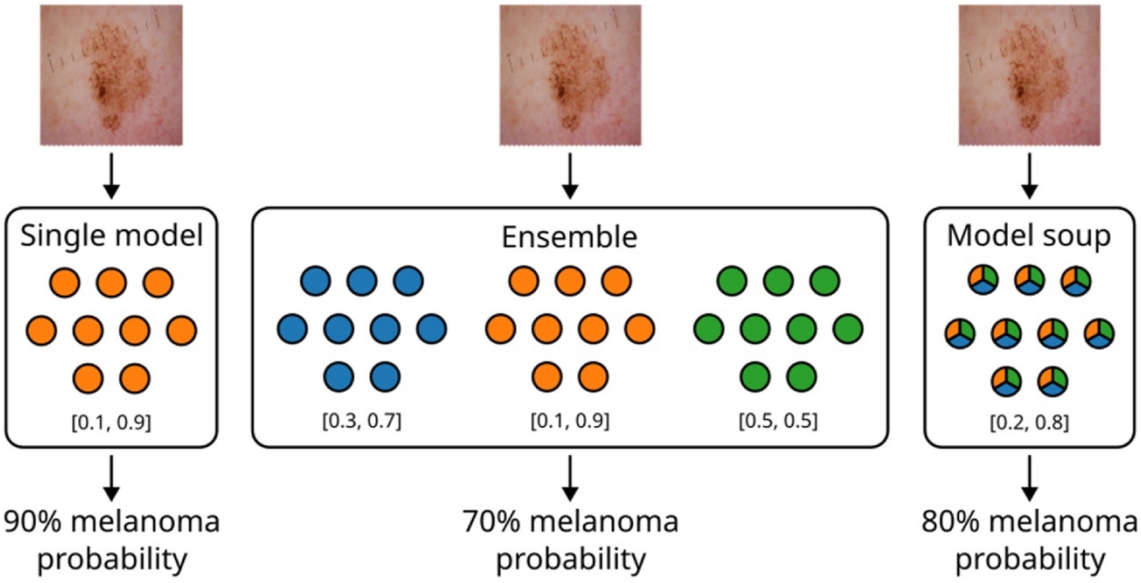
### 2.5.2 Introduction

While deep learning-based skin cancer classifiers have achieved numerous accolades in the past – such as on par or outperformance of human experts in artificial settings [22–25,27,28,63,122] – transition to the clinical setting proves difficult. Although prospective studies have shown promising results [123,124] and human experts may benefit from deep learning support [20,62,98,125], progressing research reveals various limitations. For example, studies have shown that a market-approved deep learning device has learned spurious correlations during training, leading to the association of skin markings [59] or scale bars [126] to MM. What is even more surprising is that small image changes, which occur through everyday image acquisition fluctuations or targeted adversarial attacks, can already change the classification and confidence of deep learning algorithms [61,105]. In combination, such limitations negatively affect the overall performance of deep learning algorithms.

A commonly employed method for addressing some of these problems is ensemble learning, where the predictions of multiple models are combined into one. The winning solution of the SIIM-ISIC Melanoma Classification Challenge 2020, for example, used an ensemble of 90 models [127]. Such large-scale solutions impose obvious drawbacks. For one, ensemble predictions require multiple inference passes and more computational power. However, with ever improving hardware, the severity of this limitation will continue to decrease in the future. A tougher problem is the necessity for transparency [128], which is complicated through the black box nature of deep learning algorithms [65]. Adding more 'black boxes' to a classifier is unlikely to simplify this problem.

Model soups, where the weights – not the outputs – from multiple models of the same architecture are averaged to produce a single model, present an interesting alternative to ensembles (see **Figure 10**). *Wortsman et al.* [129] have shown that soups, constructed from models fine-tuned on ImageNet, improve in- and out-of-distribution performance when compared to the best individual model.

**Figure 10. Differences between traditional single model solutions, ensembles and model soups.** During the model development phase, multiple models are usually generated by deep learning practitioners. For traditional single model solutions (left), only the best performing model is selected for inference. For ensembles (middle), multiple models are selected for inference. Each model separately classifies the input, and outputs are subsequently combined (e.g., via averaging). For model soups (right), multiple models are also selected for inference. However, the models are combined into a single model by averaging their weights. Thus, only one model classifies the input.




In this study, we therefore investigate how model soups fare with a real-world medical task in the form of skin cancer classification, as performance improvements – especially regarding generalisation – are not guaranteed to transfer across tasks [130]. To further expand on previous works, we focus our research exclusively on smaller CNN architectures pre-trained on ImageNet, instead of models pre-trained on larger, heterogeneous datasets such as CLIP [131] or ALIGN [132]. In addition, we investigate whether the model selection process for soups introduced by *Wortsman et al.* [129] can further be improved. Our analysis will focus on three clinically relevant aspects which are (1) generalisation to images from other clinics, (2) robustness against small image changes and (3) calibration. The latter measures whether the confidence of predictions match the probability of being correct for all confidence levels. This is a crucial aspect for safety critical applications in medicine, where practitioners need to have

access to reliable predictive uncertainty to correctly estimate whether the model is right or wrong [133].

### 2.5.3 Methods

**Study design**

To obtain a pool of models that can be used to construct soups and ensembles, we separately trained seven models on different image sizes, starting at 112x122 pixels and moving up to 448x448 pixels in steps of 56 pixels. This pool of models is subsequently used to construct soups – using uniform, greedy or custom soup algorithms – and ensembles. We also choose a baseline model from this pool, which is simply the model that performs best on the validation set (see **Supplementary Materials**).

We evaluate our results by defining three clinically relevant aspects which target the generalisation, robustness and calibration of our classifier.

To ensure the results are representative, we include eight CNN architectures (ResNet34/50 [134], DenseNet121/169 [135], EfficientNetB1/B3 [136] and VGG11/16 [135,137]. We also average over five-fold cross validation, so that our results are less influenced by stochastic training events and dataset differences.

Ethics approval was waived by the ethics committee of the University of Heidelberg, as images were open source and anonymous.

**Datasets**

For model training, we use a multi-source dermoscopic image dataset [110,113,138], resulting in 14,648 melanoma and nevus images (see **Supplementary Materials**).

For model evaluation, we set aside 562 images as a holdout component. Furthermore, we construct a multi-source external component from six datasets (MSK [109], DERM7PT [139], ISIC2020_SYDNEY [138], SAM [140], PH2 [111] and SKINL2 [112]), resulting in 5,678 images.

In order to evaluate model robustness against image corruptions and perturbations, we add SAM-C and SAM-P to the test set [140]. SAM-C tests how well a classifier fares with low-quality images (e.g., blurry images). SAM-P tests classifier stability in response to subtle image changes (e.g., continuous change of brightness). See **Supplementary Materials** for further details.

**Model training**

We train each model according to the training procedure described by *Ha et al.* [127], leaving all hyperparameters fixed except for image size (see **Supplementary Materials**).

**Model souping**

As each architecture is trained across seven different image sizes, the pool of potential members for a soup consists of seven models. For uniform soups, all models are included by uniformly averaging their weights. For greedy soups, models are selected according to the algorithm described by *Wortsman et al.* [129]. Models are first ranked in descending order based on their performance on a separate validation set. The first, i.e., best performing model, is always included in the soup. Subsequent models will only be added if they improve the performance of the soup on the separate validation set. The separate validation set consists of 164 MM and 398 nevi images and models were ranked based on the AUROC.

As construction of the greedy soup is heavily dependent on the separate validation set, we experiment with artificially increasing the size and variety of this set by either duplicating each image ten times via random cropping or via data augmentations from *Ha et al.* [127]. We refer to these methods as greedy cropped and greedy data augmentation. Finally, we take each of the three greedy algorithms and inverse the order of the model rankings, i.e., sort from worst to best. We refer to these methods as greedy inverse, greedy cropped inverse and greedy data augmentation inverse. Uniform and greedy ensembles are constructed in the same way, but instead of averaging weights, model outputs are averaged.

**Model evaluation**

In order to cover the three different aspects – generalisation, robustness and calibration – we use the metrics below:

To test model generalisation and calibration on the holdout and external test set, we use AUROC and expected calibration error (ECE), respectively. Since the ECE is very sensitive to the selected number of bins (n = 20), we also look at the Brier Score (BS) and negative log likelihood (NLL).

To test model performance against corruptions (SAM-C) and perturbations (SAM-P), we use the mean balanced corruption error (mBCE) and the mean flip rate (mFR), respectively. Both metrics were introduced by *Maron et al.* [140] and should be minimised. mBCE averages the balanced error rates – i.e., the opposite of balanced accuracy – across each individual corruption type in SAM-C. mFR averages the flip probabilities – i.e., the likelihood that a classifier will change its classification between two

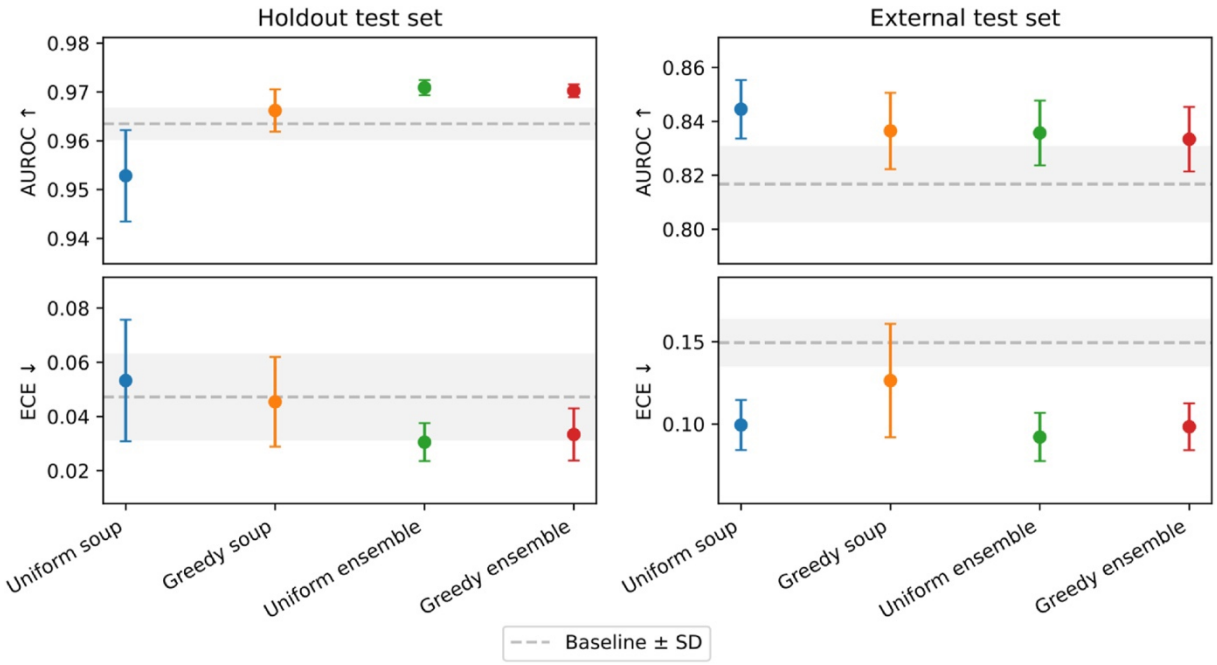
slightly different images of the same lesion – across each individual perturbation type in SAM-P.

### 2.5.4 Results

**Soups approximate ensembles**

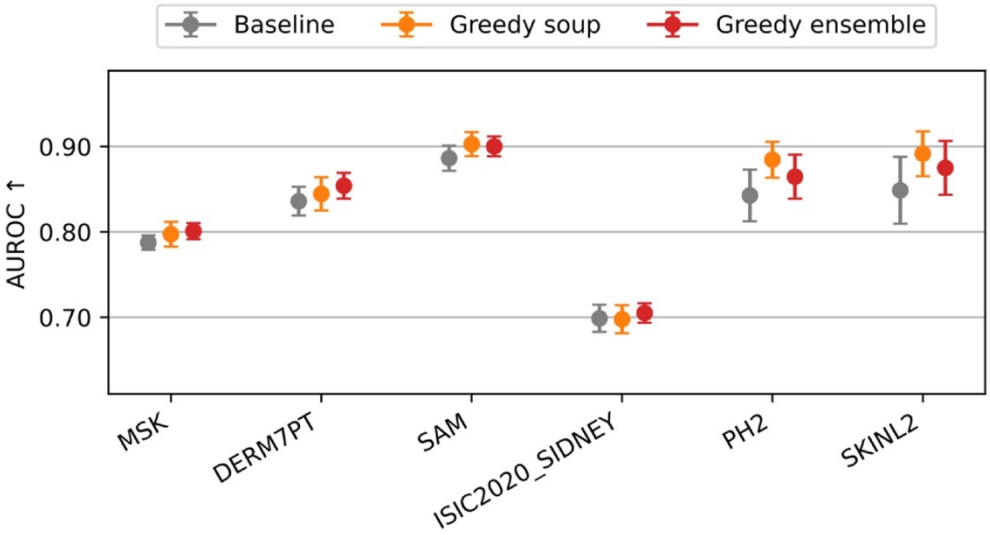
Looking at the AUROC in **Figure 11**, we observe that greedy soups on average perform better than the baseline while maintaining performance on the holdout test set. Unsurprisingly, this also holds true for the uniform and greedy ensemble, both of which outperform the greedy soup on the holdout test set, but only perform on par on the external test set. A similar trend can be observed for calibration; however, this time, both ensemble approaches clearly outperform the greedy soup. This trend stays intact when looking at the BS and NLL (see **Supplementary Materials)**. While uniform soups are better than the baseline on the external test set, they perform worse on the holdout test set, especially for AUROC. The described trends largely hold true for the individual architectures (see **Supplementary Material**).

**Figure 11. Comparing soups and ensembles across a holdout and multi-source external test set.** Each point depicts the average across eight different architectures, together with its corresponding standard deviation. The grey dashed line depicts the baseline performance, i.e., models which were not souped or ensembled. AUROC: area under the receiver operating characteristic curve, ECE: expected calibration error, SD: standard deviation.




The substantial AUROC improvement of the greedy soup and ensemble on the external test set also holds true for the majority of its individual components. Looking at **Figure 12**, one can see that except for ISIC2020_SIDNEY, greedy soups on average outperform the baseline and either approximate or outperform greedy ensembles. This is also the case for the individual architectures (see **Supplementary Materials**).

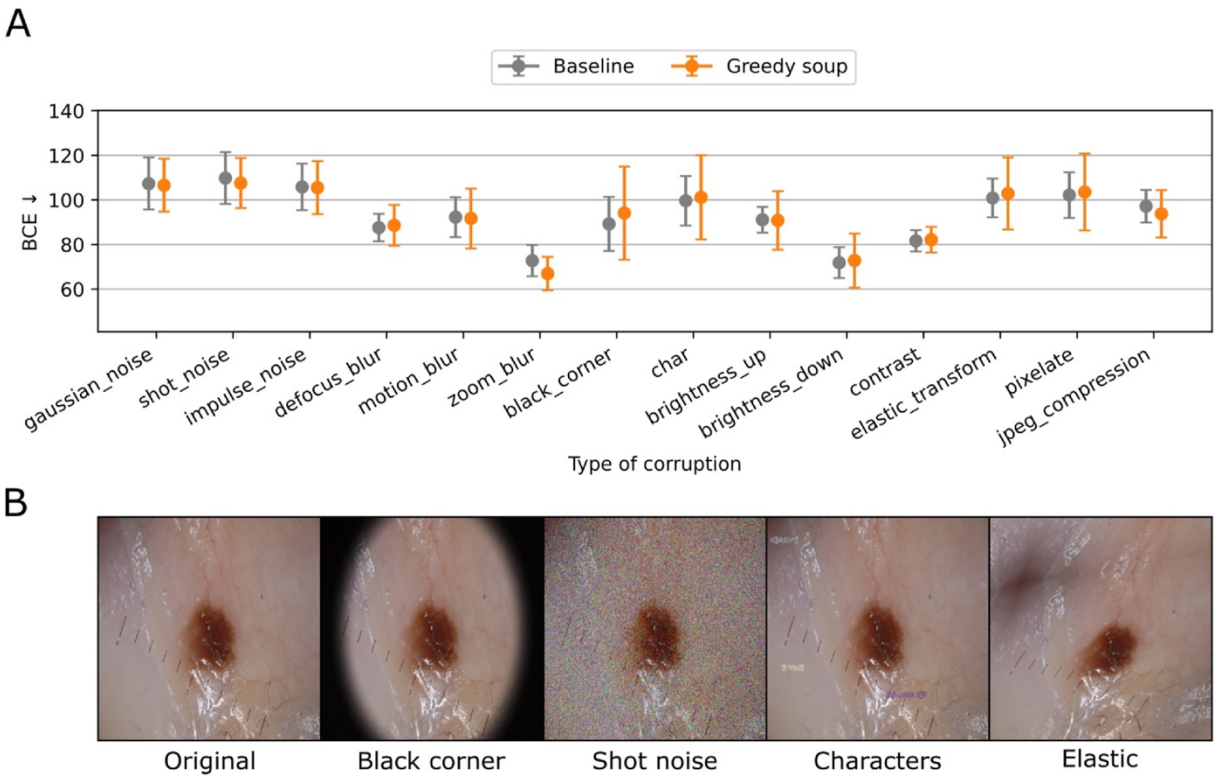
**Figure 12. Comparing AUROC of greedy soups and ensembles across six external test sets.** Each point depicts the average across eight different architectures, together with its corresponding standard deviation. Datasets on the x-axis are ordered by decreasing size. AUROC: area under the receiver operating characteristic curve.




