Diploma thesis

Assessment of a smartphone-based neural network application for the risk assessment of skin lesions under real-world conditions. A systematic retrospective comparison of detection accuracy by 3 dermatologists with digital risk assessment

submitted by

Philipp Efferl

attaining the academic degree

Doktor der gesamten Heilkunde

(Dr. med. univ.)

at the

Medical University of Graz

conducted at the

Department of Dermatology and Venereology

supervised by

Ao. Univ.-Prof. Dr. Rainer Hofmann-Wellenhof

Dr. med. univ. Teresa Maria Kränke

Graz, 23.05.2023

Affidavit

I hereby declare that the following diploma thesis has been written only by the undersigned and without any assistance from third parties. Furthermore, I confirm that no sources have been used in the preparation of this other than those indicated in the thesis itself.

Graz, 23.05.2023 *Philipp Efferl eh.*

Acknowledgment

My thanks are due to all those who have made a decisive contribution to the creation and completion of this thesis through their numerous suggestions and encouragement.
I would especially like to thank Ao. Univ. Prof. Dr. Rainer Hofmann-Wellenhof for giving me the opportunity to work on this diploma thesis by providing valuable guidance and feedback, challenging me to grow as a scientist and aspiring dermatologist.

I would like to thank Dr. med. univ. Teresa Kränke and Dr. med. univ. Katharina Tripolt-Droschl for their extraordinary willingness and work and their support in forming the Expert Panel together with Ao. Univ. Prof. Dr. Rainer Hofmann-Wellenhof, whose outstanding work in evaluating thousands of images has made a significant contribution to the scientific significance and quality of this study.
The staff at the Institute of Medical Informatics, Statistics and Documentation were also a great help to me, as they supported me in sample size calculation, statistical evaluation and study design planning.

The possibility of using the SkinScreener© device as the foundation for the feasibility of this study honors me, along with the numerous anonymous users of the app who made such a large-scale data evaluation possible in the first place by sharing their skin findings.

Finally, I would like to thank my parents, my family, and my circle of friends who have always supported me in my plans and provided advice during the intensive realization of this work.

TABLE OF CONTENTS

Abbreviations

RHW: Ao. Univ. Prof. Dr. Rainer Hofmann-Wellenhof

TK: Dr.med.univ. Teresa Kränke

KTD: Dr.med.univ. Katharina Tripolt-Droschl

PE: Philipp Efferl

CNN: Convolutional Neuronal Network

CE: Conformité Européenne

CI: Confidence Interval

NMSC: Non-melanocytic Skin Cancer

KC: Keratinocyte Carcinomas

BCC: Basal Cell Carcinoma

SCC: Squamous Cell Carcinoma

AK: Actinic Keratosis

AJCC: American Joint Committee on Cancer

HPV: Human papillomavirus infection

UVR: Ultraviolet radiation

CSE: Clinical skin examination

GP: General practitioner

PCP: Primary care provider

AI: Artificial Intelligence

RPN: Region proposal network

CLS: Classification Score layer

RGB: Red Green Blue

ID: Identification

VBA: Visual Basics for Application

FDA: Food and Drug Administration

nAI: Number of assessments for Artificial Intelligence

nRS: Number of assessments for Reference Standard

nSRS: Number of assessments for the Single Reference Standard

List of Tables

List of Figures

Zusammenfassung

Hintergrund: Die diagnostische Leistung von CNNs für die Diagnose verschiedener Arten von Hautkrebs zeigte sich in den letzten Jahren sehr vielversprechend. Der Einsatz einer Smartphone-Applikation, die Laien mit Hilfe einer integrierten künstlichen Intelligenz eine erste Einschätzung geben kann, könnte gegebenenfalls zu einer schnelleren Therapie führen.

Zielsetzung: Bis dato ist in der Literatur noch nicht über die Leistungsfähigkeit einer zertifizierten Smartphone-basierten neuronalen Netzwerkanwendung für die Analyse von makroskopischen Bildern von Hautläsionen berichtet worden, die von Laien unter realen Bedingungen aufgenommen wurden. Das Hauptziel dieser Studie ist die Bewertung der Risikobewertung durch die Smartphone-App im Vergleich zur Risikobewertung durch ein Expertengremium aus drei Dermatologen.

Methoden: Wir analysierten und verglichen die Erkennungsgenauigkeit des CE-gekennzeichneten Algorithmus der SkinScreener© Smartphone-Applikation mit der Konsensmeinung der teilnehmenden Dermatologen. Die primären Endpunkte waren Sensitivität, Spezifität und Genauigkeit für die trichotome Risikoeinschätzung (geringes, mittleres, hohes Risiko). Sekundäre Endpunkte waren interindividuelle Unterschiede in der diagnostischen Leistung der Dermatologen bei der Analyse der jeweiligen Hautläsionen als auch die Analyse der durch die KI falsch beurteilten Läsionen.

Ergebnisse: Die Leistung des CE-gekennzeichneten Smartphone-Algorithmus bei der Risikobewertung betrug 76,9% - 95% CI: {71,7% - 81,5%} für die Sensitivität und 80,9% - 95% CI: {78,5% - 83,2%} für die Spezifität. Die Gesamtgenauigkeit betrug 77,2 %. Sekundäre Endpunkte waren interindividuelle Unterschiede in der diagnostischen Leistung der Dermatologen bei der Analyse der jeweiligen Hautläsionen. Es konnte gezeigt werden, dass die Leistungsfähigkeit schlechter war als in einer zuvor durchgeführten Studie unter klinischen Bedingungen.

Schlussfolgerungen: Die Validierung von Smartphone-basierten Anwendungen wie der SkinScreener©-Anwendung in einem nicht-klinischen Umfeld kann entscheidend sein, um ausreichende Leistungsdaten für solche Anwendungen zu erhalten. Es muss ein geeigneter Referenzstandard gefunden werden, da der Goldstandard mit histopathologischer Verifizierung in einem solchen nicht-klinischen Umfeld nicht zugänglich ist. Der in Form

eines Expertengremiums verwendete Referenzstandard zeigte aus verschiedenen Gründen Schwierigkeiten, eine eindeutige Konsensmeinung zu erhalten, und wirft daher die Frage auf, wie eine solche Validierung unter Verwendung eines Expertengremiums als Referenzstandard in künftigen Studien verbessert oder durch andere Modalitäten ergänzt werden kann.

Abstract

Background: The diagnostic performance of CNNs for diagnosing different types of skin cancer has been developing promisingly in recent times. The use of a smartphone application that can give lay users a basic assessment with the help of an integrated AI may guide them to take faster therapy when necessary.

Objective: The performance of a certified smartphone-based neural network application on macroscopic images of skin lesions taken by laypersons in real-world conditions has not yet been reported and is subject to this study. The main objective of the study is to evaluate the risk-assessment accuracy of the smartphone application in comparison to a consensus opinion of a medical Expert Panel.

Methods: We analyzed the detection accuracy of the CE-marked algorithm of the SkinScreener© smartphone application with the detection by a consensus opinion/reference standard of an Expert Panel of three dermatologists at the Medical University of Graz.
The primary outcome measures were sensitivity, specificity, and accuracy for the trichotomous risk assessment (low-, medium-, and high-risk). Secondary endpoints included interindividual differences in the dermatologists' diagnostic performance of analyzing the respective skin lesions.

Results: The CE-marked smartphone algorithm's performance in risk assessment was 76.9% (CI: {71.7% - 81.5%}) for sensitivity and 80.9% (CI: {78.5% - 83.2%}) for specificity. The overall accuracy was 77.2%. As a secondary endpoint, interindividual differences in dermatologists' diagnostic performance were found to be significant, with 526 out of a total of 1428 cases not showing complete agreement. It was shown that the performance was worse than in a previously conducted study under clinical conditions.

Conclusions: Validation of smartphone-based applications such as the SkinScreener© application in a non-clinical setting can be crucial to obtaining sufficient performance data for such applications. A suitable reference standard needs to be found as the gold standard when histopathological verification is not accessible in non-clinical settings. The reference standard used in the form of a consensus opinion showed difficulties in getting a clear consensus opinion for various reasons and thus raises the question of how such a validation

using an Expert Panel as a reference standard can be improved or supplemented with other modalities in future studies.

1 Introduction

1.1 Skin cancer epidemiology

Skin cancer is the most common form of malignancy, putting a significant burden on public health concerns (1). One must differentiate between melanocytic skin cancer (cutaneous melanoma) and NMSC or, according to newer nomenclatures, KC, which comprises BCC, SCC, and AK (considered as an in-situ KC) (2).

Melanoma is reported to be among the most rapidly rising forms of cancer worldwide, with a steady increase in incidence rates among fair-skinned populations over the past four decades. This has occurred even though survival rates and the detectability of thin melanomas have enhanced. Despite melanoma accounting for less than 5% of all skin cancer lesions, it is responsible for the vast majority of associated deaths (3). Notably, melanoma has a good prognosis quoad vitam when detected early; however, the 5-year survival rate strongly varies from 92% for stage I melanomas to 16% for stage IV, based on the AJCC TNM system (4).

KCs, formerly known as NMSCs, are responsible for the majority of skin cancer cases. Unfortunately, correct epidemiologic data are lacking as adequate registration is often not standardized or even not compulsory in various countries, such as the U.S. (1). This may diminish the accurate comparison of incidences (2). The most recent data from the U.S. suggests more than 5.4 million diagnosed cases of NMSC in 2012, including invasive NMSC and in situ NMSC. In this American study, only BCC and SCC are summarized under the term KC (1). Recent data from Germany is already based on the term KC, including SCC and BCC, as well as actinic keratoses. In Germany, about 1.9 million cases of KCs have been reported. In comparison, Australia has a very high number of KCs in relation to the population (24.1 million inhabitants), with more than one million new cases annually (2).

It is assumed that incidences of BCC and SCC are increasing annually at a rate of 3.3% to 11.6% (5), estimating that incidences will have doubled in 2030 (6). Due to the previously described non-standardized recording of tumor registries, the actual incidence is probably underestimated (5), obscuring the true extent of both prevalence and incidence. In the meantime, skin cancer is described as an epidemic-like condition in the literature (7) and is also recognized as an occupational disease of outdoor occupations (2).

1.2 Skin cancer prevention

1.2.1 Primary prevention

It is common knowledge that solar ultraviolet radiation is a major risk factor for both KC (8) and melanoma (3). Interestingly, there are gender-specific differences for melanoma in the age group 20 to 49, showing that women aged < 44 years have increased incidence rates with a peak difference between 20 and 24 years. On the other hand, the highest incidences for men are reported at age > 44 years.

However, this inverse trend has not been observed in KC (3), indicating that gender-related factors (e.g., endogenous hormones) may only influence the pathogenesis of early-onset melanoma (9).

Recent data indicated a correlation between an HPV infection and the development of KC. Both immunocompromised and immunocompetent individuals show the simultaneous presence of HPV DNA in 90% and 50% of cases, respectively (10). However, recent studies which have detected certain HPV types at very low viral loads have prompted the question of their relevance and have suggested that HPV may act more as a facilitator of carcinogenesis rather than a direct causal agent, thus enabling the accumulation of DNA breaks and somatic mutations caused by UV radiation (11).

The aim of primary prevention should be the reduction of UV radiation (12). Increased awareness about the harm of excessive UVR exposure is required, indicating that the implementation of safety measures may positively affect the rate of tumorous skin disease incidence (13). However, there are concerns suggesting that limited knowledge and awareness, specifically for individuals belonging to the high-risk group (e.g., professions pursued outdoors), play a supporting role in why primary prevention lacks adequate sufficiency (14).

This emphasizes the need for adequate secondary prevention, leading to an early diagnosis and initiation of the necessary therapy.

1.2.2 Secondary Prevention

Early detection is crucial in risk reduction cases for melanomas and KC (13).

For KC, an HPV vaccine may be used as an adjuvant treatment intervention in patients with recurrent KC in the future (11). For melanoma specifically, early recognition diminishes further progression, metastasis, and local destruction and prevents patients from extended

surgical removals or possible side effects of systemic therapies for late-stage patients. Furthermore, patients with little response to specific therapies could be saved. Moreover, eventually, early recognition is likely to reduce overall care costs (15, 16).

Skin cancer is usually diagnosed with a naked-eye examination combined with dermoscopy and followed by histopathologic evaluation (16, 15). Most often, a suspected lesion is detected by SSE. Compared to SSE, CSE is superior because of earlier diagnosis and thinner melanoma when being diagnosed (17). This is due to the higher competence of dermatologists in diagnosing pigmented skin lesions compared to other professions, leading to faster detection of thinner lesions (18). Despite this, screening rates for skin cancer compared to those for breast or colorectal cancer are considerably low; missing screening guidelines for skin cancer may be a possible explanation (19).

