ORIGINAL ARTICLE

Assessment of melanoma thickness based on dermoscopy images: an open, web-based, international, diagnostic study

S. Polesie,[1,2,*] M. Gillstedt,[1,2] H. Kittler,[3] C. Rinner,[4] P. Tschandl,[3] J. Paoli[1 2]

[1]Department of Dermatology and Venereology, Institute of Clinical Sciences, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[2]Department of Dermatology and Venereology, Region Västra Götaland, Sahlgrenska University Hospital, Gothenburg, Sweden
[3]Department of Dermatology, Medical University of Vienna, Vienna, Austria
[4]Center of Medical Statistics, Informatics and Intelligent Systems (CeMSIIS), Medical University of Vienna, Vienna, Austria
*Correspondence: S. Polesie. E-mail: sam.polesie@vgregion.se

Abstract

Background Preoperative assessment of whether a melanoma is invasive or *in situ* (MIS) is a common task that might have important implications for triage, prognosis and the selection of surgical margins. Several dermoscopic features suggestive of melanoma have been described, but only a few of these are useful in differentiating MIS from invasive melanoma.

Objective The primary aim of this study was to evaluate how accurately a large number of international readers, individually as well as collectively, were able to discriminate between MIS and invasive melanomas as well as estimate the Breslow thickness of invasive melanomas based on dermoscopy images. The secondary aim was to compare the accuracy of two machine learning convolutional neural networks (CNNs) and the collective reader response.

Methods We conducted an open, web-based, international, diagnostic reader study using an online platform. The online challenge opened on 10 May 2021 and closed on 19 July 2021 (71 days) and was advertised through several social media channels. The investigation included, 1456 dermoscopy images of melanomas (788 MIS; 474 melanomas ≤1.0 mm and 194 >1.0 mm). A test set comprising 277 MIS and 246 invasive melanomas was used to compare readers and CNNs.

Results We analysed 22 314 readings by 438 international readers. The overall accuracy (95% confidence interval) for melanoma thickness was 56.4% (55.7%–57.0%), 63.4% (62.5%–64.2%) for MIS and 71.0% (70.3%–72.1%) for invasive melanoma. Readers accurately predicted the thickness in 85.9% (85.4%–86.4%) of melanomas ≤1.0 mm (including MIS) and in 70.8% (69.2%–72.5%) of melanomas >1.0 mm. The reader collective outperformed a *de novo* CNN but not a pretrained CNN in differentiating MIS from invasive melanoma.

Conclusions Using dermoscopy images, readers and CNNs predict melanoma thickness with fair to moderate accuracy. Readers most accurately discriminated between thin (≤1.0 mm including MIS) and thick melanomas (>1.0 mm).

Received: 13 January 2022; Accepted: 14 June 2022

Conflict of Interest

PT reports grants from MetaOptima and Lilly; consulting fees from Silverchair; and speaker honoraria from Lilly, Novartis and FotoFinder, all outside the submitted work. SP, MG, HK, CR and JP have no conflict of interest to declare.

Funding sources

The platform DermaChallenge is supported by independent educational sponsorship from Lilly, Indianapolis, IL, USA.

Introduction

Dermoscopy (dermatoscopy) is more accurate than naked eye examination in the diagnosis of cutaneous melanoma.[1] Several dermoscopic features suggestive of melanoma have been described,[2, 3] but only a few of these are useful in differentiating melanoma *in situ* (MIS) from invasive melanoma. Preoperative assessment of whether a melanoma is invasive or MIS is a common task that might have important implications for triage, prognosis and the selection of surgical margins.[4]

The dermoscopy images included in the manuscript (i.e. Figs S1 and S2) are deidentified and are published in agreement with the Regional Ethical Review Board in Gothenburg (approval number 283–18).

Previous studies in this field focused on the description of dermoscopy criteria that differentiate between invasive melanoma and MIS,[2, 3] but the accuracy of this approach has not been assessed in a larger setting involving multiple readers. Recently, gamification and crowdsourcing have proven useful to harness data for the assessment of skin tumours from multiple readers all over the world. The use of social media along with an increasing digitalization have facilitated the distribution of images for conducting online investigations, allowing researchers to reach numerous readers with varying backgrounds and levels of experience.

The aims of this study were to evaluate how accurately a large number of international readers, individually as well as collectively, can discriminate between MIS and invasive melanomas as well as estimate the Breslow thickness of invasive melanomas based on dermoscopy images. Finally, we aimed to compare the accuracy of two machine learning (ML) algorithms and the collective reader response.

