Journal Pre-proof

Diagnostic Performance of Augmented Intelligence with 2D and 3D Total Body Photography and Convolutional Neural Networks in a High-risk Population for Melanoma under Real-world Conditions: A New Era of Skin Cancer Screening?

Sara E. Cerminara, Phil Cheng, Lisa Kostner, Stephanie Huber, Michael Kunz, Julia-Tatjana Maul, Jette S. Böhm, Chiara F. Dettwiler, Anna Geser, Cécile Jakopović, Livia M. Stoffel, Jelissa K. Peter, Mitchell Levesque, Alexander A. Navarini, Lara V. Maul

PII: S0959-8049(23)00306-4

DOI: https://doi.org/10.1016/j.ejca.2023.112954

Reference: EJC112954

To appear in: *European Journal of Cancer*

Received date: 20 February 2023
Revised date: 13 June 2023

Accepted date: 17 June 2023

Please cite this article as: Sara E. Cerminara, Phil Cheng, Lisa Kostner, Stephanie Huber, Michael Kunz, Julia-Tatjana Maul, Jette S. Böhm, Chiara F. Dettwiler, Anna Geser, Cécile Jakopović, Livia M. Stoffel, Jelissa K. Peter, Mitchell Levesque, Alexander A. Navarini and Lara V. Maul, Diagnostic Performance of Augmented Intelligence with 2D and 3D Total Body Photography and Convolutional Neural Networks in a High-risk Population for Melanoma under Real-world Conditions: A New Era of Skin Cancer Screening?, *European Journal of Cancer,* (2023) doi:https://doi.org/10.1016/j.ejca.2023.112954

Diagnostic Performance of Augmented Intelligence with 2D and 3D Total Body Photography and Convolutional Neural Networks in a High-risk Population for Melanoma under Real-world Conditions: A New Era of Skin Cancer Screening?

Sara E. Cerminara[1], Phil Cheng[2], Lisa Kostner[1], Stephanie Huber[1], Michael Kunz[1], Julia-Tatjana Maul[2,3], Jette S. Böhm[1], Chiara F. Dettwiler[1], Anna Geser[1], Cécile Jakopović[1], Livia M. Stoffel[1], Jelissa K. Peter[1], Mitchell Levesque[2], Alexander A. Navarini[1], Lara V. Maul[1]

Author affiliations:

[1]Department of Dermatology, University Hospital of Basel, Burgfelderstrasse 101, 4055 Basel, Switzerland; sara.cerminara@usb.ch, lisa.kostner@usb.ch, stephanie.huber@usb.ch, Michael.kunz@usb.ch, jette.boehm@stud.unibas.ch, chiaraflurina.dettwiler@stud.unibas.ch, anna.geser@stud.unibas.ch, c.jakopovic@stud.unibas.ch, livia.stoffel@stud.unibas.ch, jelissa.peter@stud.unibas.ch, alexander.navarini@usb.ch, laravaleska.maul@usb.ch

[2]Department of Dermatology, University Hospital of Zurich, Rämistrasse 100, 8091 Zurich, Switzerland; phil.cheng@usz.ch, julia-tatjana.maul@usz.ch, mitchell.levesque@usz.ch

[3]Faculty of Medicine, University of Zurich, Zurich, Switzerland

Corresponding author:
Lara Valeska Maul, MD
Department of Dermatology
University Hospital Basel
Burgfelderstr. 101
4055 Basel, Switzerland
E-Mail: laravaleska.maul@usb.ch

Number of Tables: 3 (exclusive Appendices)

Number of Figures: 6 (exclusive Appendices)

Keywords: melanoma, artificial intelligence, pigmented nevus, photography augmented intelligence, 3D, 2D, total body photography, convolutional neural network, skin neoplasm, deep learning

Abstract

Background

Convolutional neural networks (CNNs) have outperformed dermatologists in classifying pigmented skin lesions under artificial conditions. We investigated, for the first time, the performance of 3D and 2D CNNs and dermatologists in the early detection of melanoma in a real-world setting.

Methods

In this prospective study, 1690 melanocytic lesions in 143 patients with high-risk criteria for melanoma were evaluated by dermatologists, 2D-FotoFinder-ATBM® and 3D-Vectra® WB360 total body photography (TBP). Excision was based on the dermatologists' dichotomous decision, an elevated CNN risk score (study-specific malignancy cut-off: FotoFinder >0.5, Vectra >5.0) and/or the second dermatologist's assessment with CNN support. The diagnostic accuracy of the 2D/3D CNN classification was compared with that of the dermatologists and the augmented intelligence based on histopathology and dermatologists' assessment. Secondary endpoints

included reproducibility of risk scores and nevus counts per patient by medical staff (gold standard) compared to automated 3D/2D TBP CNN counts.

Results

The sensitivity, specificity, and receiver operating characteristics area under the curve (ROC-AUC) for risk-score-assessments compared to histopathology of 3D-CNN with 95% confidence intervals (CI) were 90.0%, 64.6% and 0.92 (CI 0.85-1.00), respectively. While dermatologists and augmented intelligence achieved the same sensitivity (90%) and comparable classification ROC-AUC (0.91(CI 0.80-1.00), 0.88(CI 0.77-1.00)) with 3D-CNN, their specificity was superior (92.3% and 86.2%, respectively). The 2D-CNN (sensitivity:70%, specificity:40%, ROC-AUC:0.68(CI 0.46-0.90) was outperformed. The 3D-CNN showed a higher correlation coefficient for repeated measurements of 246 lesions (R=0.89) than the 2D-CNN (R=0.79). The mean nevus count per patient varied significantly (gold standard: 210 lesions; 3D-CNN: 469; 2D-CNN: 1324; $p<0.0001$).

Conclusions

Our study emphasises the importance of validating the classification of CNNs in real life. The novel 3D-CNN device outperformed the 2D-CNN and achieved comparable sensitivity with dermatologists. The low specificity of CNNs and the lack of automated counting of TBP nevi currently limit the use of augmented intelligence in clinical practice.

Clinical Trial Registration Number: NCT04605822

Introduction

The incidence of invasive melanoma in the Caucasian population has risen steadily worldwide over the past few decades [1,2]. Early detection of cutaneous melanoma is the key to promising overall survival, as prognosis is strongly correlated with tumour thickness at diagnosis [3]. Skin cancer screening in patients with dysplastic nevus syndrome or multiple nevi remains a challenge for dermatologists. In addition to the time-consuming process of tracking down all melanocytic lesions, the detection of new lesions is almost impossible without machine assistance. However, accurate documentation of patients with multiple nevi is invaluable as approximately 70% of melanomas grow de novo and 30% develop from pre-existing nevi [4]. Therefore, the European Melanoma Guideline recommends the use of total body photography (TBP) in addition to sequential digital dermoscopy (SSDI) to improve melanoma detection, especially in high-risk populations [5].

Risk prediction imaging technologies are currently changing the landscape of dermatology and appear to have great potential to improve skin cancer screening under artificial conditions [6]. Automated artificial intelligence (AI)-based three-dimensional (3D) and two-dimensional (2D) TBP devices use deep learning convolutional neural networks (CNNs) for malignancy risk assessment. The advantage of 2D and 3D CNNs is the early detection of de novo lesions and the minimisation of unnecessary biopsies [7,8]. In a meta-analysis of 46 trials, the number needed to biopsy (NNB) for melanoma decreased with TBP to an overall average of 8.6, while the accuracy of clinicians in diagnosing melanoma had a NNB of 14.8 [8,9].

For at least two decades, 2D imaging has shown certain benefits in high-risk patients: detection of melanoma in combination with SSDI at earlier stages and a lower excision rate [10]. On the market since 2017, 3D TBP offers additional benefits such as time efficiency as all images are captured at once within a few seconds, 3D avatar formation, and reduced fear of melanoma among patients and fear of missing melanoma among physicians [11,12]. Previous studies comparing deep neural networks with dermatologists have shown that, on average, the CNN outperformed dermatologists under artificial conditions with missing clinical information [13-15].
Both devices, 2D and 3D TBP implemented the ability to count nevi. A previous study from Australia investigating the accuracy of nevi counts by 3D TBP revealed a sensitivity of 79% for pigmented lesions ≥ 2 mm (resp. 84% with lesions ≥5 mm) and a specificity of 91% (resp. 91% for lesions ≥5 mm) compared to senior dermatologists (n = 3) [16].

The aim of this study was to investigate the clinical performance of novel 3D TBP, 2D TBP and dermatologist alone and augmented intelligence (collaboration between AI and physician in decision-making process) in the early detection of melanoma in individuals at high risk of melanoma in a real-world setting. We analysed the potential bias in the AI scores of the 3D and 2D devices when measured twice in a subgroup to provide robust

evidence for the management and interpretation of AI scores in clinical practice. We also investigated the accuracy of nevi counts using both 3D and 2D TBP devices to determine if they have potential for everyday melanoma screening.