### Soups improve overall robustness

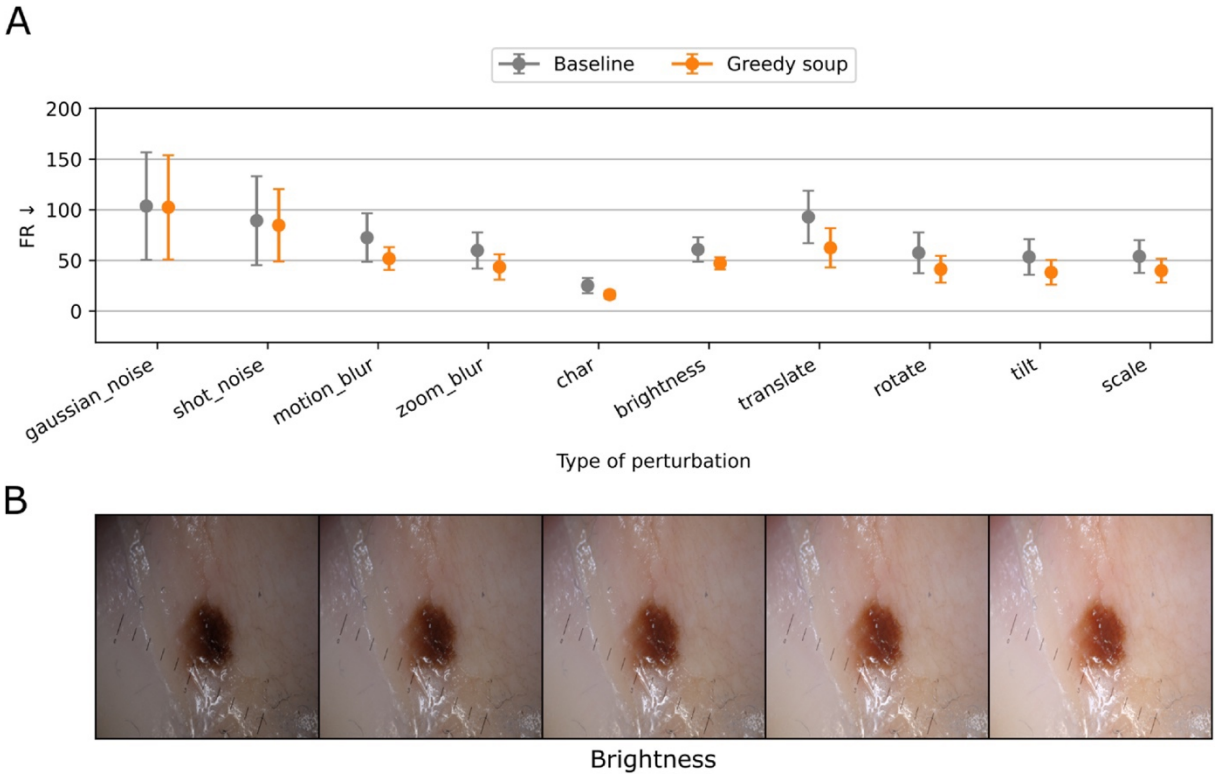
Regarding model robustness towards image corruptions, we observe that greedy soups on average have an mBCE of 93.43 ± 9.96 compared to a baseline of 93.51 ± 6.45. Thus, performance remains on par and does not improve. Looking at **Figure 13**, it becomes apparent that there is no consistent trend for the individual corruption types and that differences between the baseline and greedy soup are relatively minor.

**Figure 13. Greedy soup performance across 14 types of corruptions.** Part A: Performance is measured using the BCE which is adjusted by an AlexNet baseline, i.e., lower score is better. Each point depicts the average across eight different architectures, together with its corresponding standard deviation. Part B: Exemplary selection of the applied artificial image corruptions. For the performance of individual architectures, see **Supplementary Materials**. BCE: balanced corruption error.




When looking at model robustness towards image perturbations, we observe a substantial average improvement in mFR from 66.9 ± 21.87 to 52.77 ± 16.09. This time, improvements were consistently observed across all ten perturbation types (see **Figure 14**).

**Figure 14. Greedy soup performance across 10 types of perturbations.** Part A: Performance is measured using the FR which is adjusted by an AlexNet baseline, i.e., lower score is better. Each point depicts the average across eight different architectures, together with its corresponding standard deviation. Part B: Exemplary extract of five perturbation steps for the brightness perturbation. For the performance of individual architectures, **see Supplementary Materials.** FR: flip rate.




## Modifications on the greedy soup algorithm

From the six investigated souping algorithms (greedy, greedy cropped, greedy data augmentation, greedy inverse, cropped inverse, greedy data augmentation inverse), the three greedy soup algorithms on average add fewer models to the soup (n ~ 2.2) than the three greedy inverse soup algorithms (n ~ 6.0). Notably, the greedy inverse soups, on average, always perform worse on the holdout test set than the greedy soups (see **Table 8**). In contrast, they always outperform the greedy soups when looking at the external, corrupted and perturbed test set performance.

**Table 8. Performance comparison of modified greedy soup algorithms.** Each number depicts the average across eight different architectures, together with its corresponding standard deviation. Numbers are highlighted in blue if they are better than the greedy approach (second column to the left), white if they are equal and orange if they are worse. Bold highlighting indicates the overall best value for each row. For AUROC, higher values are better, and for mBCE/mFR, lower values are better. For the performance of individual architectures, see **Supplementary Materials**. AUROC: area under the receiver operating characteristic curve, DA: data augmentation, INV: inverse; mBCE: mean balanced corruption error, mFR: mean flip rate.

| Test set | Greedy | Greedy cropped | Greedy DA | Greedy INV | Greedy cropped INV | Greedy DA INV |
|---|---|---|---|---|---|---|
| Holdout *AUROC* ↑ | **0.966 ± 0.004** | **0.966 ± 0.005** | 0.965 ± 0.004 | 0.964 ± 0.006 | 0.964 ± 0.006 | 0.963 ± 0.007 |
| External *AUROC* ↑ | 0.836 ± 0.014 | 0.836 ± 0.013 | 0.838 ± 0.012 | **0.845 ± 0.007** | **0.845 ± 0.007** | **0.845 ± 0.008** |
| Corrupted *mBCE* ↓ | 93.43 ± 9.96 | 94.16 ± 10.44 | 93.18 ± 8.22 | 92.20 ± 6.19 | 92.52 ± 7.04 | **90.61 ± 4.97** |
| Perturbed *mFR* ↓ | 52.77 ± 16.09 | 54.22 ± 20.00 | 55.43 ± 17.93 | 50.26 ± 17.41 | **50.12 ± 19.42** | 52.25 ± 18.36 |

### 2.5.5 Discussion

We constructed model soups for eight CNN architectures and evaluated their skin cancer classification performance with respect to (1) their generalisation to images from other clinics, (2) their robustness against small image changes and (3) their calibration such that the confidences correspond closely to the actual predictive uncertainties. On average, we find that greedy soups at worst perform on par but often outperform the baseline with respect to our predefined clinical aspects. We thereby largely confirm the findings by *Wortsman et al.* [129], but translate them into a medical setting and extend them by using a variety of ImageNet-1k pre-trained CNN architectures (i.e., no pre-training on especially large and heterogeneous datasets).

This has important implications for deep learning practitioners working in the field of skin cancer classification, as souping makes use of models that are generated during

the hyperparameter-tuning stages of development. Usually, these models are discarded in favour of the best performing model.

For dermatologists, less complex solutions might be easier to interpret and integrate into an already time-pressed schedule. In the light of improving hardware and ongoing research for better interpretability methods, this aspect, however, is controversial.

**Soup performance: generalisation, robustness and calibration**

The observed, on average, better generalisation performance when looking at AUROC and ECE holds true for the majority of individual architectures. Simultaneously, both metrics do not decrease on the holdout test set. While uniform soups generalise even better than greedy soups, they perform worse than the baseline on the holdout test set. A trade-off which limits their utility. This highlights the advantages of the greedy soup, where inclusion of the best performing model practically guarantees at least on par performance on the holdout test set (with respect to the metric used to construct the soup).

However, ensembles are at least as good – if not even better – than greedy soups. This is unsurprising, as ensemble solutions are commonly employed in deep learning-based skin cancer classification. The top three solutions of the SIIM-ISIC Melanoma Classification Challenge 2020 consist of ensembles. Expecting ensemble level performance from a single model is therefore a demanding task. However, model soups combine advantages from ensembles, i.e., better performance, with advantages from single models, i.e., less computational power and better interpretability. While there are studies which investigate more interpretable ensembles [141,142], the majority of interpretability studies focus on single models, which is already a challenging task [65].

In terms of robustness, an aspect viewed under the light of image corruptions and perturbations, we observe an overall improvement. This is, however, only due to greedy soups faring better with perturbations, as no improvements are observed for corruptions. For corruptions, only five out of the eight architectures showed a lower mBCE (i.e., better result) than the baseline. This is in contrast to perturbations, where seven out of eight architectures showed a lower mFR (see **Supplementary Materials**), which explains the observed overall improvement for perturbations.

**Soup construction: dependency on the validation set**

As the greedy soup algorithm uses a separate validation set to add models to a soup, much depends on the size and variation contained within this validation set. By artificially increasing the size and variation through various augmentations, we attempt to address this limitation (i.e., greedy cropped, greedy data augmentation). This is only somewhat successful, as when looking at the holdout, external and robustness test sets, performance never consistently increases across all sets or individual architectures.

Next, we tried to increase the soup diversity by including more models. Instead of ranking the models from best to worst, we simply reverse the order (i.e., greedy inverse, greedy cropped inverse, greedy data augmentation inverse). As expected, the number of models included in the soup increases. This results in a slight drop in holdout performance, but improves average performance on the external, corrupted and perturbed test set. A downside of the greedy algorithm seems to be the relatively small number of models that are included in the soup. Reversing the ranked order circumvents this problem, but at the cost, that holdout performance is not guaranteed to be equal to the baseline anymore. However, when looking at the performance of the individual

architectures (**Supplementary Materials**), the observed overall improvement is not consistent across all architectures.

Thus, while greedy soups are dependent on a separate validation set, we only find negligible performance differences when artificially changing the size and variation of this validation set or the model ranking order. We therefore believe the standard souping algorithm to be relatively effective at soup construction and that already small holdout validation sets are sufficient.

**Limitations**

The model pool size consists of only seven models, which is in contrast to the pool sizes in *Wortsman et al.* [129], which range from 12 to 72. As we only vary image size to mimic a multi-resolution classifier similar to *Ha et al.* [127], we are constrained by the highest native resolution that is available online (600x450 pixels for HAM10000). However, by keeping the model pool small, we show that the advantages of souping even manifest themselves when the choice of models is limited.

### 2.5.6 Conclusions

Model soups are able to improve external performance for a skin cancer classification task, while retaining on par holdout performance. In addition, model souping has a positive effect on the robustness and calibration of skin cancer classifiers. While multi-model/ensemble solutions still perform better, soups are single model solutions. This distinction might be highly relevant in a clinical setting, as less complexity is usually easier to interpret.

### 2.5.7 Acknowledgement

**Role of the funding source**

This study was funded by the Federal Ministry of Health, Berlin, Germany (grant: Skin Classification Project 2; grant holder: Titus J. Brinker, 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.

**Author contribution statement**

Roman C. Maron: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - Original Draft, Validation, Visualization. Achim Hekler: Conceptualization, Methodology, Writing - Review & Editing, Project administration, Supervision, Visualization. Sarah Haggenmüller: Validation, Writing - Review & Editing, Visualization. Christof von Kalle: Resources, Writing - Review & Editing. Jochen S. Utikal: Resources, Writing - Review & Editing. Verena Müller: Resources, Writing - Review & Editing. Maria Gaiser: Resources, Writing - Review & Editing. Friedegund Meier: Resources, Writing - Review & Editing. Sarah Hobelsberger: Resources, Writing - Review & Editing. Frank F. Gellrich: Resources, Writing - Review & Editing. Mildred Sergon: Resources, Writing - Review & Editing. Axel Hauschild: Resources, Writing - Review & Editing. Lars E. French: Resources, Writing - Review & Editing. Lucie Heinzerling: Resources, Writing - Review & Editing. Justin G. Schlager: Resources, Writing - Review & Editing. Kamran Ghoreschi: Resources, Writing - Review & Editing. Max Schlaak: Resources, Writing - Review & Editing. Franz J. Hilke: Resources, Writing - Review & Editing. Gabriela Poch: Resources, Writing - Review & Editing. Sören

Korsing: Resources, Writing - Review & Editing. Carola Berking: Resources, Writing - Review & Editing. Markus V. Heppt: Resources, Writing - Review & Editing. Michael Erdmann: Resources, Writing - Review & Editing. Sebastian Haferkamp: Resources, Writing - Review & Editing. Dirk Schadendorf: Resources, Writing - Review & Editing. Wiebke Sondermann: Resources, Writing - Review & Editing. Matthias Goebeler: Resources, Writing - Review & Editing. Bastian Schilling: Resources, Writing - Review & Editing. Jakob N. Kather: Resources, Writing - Review & Editing. Stefan Fröhling: Resources, Writing - Review & Editing. Daniel B. Lipka: Resources, Writing - Review & Editing. Eva Krieghoff-Henning: Conceptualization, Writing - Review & Editing, Project administration, Supervision. Titus J. Brinker: Conceptualization, Writing - Review & Editing, Project administration, Supervision, Funding acquisition.

**Conflict of interest statement**

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

JSU is on the advisory board or has received honoraria and travel support from Amgen, Bristol Myers Squibb, GSK, Immunocore, LeoPharma, Merck Sharp and Dohme, Novartis, Pierre Fabre, Roche, outside the submitted work. FM has received travel support or/and speaker's fees or/and advisor's honoraria by Novartis, Roche, BMS, MSD and Pierre Fabre and research funding from Novartis and Roche. SH reports advisory roles for or has received honoraria from Pierre Fabre Pharmaceuticals, Novartis, Roche, BMS, Amgen and MSD outside the submitted work. Axel H. reports clinical trial support, speaker's honoraria, or consultancy fees from the following companies: Amgen, BMS, Merck Serono, MSD, Novartis, Oncosec, Philogen, Pierre Fabre, Provectus, Regeneron, Roche, OncoSec, Sanofi-Genzyme, and Sun Pharma, outside,

the submitted work. LF is on the advisory board or has received consulting/speaker honoraria from for Galderma, Janssen, Leo Pharma, Eli Lilly, Almirall, Union Therapeutics, Regeneron, Novartis, Amgen, Abbvie, UCB, Biotest, and InflaRx. MS reports advisory roles for Bristol-Myers Squibb, Novartis, MSD, Roche, Pierre Fabre, Kyowa Kirin, Immunocore and Sanofi-Genzyme. WS reports grants, speaker's honoraria or consultancy fees from medi GmbH Bayreuth, Abbvie, Almirall, Amgen, Bristol-Myers Squibb, Celgene, GSK, Janssen, LEO Pharma, Lilly, MSD, Novartis, Pfizer, Roche, Sanofi Genzyme and UCB outside the submitted work. BS reports advisory roles for or has received honoraria from Pierre Fabre Pharmaceuticals, Incyte, Novartis, Roche, BMS and MSD, research funding from BMS, Pierre Fabre Pharmaceuticals and MSD, and travel support from Novartis, Roche, BMS, Pierre Fabre Pharmaceuticals and Amgen; outside the submitted work. MG 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. TJB is the owner of Smart Health Heidelberg GmbH (Handschuhsheimer Landstr. 9/1, 69120 Heidelberg, Germany, https://smarthealth.de) which develops telemedicine mobile apps (such as AppDoc; https://online-hautarzt.net and Intimarzt; https://intimarzt.de), outside of the submitted work. 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.

