Robustness of a neural network approved for skin cancer diagnosis against variations in sequential images

Alicia Zimmermann, MD; Julia K. Winkler, MD; Katharina Kommoss, MD; Anastasia Vollmer, MD; Ferdinand Toberer, MD; Alexander Enk, MD; Albert Rosenberger, MSC; Holger A. Haenssle, MD

Department of Dermatology, University of Heidelberg, Heidelberg, Germany

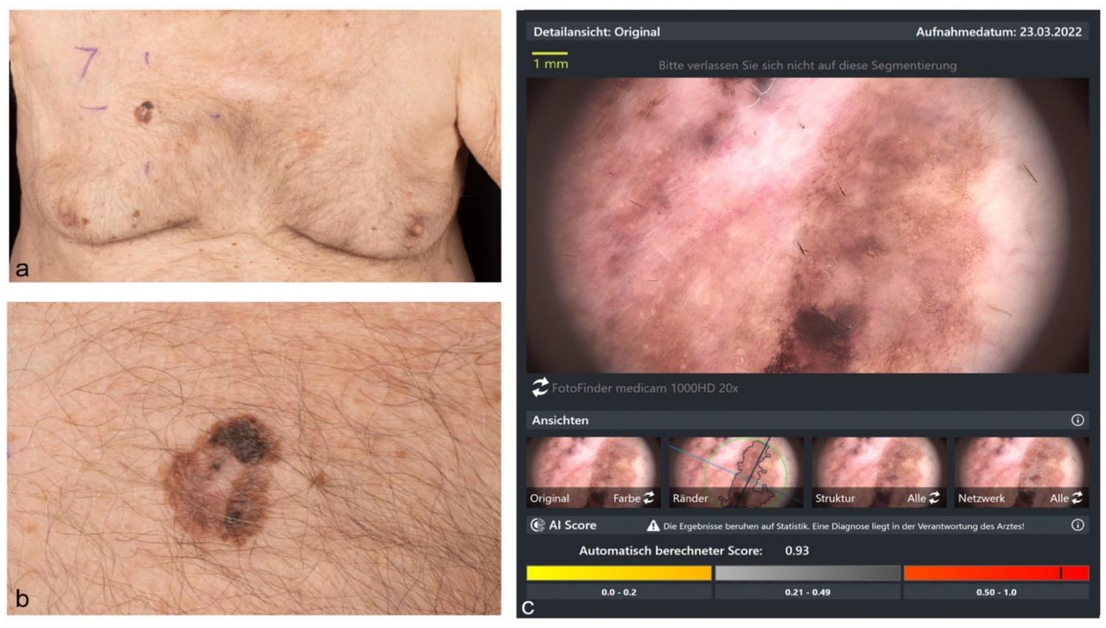

Fig. 1: Clinical overview image of a melanoma located at the right parasternal region (a), close-up image (b), dermoscopic image and CNN malignancy score of 0.95 (c).

Objective

Although convolutional neural networks (CNN) have achieved dermatologist-level performance in skin cancer classification, their robustness to subtle, real-world image variations remains uncertain. This study evaluated the robustness of an MDR class IIa-certified, commercially available CNN using sequential dermoscopic images of identical skin lesions to assess its reliability under routine clinical conditions.

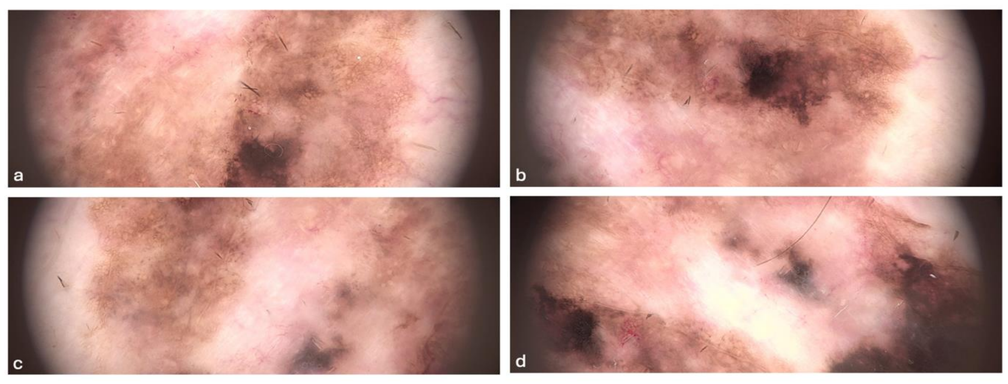

Fig. 2: Dermoscopic images of a melanoma: reference image (a), 90° rotation (b), 180° rotation (c) and 270° rotation (d).