2. Material and methods

2.1. Study design and participating population

In this prospective, single-centre, observational study performed at the Department of Dermatology of the University Hospital Basel, we recruited participants at high-risk for developing melanoma from January to August 2021 aged ≥18 years. Inclusion criteria were previous cutaneous invasive and/or in situ melanoma, melanoma-suspicious lesions, family history of melanoma, ≥100 nevi, ≥5 atypical nevi, known CDKN2A mutation and/or diagnosis of dysplastic nevus syndrome. Patients with a Fitzpatrick skin type V - VI, papillomatous nevi, pigmented palmoplantar lesions, an acute mental illness or lack of informed consent were excluded. Patients with multiple concurrent risk factors were not excluded, especially melanoma patients with additional risk factors, which may potentially amplify the likelihood of developing melanoma. The dermatologists' experience (beginners: <2 years; intermediates: 2–5 years; experts: ≥5 years) and the demographic characteristics of the patients were included in the analysis. The distribution of patients for the physicians' assessments was conducted randomly and independently to the dermatologists' experience level.

2.2. Data collection

Detailed information about our study setup was published in a first interim analysis by Jahn et al. [17]. All participants first underwent routine skin cancer screening including dermoscopy conducted by a dermatologist, who provided a dichotomous diagnosis (malignant vs benign) for each pigmented skin lesion. Patients then underwent TBP using the 3D imaging system Vectra WB360 (Canfield Scientific Inc, Parsippany, New Jersey, USA, version 4.7.1) and the 2D FotoFinder© automatic total body mapping (ATBM Master) device (FotoFinder System GmbH, Bad Birnbach, Germany, version 3.3.1.0; Appendices Table A.1). Vectra captures the TBP with 92 cameras simultaneously within a few seconds, creating a 3D avatar of the patient's skin surface. FotoFinder ATBM takes 20 images of the body in 8 assigned positions with one camera in about 10-15 minutes, resulting in a fragmented view. Skin folds and mucosal lesions are not imaged by either device. After TBP of all images, all melanocytic lesions ≥3 mm in diameter and any smaller suspicious lesions, identified by dermatologists, were captured with the corresponding digital dermoscopy cameras (VISIOMED® D200evo dermoscope, Medicam 1000®). All manually photographed lesions were identified by dermatologists together with trained medical staff. The dermoscopy camera Medicam® is directly attached to the FotoFinder device, whereas the camera of VECTRA, VISIOMED® D200evo, is connected to a separate computer, which is linked with the same software program. Each defined lesion was manually photographed using the two dermoscopy devices, VISIOMED® D200evo dermoscope and Medicam 1000®, in addition to being assessed by the dermatologists. Their deep learning based malignancy risk assessments were documented with the Dermoscopy EXplainable Intelligence (DEXI) score, trained on >66,000 photographs [18]. It categorises each lesion from 0.0-10.0 and FotoFinder's Moleanalyzer Pro from 0.0-1.0. The higher the score, the higher the risk of malignancy according to their CNN. Finally, the dermatologist reassessed all lesions with the knowledge of the AI risk assessment. In our study, we defined augmented intelligence as combination of decision-making by dermatologists plus 2D-CNN and 3D-CNN. Excision was based on the dermatologist's initial assessment, an elevated AI score considering our predefined study-specific malignancy cut-off (FotoFinder >0.5, Vectra WB360 >5.0) and/or the dermatologist's assessment knowing the AI score. All scores above the cut-off had to be increased twice to meet the requirement for biopsy. Patients received a second imaging for AI-based risk-assessment evaluation of a lesion in case of an elevated predefined cut-off-score. If the risk assessment score of a lesion was repeatedly increased twice in immediate sequence, the lesion met the criteria for an excision. Histopathology and benign classification by the combination of dermatologists, 3D/2D-TBP-CNN, and dermatologists' knowledge of AI were used as gold standards to analyse the accuracy of the scoring systems.

The total number of nevi in each patient was counted by trained medical staff including physicians, study nurses and medical students using a 4-eye principle (gold standard). Non-photographed areas of the body were excluded from the lesion count. The count was compared with the automatically generated nevi count by the 2D and 3D TBP devices.

In a subgroup analysis of 21 randomly assigned patients, all identified pigmented skin lesions were imaged twice by the CNN dermoscopy devices (VISIOMED® D200evo dermoscope, Medicam 1000®) to investigate and compare any deviation of the two AI-scoring systems in their risk assessment scores. The patients of this subgroup and their associated lesions were part of the original analysis. This was an exploratory analysis with no prior assumptions about the accuracy and precision of either device.

2.3. Statistical analysis

Comparisons of continuous variables were tested using the Wilcoxon rank test and categorical variables were tested using Fishers Exact test. Comparisons of nevus counts were tested with Students t-test. Correlations were tested using Pearson's correlation coefficient (r). A p-value less than 0.05 was considered significant. Sensitivity was calculated by TP / (TP + FN). Specificity was calculated by TN / (TN + FP). (TP is true positive, FP is false positive, TN is true negative, and FN is false negative). Receiver Operating Characteristics (ROC) analysis was used to assess the performance of FotoFinder and Vectra WB360 labels (benign and suspicious) and the results against numeric scores to histology. All analyses were performed using R (version 4.1) and visualised using ggplot2.

2.4. Ethics

The study was approved by the local Ethics Committee (2020-02482) and registered with ClinicalTrials.gov (NCT04605822).

3. Results

3.1. Patient characteristics

We included 143 high-risk patients (48.3% female, median age 56 years [22-85)]; Table 1). In total, 1,690 melanocytic skin lesions were assessed with a mean of 12 and median of 8 per patient. These lesions were all photographed manually by VISIOMED® D200evo dermoscope, Medicam 1000® and were subsequently evaluated by their CNNs, as well as by the physicians. A total of 75 pigmented skin lesions were excised.

Table 1. Characteristics of the study population, nevi counts and number of excisions

Characteristics	Overall, n = 143[1]	Patients at high-risk for melanoma, n = 69[1]	Patients with melanoma, n = 74[1]	p-value[2]
Age at visit				0.019
Mean (SD)	56 (14)	53 (15)	58 (12)	
Median (range)	56 (22 - 85)	53 (22 - 85)	58 (29 - 81)	
Sex (%)				0.4
Female	69 (48.3%)	31 (44.9%)	38 (51.4%)	
Male	74 (51.7%)	38 (55.1%)	36 (48.6%)	
Fitzpatrick skin type (%)				0.4
I	11 (7.7%)	4 (5.6%)	7 (9.7%)	
II	71 (49.7%)	32 (45.1%)	39 (54.2%)	
III	61 (42.7%)	35 (49.3%)	26 (36.1%)	
Positive family history for melanoma (%)	57 (40%)	38 (54%)	19 (26%)	<0.001
Risk profile by staging (%)				<0.001
>100 nevi, > 5 dysplastic nevi, dysplastic nevus syndrome, CDKN4A mutation and/or positive family history of melanoma	69 (48.3%)	69 (100%)	0 (0.0%)	
Melanoma in situ or cutaneous melanoma	72 (50.3%)	0 (0.0%)	72 (97.3%)	
Metastatic melanoma	2 (1.4%)	0 (0.0%)	2 (2.7%)	
AJCC 8th edition (% subgroup)**)**				<0.001
in situ	13 (9.1%)	0 (0%)	13 (17.6%)	

I		49 (34.3%)	0 (0%)	49 (66.2%)	
	IA	41 (55.4%)	0 (0%)	41 (55.4%)	
	IB	9 (12.2%)	0 (0%)	9 (12.2%)	
II		3 (2.1%)	0 (0%)	3 (4.1%)	
	IIA	3 (4.1%)	0 (0%)	3 (4.1%)	
	IIB	0 (0%)	0 (0%)	0 (0%)	
	IIC	0 (0%)	0 (0%)	0 (0%)	
III		7 (4.9%)	0 (0%)	7 (9.5%)	
	IIIA	1 (1.4%)	0 (0%)	1 (1.4%)	
	IIIB	3 (4.1%)	0 (0%)	3 (4.1%)	
	IIIC	3 (4.1%)	0 (0%)	3 (4.1%)	
	IIID	0 (0%)	0 (0%)	0 (0%)	
IV		2 (1.4%)	0 (0%)	2 (2.7%)	
Assessed nevi by both CNN devices					0.9
Mean (SD)		12 (12)	12 (11)	12 (13)	
Median (Range)		8 (1 - 63)	8 (1- 47)	8 (1 - 63)	
Sum		1,690	795	895	
Gold standard nevi count					0.021
Mean (SD)		210 (158)	234 (159)	188 (155)	
Median (Range)		189 (1 - 871)	224 (7 - 871)	160 (1 - 622)	
Sum		30,061	16,179	13,882	
2D TBP CNN nevi count					>0.9
Mean (SD)		1,324 (873)	1,293 (820)	1,352 (924)	
Median (Range)		1,164 (167 - 4,862)	1,061 (167 - 4,862)	1,178 (282 - 4,226)	
Sum		189,268	89,228	100,040	
3D TBP CNN nevi count					>0.9
Mean (SD)		469 (382)	454 (346)	483 (414)	
Median (Range)		344 (32 - 2,031)	332 (32 - 2,002)	362 (54 - 2,031)	
Sum		67,081	31,337	35,744	
Patients with nevus/nevi excision (%)		53 (37.1%)	23 (33.3%)	30 (40.5%)	0.4
Excision(s) per patient					0.7
Mean (SD)		1.42 (0.63)	1.48 (0.73)	1.37 (0.56)	
Median (Range)		1.00 (1.00 - 3.00)	1.00 (1.00 - 3.00)	1.00 (1.00 - 3.00)	
Total excisions					
Sum		75	34	41	

[1]Median (Range); n (%); Median (SD)

[2]Wilcoxon rank sum test; Pearson's Chi-squared test; Fisher's exact test

3.2. Correlation and accuracy of 2D- and 3D-TBP CNN risk classification of melanocytic lesions

We observed that 3D- and 2D-TBP CNN classified most of the lesions as benign (n=1,603, 94.9%), while 15 (0.9%) were categorized as suspicious. The remaining 72 (4.3%) lesions were classified differently by the AI systems (Figure 1, Appendices Figure A.1). We also calculated the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of 2D and 3D TBP CNN using the dermatologist assessment as the ground truth (Table 2a and 2b). 2D TBP CNN had a sensitivity of 58.3% and a specificity of 97.1%. 3D TBP CNN had a sensitivity of 66.7% and a specificity of 97.7%. When dermatologists predicted melanoma, the 2D-CNN agreed in 58.33% of cases and the 3D TBP CNN in 66.67% of cases.