Material and methods

We conducted an open, web-based, international, diagnostic reader study using the online platform DermaChallenge (Dermonaut, Medical University of Vienna, Vienna, Austria).[5, 6] The investigation adhered to the Standards for Reporting Diagnostic Accuracy (STARD).[7] We included dermoscopic images obtained at the department of Dermatology at Sahlgrenska University Hospital (Gothenburg, Sweden) between 2016 and 2020. All images depicted melanomas that were diagnosed by a dermatopathologist. The grossing procedure for melanomas at the pathology department of the Sahlgrenska University Hospital routinely includes sections (3–4 mm thickness) from the entire lesion. For each section, two slides (3 µm thickness) are obtained. Original resolution of the dermoscopic images ranged from 1600 × 1200 to 4416 × 3312 pixels. All images were standardized and transformed to 600 × 450 pixels using a previously described algorithm.[8] The images were not manually curated allowing imperfections such as light reflections and other artefacts including surgical skin markings. In total, 1456 images were included. Each challenge consisted of a random mix of 20 images with no pre-set distribution. For each image, the reader could choose one of the following three categories: 'melanoma *in situ*', 'melanoma ≤1.0 mm' or 'melanoma >1.0 mm'. The 1.0 mm thickness was selected since it is the cut-off for sentinel node biopsy in Sweden.[9] To restrict the assessment to the dermoscopy image, no clinical close-up image or other metadata was provided.

Before the challenge started, readers were provided with a short introduction on dermoscopic findings suggestive of MIS and invasive melanomas (Appendix S1). The challenge opened on 10 May 2021 and closed on 19 July 2021 (71 days). It was advertised through several social media channels including a private Facebook (Menlo Park, CA, USA) group with approximately 27 000 members (January 2022) called 'Dermatoscopy', which is hosted by members of the International Dermoscopy Society (IDS). Moreover, a newsletter notification was sent out to IDS members on 19 May 2021 (Appendix S2). Readers participating in at least three challenges were included in a lottery draw with three readers winning a dermoscopy textbook.

For the first three challenges, the readers were only presented with their final score at the end of each challenge. If the user participated >3 times, feedback was given after assessing each image along with a final score. In addition, we evaluated two convolutional neural networks (CNNs) on a randomized subset of lesions and compared it to the collective reader response for rounds 1 to 4. The first ML algorithm was a *de novo* CNN (model with no pretrained parameters) used in a previous publication by Gillstedt *et al.*[10] and the second was a fine-tuned pretrained CNN based on the ResNet-50 model.[11] A detailed description of the models are presented in Appendices S3 and S4.

Statistical analysis

All data were analysed using R.[12] In the primary analysis, we examined the proportions of correct predictions for the three included classes: MIS, melanoma ≤1.0 mm and melanoma >1.0 mm. To get the collective vote for each image, the outputs for eight readers were randomly selected (sampling with replacement) 100 times to generate a bootstrapped data set. For the second analysis, the accuracy rates for determining whether a melanoma was invasive (regardless of thickness) or MIS was compared. In a third analysis, the accuracy rates for classifying the melanomas as *in situ* or ≤1.0 mm combined vs. melanoma >1.0 mm were calculated. The readers could take the challenge an unlimited number of times. To avoid recall bias, only the first six rounds for each user were included in the analysis (i.e. max 120 single image evaluations per reader). If a user aborted a challenge before completion, all valid evaluations were considered for that specific round. Fisher's exact test was used for comparing proportions. Wilcoxon rank sum test and Kruskal–Wallis test were used for comparisons of accuracy rates between 2 and >2 groups respectively. DeLong's test for two correlated receiver operating characteristic (ROC) curves was used to compare the area under the ROC curve (AUC) between the readers and the two CNN models. All *P*-values were adjusted (P_{adj}) for multiple comparisons using the Holm method.[13] All tests were two-sided and $P_{adj} < 0.05$ was considered to be statistically significant.

Results

Of 1456 melanomas, 788 were *in situ* (54.1%) and 668 were invasive (45.9%). Among the invasive melanomas, 474 (71.0%) and 194 (29.0%) had a Breslow thickness ≤1.0 and >1.0 mm respectively. The median Breslow thickness of invasive melanomas was 0.7 mm (interquartile range [IQR] 0.5–1.2 mm). With regard to anatomic site, 715 melanomas were located on the trunk, 603 on the extremities and 138 in the head and neck region. The proportion of MIS was higher on the trunk (60.6%)

compared to the extremities (48.1%) and the head and neck region (47.1%; $P_{adj} < 0.0001$).