However, although several GPs and dermatologists offer preventive skin examinations, it is not yet commonly used by large parts of the population. Approximately 33% of Australian individuals had their skin checked by a doctor within a 12-month period (20). In Australia, individuals may seek a skin check from a GP, a skin cancer clinic (typically composed of GPs who specialize in skin cancer), or a dermatologist (20). PCPs, in general, are composed of non-physician PCPs (such as physician assistants or nurse practitioners) as well as PCP physicians, such as GPs.

PCPs are particularly important in skin cancer diagnosis, as pigmentary lesions are often detected by PCPs; e.g., the majority of tumorous skin diseases in Australia are diagnosed and treated entirely by PCPs (17). In the US, only 8% of patients had undergone a skin examination after consulting a PCP (21).

In Germany, however, nationwide population-based skin cancer screening has proven to be feasible and effective, although a decrease in melanoma mortality has not been observed in all states, and the benefit-harm ratio of skin cancer screening and its cost-effectiveness need to be further researched (22). Still, CSE is not commonly performed by GPs. Reported reasons for this include a lack of time (70%), insufficient training, and low confidence in correctly assessing skin lesions (17) and may contribute to the burden PCPs have in detecting malignant skin lesions. Interestingly, Samaran et al. found that 98% of GPs face particular difficulties in diagnosing NMSC (23), whereas the general accuracy for the detection of skin

cancer, both melanocytic and keratinocyte carcinoma, by non-specialists, such as PCPs, is reported to be just between 24% and 70% (24).

One possible solution is to involve people more in the early detection of skin cancer in order to raise their awareness of any changes. Moreover, partner-assisted skin examination, performed periodically (if possible, monthly) may also be beneficial for early detection (2).

1.3 Smartphone-based applications for skin cancer detection

Due to the almost ubiquitous use of smartphones and their availability and usage by individuals (more than 6.3 billion global smartphone users in 2021) (24, 26), such a technology could be very beneficial in the early detection of skin cancer and in reducing the burden of rising skin cancer incidences. These applications, therefore, might play a role in triaging patients (27) based on their individual risk for skin cancer.

Following the increasing use of smartphones associated with technical improvements concerning picture quality and performance, early detection of skin cancer has become possible. In this context, particular attention is paid to smartphone applications using AI that are able to recognize skin changes and categorically assess them according to their respective risk. These applications are based on algorithms that "extract and analyze criteria of skin lesions […]" (28). The algorithms later stratify the lesions into "low risk," "medium risk," or "high risk" using pattern recognition software. Based on this risk assessment, further recommendations are given.

Notably, these technologies have to fulfill two criteria: They have to have a high accuracy in detecting skin cancer as a miss-classified malignancy (e.g., melanoma) may have fatal consequences (27). Apart from that, these algorithms also must show high accuracy in detecting benign skin lesions in order to avoid unnecessary excisions and consultations.

Although these algorithms show advantages such as objective and quantitative feature extraction compared to human examination (29), smartphone images strongly depend on various external factors (e.g., zoom, angle, lighting) (26). Moreover, irrelevant features and artifacts may be hard to differentiate from important biological features. This deficiency of computational neuronal networks (CNN) is called 'brittleness.' Even small changes in image acquisition, which are of limited perceptibility to humans, can have a significant impact on the brittleness and consequently safety and diagnostic accuracy of a CNN to distinguish tumorous skin lesions (29).

The use of a smartphone for skin cancer screening is only reasonable if the corresponding compliance and trust in such a technology are present. Jutzi et al. conducted a web-based questionnaire showing that the acceptance of patients is generally high, depending on the scenario in which an AI system is used. Notably, 94% would endorse the use of AI as an assistant system at a doctor's office. 56% declared that they would also use AI applications at home, e.g., as a diagnostic smartphone app (30). Another European study showed that GPs' acceptance of such technologies is also very high; 86% stated they were in favor of using AI at least to detect NMSC (23). However, as most AI algorithms only target the diagnosis of melanoma, clinical practice requires that these technologies can also be used for SCC, an essential point to consider when developing and evaluating the algorithms (31).

Currently-available applications have two ways of working: The pictures are either evaluated by a tele-dermatologist or undergo a risk assessment by the algorithm. An even more special access would be triaging through smartphone-based apps, which, in the case of a higher risk, enable further digital clarification by a dermatologist through the implementation of a teledermatology service (25). In the literature to date, the performance of such smartphone-based AI applications for the assessment of skin lesions has only been tested in a clinical setting (32).

1.4 Current issues with smartphones-based applications:

As the use of AI technology in smartphone apps has become available to the general public, questions have arisen regarding the reliability and accuracy of these apps in providing medical diagnoses. Previous studies have caused controversy and debate by examining the diagnostic performance of smartphone apps for melanoma detection, as the accuracy of these apps has demonstrated a high level of variability (33).

The use of macroscopic images as the key limitation of the smartphone app for the classification of skin lesions can be seen as a weakness in its performance. It is essential to utilize close-up dermoscopic images under present conditions in order to achieve the most accurate diagnosis, whether it is evaluated by a physician or by an AI-based algorithm. Magnification attachments are already available for smartphones and could provide such detailed images. However, the implementation of these attachments would involve additional costs for laypersons and potential handling challenges.
This could make independent screening with the application more complicated (33).

Studies to date mainly investigate the performance claims of smartphone apps in a clinical setting. However, there are numerous implications from the literature calling for a study setting for performance evaluation in a non-clinical setting but under real conditions (25, 34).

Recommendations and requirements for conducting performance studies evaluating smartphone apps have been formulated by Du-Harpu et al. and Freeman et al. Most previous studies have failed to recruit representative samples of the general population, which is typically the target population for smartphone apps. The authors noted a relatively low prevalence of malignancy and a broad range of skin diseases in these studies (25).

Moreover, the authors noted that previous studies have only evaluated lesions or images selected and acquired by clinicians rather than lesions judged to be of concern to people using smartphone apps themselves. This bias should be addressed to assess the potential benefits of their availability and use outside the health system (25). It is noteworthy that the neural networks in previous performance studies were typically trained and evaluated on the same dataset (35).

In a previous study, the performance of the SkinScreener© application was already tested under clinical conditions with a performance of 96.4% (CI: {93.94% - 98.85%}) for sensitivity and 94.85% (CI: {92.46% - 97.23%}) for specificity for the detect algorithm that is used for risk evaluation, which is still in place for the current study. The authors concluded that the algorithms used in the study may have a positive impact on the healthcare system and reduce unnecessary visits and histological examinations. However, they also concluded that the study population lacked generalizability in terms of their risk profile, age, and skin type compared to the general population (32).

In contrast to other studies, we tried to incorporate and address the aforementioned demands and criticisms as far as possible in the study design. The clear demand for a real-world setting should be seen as a central strength in the present study since the users of the app analyzed their skin lesions with the mobile phone app on their own without any influence, and only their images and skin lesions are included.

1.5 Technical structure of the developed algorithm

The CNN developed for this study is based on a machine-learning framework using a detect algorithm that takes image data and processes them through a faster_rcnn_inception_v2_coco model. The inputs for this neuronal network are images that contain one or several skin lesions (details are given in Chapter 1.6.). Each image is processed by a bounded box, a label, considered as a suspected diagnosis that is assigned to the specific box. The matching probability varies from 0 to 100%. These probabilities are assigned to the morphological elements of the different skin lesions so that the skin lesion can be attributed to one of the three risk classes according to the suspected skin lesion.

A CNN allows input in the form of a matrix (width x height x color channels). The developed CNN has a multi-level structure comprising a convolutional layer, an RPN and a newly developed architecture (classes and bounding boxes prediction), which are described below. Figure 1 is intended to provide a graphical overview of the multi-level structures discussed in the following Chapters and to show the interactions between the components.

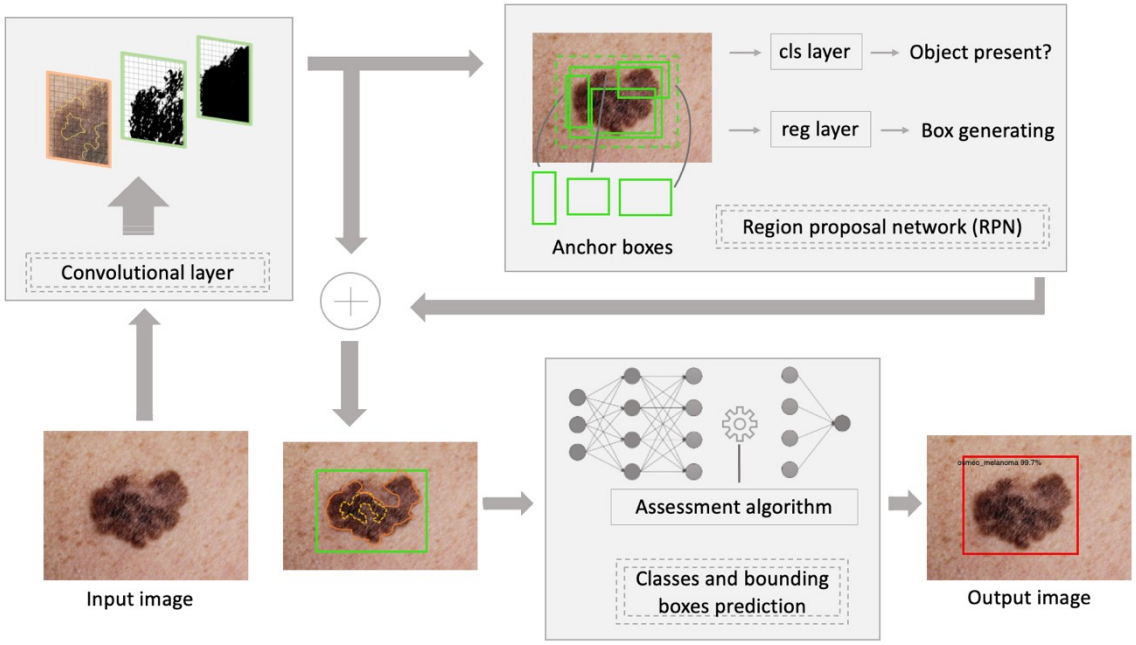


Figure 1: Overview of the technical structure of the developed algorithm

1.5.1 Convolutional layer

The convolution layer was trained to extract the appropriate features of an image. The process of convolution is a linear operation between the input data and a kernel or a so-called "filter." Applying this filter makes it possible to search for a particular type of morphological feature in that image. The output is a feature map.

For example, one filter searches for the presence of keratinized plaques that appear yellow-grey-brownish with a very texture-rich structure. Another filter scans for the presence of bleeding structures appearing reddish. By combining the results of these filters, the AI can indicate whether SCC is a possible preliminary diagnosis.

As the filter operations are run through several times, the level of abstraction of the network increases. In the first stage, only simplified structures trigger filters, such as lines; meanwhile, in the following stage, filters can only be triggered by more complex structures, such as curves or shapes. A series of dozens or even hundreds of other filters can be further developed to detect many more features within the image (36).

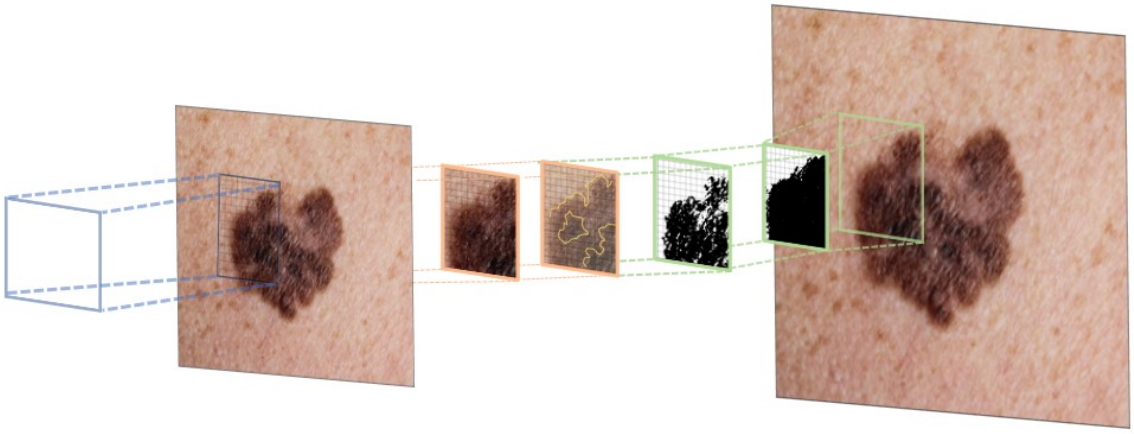


Figure 2: Illustration of the convolutional layer for image processing

1.5.2 Region Proposal Network

The RPN is a small neural network on the last feature map of the convolution layers serving for object localization by generating boxes around objects. The structure of RPN consists of a 'Classification Score layer,' indicating whether the object is present in the area or not, and a 'Regression coefficient for boxes' layer generating a bounding box around the object (37). It must be emphasized that this convolutional layer cannot interpret the lesion or object within the box.