Patients and Methods

We acquired 2,744 dermoscopic images of 385 skin lesions (80.8% benign, 19.2% malignant) at the Department of Dermatology, University Hospital Heidelberg. Sequential dermoscopic images were acquired using a digital videodermatoscope (ATBM Master, FotoFinder) under standardized conditions with naturally occurring variations in zoom (30×, 40×), rotation (90° increments), and repeated image acquisition. Sequential images of identical lesions were classified by a binary CNN (Moleanalyzer-Pro, FotoFinder Systems, Germany). Malignant lesions were confirmed histopathologically, whereas benign lesions were diagnosed by expert consensus and ≥12 months of clinical follow-up. The variability of scores was investigated using intraclass correlation coefficient (ICC), mean absolute change of scores (mac), and probability of change of predicted class.

Patient characteristics	*n*	*Lesion characteristics*	*n*
Total no. of patients	233	Total no. of lesions	385
F:M	135:98	Melanocytic:non-melanocytic lesions	334:51
Median age (range)	55 (6-92)	Benign:malignant	311:74
		Histological: expert consensus assessment	179:206

	Melanocytic lesions, n	*Non-melanocytic lesions, n*
Total no. of lesions	334	51
Benign:malignant	296:38	15:36
Mean tumor thickness (range)	1,6 mm (0,2-8 mm)	1,3 mm (0,1-3,6 mm)
Histological:expert consensus assessment	138:196	13:38

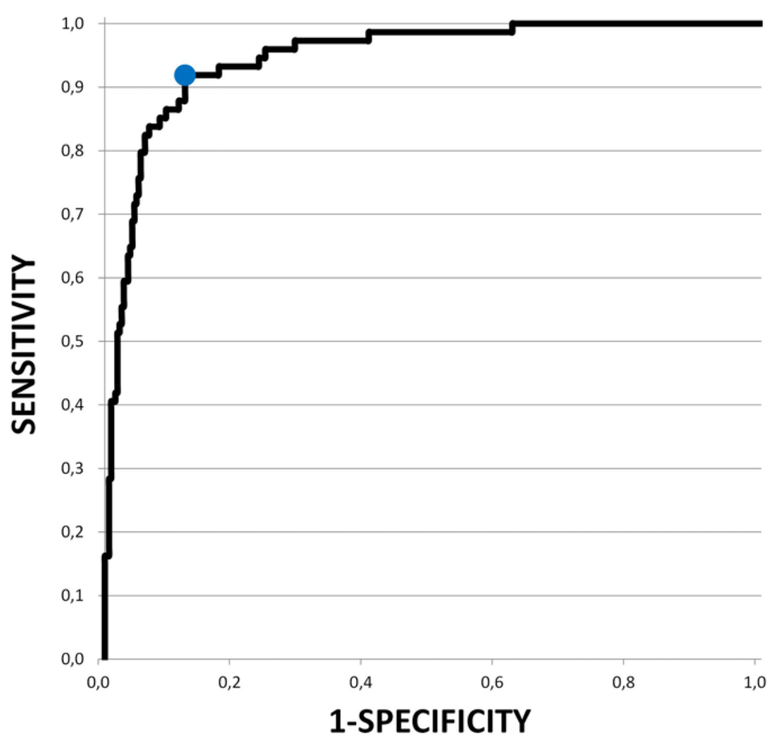

Fig. 3: ROC-AUC based on all baseline images (n = 385).

Results

On baseline dermoscopic images (n = 385), the CNN achieved a sensitivity of 91.9% (95% CI, 83.4–96.2%), a specificity of 87.8% (95% CI, 83.7–91.0%), and an AUROC of 0.947 (95% CI, 0.921–0.972). Across sequential images of identical lesions, the CNN demonstrated excellent reliability with an ICC of 0.872 (95% CI, 0.862–0.883). The mean absolute change in malignancy scores was 0.102 (95% CI, 0.090–0.115), while the probability of a change in the predicted class was 7.5% (95% CI, 5.8–9.2%), corresponding to approximately one clinically relevant classification change in every 13 sequential images. Reliability remained consistent across image rotations and zoom levels, with the highest agreement observed for repeated image acquisition.