We collected 86 562 ratings in total. For this study, we only selected the 22 314 valid readings from the first six challenges of each of the 438 readers (65.3% females, $n = 286$) from 63 countries. The majority of readers were board-certified dermatologists (53.0%, $n = 232$). The remaining readers consisted of dermatology residents (31.3%, $n = 137$), general practitioners (12.1%, $n = 53$) and others (3.7%, $n = 16$; Table 1). The median number of readings per lesion was 15 (range 5–28), the number of completed answers per reader ranged from 4 to 120 (median 40) and the accuracy rates per reader ranged from 10% to 90% (median 55.5%, IQR 48.8%–64.9%). For each round, the median score was 11 (IQR 9–13) of 20. The maximum score received for one round was 19, which was reached by one reader on one occasion.

MIS vs. melanoma ≤1.0 mm vs. melanoma >1.0 mm

The correct answer with respect to the three output categories above was given in 12 581 instances, yielding an overall accuracy of 56.4% (95% confidence interval [CI], 55.7%–57.0%; Table 2).

Table 1 Distribution of profession and experience among readers

	Level of dermoscopy experience (years)			
Profession	**<5 years**	**≥5, <10 years**	**≥10 years**	**Total**
Board-certified dermatologist	94 (41%)	95 (41%)	43 (19%)	232 (100%)
Dermatology resident	112 (82%)	16 (12%)	9 (7%)	137 (100%)
General practitioner	35 (66%)	16 (30%)	2 (4%)	53 (100%)
Other*	13 (81%)	3 (19%)	0 (0%)	16 (100%)
Total	254 (58%)	130 (30%)	54 (12%)	438 (100%)

*'Other' included the following categories: medical specialist; medical student; non-medical; nurse practitioner and non-Dermatology resident.

The mean accuracy among readers with longer experience in dermoscopy (i.e. ≥5 years of dermoscopy experience, $n = 188$) was 56.6% (95% CI 54.7%–58.4%), which was not significantly better than those with shorter experience ($n = 254$; 54.0%; 95% CI 52.4%–55.6%; $P_{adj} = 0.35$). Board-certified dermatologists performed on par with resident dermatologists, general practitioners and other participants ($P_{adj} = 1$; Fig. 1).

Accuracy rates did not depend on body site ($P_{adj} = 1$). Using the majority vote for each lesion, 974 of the 1456 lesions (66.9%, 95% CI 64.4%–69.3%) were correctly classified. The accuracy obtained during the first four rounds (19 205 lesion evaluations) were not significantly inferior to the evaluations received in rounds five to six (3109 lesion evaluations; 56.1% vs. 58.3%; $P_{adj} = 0.13$). When comparing the collective response to the individual readers, the former achieved a higher sensitivity for all three classes (Table 3).

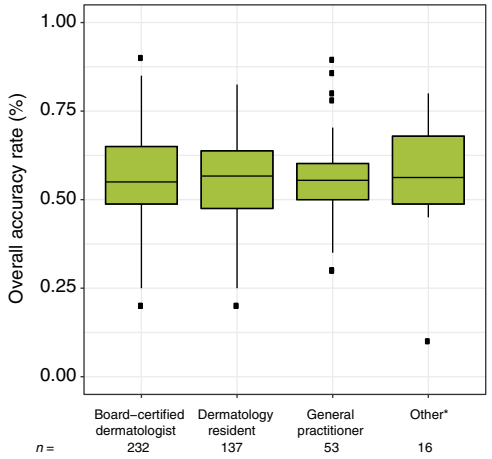

Figure 1 Accuracy rates among the different professions. *'Other' included the following categories: medical specialist; medical student; non-medical; nurse practitioner and non-Dermatology resident.