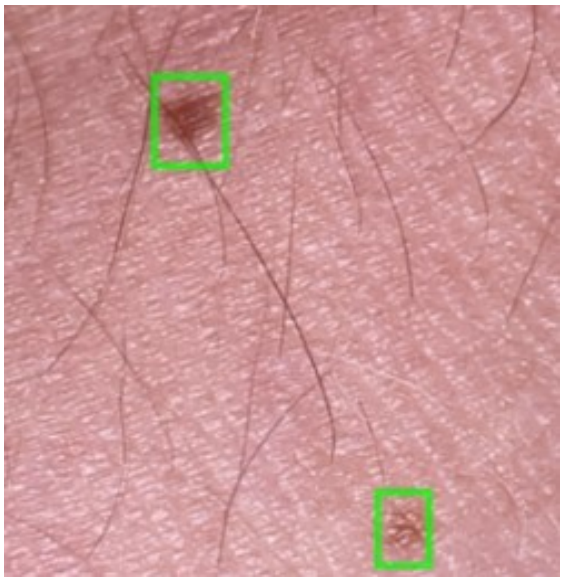


Figure 3: RPN for boxing lesions

1.5.3 Classes and Bounding Boxes Prediction

The newly developed architecture called "classes and bounding boxes prediction" takes the RGB pixel values of an image as input and calculates the label probabilities representing the suspected diagnosis for each box. Afterwards, the most likely diagnosis or label and its probability are returned. The lesion objects thus contain the label of the lesion category, the coordinates of the lesion box in the image, and a score representing the probability of the lesion that the algorithm has detected. The score is a decimal value between 0.0 and 1.0; a score of 0.0 means that the algorithm has detected the lesion object with no probability, and a score of 1.0 means that the algorithm has detected the lesion object with a probability of 100%.

In order to assign the respective diagnosis/label to one of the three risk classes, another assessment algorithm is needed. This assessment algorithm assigns the object to the corresponding risk group (low-risk, medium-risk, high-risk) depending on the percentage

diagnosis and other additional factors, as explained below. Chapter 1.5 explains in detail which diagnoses belong to the respective risk groups.

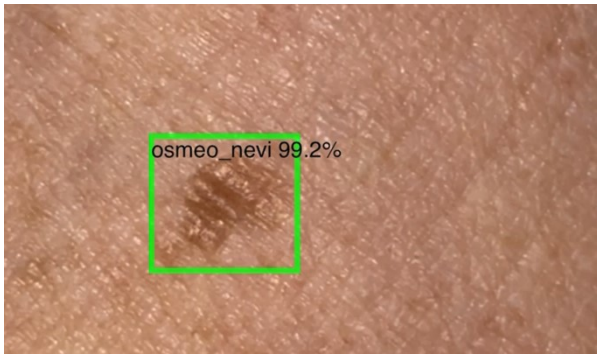


Figure 4: Assigned probability to detected lesion

The aforementioned factors for the correct assignment of the object to the correct risk group include *threshold values* which ensure that no false-positive results are obtained. For example, if the suspected diagnosis of AK falls below a certain probability threshold, this post-connected assessment algorithm does not assign it to the medium risk category.

Furthermore, these thresholds ensure that the risk score is updated to the appropriate higher-risk class (medium- or high-risk) if a disease-specific threshold of a precancerous or malignant skin lesion is exceeded. The specific values of the aforementioned thresholds are listed in Table 1 below:

Lesion type	**Threshold in %**	**Adjusted risk**
dysplastic nevus	17	medium
actinic keratosis	20	medium
basal cell carcinoma	18	high
bowens disease	21	high
melanoma	24	high

Table 1: Threshold values

Several objects can be detected within one image. An additional factor removes overlapping or crossing boxes. The removal of overlapping lesion boxes is done by considering the most relevant lesion for the risk assessment. If more than one lesion is within a box, relevance is considered as follows:

- First, only lesion objects with the highest risk are considered relevant, whereas lower-risk overlapping or intersecting lesion objects are removed.
- Second, for a lesion that can be described with different diagnoses, the highest-risk diagnosis with the highest probability score is taken into account, and the lower-risk ones are disregarded, so that the lesion with the highest risk is always identified and analyzed.

1.6 Taxonomy

A three-level classification according to the risk levels “low,” “medium,” and “high,” was chosen for categorization. The respective risk corresponds to the root nodes, whereas the diagnoses assigned to the corresponding risk category represent the leaf nodes, resulting in a tree-like taxonomy structure. The algorithm can detect 46 entities.

Non-neoplastic lesions, anatomical structures, and benign skin lesions were assigned to the low-risk category. A detailed listing of the included lesions is given in Table 2, whereas in Figure 5, a graphical overview is given:

Figure 5: Taxonomy

To improve performance, the algorithm was trained not only with images of skin lesions but also with images of healthy skin, skin injuries, anatomical structures, and other skin-like objects. Regarding anatomical structures, it should be noted that lesions on or in proximity to anatomical structures can look substantially different than normal. A high resemblance between melanoma and hematoma exists, for instance, when located under the nail. Therefore, skin lesions on or near the ear, eye, hair, mouth, nipples, genitals, fingers, toes, or navel were excluded from analysis by the algorithm, sometimes also due to the lack of sufficient image data to train the algorithm for those lesions.

The medium-risk group includes dysplastic nevi and actinic keratoses; the high-risk group includes SCC, BCC, melanoma and Morbus Bowen.

A list of all included diagnoses within the three risk classes is shown in Table 2:

Risk group	**Category**	**Diagnosis**
high-risk	malignant	basal cell carcinoma
		morbus bowen
		squamous cell carcinoma
		melanoma
medium-risk	precancerous	dysplastic nevus
		actinic keratosis
low-risk	anatomical structure	ear
		fingertips with nails
		fingertips without nails
		eye
		hair
		mouth
		nipple
		penis
		toe
		umbilicus
		vagina
	non-neoplastic	eczema
		comedo
		cyst
		dermatomycosis
		epidermal cyst
		hematoma
		jewelry
		keloid
		piercing
		psoriasis
		red macula
		red papule
		scar
		skin injury
		tattoo
		ulcus
		urticaria
		varicosis
		watch
	benign	dermal nevus
		dermatofibroma
		fibroma
		hemangioma
		hyperpigmentation
		hypopigmentation
		micro nevus

		nevus
		seborrheic keratosis
		verruca vulgaris

Table 2: Overview of diagnoses assigned to the risk groups

2 Methods

2.1 Study design

This was a retrospective single-center study at the Department of Dermatology and Venereology in Graz. The study was approved by the local ethics committee (approval number: 34-070 ex21/22 1508-2021) and was performed in accordance with the Declaration of Helsinki. The study was performed to evaluate the risk assessment of a smartphone-based neuronal network in the application by laypersons in comparison to an Expert Panel of three dermatologists.

The study design was planned in accordance with existing recommendations for optimizing statistical planning of performance studies for the evaluation of smartphone apps, as stated in Chapter 1.3. Based on previous publications of Kränke et al. (32), we defined three risk classes (green/low, yellow/medium, red/high) indicating the respective risk of a lesion being malignant and the further approach as follows:

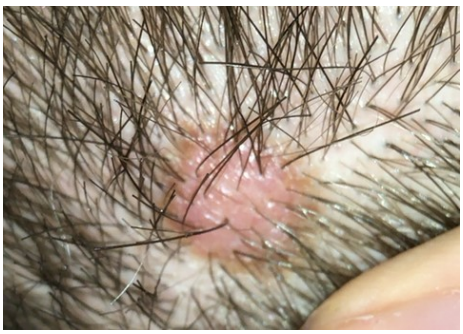
- Green: Benign, no action needed
- Yellow: Suspicious, timely dermatological examination needed
- Red: Highly suspicious, immediate dermatological examination needed

Since the used algorithm assigned a risk assessment as well as a diagnosis, we also assigned a diagnosis to the respective lesions from the selection of diagnoses recognizable for the algorithm. If the lesion did not correspond to one of these diagnoses, there was the possibility of naming a different free-text diagnosis. The dermatologists were blinded, i.e., they were only presented with the image and not the assigned assessment of the algorithm or other dermatologists. No other medical history parameters were available for the images, such as sex, age, development, location, or family history.

2.2 Participants

The participants recruited for the study were users of an application with an integrated neural network who consented to the processing of their data by agreeing to the privacy policy. The participants took images of the respective lesion(s) with their smartphone cameras. All patients reported being over 18 years of age and having a Fitzpatrick I-IV skin type. In order to fulfill data protection guidelines, data were fully anonymized and every participant got an internal ID. Hence, no further information (e.g., sex, age, development/duration of the lesion, location, or family history) could be recorded.

Besides an age <18 years and skin types V and VI, the exclusion criteria comprised all factors that could have a negative impact on a correct and valid image evaluation. These criteria are depicted in Table 3:

The user has skin type V (dark brown) or VI (darkest brown), according to Fitzpatrick
The lesion has low visual contrast to the surrounding skin area
The lesion is surrounded or covered by hair 
The skin is sunburned 
The lesion has previously been traumatized (excised/biopsied)
The surrounding skin is not intact (e.g., open wounds, ulcers, bleeding, irritation) 
The lesion is located on or next to anatomical structures (“special sites”) such as the ear, eye, genitals, hair, mouth, nails, nose, or nipples


The lesion is very close to scars or tattoos or areas partly or fully covered with opaque or glittering substances, such as make-up or any kind of skin cream 
The lesion is on mucosal surfaces, such as the lips, in the mouth or in the genital region
The lesion is in a skinfold 
The lesion is not on human skin
The scanned region is even partially covered by clothes
The lesion is not captured in focus

Table 3: List of exclusion criteria

The recruitment phase ran from 1.6.2021 to 1.8.2021.

2.3 Reader study

The skin lesions analyzed by the AI were provided in the form of jpg files with associated XML files for coding the image ID, risk class, diagnosis and percentage estimate of diagnosis. In order to systematically transfer these data into the Excel document, the data were automatically entered using a VBA program.

The cases in the Excel file were presented to the three dermatologists (TK, KTD, RHW) and they were asked to assign each lesion to a risk group and indicate their diagnosis.

As only images were collected without any further information from the participants, histopathological examination was not performed as the gold standard; therefore, we chose the consensus opinion of dermatologists as the reference standard described below. Hereinafter, we will call the reference standard chosen for us the ‘consensus opinion’ or, as an abbreviation, nRS. The consensus opinion was formed from the opinion of three clinically active dermatologists with different levels of training, experience, and specialization.

The risk classification of the algorithm of the respective lesion was deemed correct if the Expert Panel agreed with at least a two-thirds majority. In order to minimize possible differences in the interindividual assessment of the dermatologists, a joint review process was conducted if no consensus could be reached. The following exceptions to the established two-thirds majority of dermatologists were defined:

1. All lesions that were assessed by the three dermatologists with all three risk ratings, i.e., a lesion was classified as low, moderate or high risk.

2. Two out of three dermatologists rated a lesion as low-risk and the third dermatologist as high-risk.
3. Two out of the three dermatologists rated a lesion differently, while the third dermatologist did not give a rating, so no consent could be made.

Despite the existing two-thirds majority, in the case of exception number 2, a considerable dilemma between a clinically completely harmless suspicion (low-risk) by two dermatologists and the maximum finding of a malignant assessment (high-risk) by the third dermatologist should be reconsidered and discussed again among all the assessing dermatologists.

In the case of exception number 3, the lack of evaluation by one of the three dermatologists in the case of different evaluations by the other two dermatologists led to an unclear consensus opinion which required a new evaluation by all three.

At the joint review, the lesions with no consensus opinion were discussed together and not blinded to the other dermatologist's opinion but to the risk assessment by the AI. Special care was taken to ensure that, in this review process, the clinical judgment of the most experienced dermatologists did not influence the others, and vice versa. In order to ensure this, the dermatologists took turns in expressing their suspected diagnosis so that each could express their suspicions without influencing the others.

2.4 Review process with the senior dermatologist (RHW)

An additional review process was initiated to analyze lesions incorrectly assigned by the algorithm. This was done by the senior dermatologist (RHW), who was not blinded to the risk assessment and diagnosis by the algorithm in this setting.

This was performed for several reasons:

- To search for potential causes of the AI's incorrect classifications in comparison to the dermatological consensus opinion assessment
- To compare the performance (concerning both risk assessment and diagnosis) of the AI with that of the senior dermatologist.
- To investigate the performance of the AI compared with the individual assessment of each dermatologist and with the consensus opinion

For this purpose, the following procedure was chosen for selection:

- Lesions that were classified as low-risk by the algorithm, whereas the consensus opinion was medium- or high-risk
- Lesions that were classified as high-risk by the algorithm, whereas the consensus opinion was low-risk[1].

2.5 Procedure

2.5.1 Sample size calculation:

A power analysis was carried out for the sample size calculation. The sample size was estimated using the software nQuery. In order to verify the non-inferiority of the app in terms of risk assessment of lesions (sensitivity), the non-inferiority margin was set at 90% based on the FDA's requirements (38).