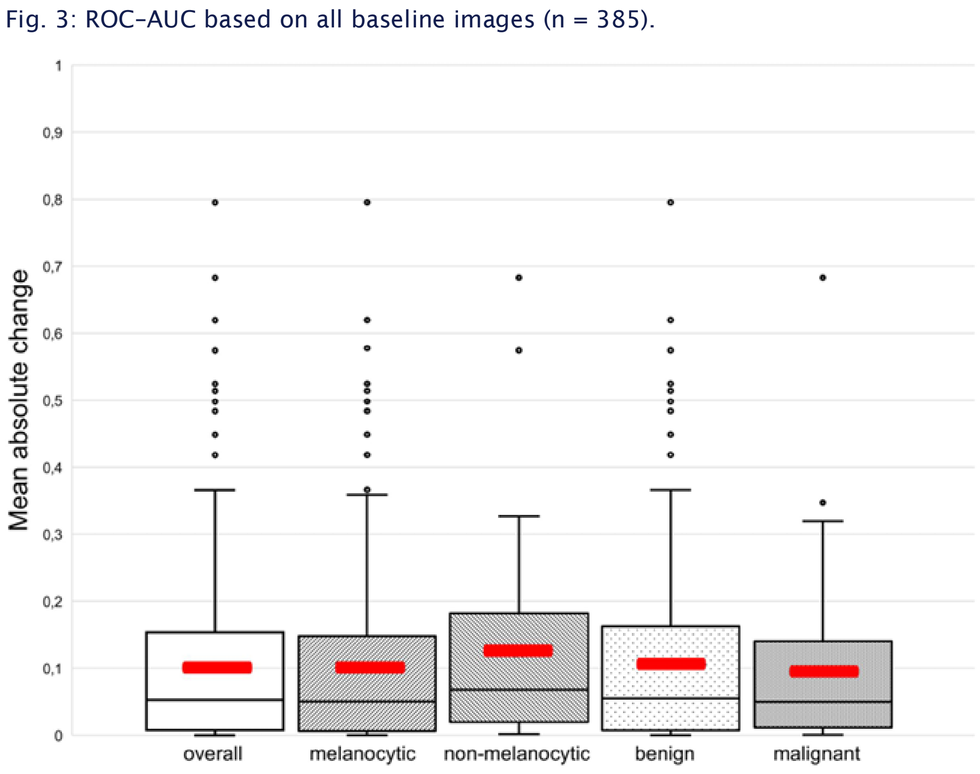

Fig. 4: Mean absolute change across all image variations.

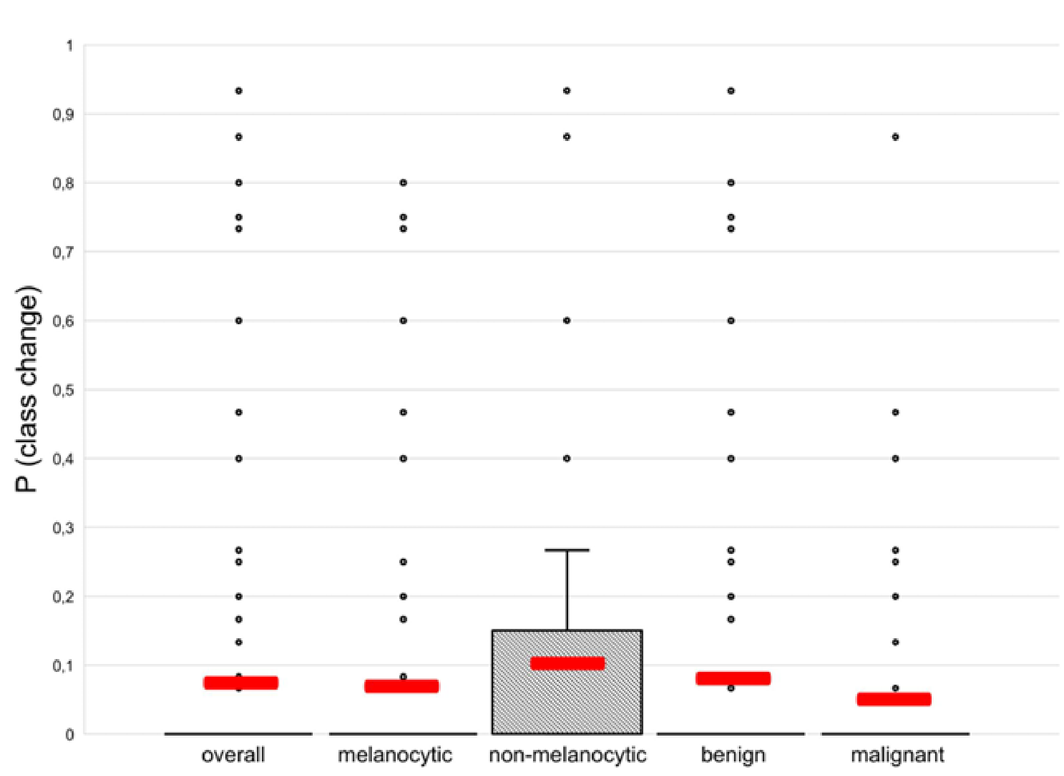

Fig. 5: Mean absolute change across all image variations.

Conclusion

The evaluated CNN demonstrated high robustness against naturally occurring variations in dermoscopic image acquisition, as reflected by a high ICC and only minor changes in malignancy scores.

Nevertheless, even subtle differences in image acquisition may occasionally result in changes of the predicted diagnostic class. These findings highlight that, despite their high reliability, CNN-based diagnostic tools should be used as decision-support systems rather than standalone diagnostic instruments, and their predictions should always be interpreted in the context of clinical assessment.

Fig. 6: Representative examples of a class change in repeated images