Table 2 Confusion matrices

	Histopathological diagnosis			
	MIS	Invasive melanoma		
Reader assessment	**MIS**	**Melanoma ≤1.0 mm**	**Melanoma >1.0 mm**	**Total**
MIS	7655 (63%)	2702 (37%)	245 (8%)	10 602 (48%)
Melanoma ≤1.0 mm	3443 (28%)	2827 (39%)	619 (21%)	6889 (31%)
Melanoma >1.0 mm	985 (8%)	1739 (24%)	2099 (71%)	4823 (22%)
MIS	7655 (63%)	2947 (29%)		10 602 (48%)
Invasive melanoma	4428 (37%)	7284 (71%)		11 712 (52%)
MIS or melanoma ≤1.0 mm combined	16 627 (86%)		864 (29%)	17 491 (78%)
Melanoma >1.0 mm	2724 (14%)		2099 (71%)	4823 (22%)
Total	12 083 (100%)	7268 (100%)	2963 (100%)	22 314 (100%)

MIS, Melanoma *in situ*.

Table 3 Comparison between the collective response and the individual reader

		Sensitivity (95% CI)	Specificity (95% CI)	Negative predictive value (95% CI)	Positive predictive value (95% CI)
Collective response (Bootstrapping)	MIS	71.9% (71.6–72.2)	69.8% (69.5–70.2)	67.8% (67.5–68.2)	73.8% (73.5–74.1)
	Melanoma ≤1.0 mm	39.0% (38.6–39.5)	78.5% (78.3–78.8)	72.7% (72.5–73.0)	46.7% (46.2–47.2)
	Melanoma >1.0 mm	78.6% (78.0–79.2)	89.0% (88.8–89.1)	96.4% (96.3–96.5)	52.3% (51.7–52.9)
Individual reader	MIS	63.4% (62.5–64.2)	71.2% (70.3–72.1)	62.2% (61.3–63.1)	72.2% (71.3–73.1)
	Melanoma ≤1.0 mm	38.9% (37.8–40.0)	73.0% (72.3–73.7)	71.2% (70.5–71.9)	41.0% (39.9–42.2)
	Melanoma >1.0 mm	70.8% (69.2–72.5)	85.9% (85.4–86.4)	95.1% (94.7–95.4)	43.5% (42.1–44.9)

CI, confidence interval; MIS, Melanoma *in situ*.

Thirty-two lesions exhibited the highest discrepancy in output (i.e. most frequent evaluation was MIS or melanoma >1.0 mm but the histopathological diagnosis was melanoma >1.0 mm and MIS respectively). Among these, 21 were MIS but were considered to be melanoma >1.0 mm by most readers. Contrarily, 11 lesions were considered as MIS by most readers but were melanoma >1.0 mm (Figs S1 and S2). In a *post hoc* analysis, the dermoscopic features of the 32 lesions above was performed (Table S1).

Invasive vs. MIS

When combining the two invasive melanoma groups (melanoma ≤1.0 mm and melanoma >1.0 mm), the overall accuracy rate for correctly classifying MIS and invasive melanomas was 63.4% (95% CI 62.5%–64.2%) and 71.0% (95% CI 70.3%–72.1%) respectively (Table 2). Applying the majority vote, 575 of 788 MIS lesions (73.0%; 95% CI 69.7%–76.0%) and 503 of 668 invasive melanomas (75.3%; 95% CI 71.8%–78.5%; $P_{adj} = 1$) were classified correctly.

Thin vs. thick melanomas

When thin melanomas (MIS and melanoma ≤1.0 mm combined, $n = 1268$) were compared to thick melanomas (melanoma >1.0 mm, $n = 194$), the overall accuracy rate was 85.9% (95% CI 85.4%–86.4%) and 70.8% (95% CI 69.2%–72.5%) respectively (Table 2). Using majority voting, 1179 of the 1268 *thin melanomas* (93.4%; 95% CI 91.9%–94.7%) and 156 of 194 thick melanomas (80.4%; 95% CI 74.1%–85.8%; $P_{adj} < 0.0001$) were classified correctly.

CNNs vs. readers' collective response

The two CNNs were evaluated on a test set comprising 523 lesions (277 MIS [53.0%] and 246 invasive melanomas [47.0%]). Among the invasive melanomas, 155 (63.0%) and 91 (37.0%) had a Breslow thickness ≤1.0 and >1.0 mm respectively. The median Breslow thickness among the invasive melanomas was 0.8 mm (IQR 0.5–1.3 mm). In terms of differentiating between invasive and MIS, the AUC for the *de novo* and pretrained CNN were 0.80 (95% CI 0.76–0.84) and 0.83 (95% CI 0.80–0.87, $P_{adj} = 0.35$) respectively. The individual readers' combined AUC was 0.85 (95% CI 0.81–0.88) which significantly outperformed the *de novo* CNN ($P_{adj} = 0.021$), but not the pretrained CNN ($P_{adj} = 1$; Figs 2 and S3).