### 2.5.8 Supplementary materials

Supplementary Materials are available at https://doi.org/10.1016/j.ejca.2022.07.002

## 2.6 Supplementary publication 3: Using mutiple real-world dermoscopic photographs of one lesion improves melanoma classification via deep learning

Achim Hekler, MSc[†,1], Roman C. Maron, MSc[†,1], Sarah Haggenmüller, MSc[1], Max Schmitt, MSc[1], Christoph Wies, MSc[1,15], Jochen S. Utikal, MD[2,3,4], Friedegund Meier, MD[5], Sarah Hobelsberger, MD[5], Frank F. Gellrich, MD[5], Mildred Sergon, MD[5], Axel Hauschild, MD[6], Lars E. French, MD[7,8], Lucie Heinzerling, MD[7,10], Justin G. Schlager, MD[7], Kamran Ghoreschi, MD[9], Max Schlaak, MD[9], Franz J. Hilke, PhD[9], Gabriela Poch, MD[9], Sören Korsing, MD[9], Carola Berking, MD[10], Markus V. Heppt, MD[10], Michael Erdmann, MD[10], Sebastian Haferkamp, MD[11], Konstantin Drexler, MD[11], Dirk Schadendorf, MD[12], Wiebke Sondermann, MD[12], Matthias Goebeler, MD[13], Bastian Schilling, MD[13], Jakob N. Kather, MD[14], Eva Krieghoff-Henning, PhD[1], Titus J. Brinker, MD[1,*]

[†]These authors contributed equally to this work.

[1] Digital Biomarkers for Oncology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany

[2] Department of Dermatology, Venereology and Allergology, University Medical Center Mannheim, Ruprecht-Karl University of Heidelberg, Mannheim, Germany

[3] Skin Cancer Unit, German Cancer Research Center (DKFZ), Heidelberg, Germany

[4] DKFZ Hector Cancer Institute at the University Medical Center Mannheim, Mannheim, Germany

[5] Skin Cancer Center at the University Cancer Center and National Center for Tumor Diseases Dresden, Department of Dermatology, University Hospital Carl Gustav Carus, Technische Universität Dresden, Germany

[6] Department of Dermatology, University Hospital (UKSH), Kiel, Germany

[7] Department of Dermatology and Allergy, University Hospital, LMU Munich, Munich, Germany

[8] Dr. Phillip Frost Department of Dermatology and Cutaneous Surgery, University of Miami, Miller School of Medicine, Miami, FL, USA

[9] Department of Dermatology, Venereology and Allergology, Charité – Universitätsmedizin Berlin, Corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany

[10] Department of Dermatology, University Hospital Erlangen, Comprehensive Cancer Center Erlangen – European Metropolitan Region Nürnberg, CCC Alliance WERA, Erlangen, Germany

[11] Department of Dermatology, University Hospital Regensburg, Regensburg, Germany

[12] Department of Dermatology, Venereology and Allergology, University Hospital Essen, Essen, Germany

[13] Department of Dermatology, Venereology and Allergology, University Hospital Würzburg and National Center for Tumor Diseases (NCT) WERA Würzburg, Germany

[14] Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany

[15] Medical Faculty, University Heidelberg, Heidelberg, Germany

* Corresponding author: Digital Biomarkers for Oncology Group, 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).

### 2.6.1 Research letter

AI-based melanoma classifiers suffer from generalizability, robustness and uncertainty estimation issues, which severely limit their usefulness in clinical practice [40,42]. A common technique which addresses these limitations is to provide the classifier with multiple views of the same lesion, which are normally created by digitally transforming the original image (a technique referred to as test-time augmentation [143]). While this artificial process has proven to be effective [61,144], the optimal parameters may vary across different classifiers and domains.

In this report, we therefore investigate if this artificial multi-view approach can be improved when the artificial images are substituted by multiple real-world images of the same lesion. To this end, we evaluate the performance of a dermoscopic image classifier for a single-view scenario using one image (Single-View), a multi-view scenario using multiple artificially-modified images per lesion (MV-Artificial) and our proposed multi-view scenario with multiple real-world images per lesion (MV-Real). To ensure clinical relevance, our analysis focuses on diagnostic accuracy, uncertainty estimation and robustness, which we measure using the AUROC, expected calibration error (ECE) and maximum confidence change (MCC), respectively. See **Supplementary**

**Methods** in the **Supplementary Materials** for details on study design, classifier implementation and statistical evaluation.

We evaluated our classifier on a prospective multi-center test set of 293 MM and 363 melanocytic nevi from 617 patients (see **Table 9** for patient characteristics). Our proposed approach (MV-Real) showed a significantly higher diagnostic accuracy (0.930; 95% CI, 0.909-0.951) compared to the Single-View (0.905; 95% CI, 0.879-0.929; $p<0.001$) and MV-Artificial approach (0.929; 95% CI: 0.908-0.948; $p<0.001$) (see **Figure 15**). While the pronounced numeric difference between our MV-Real and the Single-View approach seems clinically relevant, the relatively small improvement between the MV-Real and the MV-Artificial approach indicates no practical difference (see **Supplementary Analysis I** in the **Supplementary Materials**). However, our approach (MV-Real) showed a substantially better performance in uncertainty estimation and robustness as indicated by significantly lower ECE and MCC scores, respectively (see **Supplementary Results I and II** in the **Supplementary Materials** for details).

These findings indicate that using multiple real-world images of a lesion improves the overall performance of an AI-based melanoma classifier compared to traditional approaches. While these results were somewhat expected based on previous studies on test-time augmentation [144], the substantial outperformance of real-world versus artificial images with regard to robustness and uncertainty estimation is notable. It highlights the importance of using actual photographs for future multi-view approaches as this presumably results in a richer representation of the lesion (e.g., through different camera angles or dermoscopy modes).

Future work should investigate how the multi-view approach with real-world images can be optimized to reduce the physician's workload (for example recording a video sequence of the suspicious lesion).

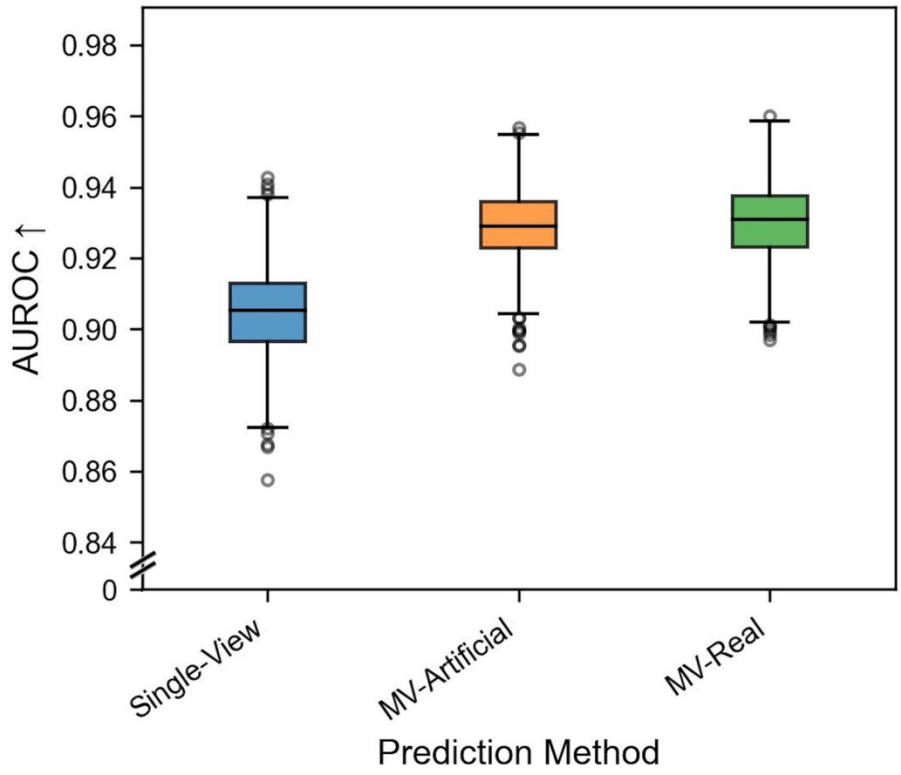
Altogether our proposed approach only requires additional photographs, is easy-to-implement and cost-effective. We therefore recommend integrating it into future clinical workflows, which make use of AI-based computer vision.

**Table 9. Patient characteristics of the study sample.** Distributions of the age at diagnosis, lesion location and lesion diameter are reported.

| | Melanoma[a] | Melanocytic nevus |
|---|---|---|
| Patient age at diagnosis (in years) | n=293 | n=363[b] |
| <35 | 7 (2.4%) | 82 (22.6%) |
| 35-54 | 47 (16.0%) | 124 (34.2%) |
| 55-74 | 124 (42.3%) | 105 (28.9%) |
| >74 | 115 (39.2%) | 52 (14.3%) |
| Lesion location | | |
| Palms/soles | 7 (2.4%) | 11 (3.0%) |
| Face/scalp/neck | 65 (22.2%) | 22 (6.1%) |
| Upper extremities | 54 (18.4%) | 36 (9.9%) |
| Lower extremities | 52 (17.7%) | 83 (22.9%) |
| Back | 72 (24.6%) | 120 (33.1%) |
| Abdomen | 17 (5.8%) | 37 (10.2%) |
| Chest | 20 (6.8%) | 40 (11.0%) |
| Buttocks | 2 (0.7%) | 9 (2.5%) |
| Genitalia | 2 (0.7%) | 4 (1.1%) |
| Unknown | 2 (0.7%) | 1 (0.3%) |
| Lesion diameter (in mm) | | |
| ≤ 3.00 | 11 (3.8%) | 63 (17.4%) |
| 3.01 to 6.00 | 27 (9.2%) | 137 (37.7%) |
| 6.01 to 9.00 | 26 (8.9%) | 66 (18.2%) |
| 9.01 to 12.00 | 60 (20.5%) | 52 (14.3%) |
| 12.01 to 15.00 | 46 (15.7%) | 18 (5.0%) |
| 15.01 to 18.00 | 13 (4.4%) | 4 (1.1%) |
| 18.01 to 21.00 | 35 (11.9%) | 8 (2.2%) |
| > 21 | 75 (25.6%) | 15 (4.1%) |

a. Including in situ tumors. b. Consisting of n=163 dysplastic and n=200 non-dysplastic nevi

**Figure 15. MV-Real outperforms both Single-View and MV-Artificial with respect to diagnostic accuracy.** The AUROC is plotted for the three investigated methods. Each box extends from the lower to the upper quartile of the 1000 bootstrap iterations, with a line at the median. In addition, whiskers and fliers indicate the range and any outliers. AUROC: area under the receiver operating characteristic curve, MV-Artificial: multiview-artificial, MV-Real: multiview-real.




### 2.6.2 Acknowledgement

**Role of the funding source**

This study was funded by the Federal Ministry of Health, Berlin, Germany (grant: Skin Classification Project 2; grant holder: Titus J. Brinker, 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.

**Author contribution statement**

Achim Hekler: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - Original Draft, Validation, Visualization, Writing - Review & Editing, Project administration, Supervision. Roman C. Maron: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - Original Draft, Validation, Visualization. Sarah Haggenmüller: Conceptualization, Data Curation, Validation, Writing - Review & Editing, Visualization, Ethics Approval. Max Schmitt: Validation, Writing - Review & Editing. Christoph Wies: Validation, Writing - Review & Editing. Jochen S. Utikal: Resources, Writing - Review & Editing. Friedegund Meier: Resources, Writing - Review & Editing. Sarah Hobelsberger: Resources, Writing - Review & Editing. Frank F. Gellrich: Resources, Writing - Review & Editing. Mildred Sergon: Resources, Writing - Review & Editing. Axel Hauschild: Resources, Writing - Review & Editing. Lars E. French: Resources, Writing - Review & Editing. Lucie Heinzerling: Resources, Writing - Review & Editing. Justin G. Schlager: Resources, Writing - Review & Editing. Kamran Ghoreschi: Resources, Writing - Review & Editing. Max Schlaak: Resources, Writing - Review & Editing. Franz J. Hilke: Resources, Writing - Review & Editing. Gabriela Poch: Resources, Writing - Review & Editing. Sören Korsing: Resources, Writing - Review & Editing. Carola Berking: Resources, Writing - Review & Editing. Markus V. Heppt: Resources, Writing - Review & Editing. Michael Erdmann: Resources, Writing - Review & Editing. Sebastian Haferkamp: Resources, Writing - Review & Editing. Konstantin Drexler: Resources, Writing - Review & Editing. Dirk Schadendorf: Resources, Writing - Review & Editing. Wiebke Sondermann: Resources, Writing - Review & Editing. Matthias Goebeler: Resources, Writing - Review & Editing. Bastian Schilling: Resources, Writing - Review & Editing. Jakob N. Kather: Writing - Review & Editing. Eva Krieghoff-Henning: Conceptualization, Writing - Review & Editing, Project

administration, Supervision. Titus J. Brinker: Conceptualization, Writing - Review & Editing, Project administration, Supervision, Funding acquisition.

**Conflict of interest statement**

Jochen S. Utikal is on the advisory board or has received honoraria and travel support from Amgen, Bristol Myers Squibb, GSK, Immunocore, LeoPharma, Merck Sharp and Dohme, Novartis, Pierre Fabre, Roche and Sanofi outside the submitted work. Friedegund Meier has received travel support and/or speaker's fees and/or advisor's honoraria by Novartis, Roche, BMS, MSD and Pierre Fabre and research funding from Novartis and Roche. Sarah Hobelsberger reports clinical trial support from Almirall and speaker’s honoraria from Almirall, UCB and AbbVie and has received travel support from the following companies: UCB, Janssen Cilag, Almirall, Novartis, Lilly, LEO Pharma and AbbVie outside the submitted work. Sebastian Haferkamp reports advisory roles for or has received honoraria from Pierre Fabre Pharmaceuticals, Novartis, Roche, BMS, Amgen and MSD outside the submitted work. Konstantin Drexler has received honoraria from Pierre Fabre Pharmaceuticals and Novartis. Axel Hauschild reports clinical trial support, speaker's honoraria, or consultancy fees from the following companies: Agenus, Amgen, BMS, Dermagnostix, Highlight Therapeutics, Immunocore, Incyte, IO Biotech, MerckPfizer, MSD, NercaCare, Novartis, Philogen, Pierre Fabre, Regeneron, Roche, Sanofi-Genzyme, Seagen, Sun Pharma and Xenthera outside the submitted work. Lars E. French is on the advisory board or has received consulting/speaker honoraria from Galderma, Janssen, Leo Pharma, Eli Lilly, Almirall, Union Therapeutics, Regeneron, Novartis, Amgen, AbbVie, UCB, Biotest and InflaRx. Max Schlaak reports advisory roles for Bristol-Myers Squibb, Novartis, MSD, Roche, Pierre Fabre, Kyowa Kirin, Immunocore and Sanofi-Genzyme. Wiebke Sondermann reports grants, speaker's honoraria, or consultancy fees from medi GmbH Bayreuth, AbbVie,

Almirall, Amgen, Bristol-Myers Squibb, Celgene, GSK, Janssen, LEO Pharma, Lilly, MSD, Novartis, Pfizer, Roche, Sanofi Genzyme and UCB outside the submitted work. Bastian Schilling reports advisory roles for or has received honoraria from Pierre Fabre Pharmaceuticals, Incyte, Novartis, Roche, BMS and MSD, research funding from BMS, Pierre Fabre Pharmaceuticals and MSD and travel support from Novartis, Roche, BMS, Pierre Fabre Pharmaceuticals and Amgen outside the submitted work. Matthias Goebeler 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. Michael Erdmann declares honoraria and travel support from Bristol-Meyers Squibb, Immunocore and Novartis outside the submitted work. Jakob N. Kather reports consulting services for Owkin, France, Panakeia, UK and DoMore Diagnostics, Norway and has received honoraria for lectures by MSD, Eisai and Fresenius. Titus J. Brinker reports owning a company that develops mobile apps (Smart Health Heidelberg GmbH, Handschuhsheimer Landstr. 9/1, 69120 Heidelberg). 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.

### 2.6.3 Supplementary materials

This paragraph contains the following sections:

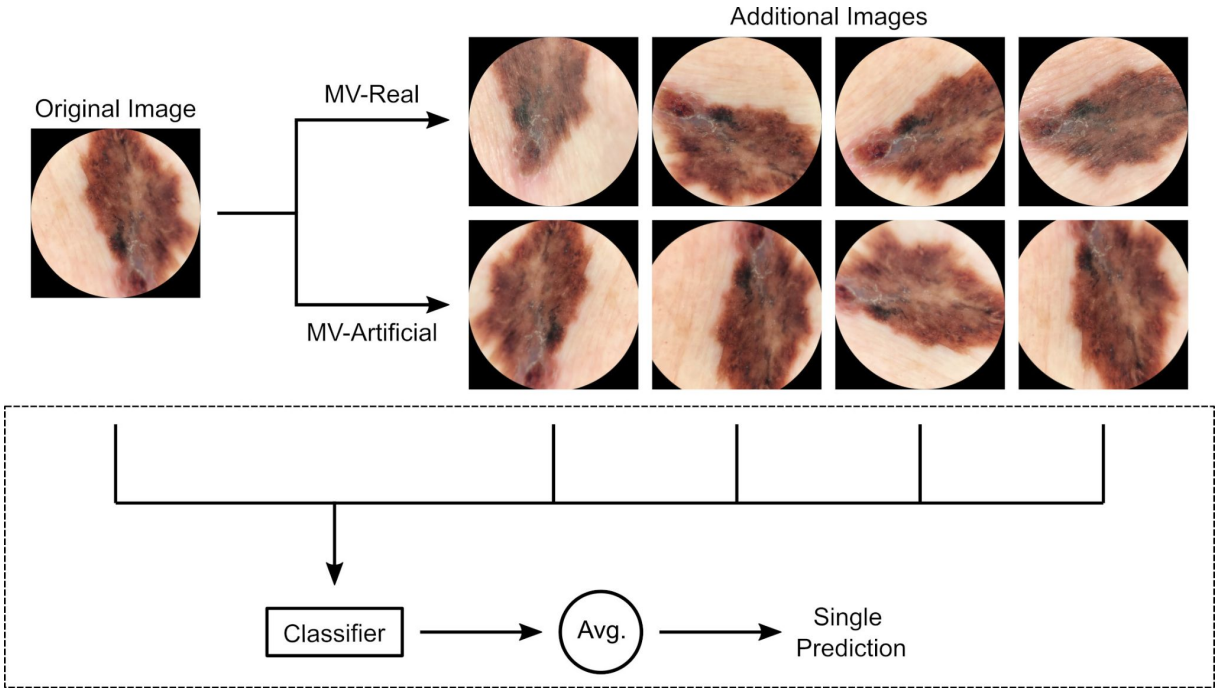
1. **Supplementary methods** where we provide a detailed description of the study design, implementation and statistical analysis.
2. **Supplementary results** where we describe in detail the results presented in the research letter.
3. **Supplementary analysis** where we present the results of additional subanalyses that did not fit within the scope of the research letter.