With a sample size of 399 lesions using an exact one-sided test (alpha=2.5%) and assuming that a sensitivity of 93.94% is observed, a statistical power of the test of >80% can be achieved. In a previous study to test the diagnostic accuracy of this tool in differentiating various benign (no-therapy) and malignant (therapy) skin lesions, conducted at the University Department of Dermatology between 11/2018 and 12/2019, a sensitivity of 96.4% (95% confidence interval 93.94%-98.85%) was observed (32). The lower limit of the confidence interval (93.94%) was used for case number planning. The resulting sample size (399 lesions) represents the number of images/lesions assigned to the "therapy" category (medium-risk and high-risk) by the Expert Panel. Assuming that this applies to approximately 27% of the images/lesions, a total sample size of 1428 images/lesions is required, which also includes the category "no therapy" (low-risk).

2.5.2 Data management

According to the sample size calculation, 1428 skin lesions had to be recruited. The measures of interest for statistical evaluation were sensitivity and specificity. In order to determine these with the highest possible precision, a weighted number of cases was used, i.e., a statistically required number of lesions from each of the three risk categories. Thus, the 1428 images were divided among the three risk categories in a ratio that was not known to the dermatologists but was needed statistically. This blinding of the examining dermatologists

[1] The classification/comparison of the medium category was deliberately not carried out here. It could be determined that only with the aid of the photo without progress pictures and clinical history, it was often not possible for the participating dermatologists to make an unambiguous classification in the medium-risk range by consensus.

was carried out to prevent their own decision from being influenced by the information about the exact quantitative distribution to the respective dignity classes.

In order to make the selection process as unbiased as possible, the following procedure was chosen.

In order to increase the statistical quality of the data collection, a large number of primary lesions (12,766) were mainly selected. From this primary dataset, the number of cases above the required number of directed cases in the three risk classes was randomly selected to obtain the secondary data set. For this purpose, a special program was written in Python that took a correspondingly large number of lesions of the respective category from the server and reduced them randomly in such a manner that the secondary dataset was finally created. This secondary dataset was then reviewed manually by the investigator (PE) and further screened out according to the inclusion and exclusion criteria in the instructions for use. This was necessary because abusively taken photos that do not comply with the instructions for use should not be included in the image database and thus should not be added to the patient collective to enable a scientifically fair comparison. However, it may be criticized that this could lead to the introduction of a selection bias (see Chapter 4.1.3).

2.5.3 Statistical Evaluation

The calculations were performed by using IBM SPSS Statistics 28. To measure inter-rater variability, Cohen's Kappa is used.

Cohen's Kappa is a method of measuring the amount of agreement between two judges when assessing something on a nominal scale (i.e., assigning it a category such as "yes" or "no"). It takes into account the proportion of units for which the raters agreed (p_o) and the proportion of units for which agreement is expected by chance (p_c). In order to calculate the agreement that is expected by chance, the chi-squared test can be used, which is equal to the row total multiplied by the column total divided by the grand total.

The coefficient (Kappa) is calculated by subtracting pc from po and dividing the result by 1-pc. The resulting proportion between 0 and 1 indicates how much agreement is due to something other than chance (39).

$$k = \frac{(pc - po)}{(1 - pc)}$$

p_o: Relative observed agreement among raters

p_c: Hypothetical probability of chance agreement

Thus, the inter-rater reliability of the dermatologists among each other on the one hand and between the AI and the respective dermatologists on the other hand can be presented. This results in a comparison between the accuracy of agreement between two human raters in each case, so that three different Kappa coefficients can be interpreted in this group (see Chapter 3.3.2). Likewise, a comparison can be made between the agreement accuracy of the artificial rater, the AI, and the respective human rater (see Chapter 3.3.3).

Inter-rater reliability is defined as "the measurement of the extent to which data collectors (raters) assign the same score to the same variable."(40). The Kappa result can be interpreted as follows:

Kappa	**Agreement**
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-0.99	Almost perfect agreement

Table 4: Classification of Kappa values

2.5.3.1 Sensitivity, Specificity and overall Accuracy calculations:

Sensitivity, Specificity and overall Accuracy calculations were performed by using IBM SPSS Statistics 28.

For the calculation of the sensitivity, a one-sided exact binomial test with an alpha α of 0.025 was used.

The unknown probability p is the sensitivity (detection of a lesion requiring therapy), with the following pairs of hypotheses being tested:

- Null hypothesis: $p \leq 90\%$
- Alternative hypothesis: $p > 90\%$

The sensitivity and specificity of the statistical measures are binary classifications, meaning that the three-stage risk classification has to be reduced to two classes. Therefore, the high- and medium-risk classes were merged, both requiring medical attention from a

dermatological perspective, whereas the low-risk class does not need any further clinical assessment.

Sensitivity describes the probability that a person with the disease will test positive accordingly. Specificity, on the other hand, refers to the probability that a disease-free person will test negative. Overall accuracy is the probability that a person will be correctly detected or tested by a screening or diagnostic test. It is the sum of the correct positive results plus the correct negative results divided by the total number of people tested, i.e. the weighted average of the sensitivity and specificity of a test (41).

2.5.3.1.1 Sensitivity

In order to calculate the sensitivity, the number of true positive high- and medium-risk lesions identified by the AI has to be compared to the total number of high- and medium-risk lesions identified by the consensus opinion. The exact Clopper-Pearson 95% CI will be used to calculate the corresponding confidence interval.

$$Sensitivity: \frac{number\ of\ correct\ high-risk\ assessment + number\ of\ correct\ medium-risk\ assessments}{total\ number\ of\ high-\ and\ medium-risk\ assessments}$$

2.5.3.1.2 Specificity

The specificity is calculated by identifying the lesions of the low-risk class in both the set of AI and the set of consensus opinions and comparing it to the total number of low-risk lesions by the consensus opinion. The confidence interval is then determined using exact Clopper-Pearson 95% CI.

$$Specificity: \frac{number\ of\ correct\ low-risk\ assessments}{total\ number\ of\ low-risk\ assessments}$$

2.5.3.1.3 Accuracy

In order to calculate the accuracy, the number of correct low-, medium-, and high-risk assessments must be compared to the total number of low-, medium-, and high-risk assessments.

$$Accuracy: \frac{number\ of\ correct\ low-risk + medium-risk +\ high-risk\ assessments}{total\ number\ of\ low-risk + medium-risk +\ high-risk\ assessments}$$

3 Results

A total of 1428 lesions were included. It is not possible to say how many patients correspond to this number of lesions, as no traceability was possible for data protection reasons.

3.1 Distribution of lesions by the algorithm – Risk assessment

A total of 1428 lesions were included; 978 lesions (69%) were assessed as low-risk by the algorithm, 372 lesions as medium-risk (26%), and 78 lesions as high-risk (5%). The assignments of lesions to the three risk classes was done by the algorithm, as shown in Figure 6.

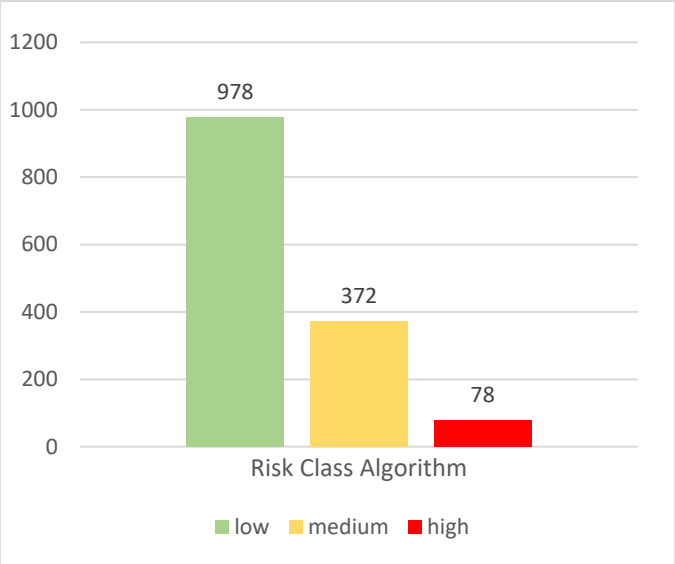


Figure 6: Distribution of lesions to the respective risk classes by the algorithm

3.2 Distribution of lesions by the algorithm – Diagnosis

The evaluation in terms of diagnosis by the algorithm is also given. The five most frequent diagnoses of the algorithm were nevus (537), dysplastic nevus (355), dermal nevus (126), verruca vulgaris (122), and seborrheic keratosis (109). All given diagnoses are shown in Table 5.

Risk group	Diagnosis	
low-risk	comedo	**4**
	cyst	1
	dermal nevus	126
	dermatofibroma	16
	eczema	1
	fibroma	23
	hemangioma	33
	hematoma	3
	hyperpigmentation	39
	hypopigmentation	3
	keloid	1
	nevus	537
	red macula	9
	red papule	14
	seborrheic keratosis	109
	skin injury	2
	urticaria	1
	verruca vulgaris	122
	varicosis	1
medium-risk	actinic keratosis	17
	dysplastic nevus	355
high-risk	melanoma	45
	squamous cell carcinoma	7
	morbus bowen	3
	basal cell carcinoma	23

Table 5: Distribution of the diagnosis assigned the algorithm

3.3 Interobserver Variability - Risk-Assessment

3.3.1 Distribution of lesions by the dermatologists

The interindividual assignment of lesions to the three risk classes, as done by the dermatologists, is shown in Figure 7.

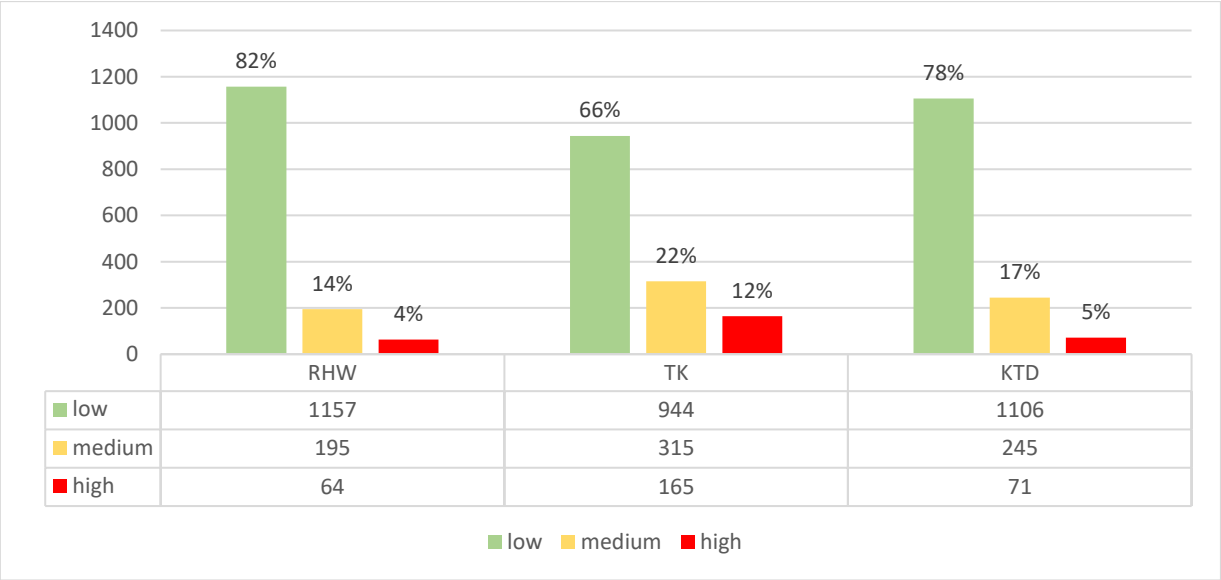


Figure 7: Distribution of the lesions to the respective risk classes by the dermatologists

3.3.2 Cohen-Kappa – inter-rater reliability between dermatologists

In our study setting, Cohen-Kappa scores ranged from 0.282 to 0.435. The comparatively low values led to the Expert Panel's need for a joint review process (see Chapter 3.5).

The agreements between the three participating dermatologists are demonstrated in Table 6:

Kappa Coefficient	RHW	KTD	TK
KTD	0.357		
TK		0.435	
RHW			0.282

Table 6: Kappa coefficient results for human inter-rater reliability

The best agreement corresponding to the highest inter-rater reliability was achieved between KTD and TK, with a score of 0.435 (moderate agreement); the lowest value between RHW and TK was 0.275 (fair agreement) and between RHW and KTD with 0.357 (fair agreement).

3.3.3 Cohen-Kappa – inter-rater reliability between AI and dermatologists

In order to determine the inter-rater reliability between the individual dermatologists and the AI as a rater, a comparison was made between the respective dermatologists and the AI in each case, which also resulted in three different kappa coefficients being interpreted. The agreements between the three participating dermatologists and AI are demonstrated in Table 7:

Kappa Coefficient	RHW	KTD	TK
AI	0.343		
AI		0.456	
AI			0.42

Table 7: Kappa coefficient results for inter-rater reliability between AI and dermatologists

The best agreement corresponding to the highest reliability was achieved between KTD and AI with 0.456 (moderate agreement) and between TK and AI with 0.42 (fair agreement); the lowest between RHW and AI with 0.343 (fair agreement).