Table 2a. Performance of lesion assessment of 3D CNN, 2D CNN, respectively dermatologists and dermatologists with AI based on their ground truth dermatologists and histopathology

	Assessment	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Ground truth: Dermatologists, n = 1,690	2D CNN	58.33	97.14	12.73	99.69
	3D CNN	66.67	97.68	17.02	99.76
Ground truth: Histopathology, n = 75	2D CNN	70.00	40.00	15.22	89.55
	3D CNN	90.00	64.62	28.12	97.67
	Dermatologists	90.00	92.31	64.29	98.36
	Dermatologists plus AI	90.00	86.15	50.00	98.25

**Note: AI = Artificial Intelligence, CNN = convolutional neural network, PPV = positive predictive value, NPV=negative predictive value*

Table 2b. Correlation of lesion classification by dermatologist compared to 2D TBP CNN and 3D TBP CNN

	Suspicious by dermatologists (n = 12)	Benign by dermatologists (n = 1,678)
2D TBP CNN		
Suspicious (n = 55)	7 (58.33%)	48 (2.86%)
Benign (n = 1,635)	5 (41.67%	1,630 (97.14%)
3D TBP CNN		
Suspicious (n = 47)	8 (66.67%)	39 (2.32%)
Benign (n = 1,643)	4 (33.33%)	1,639 (97.68%)

**Note: CNN = convolutional neural network*

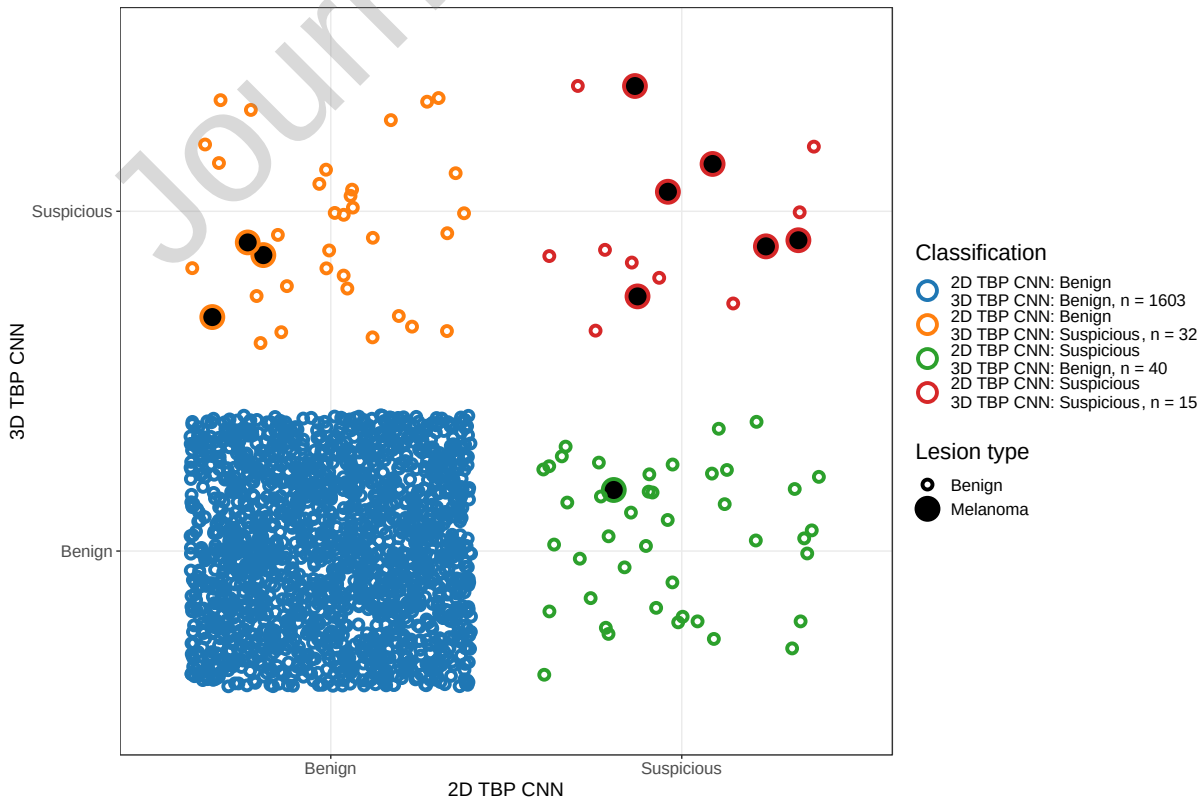

Figure 1. Overview and accuracy of the dichotomous risk classification of melanocytic skin lesions by 3D and 2D TBP CNN. Blue circles = lesions categorized as benign (n=1603) by 3D TBP CNN and 2D TBP CNN, orange circles = only 3D TBP CNN categorized lesions as suspicious (n=32), green circles = only 2D TBP CNN classified as suspicious (n=40) and red circles = suspicious lesions by both systems (n=15). Big black dot = melanoma (n=10), little circle = benign lesion.
TBP = total body photography, CNN = convolutional neural network

The 2D and 3D devices showed a low correlation coefficient for the risk assessment scores (R=0.36). The higher the scores were, the more likely the scores did not correlate (Figure 2). False-negative classified melanomas (n=4) were either in situ (n=2) or superficial spreading melanomas stage IA (n=2).

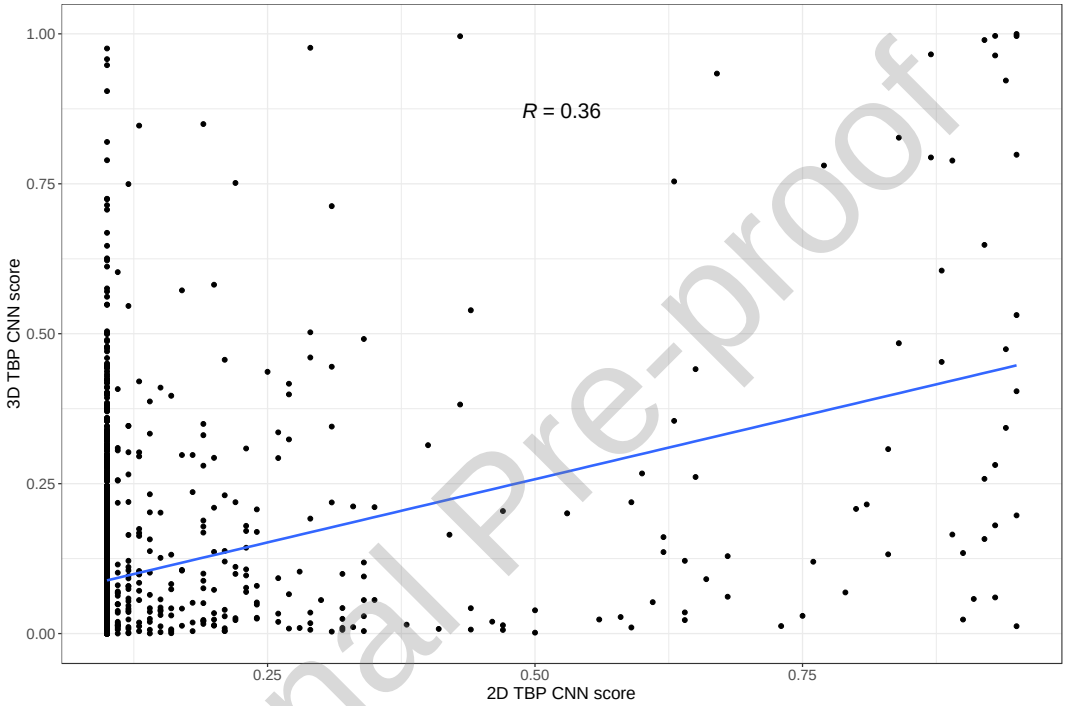

Figure 2. Correlation of risk classification scores of melanocytic skin lesions by 3D TBP CNN and 2D TBP CNN. Blue line = correlation line of both systems if classification scores matched.
Note: Overlapping measurements are seen as one dot. R = correlation coefficient, TBP = total body photography, CNN = convolutional neural network

In 75 excisions, we calculated sensitivity and specificity of the dermatologist, dermatologist + AI, 2D CNN, and 3D CNN using histology as the ground truth (Table 2, Table 3). The dermatologist, dermatologist + AI and 3D TBP CNN had all the highest sensitivity at 90%, whereas the 2D TBP CNN had a sensitivity of 70%. The dermatologist showed the highest specificity at 92.31%, followed by the dermatologist + AI at 86.15%, 3D TBP CNN with 64.62% and finally 2D TBP CNN with the lowest specificity at 40%. While the AI did not improve the sensitivity of the dermatologist assessment, it slightly decreased the specificity. Depending on the professional experience, different numbers of patients were examined (beginners: n=103 patients; intermediate dermatologists n=32 patients; expert n=8 patients). We observed an experience-dependent trend in the accuracy for the classification of lesions (Appendices Table A.2).