**Supplementary Methods**

**Study Design**

We trained a binary MM-nevus classifier on publicly available dermoscopic images and evaluated its performance on an externally collected multi-center dataset (referred to as SCP2) with respect to three clinically relevant endpoints: diagnostic accuracy, uncertainty estimation and robustness. For model prediction, we evaluated three different methods. The first method, called Single-View, represents the baseline scenario in which only one “original” image per lesion is available and the prediction is performed from that image. For the second method, referred to as multiview-artificial (MV-Artificial), the ‘original’ image is accompanied by artificially modified duplicates generated by applying various image processing techniques such as rotation, zoom and brightness to the ‘original’ image. For the final method, referred to as multiview-real (MV-Real), the ‘original’ image is accompanied by multiple real-world images (i.e., photographs taken in the clinic). At test-time, the model therefore provides a prediction for every single image, which are subsequently combined into an overall prediction (see **Supplementary Figure 12**).

**Supplementary Figure 12. Illustration of the multiview approach.** Top) For both the MV-Artificial and MV-Real methods, the model makes its final classification based on one original image accompanied by additional lesion images. For MV-Real, the additional images are actual dermoscopic photographs taken in the clinic by the physician (top row). For MV-Artificial, the additional images are artificially created from the original image by applying various image processing techniques such as rotation, zoom and brightness. Bottom) For both approaches, the classifier makes a prediction for the original and each of the additional images. All predictions are subsequently averaged into a single prediction. MV-Artificial: multiview-artificial, MV-Real: multiview-real.




The setup described above is feasible because the SCP2 dataset contains six real-world images (i.e., actual photographs) per lesion. To ensure that the comparisons and statistical tests in this study were based on the same test data for all three methods, we randomly sampled one image per lesion and labeled this image as the 'original' image (referred to as downsampling step). The remaining five images were set aside. Thus, all three prediction methods were evaluated on the same test set, with each image corresponding to a unique lesion. During test-time, the Single-View method received no further images, while MV-Artificial and MV-Real each received five additional images (artificially modified duplicates and real-world images, respectively).

### Study participants

Participants were required to be at least 18 years old and have MM-suspicious skin lesions that were excised following dermoscopic examination. The suspicious lesions should not have been previously pre-biopsied nor located near the eye or under the fingernails or toenails. Additionally, due to data privacy concerns, lesions with person-identifying features (e.g., tattoos) in their immediate vicinity were excluded from the study. All lesions were histopathologically confirmed by at least one reference pathologist at the corresponding clinic as part of routine clinical practice. In the end, only histopathologically verified MM or melanocytic nevi, recorded until October 2022, were included in this study.

### SCP2 dataset

Dermoscopic images and patient metadata (e.g., age, Fitzpatrick skin type, lesion localization and diameter) of clinically suspected MM were prospectively collected from eight university hospitals in Germany (Berlin, Dresden, Erlangen, Essen, Mannheim, Munich, Regensburg, Wuerzburg) between April 2021 and October 2022 during routine clinical care. For each lesion, a dermatologist captured six dermoscopic images during clinical examination while randomly varying the orientation/angle, position and mode of the dermatoscope (i.e., polarized or non-polarized). To minimize the effect of confounding factors, dermatologists were instructed to avoid well known artifacts (e.g., skin markings). The four hardware settings across the clinics were as follows:

- HEINE Delta30 dermatoscope with an Apple iPhone 7
- HEINE DELTAone dermatoscope with an Apple iPhone SE
- HEINE DELTAone dermatoscope with an Apple iPhone8
- HEINE IC1 dermatoscope with an Apple iPhone7

The original images were automatically cropped to exclude large parts of the black image margin which originates from dermoscopy and subsequently resized to 300x300 pixels for model training and inference (see section below). We retrospectively excluded all lesions which were histopathologically not diagnosed as MM or melanocytic nevus.

**Model training and evaluation**

We trained a CNN with a state-of-the-art ConvNeXT architecture with publicly available MM and nevus images from two well-established datasets, HAM10000 and BCN20000, containing 29,562 images (7,794 MM and 21,768 nevi). To optimize the hyperparameters of the training process, we employed a five-fold cross-validation procedure, using 20% of the training data for validation in each fold. The model architecture, the number of training epochs, the image size as well as the learning rate were optimized by maximizing the AUROC using Optuna 2.10.0. After determining the optimal hyperparameters, a final model was trained on all 29,562 images (i.e., inclusion of validation set).

At test-time, the Single-View, MV-Artificial and MV-Real approaches were used on the trained model, using the external SCP2 dataset for evaluation. Both training and inference were implemented using PyTorch 1.10.1, CUDA 11.0 and fastai 2.7.10.

**Implementation of MV-Artificial**

The MV-Artificial approach requires that the original image is duplicated *n* times and digitally modified before all images are classified by the model. In our case, the digital modifications consisted of rotation, zoom, changes in brightness and warp. Each of these modifications were applied to an image with a probability of 75%. The strength

of each modification varied as we considered five different setups: mild, moderate, strong, severe and extreme. The mild setup was considered the default setup and is simply referred to as MV-Artificial in the main manuscript. We used fastai's built-in test-time augmentation function with beta set to *None* as we wanted an unweighted average of all image predictions. The parameters for each setup are listed in **Supplementary Table 6** below.

**Supplementary Table 6. Overview about the parameters for each setup.**

| Parameter | Mild | Moderate | Strong | Severe | Extreme |
|---|---|---|---|---|---|
| flip_vert | True | True | True | True | True |
| max_rotate | 90 | 90 | 90 | 90 | 90 |
| max_zoom | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 |
| max_lightning | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 |
| max_warp | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 |
| pad_mode | zeros | zeros | zeros | zeros | zeros |

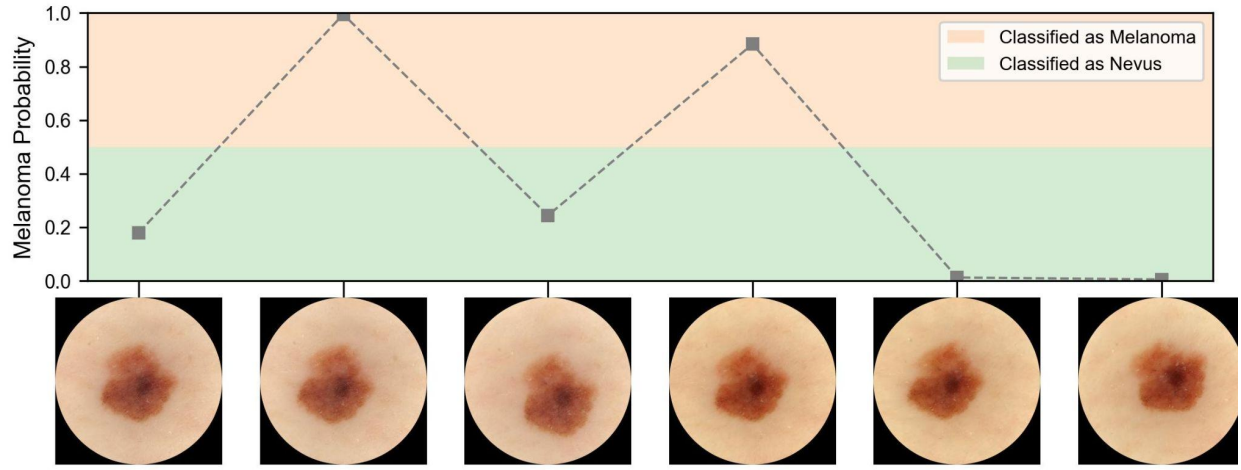
### Statistical analysis

The performance of our classifier was assessed based on three endpoints: diagnostic accuracy, uncertainty estimation and robustness. Diagnostic accuracy was measured using the AUROC, while uncertainty estimation was quantified by the ECE. A well-calibrated CNN ensures that the predicted probabilities accurately reflect the true likelihoods. For instance, if we consider all images where the model predicts a MM score of 0.6, we would expect 60% of them to be actual MM. The ECE is computed as the average difference between the predicted probabilities and the observed outcomes, with lower values indicating better-calibrated predictions.

Robustness was evaluated by analyzing the consistency of the classifier's predictions across a series of images per lesion, detecting fluctuations in the model's diagnosis

(see **Supplementary Figure 13**). We therefore computed the mean MCC, which measures the difference between the model's highest and lowest confidence scores for a series of images. Larger MMC values are worse, as the model's predictions are less consistent. As analyzing robustness requires a series of images per lesion across which to measure fluctuations, we constructed image series of either two or three images per lesion, by using the five additional images which were previously set aside during the downsampling step. However, this meant we also had to reduce the number of images used for MV-Real to three and two images respectively. To keep the comparison fair, MV-Artificial was adjusted accordingly.

**Supplementary Figure 13. Illustration of how small image changes can cause robustness issues.** The CNN-based algorithm developed in this study classifies multiple images of the same lesion, obtained from our prospective study, as either MM or nevus (as indicated by fluctuations in melanoma probability).




To reduce the impact of stochastic events, mean values for each metric were calculated using 1000 bootstrap iterations on our test sets. The corresponding 95% CIs were determined using the non-parametric percentile method. Statistical testing was conducted for all three hypotheses to identify significant differences between results with our proposed technique (i.e., MV-Real) and those with either the baseline (i.e., Single-View) or the traditional multiview technique (i.e., MV-Artificial). For each endpoint,

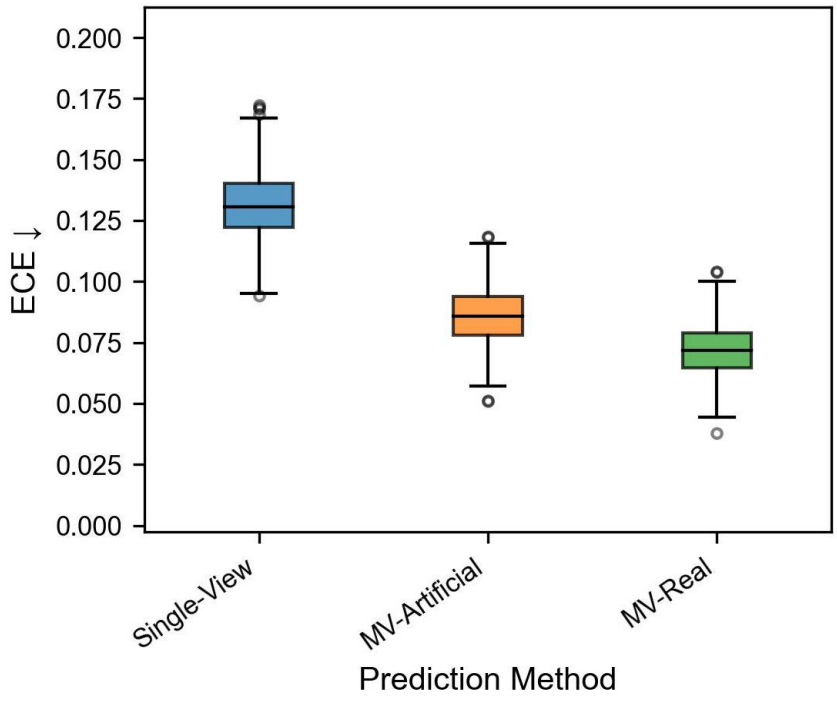
pairwise Wilcoxon signed-rank tests were used to compare the respective metrics. Significance levels of $p<0.05$ were adjusted to 0.025 according to the Bonferroni correction (m=2) which equals the expected false discovery rate. In addition, we repeated the downsampling step and all subsequent analysis steps five times in order to ensure that our findings were not based on an unfavorable sample. Statistical analysis was performed using SciPy 1.7.1

## Supplementary Results

### Part I: Uncertainty estimation

Our approach (MV-Real) showed a significantly lower ECE of 0.072 (95% CI: 0.052-0.093) than the Single-view (0.131; 95% CI, 0.105-0.159; $p<0.001$) or MV-Artificial (0.086; 95% CI: 0.064-0.110; $p<0.001$, see **Supplementary Figure 14**).

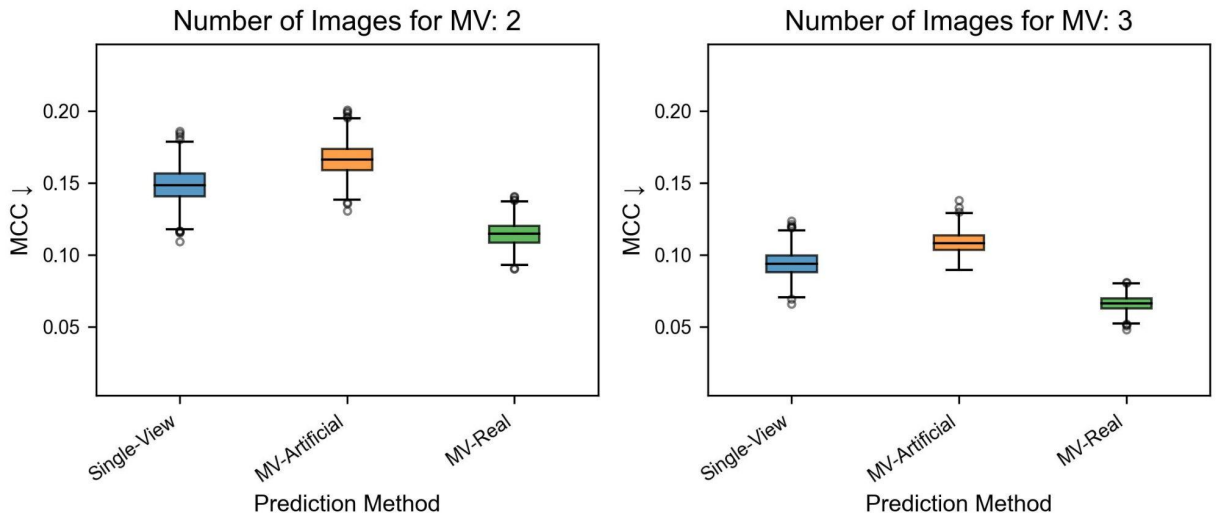
**Supplementary Figure 14. MV-Real outperforms both Single-View and MV-Artificial with respect to the uncertainty estimation.** The ECE is plotted for the three investigated methods. Each box extends from the lower to the upper quartile of the 1000 bootstrap iterations, with a line at the median. In addition, whiskers and fliers indicate the range and any outliers. ECE: expected calibration error, MV-Artificial: multiview-artificial, MV-Real: multiview-real.



### Part II: Robustness

The robustness of our classifier was analyzed across a series of either two or three images per lesion. For the series of three images, the robustness of our approach (i.e., MV-Real) improved substantially over that with single-view, as the MCC significantly decreased from 0.149 (95% CI, 0.125-0.171) to 0.115 (95% CI: 0.099-0.131; p<0.001), respectively. Similarly, robustness also improved across a series of two images, as the MMC significantly decreased from 0.094 (95% CI, 0.077-0.112) for single-view to 0.066 (95% CI: 0.056-0.076; p<0.001) for MV-Real. Surprisingly, the MV-Artificial method resulted in no robustness improvement at all, having greater MCC values than the Single-view and MV-Real approaches (see **Supplementary Figure 15**).

**Supplementary Figure 15. MV-Real outperforms both Single-View and MV-Artificial with respect to robustness.** Robustness was measured by the maximum change in the classifier's confidence (MCC) across a series of either three (left) or two (right) images. Each box extends from the lower to the upper quartile of the 1000 bootstrap iterations, with a line at the median. In addition, whiskers and fliers indicate the range and any outliers. MV-Artificial: multiview-artificial, MV-Real: multiview-real.



## Supplementary Analysis

### Part I: Replication of performance results for different test set samples

As mentioned in the **Statistical analysis** section, we repeated the downsampling step and all subsequent analysis steps five times in order to ensure that our findings were not based on an unfavorable sample.