3.4 Interobserver Variability - Diagnosis

3.4.1 Interobserver variability for the "medium" and "high" risk groups

The absolute numbers of the given diagnoses in the risk groups "medium" and "high" are given in Table 8.

Diagnosis		RHW	TK	KTD
medium-risk	dysplastic nevi	177	306	216
	actinic keratosis	11	8	18
high-risk	melanoma	40	125	52
	morbus bowen	4	6	0
	squamous cell carcinoma	1	4	2
	basal cell carcinoma	16	26	15

Table 8: Comparison of the assessments among the three dermatologists for medium- and high-risk diagnoses

3.4.2 Interobserver Variability for the risk group "low"

The most common diagnosis of all three dermatologists was "nevus" with an average score of 575 and a standard deviation of 170. All given diagnoses are shown in Table 9.

The five most frequent diagnoses of the participating dermatologist RHW were nevus (n=814), dermal nevus (n=129), seborrheic keratosis (n=73), other (n=36) and fibroma (n=24).

The five most frequent diagnoses of the participating dermatologist TK were nevus (n=472), dermal nevus (n=166), seborrheic keratosis (n=140), hemangioma (n=38) and other (n=37).

The five most frequent diagnoses of the participating dermatologist KTD were nevus (n=438), seborrheic keratosis (n=265), dermal nevus (n=195), other (n=76) and fibroma (n=36).

Diagnosis		RHW	TK	KTD
low-risk	cyst	1	0	0
	dermal nevi	129	166	195
	dermatofibroma	15	23	34
	eczema	7	2	0
	fibroma	24	20	36
	comedo	0	0	2
	hemangioma	16	38	31
	folliculitis	0	1	0
	hematoma	2	2	3
	hyperpigmentation	10	2	5
	hypopigmentation	3	0	3
	nevus	814	472	438
	other	36	37	76
	psoriasis	3	0	2
	red macula	6	0	8
	red papule	11	1	6
	pustule	0	1	0
	scar	2	1	2
	seborrheic keratosis	73	140	265
	skin injury	9	4	4
	solar lentigo	0	31	0
	tattoo	1	0	1
	ulcus	1	0	0
	verruca vulgaris	3	2	3
	varicosis	0	0	1
	micro-nevi[2]	19	0	15

Table 9: Comparison of the assessments among the three dermatologists for low-risk diagnoses

[2] The diagnosis of a micro nevus was introduced by the developers of the app. This was to take into account particularly small nevi. Since the size of a nevus is difficult to estimate without calibration or size standardisation, both for the participating dermatologists and technically for the algorithm, the diagnoses micro nevus and nevus were added together under the diagnosis nevus. However, the micro nevus diagnosis is listed separately in the overview for completeness.

3.5 Results of the joint review process

As described in Chapter 2.3, a joint review was necessary because a consensus opinion could not be reached for all lesions. The distribution of lesions for the joint review using the rules and exceptions was done as follows.

- 49 lesions were each given a different risk rating by the three dermatologists, i.e., a lesion was classified as low, medium or high risk.

- 32 lesions were classified as low risk by two of the three dermatologists, and the third dermatologist classified them as high risk.

- 4 lesions were rated differently (either low-, medium-, or high-risk) by two of the three dermatologists, while the third dermatologist did not submit an assessment.

By applying the following rules, 85 lesions (6%) were thus selected and discussed for the joint review. The joint evaluation showed that 10 of the 85 diagnoses were assessed as high-risk, 19 as medium-risk, and 56 as low-risk. The distribution of lesions before and after the joint review process is given in Figure 8.

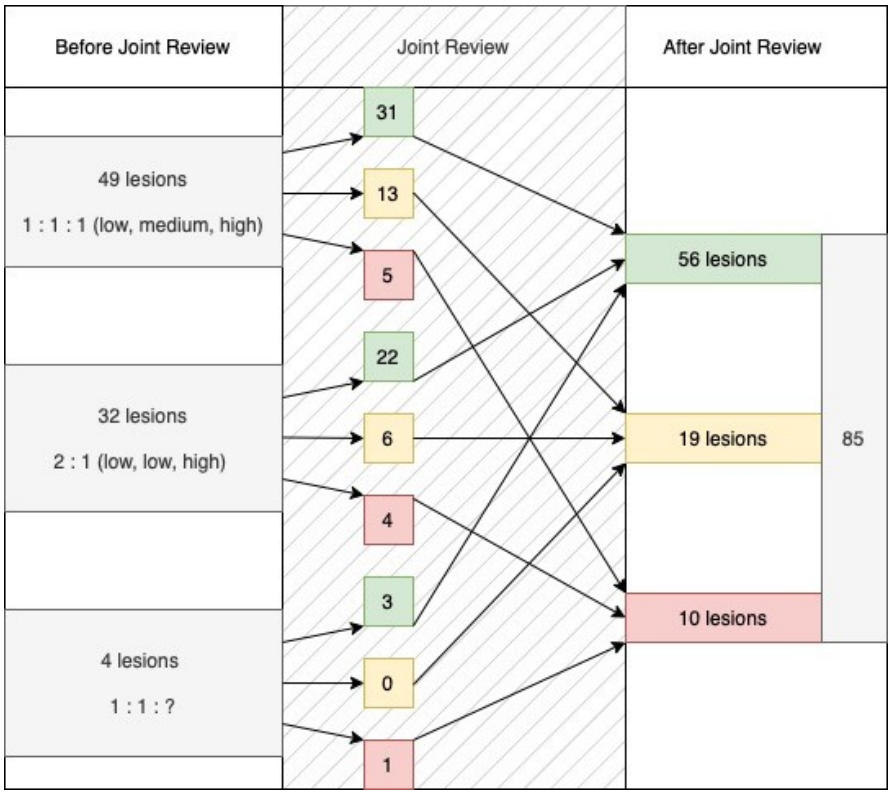


Figure 8: Illustration of the joint review process

3.6 Comparison of the algorithm with the consensus opinion

3.6.1 Comparison of risk classes assessment

Figure 9 compares the risk class score of the AI with the consensus opinion of dermatologists. The absolute values of the algorithm are shown in hatched rows, and the absolute values of the consensus opinion are shown in solid rows.

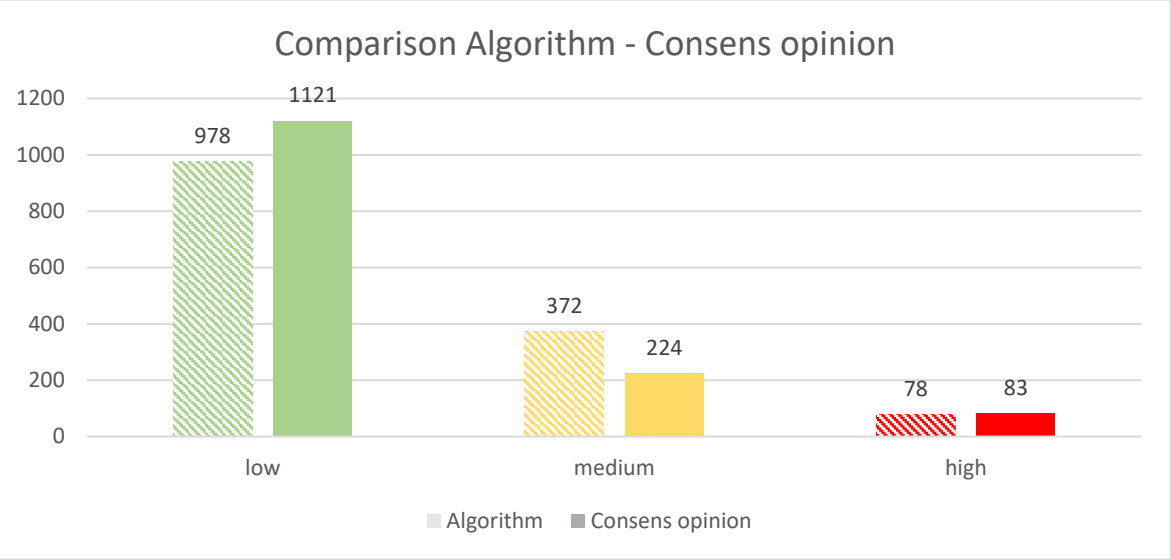


Figure 9: Comparison of lesion distribution of the algorithm compared to the consensus opinion

3.6.2 Sensitivity and specificity calculations

The following crosstabulation compares nAI with nRS as "high and medium risk" and "low risk," respectively. These numbers were the basis for the calculation of sensitivity and specificity.

		nRS		Total
		low-risk	high- and medium-risk	
nAI	low-risk	907	71	978
	high- and medium-risk	214	236	450
Total		1121	307	1428

Table 10: Crosstabulation for sensitivity/specificity calculations, based on the consensus opinion

Statistical analyses revealed a sensitivity of 76.9% % (CI: {71.7% - 81.5%}) and a specificity of 80.9% (CI: {78.5% - 83.2%}) of the investigated algorithm.

3.6.3 Overall accuracy

In order to make a general statement regarding the overall level of accuracy, risk assessments were compared in their three-level classification.

			Consensus Risk Class			
			high	low	medium	Total
Algorithm Risk class	high	Count	39	28	11	78
	low		15	907	56	978
	medium		29	186	157	372
Total			83	1121	224	1428

Table 11: Crosstabulation for accuracy calculation - based on the consensus opinion

The overall accuracy concerning the correct risk assessment was 77.2%.

3.7 Results of the review process with the senior dermatologist (RHW)

As described in Chapter 2.4, an additional review process was carried out to analyze the lesions incorrectly assigned by the algorithm. The results of this review process are presented below:

- 71 lesions were classified as low-risk by the algorithm, whereas the consensus opinion was medium- or high-risk.
- 28 lesions were classified as high-risk by the algorithm, whereas the consensus opinion was low-risk.

3.7.1 False high-risk lesions

After further analysis by RHW, 10 of the 28 lesions assessed by the expert consensus as false high-risk assessments (high-risk assessment by the algorithm) were ultimately found to be correct high-risk lesions as assessed by the algorithm.

In the case of the remaining 18 lesions, large parts of the lesions showed artifacts, reddish parts, etc., which made a correct assessment by AI difficult. According to RHW, the proportion of red tones in four of the lesions led to a strikingly high assessment as high-risk lesions. Further interpretation by RHW can be found in Chapter 4.3. No suspicion could be expressed for the remaining lesions as to why the AI assessed the respective lesion as high-risk.

Figure 10 shows four false high-risk-rated lesions with reddish parts. Figure 11 shows four examples of false high-risk rated lesions due to manipulation artifacts being displayed.

Figure 10: Overview false high-risk misjudged lesions due to reddish parts

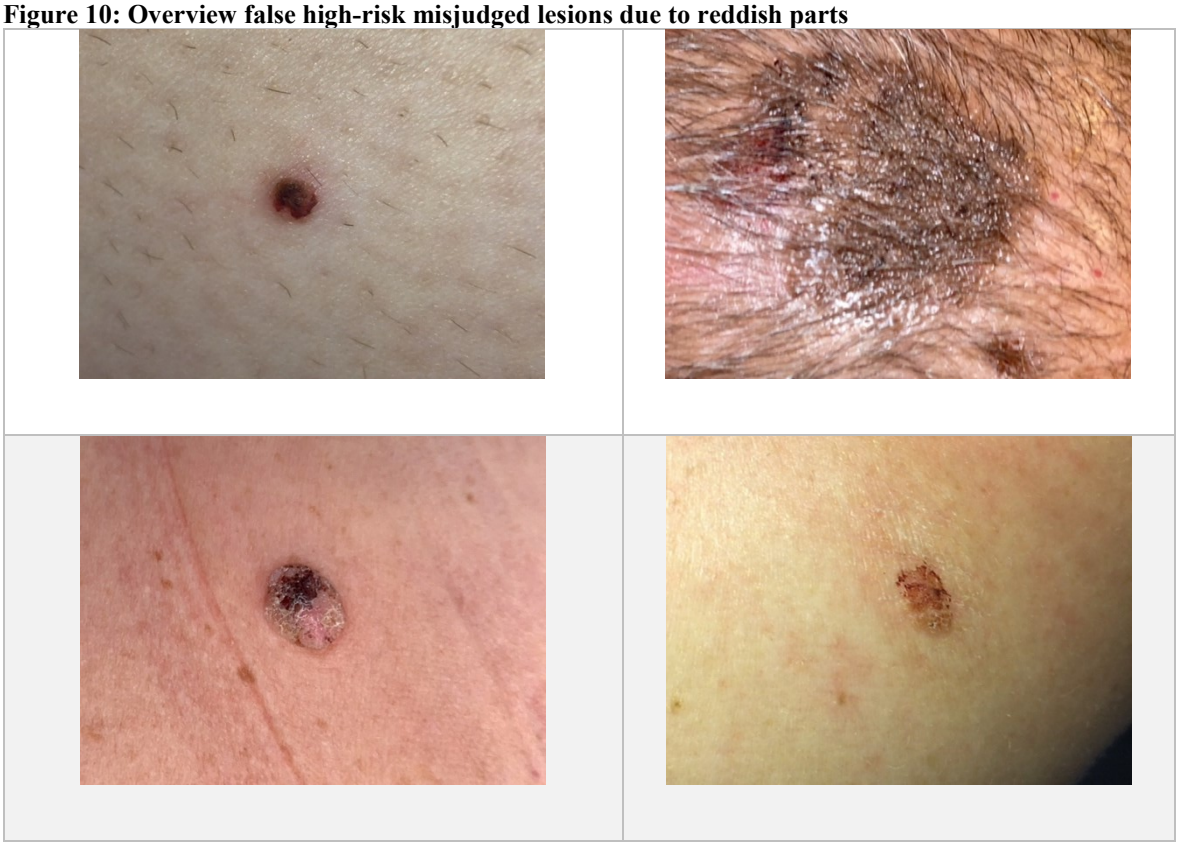


Figure 11: Overview of false high-risk misjudged lesions due to manipulation (e.g., scratching)

3.7.2 False low-risk lesions

71 lesions were classified as false low-risk by the AI compared to the dermatologists' consensus opinion.