Discussion

This study demonstrates that readers are able to discriminate between MIS, and thin (<1.0 mm) and thick (>1.0 mm) invasive melanomas with fair to moderate accuracy. Readers are more accurate when differentiating thick invasive melanomas from thin melanomas (MIS and thin invasive melanoma combined) and less accurate when they need to differentiate all three categories. Our data suggest that it is especially difficult to differentiate MIS from invasive melanomas on dermoscopy images alone. Interestingly, dermoscopy experience and profession had no significant impact for this particular classification problem. Furthermore, we show that deep learning does not outperform human readers in this task indicating a fair amount of objective ignorance.

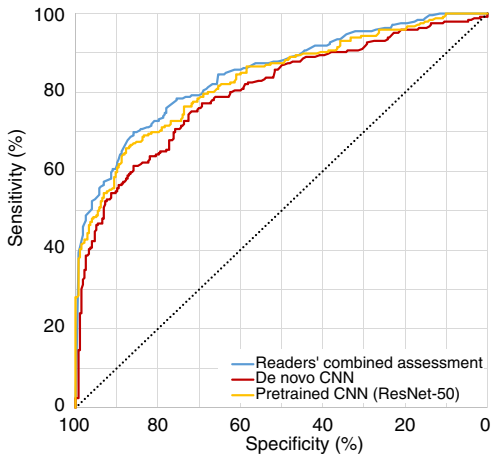

Figure 2 Performance of machine learning algorithms and readers. A *de novo* and a pretrained machine learning algorithm based on the ResNet-50 model were evaluated on a test set of 523 lesions (277 melanoma *in situ* and 246 invasive melanomas). The combined accuracy rates among the readers for these images are also displayed.

Over 20 ago, research groups focused on evaluation of melanoma thickness based on dermoscopic features.[14–16] While specific features that have a discriminatory power in assessment of melanoma thickness are important to identify, their usefulness is limited by the interobserver agreement of their presence. In two online consensus reports on dermoscopic features, the interobserver agreement only ranged from poor to fair.[17, 18] In a recent investigation, Polesie *et al.*[19] examined the interobserver agreement on a predefined set of 15 dermoscopic features stemming from the revised two-step algorithm[20] among seven dermatologists. The study included 182 melanomas (101 MIS and 81 invasive melanomas). Only two features, shiny white lines and atypical blue–white structures, exhibited moderate to substantial interobserver agreement. These dermoscopic features were also associated with melanomas >1.0 mm. Furthermore, regression/peppering was the only feature associated with thinner lesions (i.e. MIS and melanoma ≤1.0 mm combined).

Several ML models have previously been employed to predict melanoma thickness.[10, 21, 22] Using an image database consisting of 250 dermoscopic melanoma images (64 MIS; 103 melanomas <0.76 mm; 54 melanomas 0.76–1.5 mm and 29 melanomas >1.5 mm), Sáez *et al.* extracted features of the following: shape, colour, pigment network and texture. Specifically, logistic regression using initial variables and a product units model outperformed other tested models and achieved an accuracy level of 77.6% when discriminating between melanomas with a thickness over or under 0.76 mm.[21] Using the same dermoscopic data set as above, Jaworek-Korjakowska *et al.* used the pretrained VGG-19 ML model to assess melanoma thickness. Remarkably, the diagnostic accuracy rate in classifying melanomas in the three classes (i.e. MIS and melanoma <0.76 mm; melanoma 0.76–1.5 mm and melanoma >1.5 mm) was 87.2%.[22]

Our investigation has some important limitations. Fundamentally, the artificial set-up needs particular attention. This was a retrospective and academic investigation performed in an online setting where the decisions had no impact on care. As such, the set-up is contrived and has limited external validity for routine health care. In a real-life setting, physicians have access to relevant patient metadata, are able to view the lesion without dermoscopy and may touch the lesion, which might result in a better prediction of melanoma thickness. This is further supported by Carli *et al.* who demonstrated that *in vivo* dermoscopy, (i.e. combined clinical and dermoscopic examination), is more reliable than dermoscopy on photographic slides.[23]