We find that all of our reported findings were consistent across the five repeated down-samplings (see **Supplementary Tables 7a to 7e**) except for the diagnostic accuracy between MV-Artificial and MV-Real. Here, MV-Real is sometimes better, sometimes on-par and sometimes worse than MV-Artificial, indicating that there is no real practical difference between both approaches (regarding diagnostic accuracy).

**Supplementary Table 7a.**

| Metric | Single-View | MV-Artificial | MV-Real |
|---|---|---|---|
| AUROC ↑ | 0.916 (95% CI: 0.892-0.939) (p<0.001) | **0.930** (95% CI: 0.909-0.948) (p=0.003) | **0.930** (95% CI: 0.909-0.951) |
| ECE ↓ | 0.127 (95% CI: 0.102-0.151) (p<0.001) | 0.087 (95% CI: 0.066-0.110) (p<0.001) | **0.072** (95% CI: 0.052-0.093) |
| MCC (# images: 2) ↓ | 0.145 (95% CI: 0.123-0.168) (p<0.001) | 0.158 (95% CI: 0.137-0.179) (p<0.001) | **0.117** (95% CI: 0.100-0.133) |
| MCC (# images: 3) ↓ | 0.106 (95% CI: 0.087-0.125) (p<0.001) | 0.104 (95% CI: 0.090-0.120) (p<0.001) | **0.069** (95% CI: 0.058-0.080) |

**Supplementary Table 7b.**

| Metric | Single-View | MV-Artificial | MV-Real |
|---|---|---|---|
| AUROC ↑ | 0.909 (95% CI: 0.885-0.931) (p<0.001) | 0.927 (95% CI: 0.906-0.946) (p<0.001) | **0.930** (95% CI: 0.909-0.951) |
| ECE ↓ | 0.132 (95% CI: 0.107-0.158) (p<0.001) | 0.087 (95% CI: 0.066-0.109) (p<0.001) | **0.072** (95% CI: 0.052-0.093) |
| MCC (# images: 2) ↓ | 0.149 (95% CI: 0.126-0.173) (p<0.001) | 0.160 (95% CI: 0.140-0.180) (p<0.001) | **0.118** (95% CI: 0.102-0.134) |
| MCC (# images: 3) ↓ | 0.101 (95% CI: 0.083-0.120) (p<0.001) | 0.102 (95% CI: 0.088-0.119) (p<0.001) | **0.069** (95% CI: 0.059-0.081) |

**Supplementary Table 7c.**

| Metric | Single-View | MV-Artificial | MV-Real |
|---|---|---|---|
| AUROC ↑ | 0.913 (95% CI: 0.889-0.935) (p<0.001) | **0.933** (95% CI: 0.912-0.950) (p<0.001) | 0.930 (95% CI: 0.909-0.951) |
| ECE ↓ | 0.137 (95% CI: 0.111-0.165) (p<0.001) | 0.083 (95% CI: 0.060-0.107) (p<0.001) | **0.072** (95% CI: 0.052-0.093) |
| MCC (# images: 2) ↓ | 0.152 (95% CI: 0.129-0.175) (p<0.001) | 0.164 (95% CI: 0.143-0.184) (p<0.001) | **0.120** (95% CI: 0.104-0.137) |
| MCC (# images: 3) ↓ | 0.104 (95% CI: 0.085-0.124) (p<0.001) | 0.110 (95% CI: 0.095-0.125) (p<0.001) | **0.076** (95% CI: 0.064-0.087) |

**Supplementary Table 7d.**

| Metric | Single-View | MV-Artificial | MV-Real |
|---|---|---|---|
| AUROC ↑ | 0.910 (95% CI: 0.884-0.934) (p<0.001) | **0.931** (95% CI: 0.912-0.949) (p=0.09) | 0.930 (95% CI: 0.909-0.951) |
| ECE ↓ | 0.127 (95% CI: 0.102-0.151) (p<0.001) | 0.086 (95% CI: 0.065-0.108) (p<0.001) | **0.072** (95% CI: 0.052-0.093) |
| MCC (# images: 2) ↓ | 0.143 (95% CI: 0.122-0.166) (p<0.001) | 0.153 (95% CI: 0.134-0.173) (p<0.001) | **0.119** (95% CI: 0.103-0.135) |
| MCC (# images: 3) ↓ | 0.093 (95% CI: 0.076-0.111) (p<0.001) | 0.099 (95% CI: 0.085-0.114) (p<0.001) | **0.068** (95% CI: 0.058-0.078) |

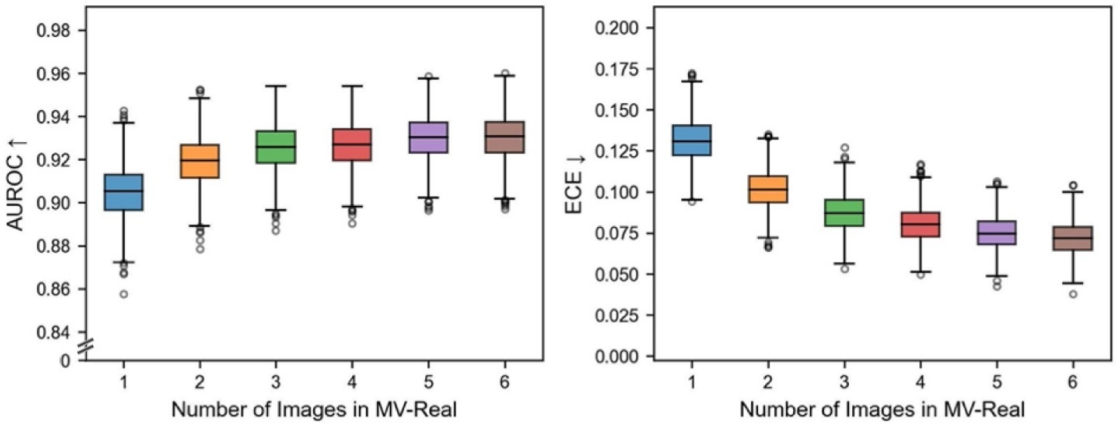
**Supplementary Table 7e.**

| Metric | Single-View | MV-Artificial | MV-Real |
|---|---|---|---|
| AUROC ↑ | 0.907 (95% CI: 0.882-0.931) (p<0.001) | **0.931** (95% CI: 0.911-0.949) (p=0.007) | 0.930 (95% CI: 0.909-0.951)) |
| ECE ↓ | 0.132 (95% CI: 0.107-0.157) (p<0.001) | 0.085 (95% CI: 0.064-0.108) (p<0.001) | **0.072** (95% CI: 0.052-0.093) |
| MCC (# images: 2) ↓ | 0.153 (95% CI: 0.131-0.176) (p<0.001) | 0.156 (95% CI: 0.135-0.177) (p<0.001) | **0.116** (95% CI: 0.101-0.133) |
| MCC (# images: 3) ↓ | 0.096 (95% CI: 0.079-0.114) (p<0.001) | 0.101 (95% CI: 0.086-0.117) (p<0.001) | **0.069** (95% CI: 0.059-0.080) |

### Part II: Influence of the number of images on MV-Real performance

While incorporating a single additional image into the classification already improves the diagnostic accuracy and uncertainty estimation of the classifier, the benefits of including additional images are even more pronounced as indicated by the trend in **Supplementary Figure 16**.

**Supplementary Figure 16. Increasing the number of images used for MV-Real improves diagnostic accuracy and uncertainty estimation.** The AUROC (diagnostic accuracy) and ECE (uncertainty estimation) are plotted for an increasing number of images used during MV-Real. Each box extends from the lower to the upper quartile of the 1000 bootstrap iterations, with a line at the median. In addition, whiskers and fliers indicate the range and any outliers. AUROC: area under the receiver operating characteristic curve, ECE: expected calibration error, MV-Real: multiview-real.



# 3 OVERALL DISCUSSION

The successful emergence of deep learning-based computer vision, combined with the increasing availability of publicly accessible skin cancer datasets (e.g., ISIC archive [145,146]), has led to the emergence of a plethora of comparative AI studies in the field of skin cancer detection. However, despite promising results, it has proven difficult to transfer these findings into clinical practice.

**Publication 1** analyzes the current state of AI research for skin cancer diagnostics, with a particular focus on studies comparing AI and human experts. It explores their actual impact and forthcoming challenges associated with the implementation of AI-systems in clinical routine by evaluating aspects such as the test setting (e.g., experimental or clinical, inclusion of metadata) or the test set characteristics (e.g., holdout or external testing). In summary, all 19 analyzed studies demonstrated at least equal AI performance compared to experienced clinicians. However, almost all studies were conducted in highly experimental settings, exclusively applying holdout testing (i.e., using unseen test data but from the same distribution, e.g., from institutions already involved in model training).

In real-world patient examinations, clinicians assess not only the skin lesion itself but take additional patient data (e.g., age, lesion localization) into consideration. Moreover, clinicians have the opportunity to review all lesions of a patient, aiming to identify the ‘ugly duckling’ as a reference point and rely on other senses (e.g., palpation). Remarkably, 18 out of the 19 analyzed comparison studies took place in highly experimental settings, where physicians were often only given a single image of the suspected skin lesion. Moreover, it is noteworthy that the vast majority of these studies recorded the

clinicians' diagnoses via online applications, thereby substantially differing from the decision-making process in clinical practice. Interestingly, in the one study where the diagnoses were recorded during clinical examination of the patient [19], the dermatologists significantly outperformed the AI algorithm. This finding highlights that no conclusions about the added value of AI for skin cancer diagnostics can be drawn solely based on experimental comparisons. Consequently, to enable meaningful insights about the practical use of AI-systems, there is a need for truly prospective studies comparing the clinicians' diagnoses in real-life patient examinations with the performance of AI-based diagnostic algorithms.

While several publicly available skin datasets exist (e.g., HAM10000, PH2 [110,111]), most comparisons between AI and human experts have not been evaluated using external test data (i.e., unseen test data from a different distribution, e.g., from an institution not involved in model training). This is particularly problematic, as previous studies have demonstrated that classifier performance usually decreases on test images from a different source than the training images [28]. Relying exclusively on holdout testing constrains the generalizability of the results, making it difficult to draw conclusions about the real-world added value of AI-systems in clinical practice. To provide more robust comparisons that account for the technical (e.g., acquisition systems, staining protocols) and/or biological (e.g., skin types, anatomical sites) heterogeneity present in clinical reality, studies employing external testing should be considered the gold standard for future research.

Moreover, the implementation of AI-systems for skin cancer diagnostics is slowed by restraint among patients and dermatologists. Therefore, **publication 2** is dedicated to investigate the criteria required for patients and dermatologists to accept AI-systems

for skin cancer diagnostics, and to assess their importance in patients' and dermatologists' decision-making process when considering the use of such systems.

While previous AI survey studies showed a positive general attitude toward AI-systems in dermatology [93,147–152], these studies have not evaluated the criteria in a multidimensional (i.e., considering multiple influencing criteria at a time) and indirect (i.e., without asking explicit and straightforward questions) manner. One-dimensional approaches may lack contextual understanding and be unable to capture complex motivations, while direct questioning can lead respondents to provide socially acceptable answers rather than express their true preferences. By employing an adaptive choice-based conjoint analysis, **publication 2** overcomes these limitations by allowing the investigation of multiple influencing criteria simultaneously and accounting for possible trade-offs from the patients' and dermatologists' perspective. In this context, participants prioritized AI-systems that go beyond diagnostic performance by providing detailed explanations that are understandable to both physicians and the patient, as well as supporting multiclass assessments instead of binary classifications (e.g., benign/malignant).

**Publication 2** found that patients demand AI-systems that provide an explanation that is comprehensive for both physicians and the patient. These findings indicate that patients may have higher expectations of AI-systems than they previously had of their treating dermatologist. While clinicians work with standardized classification frameworks, such as the ABCDE rule [8] or the 7-point checklist [153], the evaluation of skin lesions can still be subjective, particularly in borderline or atypical cases. In such instances, clinicians may have limited ability to explain their decision-making. Hence, it may no longer suffice to compete with the standard of care. Instead, it is crucial to

develop AI-systems that are traceable by patients and dermatologists to meet the evolving needs of patients.

Dermatology is a complex field, and in clinical reality, clinicians face the challenge of distinguishing between various skin diseases that share similar diagnostic features. In light of this complexity, the participating dermatologists called for AI-systems that support more refined multiclass assessments. Specifically, the binary differentiation between MM and atypical nevi, which has been the primary focus of AI research in dermatology (e.g., [31,52,53,55]), is considered insufficient from the dermatologists' perspective. Consequently, there is a need for prospective studies that evaluate the performance of AI-systems in multiclass assessments to provide a more accurate representation of clinically relevant differential diagnoses.

Current AI research predominantly focuses on diagnostic accuracy, with 'classical' metrics such as accuracy or AUROC [154,155] remaining the gold standard for comparisons studies (e.g., ISIC challenges [145,146]). However, patients and dermatologists require AI-systems that are able to explain the rationale behind their assessments and are at least somewhat understandable. Specifically, diagnostic accuracy and explainability were found to be the top influencing criteria for both patients and dermatologists, with explainability being of particular importance to patients. This growing demand for explainable AI presents a key challenge for future research, given that state-of-the-art technology does not fully explain the reasoning behind its decisions due to the AI black box phenomenon [65]. Bridging this gap necessitates the development of AI-systems that are as transparent/explainable as possible, without sacrificing diagnostic performance.

Another barrier for the successful implementation of AI-systems for skin cancer diagnostics are potential data protection concerns. The conventional development of AI-systems typically calls for large centralized datasets (known as centralized learning), requiring hospitals to transfer patient data to external institutions, often raising serious privacy concerns. To circumvent this problem, decentralized FL, where classifier development is distributed across institutions, was introduced [73,156].

**Publication 3** explores the potential of FL, as a potentially more accessible and privacy-preserving alternative, in comparison to centralized single model and ensemble (i.e., combining multiple model predictions) learning approaches. The study leverages prospective real-world distributed MM-suspicious lesion data for the binary classification of MM and nevi using histopathological whole-slide images. Altogether, FL achieved comparable performance levels to the centralized approaches, thus presenting a reliable alternative that may empower institutions to contribute to the development of AI-systems, even with limited datasets or strict data protection requirements.

While the centralized single model approach exhibited significantly better performance on the holdout test dataset (i.e., on unseen data from the same hospitals already involved in model training; AUROC of 0.9024 versus 0.8579), FL excelled with significantly better results on the external test dataset (i.e., on unseen data from another hospital not involved in model training; AUROC of 0.9126 versus 0.9045). These findings suggest that FL may not be as well suited for solving in-distribution classification problems (i.e., following the same distribution as the training data), as indicated by the inferior performance on the holdout test dataset. However, they highlight the potential benefits of FL in generalizing to out-of-distribution data, as indicated by the enhanced performance on the external test dataset (similar observations see [72,77]). This is

particularly noteworthy for the potential clinical adoption of AI-systems for skin cancer diagnostics, where dealing with biological (e.g., skin types, anatomical sites) and/or technical (e.g., acquisition systems, staining protocols) variability is inevitable.

Moreover, state-of-the-art AI algorithms for MM detection face several problems, including issues related to generalizability [28,44], robustness [40,41], and uncertainty estimation [42,43], which collectively limit their usefulness for clinical practice. Instead of learning valid decision rules (e.g., morphological characteristics) which generalize to out-of-distribution data (i.e., following a different distribution than the training data), AI algorithms often learn spurious correlations (referred to as shortcuts [101,157]) that are present in the dataset (e.g., skin markings [59] or scale bars [126]). This often leads to the development of non-generalizable or non-robust AI algorithms. Given the potentially severe consequences of lacking robustness and reliability in a clinical setting, **supplementary publications 1 to 3** explore various approaches for the development of potentially more robust diagnostic algorithms with improved generalization capabilities.

**Supplementary publication 1** investigates the brittleness (i.e., sensitivity to minor input changes, such as image rotations or scaling) of three commonly used CNN architectures (ResNet50, DenseNet121, VGG16) by evaluating their performance on images that have been modified in various ways, aiming to simulate alterations in image acquisition that may occur in clinical settings. Additionally, **supplementary publication 1** evaluates the effectiveness of three possible techniques (data augmentation, test-time augmentation [144], anti-aliased networks [106]) in addressing this issue. While all architectures exhibited brittleness on both artificial and naturally modified images, the reviewed techniques were able to reduce the brittleness to varying degrees,

especially when using artificial generated test images. These improvements were, however, less pronounced or non-existent for naturally modified images, highlighting the importance of non-artificial test datasets.

Another strategy to address these issues involves the use of ensemble solutions, where multiple model predictions are combined into one (e.g., [127]). Ensembles, however, are computationally intensive, especially for handheld devices or when remote server options are not available, and additionally may lack transparency for practitioners when compared to single model solutions. In light of this, **supplementary publication 2** presents an alternative approach that can achieve similar results while remaining straightforward to implement: the concept of constructing model soups. Here, the weights of multiple models are averaged into a single model, resulting in improved generalizability, robustness and uncertainty estimation, while still enhancing performance on the holdout test dataset. This could be a pivotal factor in achieving clinical applicability, as reduced complexity generally leads to greater transparency, thereby boosting patients' and clinicians' acceptance (see **publication 2**).