Table 3. Performance of the 3D TBP CNN, 2D TBP CNN and dermatologists with and without assistance by the AI systems in classifying pigmented skin lesions based on histology (n=75).

	Overall		Detailed subgroups			
Characteristics	**Melanoma n = 10[1]**	**Not Melanoma n = 65[1]**	Melanocytic nevus n = 22[1]	Dysplastic nevus n = 24[1]	Melanoma n = 10[1]	Others*, n = 19[1]
2D TBP CNN						

not melanoma suspicious	3 (30.0%)	26 (40.0%)	7 (31.8%)	14 (58.3%)	3 (30.0%)	5 (26.3%)
melanoma suspicious	7 (70.0%)	39 (60.0%)	15 (68.2%)	10 (41.7%)	7 (70.0%)	14 (73.7%)
3D TBP CNN						
not melanoma suspicious	1 (10.0%)	42 (64.6%)	20 (90.9%)	11 (45.8%)	1 (10.0%)	11 (57.9%)
melanoma suspicious	9 (90.0%)	23 (35.4%)	2 (9.1%)	13 (54.2%)	9 (90.0%)	8 (44.4%)
Dermatologist						
not melanoma suspicious	1 (10.0%)	60 (92.3%)	19 (86.4%)	22 (91.7%)	1 (10.0%)	19 (100%)
melanoma suspicious	9 (90.0%)	5 (7.7%)	3 (13.6%)	2 (8.3%)	9 (90.0%)	0 (0%)
Dermatologist + AI						
not melanoma suspicious	1 (10.0%)	56 (86.2%)	18 (81.8%)	20 (83.3%)	1 (10.0%)	18 (94.7%)
melanoma suspicious	9 (90.0%)	9 (13.8%)	4 (18.2%)	4 (16.7%)	9 (90.0%)	1 (5.3%)

[1]n (%)
**Note: Other lesions that were excised comprised pigmented basal cell carcinoma, lentigo solaris, seborrheic keratosis (pigmented), dermatofibroma, folliculitis with perifolliculitis, lichenoid keratosis.*
AI = Artificial Intelligence, TBP = total body photography, CNN = convolutional neural network

As the numeric scores from the 2D and 3D TBP CNN were available, we evaluated their performance against histology as the ground truth (Figure 3). 3D TBP CNN had an AUC of 0.92 (CI 0.85-1.00), whereas an optimal cut-point of 8.7 maximized sensitivity to 90% and specificity to 85%. 2D TBP CNN had an AUC of 0.68 (CI 0.46-0.90), with an optimal cut-point of 0.8 maximizing sensitivity to 70% and specificity to 77%.

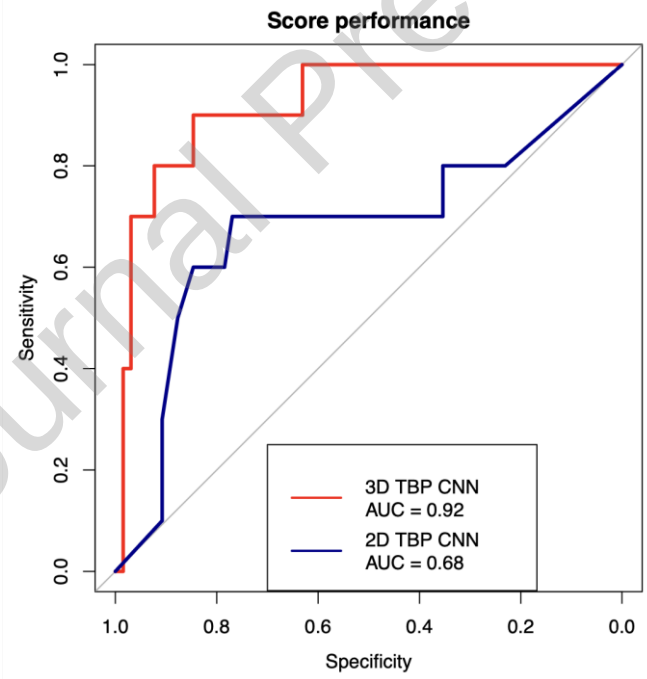

Figure 3. Receiver operating characteristic (ROC) curve of melanoma score classification by 2D TBP CNN and 3D TBP CNN compared to the gold standard histology. Score performance of 3D CNN and 2D CNN (0.0-10.0 3D TBP CNN, 0.0-1.0 2D TBP CNN). *AUC = area under the curve, AI = artificial intelligence, TBP = total body photography, CNN = convolutional neural network*

3.3. Duplicate measurements of risk scores of pigmented skin lesions by 2D- and 3D-TBP CNN

In a subgroup of 21 patients, we measured each lesion (n=246) twice with both AI systems. 3D-TBP CNN showed a higher correlation coefficient for the repeated measurements (R=0.89) than 2D-TBP CNN (R=0.79). The variation in scores was greater in higher scores in 2D- compared to 3D-TBP CNN (Figure 4).

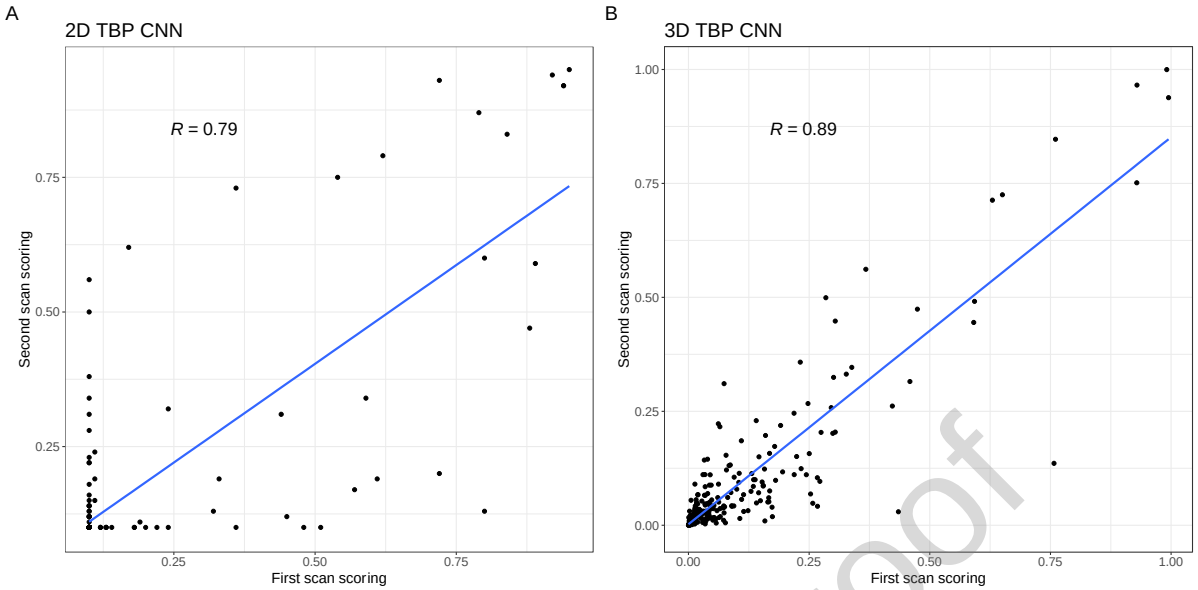

Figure 4. Duplicate measurements of risk score of pigmented skin lesions by 3D TBP CNN and 2D TBP CNN of 246 melanocytic lesions. A) 2D TBP CNN. B) 3D TBP CNN. Blue line = correlation line if both scoring of the same lesion would be the same.
Note: Overlapping measurements are seen as one dot. R = correlation coefficient, TBP = total body photography, CNN = convolutional neural network

The variation between the first and second measurement varied mostly by ≤0.1 in both systems (2D-TBP CNN 87%, 3D-TBP CNN 87.9%; Figure 5). We observed that 3D-TBP CNN showed some variation (9.7%) between >0.1 and ≤0.2. while 2D-TBP CNN revealed more variations (5.7%) of >0.2 and ≤0.4.