Another limitation is that the readers had to choose one of the three predefined categories and could not opt for the possibility of no answer when their uncertainty was too high. Moreover, we placed no emphasis on misclassification in one direction or the other, which does not reflect clinical practice. The thresholds and management decisions must be better defined in a prospective clinical trial. The goal is of course not to delay surgery. Instead, we can provide better preoperative information to the patient as the main point. While all lesions included were diagnosed by a dermatopathologist, assessment of whether the melanoma is invasive or not and measuring the tumour thickness is a challenging endeavour.[24] At the pathology department of the Sahlgrenska University Hospital, where the cases were analysed, demanding cases are usually discussed in a team to reach a consensus agreement. However, it is possible that a consensus evaluation by an external expert team of dermatopathologists would have yielded a somewhat different final diagnosis or Breslow thickness measurement. Among all responders (n = 438), only 49 (11.2%) were board-certified dermatologists with >10 years-experience. This means that the reader-set could be a bit homogeneous and with an overall low level of expertise. Since the challenge was advertised on online social media and through an open email invitation to IDS members, the response rate could not be obtained. It is also probable that most readers had a special interest in skin tumour diagnosis and were highly motivated, which may have affected the results. Nevertheless, the participating readers had varying degrees of dermoscopy experience and this did not impact their results significantly. Furthermore, even though a set of dermoscopic features generally considered suggestive of melanoma thickness were presented in the challenge introductory text, the interobserver agreement of these features was beyond the scope of this investigation.

Finally, we used 1.0 mm Breslow thickness as a discriminator between thin and thick invasive melanomas. The rationale behind this is that this tumour thickness represents the cut-off for sentinel node biopsy in Sweden. We acknowledge that recommendations on when to perform sentinel node biopsy can vary between countries.

In a subsequent investigation, we intend to organize another study that includes both a dermoscopic and a clinical close-up image for each included lesion. It is unclear whether the inclusion of a clinical close-up image will improve the accuracy rates since little is known about how readers weigh each image modality during their assessments. In future investigations, it would also be appealing to let readers systematically explain and annotate the presence of selected predefined dermoscopic criteria. It remains to be determined whether or not dermoscopy is accurate enough for preoperative triage, prognosis and optimal surgical margin selection for atypical melanocytic lesions requiring excision. Such assessments may depend on the confidence levels of the dermoscopist in specific cases.

Implementation of CNN models that can help physicians in an everyday clinical setting is still pending. Nonetheless, the results presented here pinpoints an interesting future application. As such, it would be interesting to investigate if readers assisted by the outputs of CNNs can learn and improve their accuracy rates as previously demonstrated by Tschandl *et al.*[8] However, at this stage, the ML algorithms presented in this manuscript must only be considered as an academic undertaking and are far from ready to be implemented in clinical practice.

Clearly, they will have to be evaluated in prospective clinical trials adhering to the new Medical Device Regulation (Regulation EU 2017/745) before we can approach routine health care.[25]

To summarize, our investigation underlines the inherent difficulties in correct assessment of melanoma thickness. Nonetheless, the accuracy rates were higher when discriminating between thin and thick melanomas, whereas the identification of MIS was less reliable. The wisdom of the crowd outperformed individual readers, but the diagnostic accuracy for the prediction of melanoma thickness did not correlate with the professional background nor with dermoscopy experience.

Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1 Vestergaard ME, Macaskill P, Holt PE, Menzies SW. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting. *Br J Dermatol* 2008; **159**: 669–676.

2 Lallas A, Longo C, Manfredini M *et al*. Accuracy of Dermoscopic criteria for the diagnosis of melanoma in situ. *JAMA Dermatol* 2018; **154**: 414–419.

3 Silva VP, Ikino JK, Sens MM, Nunes DH, Di Giunta G. Dermoscopic features of thin melanomas: a comparative study of melanoma in situ and invasive melanomas smaller than or equal to 1 mm. *An Bras Dermatol* 2013; **88**: 712–717.

4 Polesie S, Jergeus E, Gillstedt M *et al*. Can Dermoscopy be used to predict if a melanoma is in situ or invasive? *Dermatol Pract Concept* 2021; **11**: e2021079.

5 Medical University of Vienna Department of Dermatology. DermaChallenge. 2021. Available at: https://dermachallenge.meduniwien.ac.at [Last accessed October 5, 2021].

6 Rinner C, Kittler H, Rosendahl C, Tschandl P. Analysis of collective human intelligence for diagnosis of pigmented skin lesions harnessed by gamification via a web-based training platform: simulation reader study. *J Med Internet Res* 2020; **22**: e15597.