Another approach commonly applied to overcome the discussed challenges is to provide the classifier with multiple views of the same lesion that are typically generated by digitally transforming the original image (known as test-time augmentation [144]). **Supplementary publication 3** is dedicated to explore whether this approach can be further enhanced by substituting artificial images with multiple real-world images (i.e., photographs taken in the clinic) of the same lesion. In summary, the utilization of multiple non-artificial images per lesion yielded superior results in terms of uncertainty estimation and robustness when compared to conventional methods. This once again highlights the critical role of real-world test images (as already outlined in

**supplementary publication 1**). Actual photographs might offer a more comprehensive representation of skin lesions by capturing real-world alterations (e.g., diverse camera angles or dermoscopy modes) that are somewhat complementary to each other. Notably, including multiple real-world images is both cost-effective and easy to implement, making it a promising approach for future clinical workflows.

In summary, the integration of AI-systems into clinical routine faces challenges ranging from experimental settings that do not reflect real-world conditions to data protection concerns. Patients' and dermatologists' acceptance hinges on the demand for AI-systems to provide traceable explanations and refined multiclass assessments. FL emerges as a promising solution for data privacy, while addressing algorithmic challenges requires innovative strategies such as model soups and the incorporation of real-world test images. Consequently, the successful translation of AI-systems for skin cancer diagnostics requires a holistic approach, equally considering technological advancements, ethical considerations, and practical needs of patients and dermatologists.

# 4 SUMMARY

The present doctoral thesis aims to conduct a feasibility study on the use of AI-systems for skin cancer diagnostics to investigate the research question “How can an AI-based algorithm for image-based melanoma detection be successfully implemented in clinical practice?”.

**Publication 1** systematically analyzes the current state of AI research for skin cancer diagnostics, examining the potential clinical relevance of studies that directly compare AI performance with human experts. All 19 included comparison studies demonstrated superior or equivalent performance of AI. However, to enhance the reliability and generalizability of these results, **publication 1** calls for less artificial conditions and advocates for the use of external test data in classifier evaluation.

**Publication 2** presents the results of a prospective multicentric survey study that utilizes an adaptive choice-based conjoint analysis to investigate the criteria required for patients and dermatologists to accept AI-systems in skin cancer diagnostics. Overall, **publication 2** highlights that future AI research in dermatology must move beyond pure performance enhancements and shift its focus towards increased levels of explainability. Additionally, AI-systems that are understandable for both patients and clinicians, and are capable of differentiating among various skin disorders in a multiclass context are required to develop AI-systems that are tailored to patients’ and dermatologists’ needs, ultimately enhancing acceptance in clinical practice.

**Publication 3** develops a federated learning model for melanoma-nevus classification using prospectively-collected histopathological whole-slide images and compares its

diagnostic performance to classical centralized approaches (i.e., single model and ensemble). Altogether, **publication 3** demonstrates that federated learning presents a reliable alternative, attaining performance levels at least equivalent to centralized learning approaches while simultaneously portraying a more accessible and privacy-preserving option. Consequently, federated learning may empower institutions to contribute to the development of AI models, even with limited datasets or strict data protection rules, thereby encouraging collaboration across institutions and countries.

**Supplementary publications 1 to 3** explore various technical approaches for the development of potentially more robust diagnostic algorithms with improved generalization capabilities. In this context, particularly constructing model soups (i.e., averaging the weights of multiple models into a single model), as well as providing the classifier with multiple real-world images of the same lesion (i.e., photographs taken in the clinic), have proven to be efficient methods that positively impact the generalizability and robustness of AI-based MM detection while simultaneously being inexpensive and straightforward to implement.

Overall, the successful implementation of AI-systems for skin cancer diagnostics requires a holistic approach that equally takes into account technological advancements, ethical considerations and practical needs of patients and dermatologists.

# 5 ZUSAMMENFASSUNG

Ziel der vorliegenden Dissertation ist die Durchführung einer Machbarkeitsstudie zum Einsatz von auf künstlicher Intelligenz (KI)-basierenden Systemen in der Hautkrebsdiagnostik, um die Forschungsfrage „Wie kann ein KI-basierter Algorithmus zur bildbasierten Melanomerkennung erfolgreich in der klinischen Praxis implementiert werden?" zu untersuchen.

**Publikation 1** analysiert systematisch den aktuellen Stand der KI-Forschung im Bereich der Hautkrebsdiagnostik und untersucht die potenzielle klinische Relevanz von Studien, welche die KI-Leistung direkt mit menschlichen Experten vergleichen. Alle 19 einbezogenen Vergleichsstudien zeigten eine überlegene oder gleichwertige Leistung von KI. Um jedoch die Zuverlässigkeit sowie die Generalisierbarkeit dieser Ergebnisse zu erhöhen, fordert **Publikation 1** weniger künstliche Bedingungen und plädiert für die Verwendung von externen Testdaten bei der Evaluierung von Klassifikatoren.

**Publikation 2** präsentiert die Ergebnisse einer prospektiven multizentrischen Umfragestudie, die eine adaptive entscheidungsbasierte Conjoint-Analyse (*adaptive-choice based conjoint*) einsetzt, um die Kriterien zu untersuchen, die für Patient*innen und Dermatolog*innen erforderlich sind, um KI-Systeme in der Hautkrebsdiagnostik zu akzeptieren. Insgesamt hebt **Publikation 2** hervor, dass die zukünftige KI-Forschung in der Dermatologie über reine Leistungssteigerungen hinausgehen und sich verstärkt auf ein höheres Maß an Erklärbarkeit konzentrieren muss. Darüber hinaus sind KI-Systeme erforderlich, die sowohl für Patient*innen als auch für Kliniker*innen nachvollziehbar sind und zwischen verschiedenen Hauterkrankungen in einem Mehrklassen-Kontext unterscheiden können, um KI-Systeme zu entwickeln, die auf die Bedürfnisse

der Patient*innen und Dermatolog*innen zugeschnitten sind und somit letztendlich die Akzeptanz in der klinischen Praxis erhöhen.

**Publikation 3** entwickelt ein auf föderiertem Lernen (*federated learning*) basierendes Modell zur Klassifizierung von Melanomen und Nävi unter Verwendung von prospektiv gesammelten digitalisierten Gewebeschnitten (*whole-slide images*) und vergleicht dessen diagnostische Leistung mit klassischen Ansätzen des zentralisierten Lernens (d.h. Einzelmodell sowie Ensemble). Insgesamt zeigt **Publikation 3**, dass föderiertes Lernen (*federated learning*) eine verlässliche Alternative bietet, welche eine mindestens vergleichbare Leistung zu zentralisiert trainierten Modellen erzielt und gleichzeitig eine leichter zugängliche und datenschutzfreundliche Option darstellt. Folglich kann föderiertes Lernen (*federated learning*) Institutionen dazu befähigen, selbst mit begrenzten Datensätzen oder strengen Datenschutzvorschriften zur Entwicklung von KI-Modellen beizutragen und so die Zusammenarbeit zwischen Institutionen und Ländern fördern.

Die **ergänzenden Publikationen 1 bis 3** untersuchen verschiedene technische Ansätze für die Entwicklung von potenziell robusteren Diagnosealgorithmen mit verbesserten Generalisierungsfähigkeiten. In diesem Zusammenhang haben sich insbesondere *model soups* (d.h. die Mittelung der Gewichte mehrerer Modelle zu einem einzigen Modell) sowie die Bereitstellung von mehreren realen Bildern (d.h. in der Klinik aufgenommene Bildaufnahmen) derselben Läsion für die Klassifizierungentscheidung als effiziente Methoden erwiesen, die sich positiv auf die Generalisierbarkeit und Robustheit der KI-basierten Melanomerkennung auswirken und gleichzeitig kostengünstig und einfach zu implementieren sind.

Insgesamt erfordert die erfolgreiche Umsetzung von KI-Systemen für die Hautkrebsdiagnostik einen ganzheitlichen Ansatz, der technologische Fortschritte, ethische Gesichtspunkte und die praktischen Bedürfnisse von Patient*innen und Dermatolog*innen gleichermaßen berücksichtigt.

# 6 REFERENCES


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

[2] Arnold M, Singh D, Laversanne M, Vignat J, Vaccarella S, Meheus F, et al. Global Burden of Cutaneous Melanoma in 2020 and Projections to 2040. JAMA Dermatol 2022;158:495–503.

[3] Dzwierzynski WW. Melanoma Risk Factors and Prevention. Clin Plast Surg 2021;48:543–50.

[4] Savoye I, Olsen CM, Whiteman DC, Bijon A, Wald L, Dartois L, et al. Patterns of Ultraviolet Radiation Exposure and Skin Cancer Risk: the E3N-SunExp Study. J Epidemiol 2018;28:27–33.

[5] Cherobin ACFP, Wainstein AJA, Colosimo EA, Goulart EMA, Bittencourt FV. Prognostic factors for metastasis in cutaneous melanoma. An Bras Dermatol 2018;93:19–26.

[6] Han D, Zager JS, Shyr Y, Chen H, Berry LD, Iyengar S, et al. Clinicopathologic predictors of sentinel lymph node metastasis in thin melanoma. J Clin Oncol 2013;31:4387–93.

[7] Jitian (Mihulecea) C, Frățilă S, Rotaru M. Clinical-dermoscopic similarities between atypical nevi and early stage melanoma. Exp Ther Med 2021;22:1–5.

[8] Abbasi NR, Shaw HM, Rigel DS, Friedman RJ, McCarthy WH, Osman I, et al. Early diagnosis of cutaneous melanoma: revisiting the ABCD criteria. JAMA 2004;292:2771–6.

[9] 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.

[10] 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.

[11] Lüken F, Batz D, Kutschmann M. Evaluation der Screeninguntersuchungen auf Hautkrebs gemäß Krebsfrüherkennungs-Richtlinie des Gemeinsamen Bundesausschusses. Abschlussberichte 2011 – 2013. BQS. https://www.g-ba.de/downloads/17-98-4300/2016-12-16_BQS-HKS-Abschlussbericht-2011-2013.pdf (accessed September 15, 2023).

[12] Veit Ch, Lüken F, Melsheimer O. Evaluation der Screeninguntersuchungen auf Hautkrebs gemäß Krebsfrüherkennungs-Richtlinie des Gemeinsamen Bundesausschusses. Abschlussberichte 2009 – 2010. BQS. https://www.g-ba.de/downloads/17-98-3907/2015-03-11_BQS_HKS-Abschlussbericht-2009-2010.pdf (accessed September 15, 2023).

[13] 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.

[14] Elmore JG, Barnhill RL, Elder DE, Longton GM, Pepe MS, Reisch LM, et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. BMJ 2017;357:j2813.

[15] 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.

[16] Kutzner H, Jutzi TB, Krahl D, Krieghoff-Henning EI, Heppt MV, Hekler A, et al. Overdiagnosis of melanoma – causes, consequences and solutions. J Dtsch Dermatol Ges 2020;18:1236–43.

[17] Haggenmüller S, Maron RC, Hekler A, Utikal JS, Barata C, Barnhill RL, et al. Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. Eur J Cancer 2021;156:202–16.

[18] 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.

[19] 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 Medicine 2020;17:e1003381. https://doi.org/10.1371/journal.pmed.1003381.

[20] 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.
[21] 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.
[22] 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.
[23] 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 Cancer 2019;111:30–7.
[24] 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 Cancer 2019;111:148–54.
[25] 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 Cancer 2019;119:57–65.
[26] 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 Cancer 2019;113:47–54.
[27] 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.
[28] 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.
[29] 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.
[30] 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.
[31] 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.
[32] 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 Cancer 2021;144:192–9.
[33] 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.
[34] 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.
[35] 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.
[36] Dick V, Sinz C, Mittlböck M, Kittler H, Tschandl P. Accuracy of Computer-Aided Diagnosis of Melanoma: A Meta-analysis. JAMA Dermatol 2019;155:1291–9.
[37] Gillespie N, Lockey S, Curtis C. Trust in artificial Intelligence: a five country study. Brisbane, Australia: The University of Queensland; 2021. https://doi.org/10.14264/e34bfa3.
[38] Liopyris K, Gregoriou S, Dias J, Stratigos AJ. Artificial Intelligence in Dermatology: Challenges and Perspectives. Dermatol Ther 2022;12:2637–51.

[39] Federal Office for Information Security. Security of AI-Systems: Fundamentals - Provision or use of external data or trained models. Bundesamt Für Sicherheit in Der Informationstechnik 2022. https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Publications/Studies/KI/P464_Provision_use_external_data_trained_models.html (accessed July 11, 2023).
[40] Azulay A, Weiss Y. Why do deep convolutional networks generalize so poorly to small image transformations? arXiv [csCV] 2018.
[41] Alcorn MA, Li Q, Gong Z, Wang C, Mai L, Ku W-S, et al. Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects. arXiv [csCV] 2018.
[42] Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. arXiv [csLG] 2017.
[43] Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, et al. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. arXiv [statML] 2019.
[44] 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.
[45] Mahbod A, Schaefer G, 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.
[46] 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. 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2016, p. 1373–6.
[47] 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.
[48] 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.
[49] Hart SN, Flotte W, Norgan AP, Shah KK, Buchan ZR, Mounajjed T, et al. Classification of Melanocytic Lesions in Selected and Whole-Slide Images via Convolutional Neural Networks. J Pathol Inform 2019;10:5.
[50] 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.
[51] 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 Cancer Res 2020;26:1126–34.
[52] 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 Cancer 2019;119:11–7.
[53] 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.
[54] 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;78:270–7.e1.
[55] 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.
[56] 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.
[57] 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 Cancer 2021;149:94–101.

[58] Li W, Zhuang J, Wang R, Zhang J, Zheng W-S. Fusing Metadata and Dermoscopy Images for Skin Disease Diagnosis. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, p. 1996–2000.
[59] 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/jamadermatol.2019.1735.
[60] 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.
[61] 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 Cancer 2021;145:81–91.
[62] 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.
[63] 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 Cancer 2019;120:114–21.
[64] Sawtooth Software, Inc. ACBC Technical Paper. Technical Paper Series 2014.
[65] Hauser K, Kurz A, Haggenmüller S, Maron RC, von Kalle C, Utikal JS, et al. Explainable artificial intelligence in skin cancer recognition: A systematic review. Eur J Cancer 2022;167:54–69.
[66] McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature 2020;577:89–94.
[67] Bulten W, Kartasalo K, Chen P-HC, Ström P, Pinckaers H, Nagpal K, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nat Med 2022;28:154–63.
[68] Mei X, Lee H-C, Diao K-Y, Huang M, Lin B, Liu C, et al. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. Nat Med 2020;26:1224–8.
[69] Muti HS, Heij LR, Keller G, Kohlruss M, Langer R, Dislich B, et al. Development and validation of deep learning classifiers to detect Epstein-Barr virus and microsatellite instability status in gastric cancer: a retrospective multicentre cohort study. The Lancet Digital Health 2021;3:e654–64.
[70] Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25:1301–9.
[71] Echle A, Grabsch HI, Quirke P, van den Brandt PA, West NP, Hutchins GGA, et al. Clinical-Grade Detection of Microsatellite Instability in Colorectal Tumors by Deep Learning. Gastroenterology 2020;159:1406–16.e11.
[72] Warnat-Herresthal S, Schultze H, Shastry KL, Manamohan S, Mukherjee S, Garg V, et al. Swarm Learning for decentralized and confidential clinical machine learning. Nature 2021;594:265–70.
[73] Li Y, Chen C, Liu N, Huang H, Zheng Z, Yan Q. A blockchain-based decentralized federated learning framework with committee consensus. IEEE Netw 2021;35:234–41.
[74] Bdair T, Navab N, Albarqouni S. Semi-Supervised Federated Peer Learning for Skin Lesion Classification 2021.
[75] Agbley BLY, Li J, Haq AU, Bankas EK, Ahmad S, Agyemang IO, et al. Multimodal melanoma detection with federated learning. 2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), IEEE; 2021. https://doi.org/10.1109/iccwamtip53232.2021.9674116.
[76] Adnan M, Kalra S, Cresswell JC, Taylor GW, Tizhoosh HR. Federated learning and differential privacy for medical image analysis. Sci Rep 2022;12:1–10.
[77] Dayan I, Roth HR, Zhong A, Harouni A, Gentili A, Abidin AZ, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. Nat Med 2021;27:1735–43.
[78] Saldanha OL, Quirke P, West NP, James JA, Loughrey MB, Grabsch HI, et al. Swarm learning for decentralized artificial intelligence in cancer histopathology. Nat Med 2022;28:1232–9.