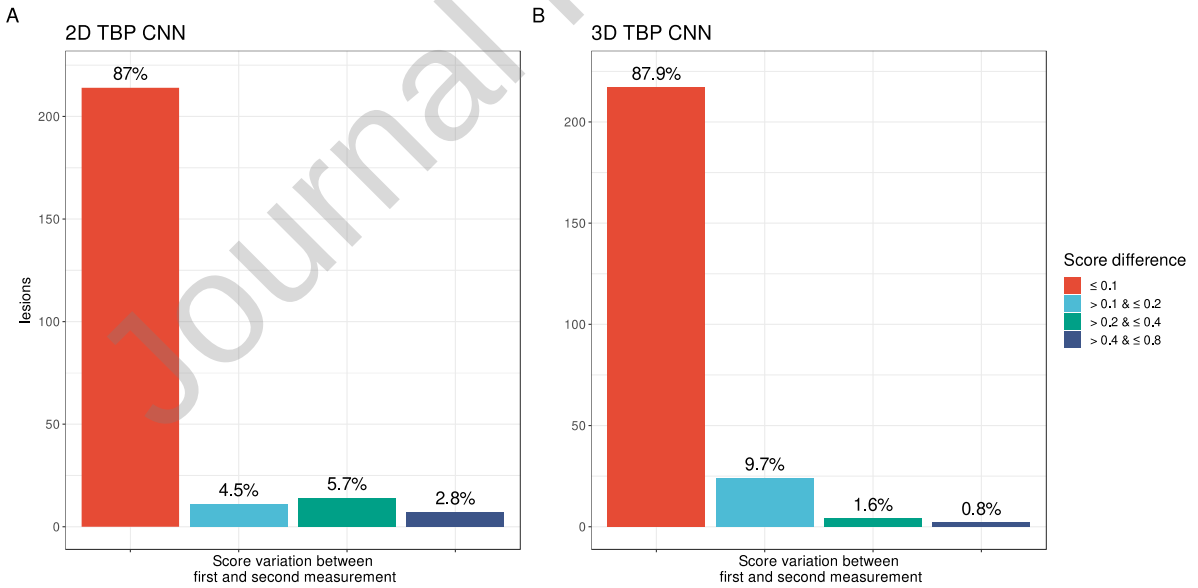

Figure 5. Score differences of the first and second scan of 246 melanocytic nevi of 2D TBP CNN and 3D TBP CNN. Variation of the scores of twice-recorded lesions. A) 2D TBP CNN: 87% of the repeated measures fall into the ≤0.1 range (n=210), 4.5% in the >0.1 & ≤0.2 range (n=11), 5.7% in the >0.2 & ≤0.4 range (n=14) and 2.8% in the >0.4 & ≤0.8 range (n=7), B) 3D TBP CNN: 87.9% of the repeated measures fall into the ≤0.1 range (n=217), 9.7% in the >0.1 & ≤0.2 range (n=24), 1.6% in the >0.2 & ≤0.4 range (n=4) and 0.8% in the >0.4 & ≤0.8 range (n=2). *TBP = total body photography, CNN = convolutional neural network*

3.4. Comparison of automated total nevi count by 3D- and 2D-TBP CNN with gold standard

We observed significant differences in the total nevi count per patient between all three methods (p<0.0001). The medical staff counted a mean of 210 lesions, while 3D-TBP CNN counted more than the double of lesions (n=469) and 2D-TBP CNN counted 6.3 times more lesions (n=1,324; Figure 6).

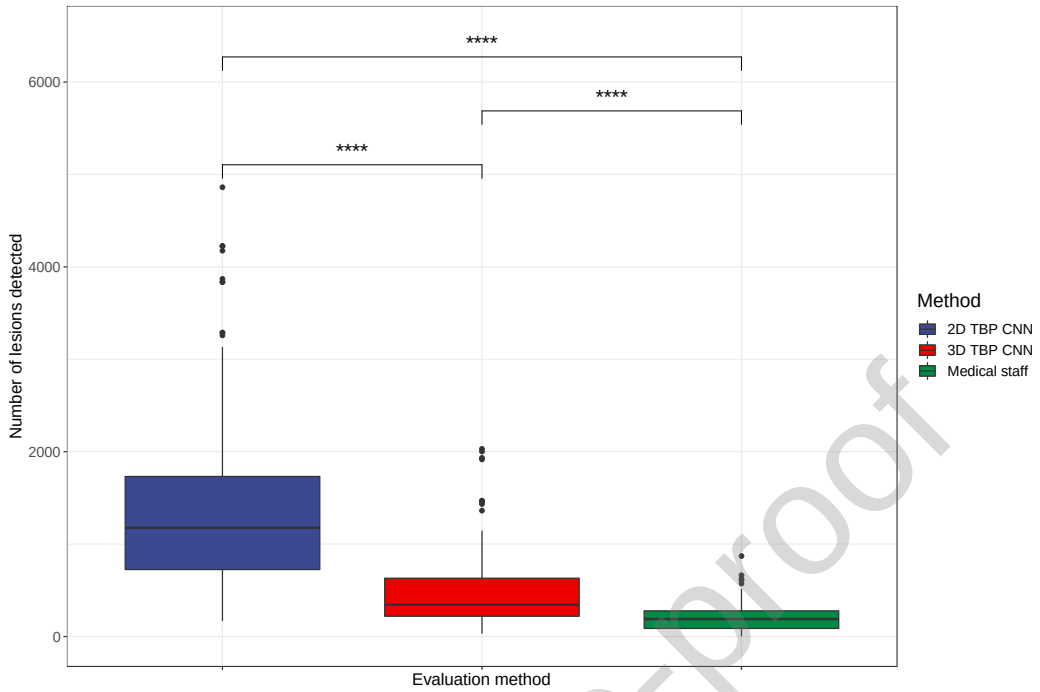

Figure 6. Total count of all melanocytic lesions per patient: Comparison of automated count by 3D TBP CNN and 2D TBP CNN with gold standard (medical staff). Mean 2D TBP CNN lesion count of 1,324 (median = 1164, range 167 - 4862), 3D TBP CNN average lesion count of 469 (median = 344, range 32 - 2031) and mean medical staff count of 210 (median = 189, range 1-871 lesions). ****p<0.0001 *TBP = total body photography, CNN = convolutional neural network*

3D TBP CNN showed a correlation of 0.47 compared with the gold standard, whereas 2D TBP CNN a coefficient of 0.54 (Appendices Figure A.2). Both AI devices revealed a high correlation coefficient of 0.91.

Discussion

Despite the advantages of AI in experimental classification of melanocytic lesions, there is a lack of real-world data on the potential of clinician-AI collaboration in melanoma detection in daily clinical practice [19]. We investigated a head-to-head comparison of the performance of 2D and 3D CNNs alone and in combination with dermatologists in the detection of melanoma in a real-world setting. The novel 3D CNN's classification of dermoscopic images of melanocytic origin and coefficient for repeated measurement of risk scores were superior to the well-established 2D CNN. Under real-world conditions, the diagnostic accuracy of dermatologists without AI support was comparable to 3D CNN, but augmented intelligence was inferior to human intelligence and 3D CNN alone. Automated nevi counting by 3D and 2D TBP CNNs still has major weaknesses.

Diagnostic performance of 3D and 2D CNNs in real-world clinical practice

The benefits of CNN-assisted classification in the early detection of melanoma in experimental settings have been demonstrated in previous studies, where CNNs performed as well as or better than dermatologists [13,20-24]. Comparing to histopathology as ground truth, the novel 3D CNN showed superior sensitivity compared to the 2D CNN device in our study and performance at the same level as dermatologists. In terms of specificity, both 3D and 2D CNN were outperformed by dermatologists. The implementation of current 2D and 3D CNN in clinical routine would not only lead to unnecessary excisions due to low specificity, but also to patient uncertainty and anxiety. Critics also say AI is not yet a viable replacement for dermatologists [21,25]. By using the dermatologists as ground truth on all assessed lesions, both devices demonstrated a lower sensitivity but a higher specificity. This suggests that AI algorithms identify more lesions as false negative and are therefore not as good as dermatologist at identifying suspicious lesions. Only by using histopathology as a second ground truth we distinguish between the accuracy in labelling suspicious lesions and detecting melanoma. Thus, the devices are still good at detecting melanoma, while the labelling of suspicious lesions still needs improvement to perform similar as dermatologists. Nevertheless, benign lesions were detected most accurately compared to physicians, resulting in high sensitivity. As previously demonstrated through a web-based user interface for comparing 3 forms of output from the CNN as decision support to human raters, the clinical diagnostic accuracy of physicians can be improved with AI support in certain scenarios in contrast to the sole opinion of either physicians or AI [24]. However, when AI results are

flawed, they can also mislead clinicians including experts in a data set analysis, the use of 2D CNN in ROC curve analysis resulted in significantly greater specificity (82.5%) compared to dermatologists (71.3%), even when dermatologists received more clinical knowledge (75.7%) [13]. Additionally, the 2D CNN was found to perform on the same level as dermatologists, even with a broader range of diagnoses [20]. A cross-sectional retrospective reader study found that the highest accuracy was achieved by the collective human intelligence of multiple experts (80%), compared to dermatologists alone (75.7%) or 2D CNN (70%) [26]. Despite that dermatologists with additional clinical information compared to a market-approved 2D CNN performed on the same level regarding the diagnostic accuracy [20], instead of being outperformed by AI under artificial conditions when analysing dermatoscopic images without clinical context [13], our data revealed that in real-world setting 3D CNN performed similarly to dermatologists, while 2D CNN was exceeded. In contrast to our results, a recent prospective study by Winkler et al. even showed that dermatologists, especially less experienced ones, performed better after the AI knowledge, improving their sensitivity and specificity (84.2% improved to 100% and 72.1% to 83.7%, respectively) in assessing exclusively suspicious lesions. Interestingly, the specificity of CNN alone was better (88.9%) and deteriorated slightly with dermatologists (83.7%), leaving room for improvement [27]. A Canadian, similarly conducted study highlighted the high sensitivity and specificity (88.1% and 78.8%) of FotoFinder, (Moleanalyzer Pro) compared to histopathology as gold standard, showing the best results compared to other methods used such as MelaFind, Tuebinger Moleanalyzer by FotoFinder, Verisante Aura and Teledermoscopic diagnosis. In this study dermatologists (n=2) showed a higher sensitivity of 96.6% but a lower specificity (32.2%) than Moleanalyzer Pro [28]. Due to the novelty of the 3D CNN device, the performance of its specific AI scores has not been investigated so far. We suggest further technological improvements of algorithms for malignancy assessments prior to clinical implementation.