After further analysis by the RHW, 22 of the 71 lesions that had been classified as false low-risk by the expert consensus (low-risk assessment by the algorithm) ultimately turned out to be correct low-risk lesions according to the algorithm assessment.

Of these 22 lesions, nine were diagnosed as black nevi/red nevi by RHW. One lesion was classified as a congenital nevus. The remaining 12 lesions were described as benign nevi with no specific description. The detailed and unblinded assessment by RHW as to why these lesions could be classified as being clearly benign is described in Chapter 4.2. Figure 12 shows four examples of these black nevi.

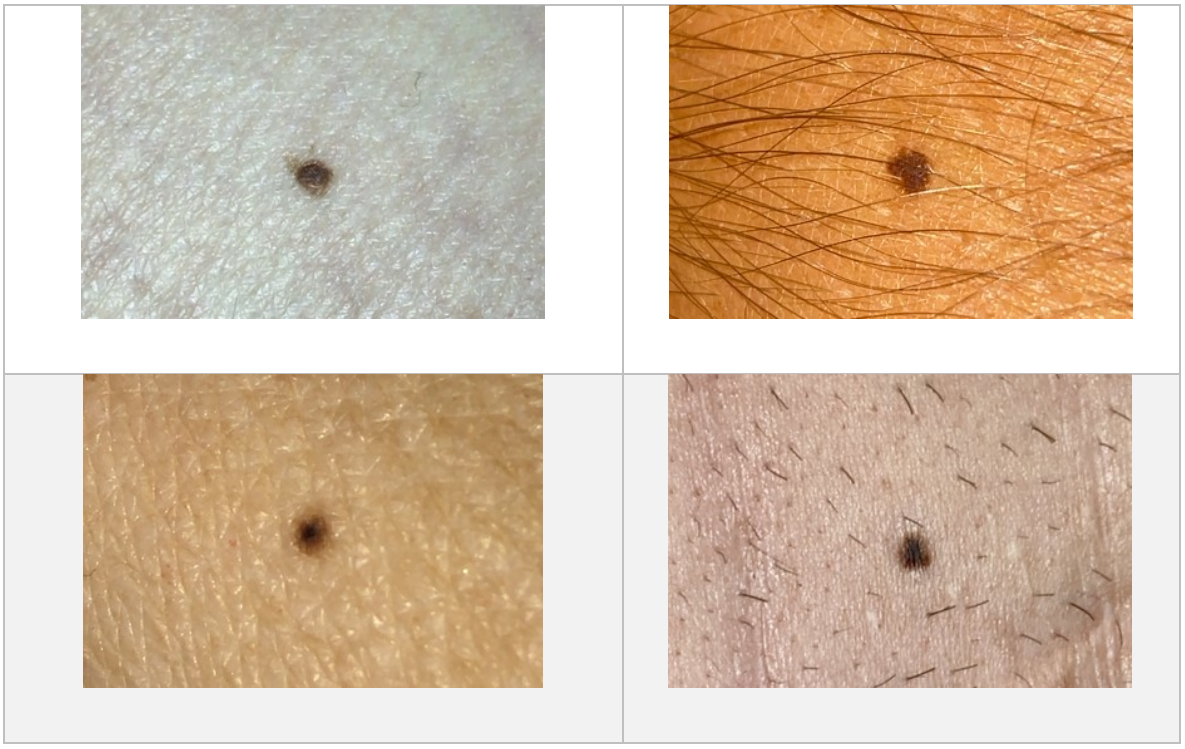


Figure 12: Overview of false low-risk judged lesions rated as black nevi

3.8 Comparison of the algorithm with the senior dermatologist´s opinion – Risk class

The individual evaluation of the senior dermatologist (RHW) was chosen as an alternative reference standard to the consensus opinion. This choice was made due to the senior dermatologist's decades of experience in dermato-oncology and dermoscopy.

For this purpose, the sensitivity and specificity were calculated with the Clopper-Pearson confidence intervals.

The following crosstabulation compares the number of assessments (n) for AI (nAI) and the opinion of the senior dermatologist as the single reference standard (nSRS) in two categories: “high and medium risk” and “low risk*”* respectively. These numbers were the basis for the following calculation of sensitivity and specificity.

		nSRS		Total
		low-risk	high- and medium-risk	
nAI	low-risk	893	78	971
	high- and medium-risk	264	181	445
Total		1157	259	1416

Table 12: Crosstabulation for sensitivity/specificity calculation - based on the single reference standard

Statistical analyses revealed a sensitivity of 69.9% (CI 63.9% - 75.4%) and a specificity of 77.2% (CI 74,7% - 79,6%) of the investigated algorithm when not compared to the consensus opinion but the single reference standard of RHW, indicating the weaker performance of the algorithm.

3.9 Comparison of the misjudged lesions by the AI before and after consensus opinion

In this Chapter, the lesions judged to be incorrect by the consensus opinion are additionally placed in relation to the individual assessments by the dermatologists before a consensus was established.

3.9.1 Comparison of false high-risk lesions by AI before and after consensus opinion

Table 13 shows the 28 false high-risk rated lesions by the AI in comparison to the evaluation by each dermatologist and the consensus opinion.

ID	Risk Class RHW	Risk Class TK	Risk Class KTD	Consensus opinion dermatologists	Risk Class Algorithm
249125	medium	low	**high**	low	high
249179	**high**	low	low	low	high
271789	low	low	**high**	low	high
279633	medium	**high**	low	low	high
281755	medium	low	**high**	low	high
283012	low	**high**	low	low	high
291743	low	**high**	medium	low	high
303507	low	low	**high**	low	high
306590	low	**high**	medium	low	high
309850	**high**	low	low	low	high
301468	low	**high**	medium	low	high
243422	low	low	low	low	high
247326	medium	low	low	low	high
247950	low	low	medium	low	high
252936	low	low	low	low	high
264703	low	low	low	low	high
271398	low	low	low	low	high
271789	low	low	**high**	low	high
273516	low	low	low	low	high
278258	low	low	low	low	high
282515	low	low	low	low	high
282575	low	low	low	low	high
282728	low	low	low	low	high
283012	low	**high**	low	low	high
286448	low	medium	low	low	high
288820	low	low	low	low	high
289850	low	low	medium	low	high
297367	low	low	low	low	high
302442	low	low	low	low	high
307644	medium	low	low	low	high
309850	**high**	low	low	low	high

Table 13: Overview of false high-risk rated lesions in comparison with consensus opinion and individual dermatologists' assessment

The majority of these cases had at least one of the three dermatologists assigning a risk rating higher than the consensus opinion assessment.

In 14 of these lesions, at least one of the three dermatologists had the same **high**-risk assessment as the AI.

In 11 of these lesions, a medium-risk assessment by one of the three dermatologists was given.

3.9.2 Comparison of false low-risk lesions by AI after and before consensus opinion

Below, Table 14 shows the 71 false low-risk rated lesions by the AI in comparison to the evaluation by each dermatologist and the consensus opinion.

ID	Risk Class RHW	Risk Class TK	Risk Class KTD	Consensus opinion dermatologists	Risk Class Algorithm
245038	medium	medium	**low**	medium	low
308410	medium	medium	**low**	medium	low
283853	medium	**low**	medium	medium	low
303816	medium	medium	**low**	medium	low
276531	medium	medium	**low**	medium	low
303576	high	high	high	high	low
272320	**low**	high	high	high	low
258449	high	**low**	high	high	low
309379	**low**	high	low	medium	low
273353	high	high	**low**	high	low
308443	**low**	medium	medium	medium	low
307467	medium	medium	**low**	medium	low
309049	medium	medium	high	medium	low
268386	**low**	medium	medium	medium	low
273498	medium	medium	**low**	medium	low
280228	**low**	medium	medium	medium	low
304477	**low**	medium	medium	medium	low
280179	**low**	medium	medium	medium	low
306844	medium	medium	**low**	medium	low
281317	high	high	high	high	low
308383	**low**	medium	medium	medium	low
281384	high	high	**low**	high	low
245312	medium	high	high	high	low
290235	medium	medium	**low**	medium	low
281195	medium	medium	**low**	medium	low
308455	**low**	medium	medium	medium	low
281581	high	high	high	high	low

289039	**low**	medium	medium	medium	low
244972	medium	medium	**low**	medium	low
251931	**low**	medium	medium	medium	low
267144	medium	medium	medium	medium	low
307252	medium	high	**low**	medium	low
306454	medium	medium	**low**	medium	low
268338	medium	medium	**low**	medium	low
247141	**low**	medium	medium	medium	low
309424	medium	medium	**low**	medium	low
305343	medium	**low**	high	medium	low
274262	**low**	medium	medium	medium	low
307473	medium	medium	**low**	medium	low
301468	**low**	high	medium	high	low
287016	high	high	**low**	high	low
280989	medium	high	high	high	low
304876	medium	medium	**low**	medium	low
281228	medium	medium	**low**	medium	low
302089	medium	high	high	high	low
268617	medium	medium	**low**	medium	low
281785	high	high	**low**	high	low
265284	medium	high	high	high	low
270197	medium	medium	**low**	medium	low
283334	**low**	medium	medium	medium	low
282927	**low**	medium	medium	medium	low
281198	medium	medium	**low**	medium	low
280543	**low**	medium	medium	medium	low
272631 271380	**low**	medium	medium	medium	low
271380	**low**	medium	medium	medium	low
270744	**low**	medium	medium	medium	low
268159	**low**	medium	medium	medium	low
267844	**low**	medium	medium	medium	low
266668	**low**	medium	medium	medium	low
265711	**low**	medium	medium	medium	low
264787	**low**	medium	medium	medium	low
263764	**low**	medium	medium	medium	low
263621	**low**	medium	medium	medium	low
261551	medium	high	medium	medium	low
260414	**low**	medium	medium	medium	low
254360	**low**	medium	medium	medium	low
244834	**low**	medium	medium	medium	low
242843	**low**	medium	medium	medium	low

Table 14: Overview of false low-risk rated lesions in comparison with consensus opinion and individual dermatologists' assessment

Of the total of 71 false negative lesions, only in 10 cases did two-thirds of the dermatologists (or all three) agree that the lesion was medium or high risk.

- In six cases, two-thirds of the dermatologists agreed that the lesion was of medium or high risk (see Figure 13)
- In three cases, all three dermatologists agreed that the lesion was of high risk (see Figure 14)
- In one case, all three dermatologists agreed that the lesion was of medium risk (see Figure 15)

Thus, these cases are especially important to discuss, as an incorrect evaluation of low-risk lesions through AI can be a considerable hazard if the lesions turn out to be high-risk, which is discussed in Chapter 4.4.