[79] Federated learning for computational pathology on gigapixel whole slide images. Med Image Anal 2022;76:102298.
[80] Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ 2015;351. https://doi.org/10.1136/bmj.h5527.
[81] Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. QuPath: Open source software for digital pathology image analysis. Sci Rep 2017;7:16878.
[82] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework 2019. https://doi.org/10.48550/arXiv.1907.10902.
[83] Bergstra J, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization n.d. https://proceedings.neurips.cc/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf (accessed March 4, 2023).
[84] Smith LN. A disciplined approach to neural network hyper-parameters: Part 1 -- learning rate, batch size, momentum, and weight decay 2018. https://doi.org/10.48550/arXiv.1803.09820.
[85] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, d\textquotesingle Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc.; 2019.
[86] Howard J, Gugger S. Fastai: A Layered API for Deep Learning. Information 2020;11:108. https://doi.org/10.3390/info11020108.
[87] Efron B, Tibshirani RJ. An Introduction to the Bootstrap. CRC Press; 1994.
[88] McMahan HB, Moore E, Ramage D, Hampson S, Arcas BA y. Communication-Efficient Learning of Deep Networks from Decentralized Data 2016. https://doi.org/10.48550/arXiv.1602.05629.
[89] Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med 2019;25:1054–6.
[90] Maji D, Santara A, Mitra P, Sheet D. Ensemble of Deep Convolutional Neural Networks for Learning to Detect Retinal Vessels in Fundus Images 2016. https://doi.org/10.48550/arXiv.1603.04833.
[91] [No title] n.d. https://www.leitlinienprogramm-onkologie.de/fileadmin/user_upload/Downloads/Leitlinien/Melanom/Melanom_Version_3/LL_Melanom_Langversion_3.3.pdf (accessed August 29, 2023).
[92] Kairouz P, Brendan McMahan H, Avent B, Bellet A, Bennis M, Bhagoji AN, et al. Advances and Open Problems in Federated Learning. 2021.
[93] Jutzi TB, Krieghoff-Henning EI, Holland-Letz T, Utikal JS, Hauschild A, Schadendorf D, et al. Artificial Intelligence in Skin Cancer Diagnostics: The Patients' Perspective. Front Med 2020;7:233.
[94] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 2016;316:2402–10.
[95] Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. JAMA 2017;318:2199–210.
[96] Tseng H-H, Wei L, Cui S, Luo Y, Ten Haken RK, El Naqa I. Machine Learning and Imaging Informatics in Oncology. Oncology 2020;98:344–62.
[97] Liu Y, Kohlberger T, Norouzi M, Dahl GE, Smith JL, Mohtashamian A, et al. Artificial Intelligence-Based Breast Cancer Nodal Metastasis Detection: Insights Into the Black Box for Pathologists. Arch Pathol Lab Med 2019;143:859–68.
[98] Maron RC, Utikal JS, Hekler A, Hauschild A, Sattler E, Sondermann W, et al. Artificial Intelligence and Its Effect on Dermatologists' Accuracy in Dermoscopic Melanoma Image Classification: Web-Based Survey Study. J Med Internet Res 2020;22:e18091.
[99] Sondermann W, Utikal JS, Enk AH, Schadendorf D, Klode J, Hauschild A, et al. Prediction of melanoma evolution in melanocytic nevi via artificial intelligence: A call for prospective data. Eur J Cancer 2019;119:30–4.

[100] Sies K, Winkler JK, Fink C, Bardehle F, Toberer F, Buhl T, et al. Past and present of computer-assisted dermoscopic diagnosis: performance of a conventional image analyser versus a convolutional neural network in a prospective data set of 1,981 skin lesions. Eur J Cancer 2020;135:39–46.
[101] Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller K-R. Unmasking Clever Hans predictors and assessing what machines really learn. Nat Commun 2019;10:1096.
[102] Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generali zation performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 2018;15:e1002683.
[103] Schmitt M, Maron RC, Hekler A, Stenzinger A, Hauschild A, Weichenthal M, et al. Hid den Variables in Deep Learning Digital Pathology and Their Potential to Cause Batch Effects: Prediction Model Study. J Med Internet Res 2021;23:e23436.
[104] Heaven D. Why deep-learning AIs are so easy to fool. Nature 2019;574:163–6.
[105] Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science 2019;363:1287–9.
[106] Zhang R. Making Convolutional Networks Shift-Invariant Again. arXiv [csCV] 2019.
[107] Fawzi A, Frossard P. Manitest: Are classifiers really invariant? arXiv [csCV] 2015.
[108] Engstrom L, Tran B, Tsipras D, Schmidt L, Madry A. Exploring the Landscape of Spatial Robustness. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th International Conference on Machine Learning, vol. 97, PMLR; 09--15 Jun 2019, p. 1802–11.
[109] Gutman D, Codella NCF, Celebi E, Helba B, Marchetti M, Mishra N, et al. Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). arXiv [csCV] 2016.
[110] Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 2018;5:180161.
[111] Mendonca T, Ferreira PM, Marques JS, Marcal ARS, Rozeira J. $PH^2$ - A dermoscopic image database for research and benchmarking. 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 2013:5437–40.
[112] de Faria SMM, Henrique M, Filipe JN, Pereira PMM, Tavora LMN, Assuncao PAA, et al. Light Field Image Dataset of Skin Lesions. Conf Proc IEEE Eng Med Biol Soc 2019;2019:3905–8.
[113] Combalia M, Codella NCF, Rotemberg V, Helba B, Vilaplana V, Reiter O, et al. BCN20000: Dermoscopic Lesions in the Wild. arXiv [eessIV] 2019.
[114] Bi L, Kim J, Ahn E, Feng D. Automatic Skin Lesion Analysis using Large-scale Dermoscopy Images and Deep Residual Networks 2017.
[115] Li X, Shen X, Zhou Y, Wang X, Li T-Q. Classification of breast cancer histopathological images using interleaved DenseNet with SENet (IDSNet). PLoS One 2020;15:e0232127.
[116] Kather JN, Krisam J, Charoentong P, Luedde T, Herpel E, Weis C-A, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS Med 2019;16:e1002730.
[117] Kather JN, Heij LR, Grabsch HI, Loeffler C, Echle A, Muti HS, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. Nat Cancer 2020;1:789–99.
[118] Li X, Zhang S, Zhang Q, Wei X, Pan Y, Zhao J, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. Lancet Oncol 2019;20:193–201.
[119] Lee J, Won T, Lee TK, Lee H, Gu G, Hong K. Compounding the Performance Improvements of Assembled Techniques in a Convolutional Neural Network 2020.
[120] Lopes RG, Yin D, Poole B, Gilmer J, Cubuk ED. Improving Robustness Without Sacrificing Accuracy with Patch Gaussian Augmentation 2019.
[121] Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. arXiv [csCV] 2017.

[122] 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 Cancer 2019;113:47–54.
[123] Phillips M, Marsden H, Jaffe W, Matin RN, Wali GN, Greenhalgh J, et al. Assessment of Accuracy of an Artificial Intelligence Algorithm to Detect Melanoma in Images of Skin Lesions. JAMA Netw Open 2019;2:e1913436.
[124] MacLellan AN, Price EL, Publicover-Brouwer P, Matheson K, Ly TY, Pasternak S, et al. The Use of Non-Invasive Imaging Techniques in the Diagnosis of Melanoma: A Prospective Diagnostic Accuracy Study. J Am Acad Dermatol 2020. https://doi.org/10.1016/j.jaad.2020.04.019.
[125] Ba W, Wu H, Chen WW, Wang SH, Zhang ZY, Wei XJ, et al. Convolutional neural network assistance significantly improves dermatologists' diagnosis of cutaneous tumours using clinical images. Eur J Cancer 2022;169:156–65.
[126] Winkler JK, Sies K, Fink C, Toberer F, Enk A, Abassi MS, et al. Association between different scale bars in dermoscopic images and diagnostic performance of a market-approved deep learning convolutional neural network for melanoma recognition. Eur J Cancer 2021;145:146–54.
[127] Ha Q, Liu B, Liu F. Identifying Melanoma Images using EfficientNet Ensemble: Winning Solution to the SIIM-ISIC Melanoma Classification Challenge. arXiv [csCV] 2020.
[128] White Paper on Artificial Intelligence: A European Approach to Excellence and Trust. 2020.
[129] Wortsman M, Ilharco G, Gadre SY, Roelofs R, Gontijo-Lopes R, Morcos AS, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time 2022.
[130] Gulrajani I, Lopez-Paz D. In Search of Lost Domain Generalization 2020.
[131] Radford A, Kim 1. Jong Wook, Hallacy 1. Chris, Ramesh A, Goh G, Agarwal S, et al. CLIP: CLIP (Contrastive Language-Image Pretraining), Predict the most relevant text snippet given an image. Github; n.d.
[132] Jia C, Yang Y, Xia Y, Chen Y-T, Parekh Z, Pham H, et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. arXiv [csCV] 2021.
[133] Kurz A, Hauser K, Mehrtens HA, Krieghoff-Henning E, Hekler A, Kather JN, et al. Uncertainty Estimation in Medical Image Classification: Systematic Review. JMIR Med Inform 2022;10:e36427.
[134] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, p. 770–8.
[135] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, p. 4700–8.
[136] Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv [csLG] 2019.
[137] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv [csCV] 2014.
[138] Rotemberg V, Kurtansky N, Betz-Stablein B, Caffery L, Chousakos E, Codella N, et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Scientific Data 2021;8:1–8.
[139] Kawahara J, Daneshvar S, Argenziano G, Hamarneh G. 7-Point Checklist and Skin Lesion Classification using Multi-Task Multi-Modal Neural Nets. IEEE J Biomed Health Inform 2018. https://doi.org/10.1109/JBHI.2018.2824327.
[140] Maron RC, Schlager JG, Haggenmüller S, von Kalle C, Utikal JS, Meier F, et al. A benchmark for neural network robustness in skin cancer classification. Eur J Cancer 2021;155:191–9.
[141] Shorfuzzaman M. An explainable stacked ensemble of deep learning models for improved melanoma skin cancer detection. Multimedia Systems 2021;28:1309–23.

[142] Wei L, Ding K, Hu H. Automatic Skin Cancer Detection in Dermoscopy Images Based on Ensemble Lightweight Deep Learning Network n.d. https://doi.org/10.1109/ACCESS.2020.2997710 (accessed November 16, 2023).
[143] Shanmugam D, Blalock D, Balakrishnan G, Guttag J. Better Aggregation in Test-Time Augmentation. arXiv [csCV] 2020.
[144] Hekler A, Brinker TJ, Buettner F. Test Time Augmentation Meets Post-hoc Calibration: Uncertainty Quantification under Real-World Conditions. AAAI 2023;37:14856–64.
[145] Codella N, Rotemberg V, Tschandl P, Celebi ME, Dusza S, Gutman D, et al. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC) 2019.
[146] Codella NCF, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza SW, et al. Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC) 2017.
[147] Oh S, Kim JH, Choi S-W, Lee HJ, Hong J, Kwon SH. Physician Confidence in Artificial Intelligence: An Online Mobile Survey. J Med Internet Res 2019;21:e12422.
[148] Polesie S, Gillstedt M, Kittler H, Lallas A, Tschandl P, Zalaudek I, et al. Attitudes towards artificial intelligence within dermatology: an international online survey. Br J Dermatol 2020;183:159–61.
[149] Shen C, Li C, Xu F, Wang Z, Shen X, Gao J, et al. Web-based study on Chinese dermatologists' attitudes towards artificial intelligence. Ann Transl Med 2020;8:698.
[150] Haggenmüller S, Krieghoff-Henning E, Jutzi T, Trapp N, Kiehl L, Utikal JS, et al. Digital Natives' Preferences on Mobile Artificial Intelligence Apps for Skin Cancer Diagnostics: Survey Study. JMIR Mhealth Uhealth 2021;9:e22909.
[151] Fink C, Uhlmann L, Hofmann M, Forschner A, Eigentler T, Garbe C, et al. Patient acceptance and trust in automated computer-assisted diagnosis of melanoma with dermatofluoroscopy. J Dtsch Dermatol Ges 2018;16:854–9.
[152] Nelson CA, Pérez-Chada LM, Creadore A, Li SJ, Lo K, Manjaly P, et al. Patient Perspectives on the Use of Artificial Intelligence for Skin Cancer Screening: A Qualitative Study. JAMA Dermatol 2020;156:501–12.
[153] Walter FM, Toby Prevost A, Vasconcelos J, Hall PN, Burrows NP, Morris HC, et al. Using the 7-point checklist as a diagnostic aid for pigmented skin lesions in general practice: a diagnostic validation study. Br J Gen Pract 2013;63:e345–53.
[154] The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 1997;30:1145–59.
[155] Šimundić A-M. Measures of Diagnostic Accuracy: Basic Definitions. EJIFCC 2009;19:203–11.
[156] Warnat-Herresthal S, Schultze H, Shastry KL, Manamohan S, Mukherjee S, Garg V, et al. Swarm Learning for decentralized and confidential clinical machine learning. Nature 2021;594:265–70.
[157] Geirhos R, Jacobsen J-H, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. Nature Machine Intelligence

# 7 TABULAR APPENDIX

**Tabular Appendix 1. Detailed overview of the comprehensive search strategy.**

| | | Filters used | | |
|---|---|---|---|---|
| (Combinations of) search terms | Retrieval Date | 2017-2021 | English* | Comparative Studies* |
| Skin Cancer Classification AND Convolutional Neural Networks | 01/18/2021 | **x** | **x** | |
| Melanoma Detection AND Convolutional Neural Networks | 01/18/2021 | **x** | **x** | |
| Malignant Melanoma AND Convolutional Neural Networks | 01/18/2021 | **x** | **x** | |
| Skin Cancer Classification AND Convolutional Neural Networks | 02/17/2021 | **x** | **x** | **x** |
| Skin Cancer AND Convolutional Neural Networks | 02/17/2021 | **x** | **x** | **x** |
| Skin Cancer Classification | 02/17/2021 | **x** | **x** | **x** |
| Melanoma Detection AND Convolutional Neural Networks | 02/17/2021 | **x** | **x** | **x** |
| Malignant Melanoma AND Convolutional Neural Networks | 02/17/2021 | **x** | **x** | **x** |
| Skin Cancer Classification AND Histopathology AND Convolutional Neural Networks | 02/17/2021 | **x** | **x** | |
| Melanoma AND Histopathology AND Convolutional Neural Networks | 02/17/2021 | **x** | **x** | |
| Skin Cancer Classification AND Histopathology AND Deep Learning | 02/17/2021 | **x** | **x** | |
| Whole slide imaging AND convolutional neural networks | 02/17/2021 | **x** | **x** | |
| Whole slide imaging AND skin cancer | 02/17/2021 | **x** | **x** | |
| Melanoma AND Histopathology AND Deep Learning | 02/17/2021 | **x** | **x** | |

*Only possible for PubMed and Medline.