Augmented intelligence in classifying melanocytic lesions

Overall, dermatologists' performance deteriorated after AI collaboration (AUC-ROC without AI 0.91 vs. with AI 0.88). Although our data did not support the benefit of augmented intelligence, we suspect that a larger study population might reveal possible advantages in decision making, especially for beginners. For example, a recent South Korean real-world study showed that untrained non-dermatology residents improved their accuracy in detecting melanoma with the benefit of AI (with AI 53.9% vs. without AI 43.8%) [29], supporting the concept of augmented intelligence. Our results indicated that the diagnostic accuracy of dermatologists improved with increasing professional experience, highlighting the need for AI support, especially for newcomers.

Nevertheless, clinical differentiation of dysplastic nevi remains a challenge even for trained dermatologists. Our data highlight that 2D and 3D CNN were outperformed by dermatologists in terms of true positive rates for dysplastic nevi. Further data and advances in CNN are needed to improve augmented intelligence for the detection of such difficult, clinically and histologically challenging, borderline lesions to avoid unnecessary biopsies.

Reproducibility of 3D and 2D CNN risk score measurements

We detected that 3D CNN had a higher correlation coefficient (0.89) between the two repeated measurements than 2D CNN (0.79). Factors influencing the deep learning-based algorithm include camera rotation, scarring, or even positioning of hair or skin markings, highlighting the need for consistent lesion imaging [30,31]. Our findings highlight the importance of identifying the bias for the first and second measurements. There is a lack of transparency in AI datasets and algorithms for skin disease [32]. Our data emphasise that deep learning algorithms need to be further optimised to avoid factors such as rotation, camera pressure and exposure intensity affecting the repeatability of the results.

Nevus counts using 3D and 2D TBP

AI-assisted nevi counting could help improve detection of de novo lesions, known as drivers, in 2/3 of melanomas [4]. However, there is no established gold standard for counting nevi [33]. Previous studies have shown a high degree of subjectivity and lack of comparability due to differences in the size of nevi counted (>2mm, >3mm, >5mm) and the standard method of counting (physician, trained researcher, self-report) [33]. In a recent Australian study, a 97% correlation between 3D CNN counts and the gold standard (senior dermatologist) was reported [16]. In contrast, we observed significant differences between gold standard and 3D CNN nevus counts. Although the 3D CNN outperformed the 2D CNN in automated detection, both devices still have significant weaknesses for use in everyday clinical practice by overcounting lesions. We observed significant differences in the performance of the devices and their TBP capturing ability, which can be attributed to technical factors. These differences led to significant variations in the counting of nevi. We noticed that the 2D TBP counted naevi multiple times, due to overlapping areas in the pictures, whereas the 3D TBP avoided this issue by creating a single, complete representation of the skin surface. In agreement with Betz-Stablein et al, we observed that the CNN miscounted seborrhoeic keratoses especially in the elderly population [16]. Interestingly, both TBP CNN devices appear to correlate (R=0.91), possibly reflecting similar machine learning-based challenges in macroscopic differentiation between nevi and other pigmented skin lesions such as lentigines. We recognized a possible correlation between the falsely detected high nevi count and median age of 56 years in our study population which may be due to the presence of numerous benign skin lesions such as seborrheic keratoses, haemangiomas, verrucae and especially

solar lentigines. We assume that our data might differ from the Australian study with other demographics despite similar median ages (56 vs. 57 years) because of Australians' early sun protection campaign since the 1980s, resulting in reduced sun damage [34]. Automated nevi counting is currently underpowered for monitoring high-risk patients in clinical practice and requires algorithmic adjustments before clinical application. Future research projects are needed to investigate the reproducibility of nevi counts and the correct nevi recognition by both TBP devices. In order to achieve accurate results, each device would need to perform multiple TBP per patient to gather additional data on the reliability and accuracy of the nevi count executed by CNN. We anticipate that the time-saving image documentation provided by TBP has the potential to be highly beneficial in supporting dermatologists in the future.

Strengths and limitations

A strength of our study is that the diagnostic accuracy of 3D, 2D CNN, dermatologists and augmented intelligence is addressed in a real-world setting, including a large number of 1,690 lesions. Due to certain limitations, generalisation of the results should be considered with caution. Presently, our reliance is primarily on non-clinical studies, highlighting the need for more clinical research to evaluate the AI score performance of both devices. Despite the high number of melanocytic lesions included, the number of melanomas was relatively low, reflecting a typical limitation of prospective studies. A small test set size of our performance evaluation (n=75) may produce biased and non-generalizable result; as such, we intend to conduct an analysis following 3 years of prospective data collection. As only patients with a Caucasian genetic background were included, this study may not provide comparable results in other skin types. Since this study was conducted at a single centre and the combination of Vectra WB360 and FotoFinder ATBM was not widely available at the beginning of the trial, it is important to exercise caution when generalizing the findings, especially with respect to demographics. Additionally, it is worth noting that the novel Vectra WB360 was only available 13 times worldwide, and only one was located in Switzerland at the time of the study. The limited access to Vectra WB360 can be attributed due to its current cost and size. A limitation of FotoFinder ATBM is the overcounting of nevi due to the relevant overlap in the captured images. Additionally, the use of 2D TBP is more time-consuming than 3D TBP as it requires obtaining all pictures of different body regions to create a complete TBP. Both devices are larger and more expensive compared to the commercially available handheld dermoscopes. Since all patients in our study received only one TBP per device, statements on the reproducibility of automatic naevi counting are not possible, as multiple TBPs would be necessary. To achieve our sample size, dermatologists with different professional experience levels were involved in this trial. However, beginners have examined proportionally more patients than experts. This bias is attributed to the fact that larger hospitals tend to have a greater proportion of younger, less experienced clinicians leading to an imbalance in the distribution of patients.

Conclusions

Our study emphasises the importance of real-world validation of AI algorithms and highlights the potential of 3D CNN in melanoma detection. We revealed that the novel 3D CNN device outperformed 2D CNN in the classification of melanocytic lesions and in the reproducibility of the scores. 3D CNN demonstrated its usefulness in practice by achieving comparable sensitivity with dermatologists. Current limitations for the use of AI in clinical practice include the low specificity of CNNs, the inferiority of augmented intelligence, and the deficient AI-assisted TBP nevus counting. Overall, we believe that augmented intelligence still has a promising potential for less experienced dermatologists and non-specialists in daily practice.

Acknowledgements
The authors thank all participants who made this study possible. LVM gratefully acknowledges support from the Research Foundation for Young Researchers Grant of the University of Basel, Switzerland and by the Voluntary Academic Society Grant, Basel, Switzerland.

Funding
This research project was fully funded by the Department of Dermatology, as well as in minor part by Research Foundation for Young Researchers Grant of the University of Basel, Switzerland, and by the Voluntary Academic Society Grant, Basel, Switzerland. Companies producing the tested products did not have the opportunity to comment or influence the manuscript.

Author Contributions
Conceptualization: LVM, AAN and SEC; Data curation: SEC, LK, SH, MK, JSB, CFD, AG, CJ, LS, LVM; Formal analysis: PC, SEC and LVM; Funding acquisition: LVM and AAN; Investigation: SEC, LK, SH, MK, JSB, CFD, AG, CJ, LS, LVM; Methodology: SEC, PC, LVM; Project administration: LVM; Resources: LVM; Supervision: LVM; Visualization: PC, SEC and LVM; Roles/Writing - original draft: SEC, PC and LVM. Writing - review & editing: SEC, LK, SH, MK, JTM, JSB, CFD, AG, CJ, LS, JKP, ML, AAN, LVM. All authors have read and agreed to the published version of the manuscript.