7 Bossuyt PM, Reitsma JB, Bruns DE *et al*. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015; **351**: h5527.

8 Tschandl P, Rinner C, Apalla Z *et al*. Human-computer collaboration for skin cancer recognition. *Nat Med* 2020; **26**: 1229–1234.

9 Swedish guidelines for malignant melanoma [*In Swedish*], Version 6.0, Last updated: December 14, 2021. Available at: https://kunskapsbanken.cancercentrum.se/globalassets/cancerdiagnoser/hud/vardprogram/nationellt-vardprogram-malignt-melanom.pdf.

10 Gillstedt M, Hedlund E, Paoli J, Polesie S. Discrimination between invasive and in situ melanomas using a convolutional neural network. *J Am Acad Dermatol* 2022; **86**: 647–649.

11 He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ 2016: 770–778.

12 R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2021, Available at:. https://www.R-project.org/

13 Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat* 1979; **6**: 65–70.

14 Stante M, De Giorgi V, Cappugi P, Giannotti B, Carli P. Non-invasive analysis of melanoma thickness by means of dermoscopy: a retrospective study. *Melanoma Res* 2001; **11**: 147–152.

15 Carli P, de Giorgi V, Palli D, Giannotti V, Giannotti B. Preoperative assessment of melanoma thickness by ABCD score of dermatoscopy. *J Am Acad Dermatol* 2000; **43**: 459–466.

16 Argenziano G, Fabbrocini G, Carli P, De Giorgi V, Delfino M. Clinical and dermatoscopic criteria for the preoperative evaluation of cutaneous melanoma thickness. *J Am Acad Dermatol* 1999; **40**: 61–68.

17 Argenziano G, Soyer HP, Chimenti S *et al*. Dermoscopy of pigmented skin lesions: results of a consensus meeting via the internet. *J Am Acad Dermatol* 2003; **48**: 679–693.

18 Carrera C, Marchetti MA, Dusza SW *et al*. Validity and reliability of Dermoscopic criteria used to differentiate nevi from melanoma: a web-based international Dermoscopy society study. *JAMA Dermatol* 2016; **152**: 798–806.

19 Polesie S, Sundback L, Gillstedt M *et al*. Interobserver agreement on Dermoscopic features and their associations with in situ and invasive cutaneous melanomas. *Acta Derm Venereol* 2021; **101**: adv00570.

20 Marghoob AA, Braun R. Proposal for a revised 2-step algorithm for the classification of lesions of the skin using dermoscopy. *Arch Dermatol* 2010; **146**: 426–428.

21 Saez A, Sanchez-Monedero J, Gutierrez PA, Hervas-Martinez C. Machine learning methods for binary and multiclass classification of melanoma thickness from Dermoscopic images. *IEEE Trans Med Imaging* 2016; **35**: 1036–1045.

22 Jaworek-Korjakowska J, Kleczek P, Gorgon M. Melanoma Thickness Prediction Based on Convolutional Neural Network with VGG-19 Model Transfer Learning. IEEE, Piscataway, NJ, 2019.

23 Carli P, De Giorgi V, Argenziano G, Palli D, Giannotti B. Pre-operative diagnosis of pigmented skin lesions: in vivo dermoscopy performs better than dermoscopy on photographic images. *J Eur Acad Dermatol Venereol* 2002; **16**: 339–346.

24 Elmore JG, Barnhill RL, Elder DE *et al*. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. *BMJ* 2017; **357**: j2813.

25 Malvehy J, Ginsberg R, Sampietro-Colom L, Ficapal J, Combalia M, Svedenhag P. New regulation of medical devices in the EU: impact in dermatology. *J Eur Acad Dermatol Venereol* 2022; **36**: 360–364.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Figure S1 Lesions considered as invasive >1.0 mm by the majority of readers but were melanoma *in situ*.

Figure S2 Lesions considered as melanoma *in situ* by the majority of readers but were invasive >1.0 mm.

Table S1 Dermoscopic descriptions of Supplementary Figures 1 and 2.

Figure S3 Comparison between the two CNNs and readers' combined assessment.

Appendix S1 Text and images displayed to the readers before challenge initiation.

Appendix S2 Newsletter sent to members of International Dermoscopy society.

Appendix S3 *De novo* CNN.

Appendix S4 Pretrained CNN.

Appendix S5 Standards for Reporting of Diagnostic Accuracy (STARD) Checklist.