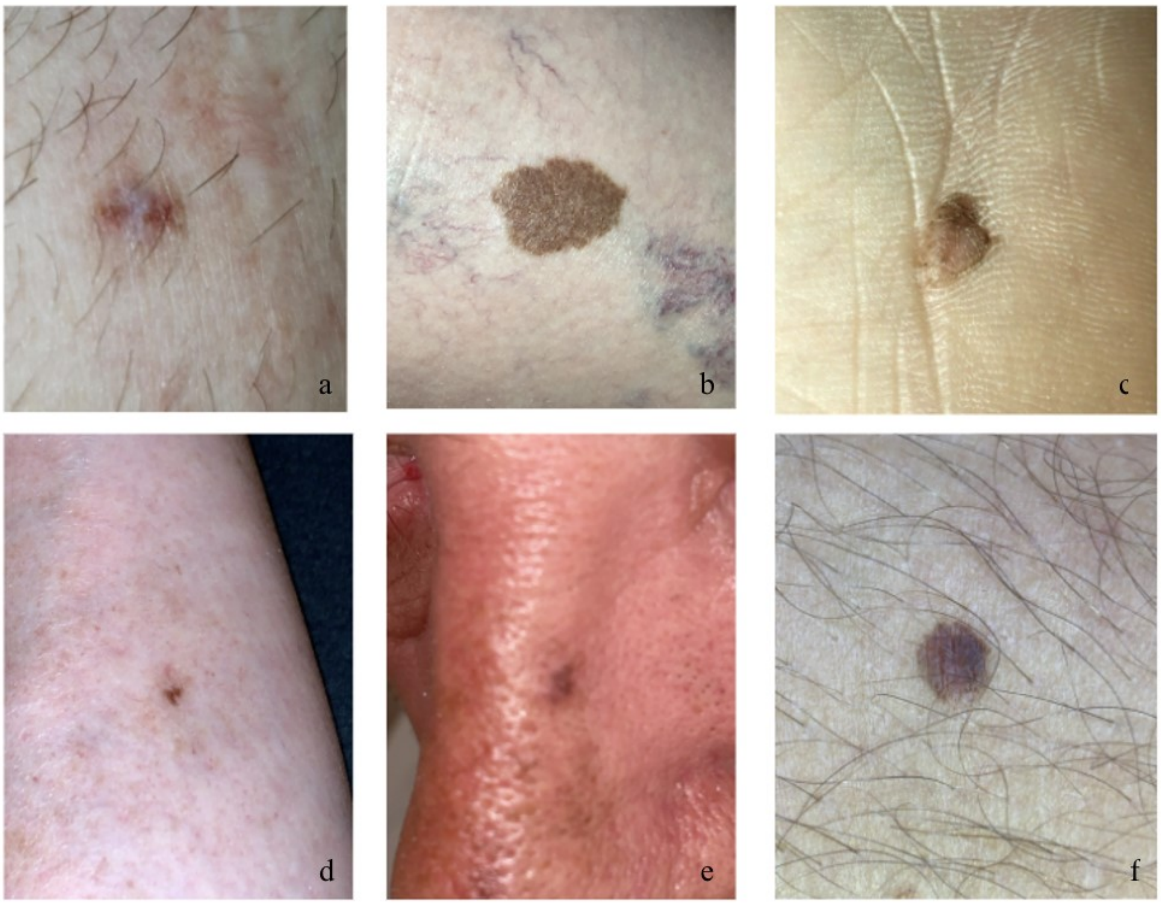


Figure 13: Overview of lesions that have a two-thirds majority for supposedly false medium- or high-risk lesions

The lesions shown in Figure 13 were assessed as high-risk by a two-thirds majority (2/3).

a) Two dermatologists suspected a melanoma and one a dysplastic nevus, while the AI suspected a dermal nevus.
b) Two dermatologists suspected a dysplastic nevus and one a melanoma, while the AI diagnosed a nevus.

c) Two dermatologists suspected a melanoma and one a dysplastic nevus, whereas AI suspected a dermal nevus.
d) Two dermatologists suspected a dysplastic nevus and one a melanoma, while the AI diagnosed a nevus.
e) One dermatologist suspected a BCC, the second a pigmented AK and the third a lentigo maligna, whereas AI suspected a red papule.
f) Two dermatologists suspected a melanoma and one a dysplastic nevus, while the AI diagnosed a nevus.

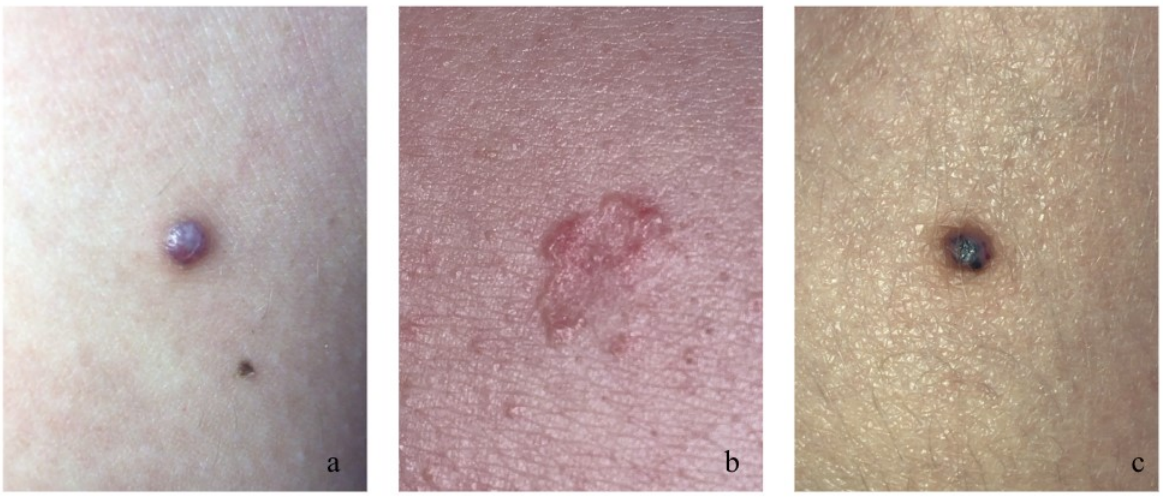


Figure 14: Overview of lesions that have a three thirds majority for supposedly false high-risk lesions

The lesions shown in Figure 14 were assessed as high-risk lesions by all dermatologists (3/3).

a) All dermatologists suspected a melanoma, while the AI suspected a hemangioma.
b) All dermatologists suspected a BCC, while the AI suspected a red macula.
c) Two dermatologists suspected a BCC and one a melanoma, while the AI diagnosed hemangioma.

Figure 15: Overview of lesions that have a three thirds majority for supposedly false medium-risk lesions

In Figure 15, all dermatologists suspected a dysplastic nevus, while the AI suspected a seborrheic keratosis.

In the remaining 61 cases, at least one of the three dermatologists assigned the lesion to the low-risk group.

4 Discussion

The overall aim of the study was to compare the detection accuracy of the AI-based app SkinScreener® to a dermatological consensus opinion by using images from smartphone application users in a real-world setting. The literature suggests numerous implications, calling for a study setting in such a non-clinical setting. Freeman et al. raised concerns in their systematic review that the studies for evaluating smartphone apps are representative of real-life use (25). The lack of real-world studies also makes it difficult to objectify the results of such smartphone apps, as noted by Du-Harpu et al. (35) and Tschandl et al., clearly calling for AI-based systems to be tested for performance under real conditions in the future (34). This clear demand in the literature was part of the rationale for conducting this study. The comparison with a consensus opinion of an Expert Panel as a reference standard, consisting of three clinically active dermatologists with different levels of experience, created different insights in the comparison of human and machine evaluation on the one hand. On the other hand, our findings also bring new insights for future performance studies as well as for teledermatology as a whole. These aspects will be highlighted in the discussion using the data we have generated.

4.1 Diagnostic accuracy and potential consequences of the smartphone app SkinScreener

In this performance study of the CE-certified app SkinScreener®, which is based on a neural network algorithm, we found a lower diagnostic accuracy compared to the preliminary clinical study with a sensitivity of 76.9% (CI 71.7-81.5) compared to 96.4% (CI 93.94-98.85) and specificity of 80.9% (CI 78.5-83.2) compared to 94.85% (CI 92.46-97.23) as found in the previous study (32). The question is to what extent this clinical reassessment of the algorithm's performance compares to the previous study.

A direct comparison of these two different performance surveys must be viewed critically, as they are based on different study designs. Due to the main goal of establishing a real-world study design, no histopathological data could be used here to validate the assessment that was at least in part available in the first study. The chosen reference standard as a consensus opinion or read study is, by means of literature, a well-established one. Nevertheless, we conclude that the reference standard we chose as a consensus opinion should be more carefully planned in future studies, which will be described in Chapter 4.2.

In addition, the study population is presumably different, as the patients included in the first study were exclusively a highly selected population (mainly patients at high risk of developing any type of skin cancer), according to the authors. Also, the authors noted that in the initial study, the average age group corresponded to 65 years. In addition, the population in the study by Kränke et al. also had significantly more high-risk cases (n=196) than our study population (n=83) (32).
Anamnestic parameters were not available in our study population for privacy reasons, but it can be assumed that the app was used by broader populations in the present study, both in age distribution and in a more unselected composition compared to the previous clinical study.

4.2 Strengths and limitations

The strengths of this study lie not only in being set in the real world but also in considering expressed criticisms of previous studies and recommendations for better planning and establishing such performance studies.

Certain requirements for conducting performance studies to evaluate smartphone apps have been taken into account. As demanded by Du-Harpu et al. and Freeman et al., many previous studies were unable to get an appropriate sample of people who would often use smartphone apps, the population for which tend to have low numbers of serious skin conditions and many different kinds of skin issues (25,35). In order to minimize these biases, we tried to employ the present study design, as the selection of users is not filtered apart from exclusion criteria. Therefore, it is reasonable to assume that a large heterogeneous patient population with respective digital smartphone skills is present.

Furthermore, it should be mentioned that laypersons also took the images used in our study by themselves with corresponding smartphone cameras which may vary in their image quality depending on the smartphone used. Nonetheless, it is crucial for conducting a study to use such images, which may appear to be of suboptimal quality. This is especially true in the required real-world setting (25).

In addition, previous studies have only rated lesions or images selected and taken by clinicians rather than those taken by users using the smartphone apps. Although individual user behavior can only be estimated, it can be assumed that the selection of skin lesions was at least not dependent on the influence of clinicians. However, we must also acknowledge

here that we also considered the exclusion criteria for the use of the app, which also introduced a bias, making the selection process not entirely uninfluenced. These exclusion criteria (see Chapter 2.2) are specified by the manufacturer of the SkinScreener device based on the requirements of a lesion to be analyzed.

Preliminary studies have shown that neural networks are usually trained and tested on the same dataset (33). With the used SkinScreener®, the distinction between training and test datasets was made during algorithm development.

Furthermore, a major strength of the application used and the analysis of this study is the evaluation of pigmented and non-pigmented skin cancers and precancerous lesions. In the literature, largely, only pigmented skin cancers are investigated in relation to AI and smartphone applications. This is because the SkinScreener® app is approved for both skin cancer entities; thus, we have included both in our analysis without influencing the selection of lesions.

One of the limitations of our study is the consensus opinion. The consensus opinion that was considered the reference standard was the only way to establish a standard in the chosen study setting. The definitive histopathological assessment was missing, as app users do not share data on the further progression of their scanned skin lesions with the developers. Thus, questions arise as to whether matching with Expert Panels alone is sufficient for assessment and comparison.

The literature points out that the selection of physicians who participate in a reader study often includes those physicians who have specific expertise in the respective medical specialty; thus, the corresponding test accuracy of the physicians in their respective specialty is very high. However, this leads to the fact that this recognition accuracy of the specialists cannot be transferred to other doctors in the same field who, for example, have specialized more broadly. As an example from radiology, including radiologists with more than 10 years of professional experience could lead to an underestimation of the impact of the new technology; in contrast, a reader study based only on residents could lead to an overestimation of the evaluation of the new technology (42).

The considerable differences in the individual ratings of the images to the three-level risk categories, which can also be illustrated by the comparatively low Cohen's Kappa values,

raise the question of the extent to which the chosen composition of the Reader Study could have been improved and what could be the cause of the different perception and rating of the lesions.

On the one hand, the different clinical experiences and the partly different levels of training of dermatologists should at least have created a certain generalizability of the consensus opinion. However, this has also led to different interpretations of risk assessment (low, medium, and high) and diagnosis.

It should be noted that the digital images in our study were not accompanied by medical history. As only macroscopic images were included, additional dermoscopic images were not available, which could have led to a higher diagnostic certainty and a better agreement between dermatologists. In addition, we only investigated the difference in assessment in terms of risk class and not at the level of diagnosis. We conclude that the composition of the Expert Panel can be improved in future studies, possibly by incorporating more dermatologists in the cases that caused disagreement.

It can be speculated that our results could also have significance for teledermatology in general. The partly high differences in the simple three-stage risk assessment could show that pure image data for the assessment of skin lesions must be regarded more than critically; thus, the importance of the synopsis of visual information, paired with an appropriate anamnesis and further diagnostic instruments, is once more clarified and should be considered for teledermatology.

We must point out that despite the inclusion of pigmented and non-pigmented skin cancers and precancerous lesions, certain skin cancers such as amelanotic melanoma, Merkel cell carcinoma, or cutaneous sarcoma cannot be detected by the AI used, which naturally also raises the question of the use of such systems if certain entities are not taken into account.