References

1. Arnold, M.; Holterhues, C.; Hollestein, L.M.; Coebergh, J.W.; Nijsten, T.; Pukkala, E.; Holleczek, B.; Tryggvadottir, L.; Comber, H.; Bento, M.J.; et al. Trends in incidence and predictions of cutaneous melanoma across Europe up to 2015. *J Eur Acad Dermatol Venereol* **2014**, *28*, 1170-1178, doi:10.1111/jdv.12236.
2. Erdmann, F.; Lortet-Tieulent, J.; Schuz, J.; Zeeb, H.; Greinert, R.; Breitbart, E.W.; Bray, F. International trends in the incidence of malignant melanoma 1953-2008--are recent generations at higher or lower risk? *Int J Cancer* **2013**, *132*, 385-400, doi:10.1002/ijc.27616.
3. Gershenwald, J.E.; Scolyer, R.A.; Hess, K.R.; Sondak, V.K.; Long, G.V.; Ross, M.I.; Lazar, A.J.; Faries, M.B.; Kirkwood, J.M.; McArthur, G.A.; et al. Melanoma staging: Evidence-based changes in the American Joint Committee on Cancer eighth edition cancer staging manual. *CA Cancer J Clin* **2017**, *67*, 472-492, doi:10.3322/caac.21409.
4. Pampena, R.; Kyrgidis, A.; Lallas, A.; Moscarella, E.; Argenziano, G.; Longo, C. A meta-analysis of nevus-associated melanoma: Prevalence and practical implications. *J Am Acad Dermatol* **2017**, *77*, 938-945 e934, doi:10.1016/j.jaad.2017.06.149.
5. Garbe, C.; Amaral, T.; Peris, K.; Hauschild, A.; Arenberger, P.; Basset-Seguin, N.; Bastholt, L.; Bataille, V.; Del Marmol, V.; Dreno, B.; et al. European consensus-based interdisciplinary guideline for melanoma. Part 1: Diagnostics: Update 2022. *Eur J Cancer* **2022**, *170*, 236-255, doi:10.1016/j.ejca.2022.03.008.
6. Ba, W.; Wu, H.; Chen, W.W.; Wang, S.H.; Zhang, Z.Y.; Wei, X.J.; Wang, W.J.; Yang, L.; Zhou, D.M.; Zhuang, Y.X.; et al. Convolutional neural network assistance significantly improves dermatologists' diagnosis of cutaneous tumours using clinical images. *Eur J Cancer* **2022**, *169*, 156-165, doi:10.1016/j.ejca.2022.04.015.
7. Feit, N.E.; Dusza, S.W.; Marghoob, A.A. Melanomas detected with the aid of total cutaneous photography. *Br J Dermatol* **2004**, *150*, 706-714, doi:10.1111/j.0007-0963.2004.05892.x.
8. Truong, A.; Strazzulla, L.; March, J.; Boucher, K.M.; Nelson, K.C.; Kim, C.C.; Grossman, D. Reduction in nevus biopsies in patients monitored by total body photography. *J Am Acad Dermatol* **2016**, *75*, 135-143 e135, doi:10.1016/j.jaad.2016.02.1152.
9. Ji-Xu, A.; Dinnes, J.; Matin, R.N. Total body photography for the diagnosis of cutaneous melanoma in adults: a systematic review and meta-analysis. *Br J Dermatol* **2021**, *185*, 302-312, doi:10.1111/bjd.19759.
10. Salerni, G.; Carrera, C.; Lovatto, L.; Puig-Butille, J.A.; Badenas, C.; Plana, E.; Puig, S.; Malvehy, J. Benefits of total body photography and digital dermatoscopy ("two-step method of digital follow-up") in the early diagnosis of melanoma in patients at high risk for melanoma. *J Am Acad Dermatol* **2012**, *67*, e17-27, doi:10.1016/j.jaad.2011.04.008.
11. Canfield. Canfield Launches Commercial Version of VECTRA WB360 3D Whole Body Imaging Solution. Available online: https://www.canfieldsci.com/in-the-news/stories/canfield-launches-

commercial-version-of-vectra-wb360---worlds-first-3d-whole-body-imaging-solution (accessed on 15.05.2023

12. Rayner, J.E.; Laino, A.M.; Nufer, K.L.; Adams, L.; Raphael, A.P.; Menzies, S.W.; Soyer, H.P. Clinical Perspective of 3D Total Body Photography for Early Detection and Screening of Melanoma. *Front Med (Lausanne)* **2018**, *5*, 152, doi:10.3389/fmed.2018.00152.
13. Haenssle, H.A.; Fink, C.; Schneiderbauer, R.; Toberer, F.; Buhl, T.; Blum, A.; Kalloo, A.; Hassen, A.B.H.; Thomas, L.; Enk, A.; et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol* **2018**, *29*, 1836-1842, doi:10.1093/annonc/mdy166.
14. Brinker, T.J.; Hekler, A.; Enk, A.H.; Klode, J.; Hauschild, A.; Berking, C.; Schilling, B.; Haferkamp, S.; Schadendorf, D.; Holland-Letz, T.; et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *Eur J Cancer* **2019**, *113*, 47-54, doi:10.1016/j.ejca.2019.04.001.
15. Marchetti, M.A.; Codella, N.C.F.; Dusza, S.W.; Gutman, D.A.; Helba, B.; Kalloo, A.; Mishra, N.; Carrera, C.; Celebi, M.E.; DeFazio, J.L.; et al. Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. *J Am Acad Dermatol* **2018**, *78*, 270-277 e271, doi:10.1016/j.jaad.2017.08.016.
16. Betz-Stablein, B.; D'Alessandro, B.; Koh, U.; Plasmeijer, E.; Janda, M.; Menzies, S.W.; Hofmann-Wellenhof, R.; Green, A.C.; Soyer, H.P. Reproducible Naevus Counts Using 3D Total Body Photography and Convolutional Neural Networks. *Dermatology* **2021**, 1-8, doi:10.1159/000517218.
17. Jahn, A.S.; Navarini, A.A.; Cerminara, S.E.; Kostner, L.; Huber, S.M.; Kunz, M.; Maul, J.T.; Dummer, R.; Sommer, S.; Neuner, A.D.; 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 (Basel)* **2022**, *14*, doi:10.3390/cancers14153829.
18. Canfield. Canfield Scientific Premieres New AI Solutions Including the Most Advanced Grading System for Suspicious Lesions – DEXI™ and the Fastest and Most Convenient Hair Consultation Solution – HairMetrix™. **2019**.
19. Taylor, M.; Liu, X.; Denniston, A.; Esteva, A.; Ko, J.; Daneshjou, R.; Chan, A.W.; Spirit, A.I.; Group, C.-A.W. Raising the Bar for Randomized Trials Involving Artificial Intelligence: The SPIRIT-Artificial Intelligence and CONSORT-Artificial Intelligence Guidelines. *J Invest Dermatol* **2021**, *141*, 2109-2111, doi:10.1016/j.jid.2021.02.744.
20. Haenssle, H.A.; Fink, C.; Toberer, F.; Winkler, J.; Stolz, W.; Deinlein, T.; Hofmann-Wellenhof, R.; Lallas, A.; Emmert, S.; Buhl, T.; et al. Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. *Ann Oncol* **2020**, *31*, 137-143, doi:10.1016/j.annonc.2019.10.013.
21. Haggenmuller, S.; Maron, R.C.; Hekler, A.; Utikal, J.S.; Barata, C.; Barnhill, R.L.; Beltraminelli, H.; Berking, C.; Betz-Stablein, B.; Blum, A.; et al. Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. *Eur J Cancer* **2021**, *156*, 202-216, doi:10.1016/j.ejca.2021.06.049.
22. Phillips, M.; Marsden, H.; Jaffe, W.; Matin, R.N.; Wali, G.N.; Greenhalgh, J.; McGrath, E.; James, R.; Ladoyanni, E.; Bewley, A.; et al. Assessment of Accuracy of an Artificial Intelligence Algorithm to Detect Melanoma in Images of Skin Lesions. *JAMA Netw Open* **2019**, *2*, e1913436, doi:10.1001/jamanetworkopen.2019.13436.
23. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115-118, doi:10.1038/nature21056.
24. Tschandl, P.; Rinner, C.; Apalla, Z.; Argenziano, G.; Codella, N.; Halpern, A.; Janda, M.; Lallas, A.; Longo, C.; Malvehy, J.; et al. Human-computer collaboration for skin cancer recognition. *Nat Med* **2020**, *26*, 1229-1234, doi:10.1038/s41591-020-0942-0.
25. Lallas, A.; Argenziano, G. Artificial intelligence and melanoma diagnosis: ignoring human nature may lead to false predictions. *Dermatol Pract Concept* **2018**, *8*, 249-251, doi:10.5826/dpc.0804a01.
26. Winkler, J.K.; Sies, K.; Fink, C.; Toberer, F.; Enk, A.; Abassi, M.S.; Fuchs, T.; Blum, A.; Stolz, W.; Coras-Stepanek, B.; et al. Collective human intelligence outperforms artificial intelligence in a skin lesion classification task. *J Dtsch Dermatol Ges* **2021**, *19*, 1178-1184, doi:10.1111/ddg.14510.
27. Winkler, J.K.; Blum, A.; Kommoss, K.; Enk, A.; Toberer, F.; Rosenberger, A.; Haenssle, H.A. Assessment of Diagnostic Performance of Dermatologists Cooperating With a Convolutional Neural Network in a Prospective Clinical Study: Human With Machine. *JAMA Dermatol* **2023**, doi:10.1001/jamadermatol.2023.0905.