**Tabular Appendix 2. Training and test procedure of AI-based models for the classification of dermoscopic images.**

| study | architecture | ensemble y/n | scope of the training data | origin of the training data | ground truth of the training data | metadata y/n | classification task | scope of the test data | origin of the test data (Holdout/OOD) | ground truth of the test data (*in case of OOD)* |
|---|---|---|---|---|---|---|---|---|---|---|
| *Brinker et al.* [26] | pre-trained ResNet50 | n | 13 637 images*, randomly selected out 20 635 images available at ISIC (after excluding the test set) | ISIC image archive | melanoma: histology-proven<br>nevi: histology-proven (~24%), expert consensus (~54%), or e.g., series of images with no change over time (~22%). | n | binary: melanoma/ melanocytic nevi | 100 images, randomly selected out of 20 735 images available at ISIC | ISIC image archive (Holdout) | / |
| *Brinker et al.* [52] | pre-trained ResNet50 | n | 4 204 images | ISIC image archive | histology-proven | n | binary: melanoma/ melanocytic nevi | 804 images in total | ISIC image archive (Holdout) | / |
| *Yu et al.* [53] | modified, pre-trained VGG-16 | n | 724 images in total, 362 images per subset | Severance Hospital in the Yonsei University Health System, Seoul, Korea, Dongsan Hospital in the Keimyung University Health System, Daegu, Korea | histology-proven | n | binary: melanoma/ melanocytic nevi | 724 images in total, 362 images per subset (2-fold cross-testing) | Severance Hospital in the Yonsei University Health System, Seoul, Korea (Holdout),<br>Dongsan Hospital in the Keimyung University Health System, Daegu, Korea (Holdout) | / |
| *Marchetti et al.* [54] | fusion of the algorithms from the ISBI 2016 challenge | y | 900 images | ISBI 2016 challenge, ISIC image archive | melanoma: histology-proven<br>nevi/lentigines: majority histology-proven,162 nevi were reviewed by ≥2 dermatologists | n | binary: malignant/ benign;<br>biopsy/ observation or reassurance | 379 images in total | ISBI 2016 challenge, ISIC image archive (Holdout) | / |
| *Marchetti et al.* [55] | top challenge algorithm from the ISBI 2017 challenge | n | 2 150 images* | ISBI 2017 challenge, ISIC image archive | not specified | n | binary: melanoma/ non-melanoma;<br>biopsy/ observation | 600 images | ISIC image archive (Holdout) | / |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Haenssle et al.* [30] | pre-trained, modified Google Inception v4 | n | not specified | cooperating dermatologists, ISIC image archive | histology-proven; in case of non-excised lesions: diagnosed by experienced dermatologists and/or follow-up examinations | n | binary: melanoma/ melanocytic nevi; excision or short-term follow/no action | 300 images | Department of Dermatology, University of Heidelberg, Germany (OOD) | histology-proven or follow-up examination |
| *Haenssle et al.* [31] | pre-trained, modified Google Inception v4 (Moleanalyzer Pro, FotoFinder Systems, Bad Birnbach, Germany) | n | not specified | cooperating dermatologists, ISIC image archive | histology-proven; in case of non-excised lesions: diagnosed by experienced dermatologists and/or follow-up examinations | n | binary: (pre)malignant/ benign; excision or treatment/ follow up or no action | 2 711 images in total<br>1) 100 images<br>2) 1 100 images<br>3) 1 511 images | 1) Department of Dermatology, University of Heidelberg, Germany (OOD)<br>2) MSK-1 data set (OOD)<br>3) ISIC-2018 challenge data set (OOD) | 1) histology-proven or follow-up examination<br>2) not specified<br>3) not specified |
| *Haenssle et al.* [32] | pre-trained, modified Google Inception v4 (Moleanalyzer Pro, FotoFinder Systems, Bad Birnbach, Germany) | n | not specified | cooperating dermatologists, ISIC image archive | histology-proven; in case of non-excised lesions: diagnosed by experienced dermatologists and/or follow-up examinations | n | binary: malignant/ benign; excision or treatment/ follow up or no action | 4 932 images<br>1) 100 images<br>2) 240 images<br>3) 1 511 images<br>4) 1 100 images<br>5) 1 981 images | 1) Department of Dermatology, University of Heidelberg, Germany (OOD)<br>Department of Dermatology Hospital Thalkirchner Street, Munich, Germany (OOD)<br>Department of Dermatology, Medical University Graz, Austria (OOD)<br>First Department of Dermatology, Aristotle University, Thessaloniki, Greece (OOD)<br>Dermatology Office based clinic of Dermatology, Konstanz, Germany (OOD)<br>2) Primary skin cancer clinic in Queensland, Australia (OOD)<br>3) ISIC-2018 challenge data set (OOD)<br>4) MSK-1 data set (OOD)<br>5) Prospective data set, acquired during 15 years of follow-up examinations (OOD) | 1) histology-proven (98%), unremarkable follow-up >2 years (2%)<br>2) histology-proven<br>3) not specified<br>4) not specified<br>5) histology-proven (~40%, including all malignant lesions), expert consensus plus an unremarkable follow-up >2 years (~60%) |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Tschandl et al.* [28] | 139 challenge algorithms from the ISBI 2018 challenge | n | 10 015 images | HAM 10000 data set, ISBI 2018 challenge, ISIC image archive | histology-proven (>50% of all lesions), sequential dermatoscopic imaging without changes, expert consensus | n | multiclass (7) | 1 511 images in total<br>1) 1 195 images<br>2) 316 images | 1) HAM10000 data set, ISBI 2018 challenge, ISIC image archive (Holdout)<br>2) images from Turkey, New Zealand, Sweden and Argentina (OOD) | not specified |
| *Maron et al.* [25] | pre-trained ResNet50 | n | 11 444 images | HAM10000 data set, additional images from the ISIC image archive | 6 390 images histology-proven | n | binary: malignant/ benign;<br>multiclass (5) | 300 images | HAM10000 data set (Holdout) | / |
| *Tschandl et al.* [29] | cCNN, Inception V3 and ResNet50 | y | 8 235 dermoscopic images and 6 458 close-up images* | Primary skin cancer clinic in Queensland, Australia | histology-proven | n | multiclass (8) | 2 072 dermoscopic and clinical close-up images | Primary skin cancer clinic in Queensland, Australia (Holdout),<br>Department of Dermatology of the Medical University of Vienna, Austria (OOD),<br>additional images from dermatologists from Sweden, Italy, Austria, France, Turkey, German (OOD) | histology-proven |

*Specification includes training and validation set.

metadata (additional information beyond image input, e.g., age, gender, localization of the suspicious lesion)

y (yes)

n (no)

OOD (out-of-distribution)

**Tabular Appendix 3. Training and test procedure of AI-based models for the classification of clinical images.**

| study | architecture | ensemble y/n | scope of the training data | origin of the training data | ground truth of the training data | metadata y/n | classification task | scope of the test data | origin of the test data (Holdout/OOD) | ground truth of the test data (*in case of OOD)* |
|---|---|---|---|---|---|---|---|---|---|---|
| *Fujisawa et al.* [18] | pre-trained GoogleNet | n | 4 867 images of 14 disorders from 1 842 patients | University of Tsubuka Hospital, Japan | histology-proven, except for cases of congenital melanocytic nevus, nevus spilus, and lentigo simplex | n | 1) binary: malignant/benign<br>2) multiclass (14) | 1 142 images of 14 disorders from 454 patients | University of Tsubuka Hospital, Japan (Holdout) | / |
| *Jinnai et al.* [21] | (fast region-based) FRCNN with VGG-16 as backbone | n | 4 732 images of 6 disorders from 2 885 patients | Dermatologic Oncology in the National Cancer Center Hospital, Tokyo, Japan | malignant lesions: histology-proven<br>benign lesions: clinical diagnosis with dermoscopy or histology-proven | n | 1) binary: malignant/benign<br>2) multiclass (6) | 1 114 images of 6 disorders from 666 patients (randomly selected one images per patient for testing) | Dermatologic Oncology in the National Cancer Center Hospital, Toky, Japan (Holdout) | / |
| *Han et al.* [20] | pre-trained SENet, SE-ResNet-50, VGG-19 | y | 220 680 images of 174 disorders | ASAN, MEDNODE, Web, Normal data set | not specified | n | 1) binary: malignant/benign<br>2) multiclass (134) | 1) 1 300 images of 10 disorders<br>2) 2 201 images of 134 disorders | 1) Edinburgh data set (OOD)<br>2) SNU data set (OOD) | 1) histology-proven<br>2) not specified |
| *Han et al.* [22] | pre-trained ResNet-152 | n | 19 398 images<br>1) 15 408 images of 12 disorders<br>2) 170 melanoma/nevus<br>3) 3 820 images | 1) ASAN data set<br>2) MEDNODE data set<br>3) Atlas data set (from several dermatologic atlas sites) | not specified | n | multiclass (12) | 2 728 images<br>1) 1 300 images of 10 disorders<br>2) 152 BCC images<br>3) 1 276 images of 12 disorders | 1) Edinburgh data set (OOD)<br>2) Hallym data set (OOD)<br>3) Asan data set (Holdout) | histology-proven |
| *Brinker et al.* [24] | pre-trained ResNet50 | n | 13 637 dermoscopic images*, randomly selected out 20 735 images available at ISIC | HAM10000, ISIC image archive | melanoma: histology-proven<br>nevi: histology-proven (~24%), expert consensus (~54%), or e.g., series of images with no change over time (~22%). | n | binary: melanoma/ melanocytic nevi | 100 images | MClass-Benchmark obtained from the MED-NODE database (OOD) | melanoma: histology-proven; nevi: expert consensus |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Han et al.* [19] | SENet and SE-ResNeXt-50 trained with a region-based CNN (RCNN) | y | 1 106 886 clinical images of 178 disorders | Asan Medical Center, various websites | histology-proven | n | 1) binary: malignant/benign<br>2) multiclass (32) | 1) 1 300 images of 10 disorders<br>2) 40 331 images from 10 426 patients of 43 disorders | 1) Edinburgh data set (OOD)<br>2) Department of Dermatology, Severance Hospital in Seoul, Korea (OOD) | 1) histology-proven<br>2) not specified |

*Specification includes training and validation set.

metadata (additional information beyond image input, e.g., age, gender, localization of the suspicious lesion)

y (yes)

n (no)

OOD (out-of-distribution)

**Tabular Appendix 4. Training and test procedure of AI-based models for the classification of histopathological images.**

| study | architecture | ensemble y/n | scope of the training data | origin of the training data | ground truth of the training data | metadata y/n | classification task | scope of the test data | origin of the test data (Holdout/OOD) | ground truth of the test data (*in case of OOD)* |
|---|---|---|---|---|---|---|---|---|---|---|
| *Hekler et al.* [33] | pre-trained ResNet50 | n | 595 digitised H&E slides, image sections of the epithelium (0.06% of the whole slide on average) with a 10-fold magnification were randomly cropped (one crop per slide) | Dermatohistopathologic Institute Dr. D. Krahl, Heidelberg, Germany | expert histopathologist with more than 20 years of experience in accordance with current guidelines | n | binary: melanoma/ melanocytic nevi | 100 digitised H&E slides | Dermatohistopathologic Institute Dr. D. Krahl, Heidelberg, Germany (Holdout) | / |
| *Brinker et al.* [34] | pre-trained ResNeXt50 | y | 100 digitised H&E slides (5 folds, consisting of 80 images each) | routine files of 2 expert board-certified dermatopathologists from Friedrichshafen, Germany | panel of 2 experienced dermatopathologists according to the standard practice | n | binary: melanoma/ melanocytic nevi | 100 digitised H&E slides (5 folds, consisting of 20 images each) | 1) routine files of 2 expert board-certified dermatopathologists from Friedrichshafen, Germany (Holdout, 5-fold cross-testing)<br>2) Dermatohistopathologic Institute Dr. D. Krahl, Heidelberg, Germany (OOD) | expert histopathologist with more than 20 years of experience in accordance with current guidelines |

metadata (additional information beyond image input, e.g., age, gender, localization of the suspicious lesion)

y (yes)

n (no)

OOD (out-of-distribution)

**Tabular Appendix 5. Detailed specification of the classes used for multiclass classification tasks.**

| study | number of classes | classes |
|---|---|---|
| *Tschandl et al*. [28] | 7 | malignant melanoma<br>intraepithelial carcinoma including actinic keratosis and Bowen's disease<br>basal cell carcinoma<br>benign keratinocytic lesions including solar lentigo, seborrheic keratosis and lichen planus-like keratosis<br>dermatofibroma<br>melanocytic nevus<br>vascular lesions |
| *Maron et al.* [25] | 5 | malignant melanoma<br>actinic keratosis, intraepithelial carcinoma, Bowen's disease, squamous cell carcinoma<br>basal cell carcinoma<br>benign keratosis including seborrheic keratosis, solar lentigo and lichen planus-like keratosis<br>melanocytic nevi |
| *Tschandl et al.* [29] | 8 | actinic keratosis and intraepithelial carcinoma (also known as Bowen's disease)<br>basal cell carcinoma (all subtypes)<br>benign keratosis like lesions (including solar lentigo, seborrheic keratosis, and lichen planus–like keratosis)<br>dermatofibroma<br>malignant melanoma<br>invasive squamous cell carcinoma and keratoacanthoma<br>benign sebaceous neoplasms<br>benign hair follicle tumors |
| *Fujisawa et al.* [18] | 14 | malignant melanoma<br>squamous cell carcinoma<br>Bowen's disease<br>actinic keratosis<br>basal cell carcinoma<br>nevus cell nevus<br>blue nevus<br>congenital melanocytic nevus<br>Spitz nevus<br>sebaceus nevus<br>poroma<br>seborrheic keratosis<br>nevus spilus<br>lentigo simplex |
| *Jinnai et al.* [21] | 6 | malignant melanoma<br>basal cell carcinoma<br>nevus<br>seborrheic keratosis |

| | | |
|---|---|---|
| | | senile lentigo<br>hematoma/hemangioma |
| *Han et al.* [20] | 134 | A complete listing has been omitted due to space restrictions. For further information, see [26]. |
| *Han et al.* [22] | 12 | malignant melanoma<br>basal cell carcinoma<br>squamous cell carcinoma<br>intraepithelial carcinoma<br>actinic keratosis<br>seborrheic keratosis<br>melanocytic nevus<br>lentigo<br>pyogenic granuloma<br>hemangioma<br>dermatofibroma<br>wart |
| *Han et al.* [19] | 32 | A complete listing has been omitted due to space restrictions. For further information, see [28]. |

**Tabular Appendix 6. The STARD 2015 list.**

| Topic | No | Item | Section in Manuscript |
|---|---|---|---|
| **Title or abstract** | | | |
| | 1 | Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC) | Abstract |
| **Abstract** | | | |
| | 2 | Structured summary of study design, methods, results, and conclusions | Abstract |
| **Introduction** | | | |
| | 3 | Scientific and clinical background, including the intended use and clinical role of the index test | Abstract, Introduction |
| | 4 | Study objectives and hypotheses | Abstract, Introduction |
| **Methods** | | | |
| Study Design | 5 | Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study) | Patient Cohorts and Slide Acquisition |
| Participants | 6 | Eligibility criteria | Patient Cohorts and Slide Acquisition |
| | 7 | On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry) | Patient Cohorts and Slide Acquisition |
| | 8 | Where and when potentially eligible participants were identified (setting, location, and dates) | Patient Cohorts and Slide Acquisition |
| | 9 | Whether participants formed a consecutive, random, or convenience series | Patient Cohorts and Slide Acquisition |
| Test methods | 10a | Index test, in sufficient detail to allow replication | Whole Slie Image Preprocessing, Model Development, Statistical Analysis |

| | | | |
|---|---|---|---|
| | 10b | Reference standard, in sufficient detail to allow replication | Whole Slide Image Preprocessing, Model Development, Statistical Analysis |
| | 11 | Rationale for choosing the reference standard (if alternatives exist) | Introduction |
| | 12a | Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory | not applicable |
| | 12b | Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory | not applicable |
| | 13a | Whether clinical information and reference standard results were available to the performers or readers of the index test | not applicable |
| | 13b | Whether clinical information and index test results were available to the assessors of the reference standard | not applicable |
| **Analysis** | 14 | Methods for estimating or comparing measures of diagnostic accuracy | Statistical Analysis |
| | 15 | How indeterminate index test or reference standard results were handled | not applicable |
| | 16 | How missing data on the index test and reference standard were handled | Patient Cohorts and Slide Acquisition, Number of Eligible Slides and Patients |
| | 17 | Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory | Statistical Analysis |
| | 18 | Intended sample size and how it was determined | Patient Cohorts and Slide Acquisition, Number of Eligible Slides and Patients |
| **Results** | | | |
| Participants | 19 | Flow of participants, using a diagram | Number of Eligible Slides and Patients |

| | | | |
|---|---|---|---|
| | 20 | Baseline demographic and clinical characteristics of participants | Patient Characteristics and Differences among Datasets |
| | 21a | Distribution of severity of disease in those with the target condition | Patient Characteristics and Differences among Datasets |
| | 21b | Distribution of alternative diagnoses in those without the target condition | Patient Characteristics and Differences among Datasets |
| | 22 | Time interval and any clinical interventions between index test and reference standard | not applicable |
| Test results | 23 | Cross tabulation of the index test results (or their distribution) by the results of the reference standard | Comparison of Federated Learning with Other Approaches |
| | 24 | Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals) | Comparison of Federated Learning with Other Approaches |
| | 25 | Any adverse events from performing the index test or the reference standard | not applicable |
| **Discussion** | | | |
| | 26 | Study limitations, including sources of potential bias, statistical uncertainty, and generalisability | Discussion |
| | 27 | Implications for practice, including the intended use and clinical role of index text | Discussion |
| **Other information** | | | |
| | 28 | Registration number and name of registry | not applicable |
| | 29 | Where the full study protocol can be accessed | not applicable |
| | 30 | Sources of funding and other support; role of funders | Role of Funding Source |

# 8 CURRICULUM VITAE

PERSONALIEN

| | |
|---|---|
| Name, Vorname: | Haggenmüller, Sarah |
| Geburtsdatum: | 30.11.1995 |
| Geburtsort: | Ulm |

SCHULISCHER WERDEGANG

| | |
|---|---|
| 2006 bis 2014 | Illertal-Gymnasium, Vöhringen |
| 27.06.2014 | Erwerb der Allgemeinen Hochschulreife (1,2) |

UNIVERSITÄRER WERDEGANG

| | |
|---|---|
| 2014 bis 2017 | Duales Studium Fachrichtung BWL-Industrie an der Dualen Hochschule Baden-Württemberg, Heidenheim |
| 2017 | Bachelorarbeit im Bereich Change Management und Digitalisierung<br>Thema: Die Uzin Utz AG im Wandel – Entwicklung eines Change Kommunikationskonzepts für den Relaunch des Webshops als digitales Hilfsmittel im Außendienst der Marke UZIN (1,0) |
| 30.09.2017 | Bachelor of Arts (1,2) |
| 2018 bis 2020 | Master of Advanced Management an der University of Applied Science, Neu-Ulm |
| 2020 | Masterarbeit im Bereich Marktforschung und Künstliche Intelligenz<br>Thema: Artificial Intelligence for Early Skin Cancer Detection (1,0) |
| 22.06.2020 | Master of Science (1,1) |

# 9 ACKNOWLEDGEMENT

An meine Familie, die Studienstiftung des deutschen Volkes und Kollegen, die zu Freunden wurden, in ewiger Dankbarkeit.