4.3 Findings of the dermatological re-evaluation

As noted in Chapter 3.7.1, some (10 out of 28) of the false positive lesions were incorrectly assessed as false positives relative to the consensus opinion, according to RHW. The other part (18 out of 28) included lesions that were classified as correct false positives due to other circumstances, which are discussed below. At least seven lesions were so influenced by irritation from the app users, such as scratching or similar, that the AI recognized them as

high-risk. These scratching artifacts can lead to AI misjudging a supposedly malignant-looking lesion. However, it should be noted that this type of manipulation is considered an exclusion criterion or contraindication for the use of the app. When analyzing other false positive lesions, it was noticed that the proportion of red tones in certain lesions had led to a strikingly high rating as high-risk lesions. In most cases, these were misjudged as BCC, although they were benign in character. Insufficient training of the AI with such lesions could be cited as a possible cause, requiring further specific training of the used algorithm to differentiate such lesions.

We also analyzed the supposedly false-negative lesions. As described in Chapter 3.6, 22 of the 71 lesions can be classified as clearly benign, contrary to the consensus opinion of a moderate or high-risk lesion. 9 of these 22 lesions re-evaluated during the re-evaluation were consistent with a diagnosis of black naevus/red naevus after further visual analysis. RHW pointed out that the high pigmentation of the lesions had probably caused confusion among the other dermatologists, which may have led to an incorrect risk assessment.

4.4 Chances and Risks of AI in Dermatology

In general, the question arises as to the justifiability of the use of such systems for widespread use by laypersons. In our study, we were able to find at least three lesions that AI incorrectly assessed and where the question arises as to how these lesions have or have not cleared further. To what extent can such systems bring an epidemiological benefit or, on the other hand, also represent a danger as well as an additional burden for our health system where it is known that certain cases of skin cancer are misidentified? In order to answer these questions, large-scale studies will be necessary in the future to create evidence here.

Of course, the use of these products also raises questions of liability law as to who is ultimately responsible and who is not.

In principle, there is an obligation to take out liability insurance for labeled medical devices, which provides the user with financial protection in the event of causality leading to a misjudgment of the algorithm. Nevertheless, the apps on the market with CE certification point out that their use does not make dermatological examinations redundant since the manufacturers explicitly communicate this in their instructions.

Whether the common user also handles this in practice and does not use the app for a mole check (for example) without following the manufacturer's instructions must be considered. Is it then the manufacturer who is liable if the user does not have the supposedly malignant lesion diagnosed and treated in time by a dermatologist due to a wrongly negative assessment of the lesion by an AI? Is it the user's responsibility to consult a dermatologist independently of the recommended skin cancer check? What is the role of the dermatologist and their liability if they biopsy the lesion assessed as suspicious by the AI at the patient's request? What is the physician's comparative liability (whether a GP or dermatologist) should they be mistaken in their assessment, although an AI implies otherwise?

In any case, the manufacturers must be held responsible because they can possibly relieve and massively strain the healthcare system and its resilience through the widespread use of their technologies. However, researchers can take a common path with and support manufacturers from the academic side to be able to improve their technologies, as digitalization does not stop with medicine, and the emergence of these technologies will be deployed.

4.5 Synopsis

The use of AI solutions for triaging skin lesions by lay users has the potential to be useful. Yet it should be noted that the use of macroscopic AI solutions is presumably not on par with conventional diagnostics at this stage. However, dermoscopic AI solutions have already surpassed human dermoscopic evaluation. It's worth noting that opportunity also lies in the intensive training of the algorithms with high-quality images so that laypersons can use this technology as an evidence-based and widely effective triaging tool.

However, the implementation and use of these solutions may make a significant contribution as an augmentation tool and decision support for dermatologists and GPs. The comparison between man and machine, which is often exaggerated in medicine and called for in the AI research and development fields, is of particular interest in dermatology since the visual assessment by neural networks may prove successful. The increasing depth of abstraction by neuronal networks largely exceeds what is perceptible to the human eye. However, the ultimate comparison between the assessment of human intelligence and artificial intelligence is only meaningful and possible if a direct comparison is made in terms of how humans systematically analyze skin changes. The first voices are being raised in the literature that call for an equal comparison in which the AI should also be fed with the corresponding

anamnesis in addition to the image information and must take this into account in the diagnosis in order to have a realistic comparison to a genuine human assessment (35).

We will continue to see where this technology could improve our medical decision-making process in the future as a technical second opinion or issued differential diagnosis, but where human intelligence and freedom of decision will continue to have their justification.

Bibliography

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. CA Cancer J Clin. 2021 Jan;71(1):7–33.
2. Zink A. Trends in the treatment and prevention of keratinocyte carcinoma (non-melanoma skin cancer). Curr Opin Pharmacol. 2019 Jun;46:19–23.
3. Nikolaou V, Stratigos AJ. Emerging trends in the epidemiology of melanoma. Br J Dermatol. 2014 Jan;170(1):11–9.
4. Keung EZ, Gershenwald JE. The eighth edition American Joint Committee on Cancer (AJCC) melanoma staging system: implications for melanoma treatment and care. Expert Rev Anticancer Ther. 2018 Aug 3;18(8):775–84.
5. Rudolph C, Schnoor M, Eisemann N, Katalinic A. Incidence trends of nonmelanoma skin cancer in Germany from 1998 to 2010: Incidence trends of nonmelanoma skin cancer. JDDG J Dtsch Dermatol Ges. 2015 Aug;13(8):788–97.
6. Leiter U, Keim U, Eigentler T, Katalinic A, Holleczek B, Martus P, et al. Incidence, Mortality, and Trends of Nonmelanoma Skin Cancer in Germany. J Invest Dermatol. 2017 Sep 1;137(9):1860–7.
7. Donaldson MR, Coldiron BM. No End in Sight: The Skin Cancer Epidemic Continues. Semin Cutan Med Surg. 2011 Mar;30(1):3–5.
8. Nehal KS, Bichakjian CK. Update on Keratinocyte Carcinomas. Longo DL, editor. N Engl J Med. 2018 Jul 26;379(4):363–74.
9. Liu F, Bessonova L, Taylor TH, Ziogas A, Meyskens FL, Anton-Culver H. A unique gender difference in early onset melanoma implies that in addition to ultraviolet light exposure other causative factors are important. Pigment Cell Melanoma Res. 2013 Jan;26(1):128–35.
10. Chahoud J, Semaan A, Chen Y, Cao M, Rieber AG, Rady P, et al. Association Between β-Genus Human Papillomavirus and Cutaneous Squamous Cell Carcinoma in Immunocompetent Individuals—A Meta-analysis. JAMA Dermatol. 2016 Dec 1;152(12):1354.
11. Neagu N, Dianzani C, Venuti A, Bonin S, Voidăzan S, Zalaudek I, et al. The role of HPV in keratinocyte skin cancer development: A systematic review. J Eur Acad Dermatol Venereol. 2023 Jan;37(1):40–6.
12. Narayanan DL, Saladi RN, Fox JL. Review: Ultraviolet radiation and skin cancer: UVR and skin cancer. Int J Dermatol. 2010 Aug 30;49(9):978–86.
13. Lopez AT, Carvajal RD, Geskin L. Secondary Prevention Strategies for Nonmelanoma Skin Cancer. Oncol Williston Park N. 2018 Apr 15;32(4):195–200.
14. Centers for Disease Control and Prevention (CDC). Sunburn and sun protective behaviors among adults aged 18-29 years--United States, 2000-2010. MMWR Morb Mortal Wkly Rep. 2012 May 11;61(18):317–22.
15. Petrie T, Samatham R, Witkowski AM, Esteva A, Leachman SA. Melanoma Early Detection: Big Data, Bigger Picture. J Invest Dermatol. 2019 Jan;139(1):25–30.
16. Dorrell DN, Strowd LC. Skin Cancer Detection Technology. Dermatol Clin. 2019 Oct;37(4):527–36.
17. Loescher LJ, Janda M, Soyer HP, Shea K, Curiel-Lewandrowski C. Advances in Skin Cancer Early Detection and Diagnosis. Semin Oncol Nurs. 2013 Aug;29(3):170–81.
18. Carli P, De Giorgi V, Palli D, Maurichi A, Mulas P, Orlandi C, et al. Dermatologist Detection and Skin Self-examination Are Associated With Thinner Melanomas: Results From a Survey of the Italian Multidisciplinary Group on Melanoma. Arch Dermatol [Internet]. 2003 May 1 [cited 2023 Mar 17];139(5). Available from: http://archderm.jamanetwork.com/article.aspx?doi=10.1001/archderm.139.5.607
19. Lakhani NA, Saraiya M, Thompson TD, King SC, Guy GP. Total body skin examination for skin cancer screening among U.S. adults from 2000 to 2010. Prev Med.

2014 Apr;61:75–80.
20. Reyes-Marcelino G, Tabbakh T, Espinoza D, Sinclair C, Kang YJ, McLoughlin K, et al. Prevalence of skin examination behaviours among Australians over time. Cancer Epidemiol. 2021 Feb;70:101874.
21. LeBlanc WG, Vidal L, Kirsner RS, Lee DJ, Caban-Martinez AJ, McCollister KE, et al. Reported skin cancer screening of US adult workers. J Am Acad Dermatol. 2008 Jul;59(1):55–63.
22. Katalinic A, Waldmann A, Augustin M, Breitbart E, Eisemann N. Evidenz für ein Hautkrebsscreening. Onkol. 2014 Jun;20(6):535–42.
23. Samaran R, L'Orphelin JM, Dreno B, Rat C, Dompmartin A. Interest in artificial intelligence for the diagnosis of non-melanoma skin cancer: a survey among French general practitioners. Eur J Dermatol. 2021 Aug;31(4):457–62.
24. Liu Y, Jain A, Eng C, Way DH, Lee K, Bui P, et al. A deep learning system for differential diagnosis of skin diseases. Nat Med. 2020 Jun;26(6):900–8.
25. Freeman K, Dinnes J, Chuchu N, Takwoingi Y, Bayliss SE, Matin RN, et al. Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. BMJ. 2020 Feb 10;368:m127.
26. 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 Feb 2;542(7639):115–8.
27. Chao E, Meenan CK, Ferris LK. Smartphone-Based Applications for Skin Monitoring and Melanoma Detection. Dermatol Clin. 2017 Oct;35(4):551–7.
28. Wurm EM, Curchin CE, Soyer HP. Recent advances in diagnosing cutaneous melanomas. F1000 Med Rep [Internet]. 2010 Jun 23 [cited 2021 Mar 21];2. Available from: https://facultyopinions.com/prime/reports/m/2/46/
29. 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 Mar 1;145:81–91.
30. 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 Jun 2;7:233.
31. Jones OT, Matin RN, van der Schaar M, Prathivadi Bhayankaram K, Ranmuthu CKI, Islam MS, et al. Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review. Lancet Digit Health. 2022 Jun;4(6):e466–76.
32. Kränke T, Tripolt-Droschl K, Röd L, Hofmann-Wellenhof R, Koppitz M, Tripolt M. New AI-algorithms on smartphones to detect skin cancer in a clinical setting—A validation study. Hammad M, editor. PLOS ONE. 2023 Feb 15;18(2):e0280670.
33. Jahn AS, Navarini AA, Cerminara SE, Kostner L, Huber SM, Kunz M, et al. Over-Detection of Melanoma-Suspect Lesions by a CE-Certified Smartphone App: Performance in Comparison to Dermatologists, 2D and 3D Convolutional Neural Networks in a Prospective Data Set of 1204 Pigmented Skin Lesions Involving Patients' Perception. Cancers. 2022 Aug 7;14(15):3829.
34. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human–computer collaboration for skin cancer recognition. Nat Med. 2020 Aug;26(8):1229–34.
35. Du-Harpur X, Watt FM, Luscombe NM, Lynch MD. What is AI? Applications of artificial intelligence to dermatology. Br J Dermatol. 2020 Sep;183(3):423–30.
36. Ke Q, Liu J, Bennamoun M, An S, Sohel F, Boussaid F. Computer Vision for Human–Machine Interaction. In: Computer Vision for Assistive Healthcare [Internet]. Elsevier; 2018 [cited 2021 Apr 7]. p. 127–45. Available from:

https://linkinghub.elsevier.com/retrieve/pii/B9780128134450000058
37. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. ArXiv150601497 Cs [Internet]. 2016 Jan 6 [cited 2021 Apr 6]; Available from: http://arxiv.org/abs/1506.01497
38. U.S. Department of Health and Human Services, Food and Drug Administration. Non-Inferiority Clinical Trials to Establish Effectiveness Guidance for Industry [Internet]. 2016 [cited 2023 Mar 18]. Available from: https://www.fda.gov/media/78504/download
39. Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas. 1960 Apr;20(1):37–46.
40. McHugh ML. Interrater reliability: the kappa statistic. Biochem Medica. 2012;22(3):276–82.
41. Alberg AJ, Park JW, Hager BW, Brock MV, Diener-West M. The use of "overall accuracy" to evaluate the validity of screening or diagnostic tests. J Gen Intern Med. 2004 May;19(5):460–5.
42. Gennaro G. The "perfect" reader study. Eur J Radiol. 2018 Jun;103:139–46.