28. MacLellan, A.N.; Price, E.L.; Publicover-Brouwer, P.; Matheson, K.; Ly, T.Y.; Pasternak, S.; Walsh, N.M.; Gallant, C.J.; Oakley, A.; Hull, P.R.; et al. The use of noninvasive imaging techniques in the diagnosis of melanoma: a prospective diagnostic accuracy study. *J Am Acad Dermatol* **2021**, *85*, 353-359, doi:10.1016/j.jaad.2020.04.019.
29. Han, S.S.; Kim, Y.J.; Moon, I.J.; Jung, J.M.; Lee, M.Y.; Lee, W.J.; Won, C.H.; Lee, M.W.; Kim, S.H.; Navarrete-Dechent, C.; et al. Evaluation of Artificial Intelligence-Assisted Diagnosis of Skin Neoplasms: A Single-Center, Paralleled, Unmasked, Randomized Controlled Trial. *J Invest Dermatol* **2022**, *142*, 2353-2362 e2352, doi:10.1016/j.jid.2022.02.003.
30. Winkler, J.K.; Fink, C.; Toberer, F.; Enk, A.; Deinlein, T.; Hofmann-Wellenhof, R.; Thomas, L.; Lallas, A.; Blum, A.; Stolz, W.; et al. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatol* **2019**, *155*, 1135-1141, doi:10.1001/jamadermatol.2019.1735.
31. Maron, R.C.; Hekler, A.; Krieghoff-Henning, E.; Schmitt, M.; Schlager, J.G.; Utikal, J.S.; Brinker, T.J. Reducing the Impact of Confounding Factors on Skin Cancer Classification via Image Segmentation: Technical Model Study. *J Med Internet Res* **2021**, *23*, e21695, doi:10.2196/21695.
32. Daneshjou, R.; Smith, M.P.; Sun, M.D.; Rotemberg, V.; Zou, J. Lack of Transparency and Potential Bias in Artificial Intelligence Data Sets and Algorithms: A Scoping Review. *JAMA Dermatol* **2021**, *157*, 1362-1369, doi:10.1001/jamadermatol.2021.3129.
33. Lawson, D.D.; Moore, D.H., 2nd; Schneider, J.S.; Sagebiel, R.W. Nevus counting as a risk factor for melanoma: comparison of self-count with count by physician. *J Am Acad Dermatol* **1994**, *31*, 438-444, doi:10.1016/s0190-9622(94)70207-1.
34. Tabbakh, T.; Volkov, A.; Wakefield, M.; Dobbinson, S. Implementation of the SunSmart program and population sun protection behaviour in Melbourne, Australia: Results from cross-sectional summer surveys from 1987 to 2017. *PLoS Med* **2019**, *16*, e1002932, doi:10.1371/journal.pmed.1002932.
35. FotoFinder. Informationen zur Leistung und Sicherheit der Produkte. Available online: https://www.fotofinder.de/fileadmin/user_upload/FotoFinder_Systems_Informationsbroschüre_zur_Leistung_und_Sicherheit_DE_Okt_2021.pdf (accessed on 15.05.2023
36. Canfield. 3D-Bildgebungssysteme für Ganzkörperaufnahmen. Available online: https://www.canfieldsci.com/imaging-systems/vectra-wb360-imaging-system/ (accessed on
37. Canfield. VECTRA 3D. Available online: https://surfaceimaging.co.uk/wp-content/uploads/2020/12/VECTRA-WB360-brochure.pdf (accessed on 15.05.2023

Figure legend

Table 1. Characteristics of the study population, nevi counts and number of excisions

Figure 1. Overview and accuracy of the dichotomous risk classification of melanocytic skin lesions by 3D and 2D TBP CNN. Blue circles = lesions categorized as benign (n=1603) by 3D TBP CNN and 2D TBP CNN, orange circles = only 3D TBP CNN categorized lesions as suspicious (n=32), green circles = only 2D TBP CNN classified as suspicious (n=40) and red circles = suspicious lesions by both systems (n=15). Big black dot = melanoma (n=10), little circle = benign lesion.
TBP = total body photography, CNN = convolutional neural network

Figure 2. Correlation of risk classification scores of melanocytic skin lesions by 3D TBP CNN and 2D TBP CNN. Blue line = correlation line of both systems if classification scores matched.
Note: Overlapping measurements are seen as one dot. R = correlation coefficient, TBP = total body photography, CNN = convolutional neural network

Table 2a. Performance of lesion assessment of 3D CNN, 2D CNN, respectively dermatologists and dermatologists with AI based on their ground truth dermatologists and histopathology

Table 2b. Correlation of lesion classification by dermatologist compared to 2D TBP CNN and 3D TBP CNN

Table 3. Performance of the 3D TBP CNN, 2D TBP CNN and dermatologists with and without assistance by the AI systems in classifying pigmented skin lesions based on histology (n=75).

Figure 3. Receiver operating characteristic (ROC) curve of melanoma score classification by 2D TBP CNN and 3D TBP CNN compared to the gold standard histology. Score performance of 3D CNN and 2D CNN (0.0-10.0 3D TBP CNN, 0.0-

1.0 2D TBP CNN). *AUC = area under the curve, AI = artificial intelligence, TBP = total body photography, CNN = convolutional neural network*

Figure 4. Duplicate measurements of risk score of pigmented skin lesions by 3D TBP CNN and 2D TBP CNN of 246 melanocytic lesions. A) 2D TBP CNN. B) 3D TBP CNN. Blue line = correlation line if both scoring of the same lesion would be the same.
Note: Overlapping measurements are seen as one dot. R = correlation coefficient, TBP = total body photography, CNN = convolutional neural network

Figure 5. Score differences of the first and second scan of 246 melanocytic nevi of 2D TBP CNN and 3D TBP CNN. Variation of the scores of twice-recorded lesions. A) 2D TBP CNN: 87% of the repeated measures fall into the ≤0.1 range (n=210), 4.5% in the >0.1 & ≤0.2 range (n=11), 5.7% in the >0.2 & ≤0.4 range (n=14) and 2.8% in the >0.4 & ≤0.8 range (n=7), B) 3D TBP CNN: 87.9% of the repeated measures fall into the ≤0.1 range (n=217), 9.7% in the >0.1 & ≤0.2 range (n=24), 1.6% in the >0.2 & ≤0.4 range (n=4) and 0.8% in the >0.4 & ≤0.8 range (n=2). *TBP = total body photography, CNN = convolutional neural network*

Figure 6. Total count of all melanocytic lesions per patient: Comparison of automated count by 3D TBP CNN and 2D TBP CNN with gold standard (medical staff). Mean 2D TBP CNN lesion count of 1,324 (median = 1164, range 167 - 4862), 3D TBP CNN average lesion count of 469 (median = 344, range 32 - 2031) and mean medical staff count of 210 (median = 189, range 1-871 lesions). ****p<0.0001 *TBP = total body photography, CNN = convolutional neural network*

CRediT author statement - Author Contributions

All authors have read and agreed to the published version of the manuscript. **Sara E. Cerminara**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Roles/Writing - original draft, Writing - review & editing. **Phil Cheng**: Formal analysis, Methodology, Visualization, Roles/Writing – original. **Lisa Kostner**: Data curation, Investigation, Writing - review & editing. **Stephanie Huber**: Data curation, Investigation, Writing - review & editing. **Michael Kunz**: Data curation, Investigation, Writing - review & editing. **Julia-Tatjana Maul**: Writing - review & editing. **Jette S. Böhm**: Data curation, Investigation, Writing - review & editing. **Chiara F. Dettwiler**: Data curation, Investigation, Writing - review & editing. **Anna Geser**: Data curation, Investigation, Writing - review & editing. **Cécile Jakopović**: Data curation, Investigation, Writing - review & editing. **Livia M. Stoffel**: Data curation, Investigation, Writing - review & editing. **Jelissa K. Peter**: Writing - review & editing. **Mitchell Levesque**: Writing - review & editing. **Alexander A. Navarini**: Conceptualization, Funding acquisition, Writing - review & editing. **Lara V. Maul**: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Visualization, Roles/Writing – original, Writing - review & editing

Declaration of interests

☐ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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

Conflict of Interest

SEC has no conflict of interest.

PC has no conflict of interest.

LK has received speaking fees for a presentation sponsored by Boehringer Ingelheim.

SH has no conflict of interest.

MK has received speaking fees from Almirall and Sanofi outside of the current work

J-TM has served as advisor and/or received speaking fees and/or participated in clinical trials sponsored by AbbVie, Almirall, Amgen, BMS, Celgene, Eli Lilly, LEO Pharma, Janssen-Cilag, MSD, Novartis, Pfizer, Pierre Fabre, Roche, Sanofi, UCB.

JSB has no conflict of interest.

CFD has no conflict of interest.

AG has no conflict of interest.

CJ has no conflict of interest.

LMS has no conflict of interest.

JKP has no conflict of interest.

ML has received research funding unrelated to the manuscript from Roche, Novartis, Molecular Partners, and Oncobit.

AAN declares being a consultant and advisor and/or receiving speaking fees and/or grants and/or served as an investigator in clinical trials for AbbVie, Almirall, Amgen, Biomed, BMS, Boehringer Ingelheim, Celgene, Eli Lilly, Galderma, GSK, LEO Pharma, Janssen-Cilag, MSD, Novartis, Pfizer, Pierre Fabre Pharma, Regeneron, Sandoz, Sanofi, and UCB.

LVM has served as advisor and/or received speaking fees and/or participated in clinical trials sponsored by Almirall, Amgen, Eli Lilly, MSD, Novartis, Pierre Fabre, Roche, and Sanofi outside of the current work.

Highlights

- Real-world comparison of dermatologists and AI in melanoma detection.
- Sensitivity of 3D CNN vs. dermatologists comparable in classification.
- 2D CNN outperformed in classifying melanocytic lesions by 3D CNN/dermatologists.
- Low specificity by augmented intelligence in real-life setting.
- CNN nevi counting still lacking utility for clinical practice due to low correlation.