Depósito de Investigación
Universidad de Sevilla

Depósito de investigación de la Universidad de Sevilla

https://idus.us.es/

Esta es la versión aceptada del artículo publicado en:

This is a accepted manuscript of a paper published in:

Clinical and Experimental Dermatology (2023): July 2023

DOI: https://doi.org/10.1093/ced/llad107

Copyright: © The Author(s) 2023. Published by Oxford University Press on behalf of British Association of Dermatologists. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

El acceso a la versión publicada del artículo puede requerir la suscripción de la revista.

Access to the published version may require subscription.

"This is a pre-copyedited, author-produced version of an article accepted for publication in *Clinical and Experimental Dermatology* following peer review. The version of record Juan-Carlos Hernández-Rodríguez, Lourdes Durán-López, Juan P Domínguez-Morales, Juan Ortiz-Álvarez, Julián Conejo-Mir, Jose-Juan Pereyra-Rodriguez, Prediction of melanoma Breslow thickness using deep transfer learning algorithms*, Clinical and Experimental Dermatology, Volume 48, Issue 7, July 2023, Pages 752–758,*, is available online at: https://academic.oup.com/ced/article/48/7/752/7086611, https://doi.org/10.1093/ced/llad107".

Abbreviated Abstract

We aim with this original article to develop, validate, and compare three deep transfer learning algorithms to predict between in situ or invasive melanoma and < 0.8 or â‰¥ 0.8 millimetres of Breslow thickness.

For Peer Review

Title page

Title: Prediction of melanoma Breslow thickness using Deep Transfer Learning Algorithms

Running head: Deep Transfer Learning and melanoma

Wordcount: 1,998

Tables: 2

Figures: 4

Authors:

Juan-Carlos Hernández-Rodríguez, MD, MSc[1, 2]. https://orcid.org/0000-0003-2525-4069

Lourdes Durán-López, PhD[3]. https://orcid.org/0000-0002-5849-8003

Juan P. Domínguez-Morales, PhD[3]. https://orcid.org/0000-0002-5474-107X

Juan Ortiz-Álvarez, MD[1]. https://orcid.org/0000-0003-2134-3902

Julián Conejo-Mir, MD, PhD[1, 2]. https://orcid.org/0000-0001-9108-9538

Jose-Juan Pereyra-Rodriguez, MD, PhD[1, 2]. https://orcid.org/0000-0001-6843-5877

Author's information

1. Department of Dermatology, Virgen del Rocio University Hospital. 41013 Seville, Spain.
2. Department of Medicine, Faculty of Medicine, University of Seville, 41009, Seville, Spain.
3. Robotics and Technology of Computers Laboratory, University of Seville, 41012 Seville, Spain

Correspondence to:

Juan-Carlos Hernández-Rodríguez.

Department of Dermatology. Virgen del Rocio University Hospital.

Av. Manuel Siurot S/N. 41013 Seville, Spain.

For Peer Review

Email: j.carlos.her.rod@gmail.com; Phone: +34 955 013 166

Funding: Authors declared not receiving specific grant for this study from any public or private agency or not-for-profit sectors.

Conflict of interest: Authors of this paper declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval:

The Andalusian Review Board and Ethics Committee Virgen Macarena-Virgen del Rocio Hospitals approved the study protocol (ID 0096-N-20).

Bullet statements:

'What is already known about this topic?'

- Previous studies assessed the prediction of CNN for the comparison between in situ or invasive melanoma using ResNet50 and *de novo* CNN.

'What does this study add?'

- This study compared three DTL pretrained CNN and dermatologist performance predicting in situ versus invasive melanoma and < or ≥ 0.8 millimetres of Breslow thickness.
- DTL could be an ancillary aid to support dermatologists' decision in the near future.

For Peer Review

Abstract

Background: The distinction between in situ (MIS) or invasive melanoma is challenging even for expert dermatologists. The use of pretrained convolutional neural networks (CNNs) as ancillary decision systems needs further research.

Objective: To develop, validate and compare three deep transfer learning algorithms to predict between MIS or invasive melanoma and < or ≥ 0.8 millimetres of Breslow thickness (BT).

Methods: A dataset of 1,315 dermoscopic images of histopathologically confirmed melanomas was created from Virgen del Rocio University Hospital and open repositories of the ISIC archive and Polesie et al. The images were labelled as MIS or invasive melanoma and < or ≥ 0.8 millimetres of BT. We conducted three trainings, and overall means for ROC curves, sensitivity, specificity, positive and negative predictive value, and balanced diagnostic accuracy outcomes were evaluated on the test set with ResNetV2, EfficientNetB6, and InceptionV3. The results of ten dermatologists were compared with the algorithms. Grad-CAM gradient maps were generated, highlighting relevant areas considered by the CNNs within the images.

Results: EfficientNetB6 achieved the highest diagnostic accuracy for the comparison between MIS and invasive melanoma, and < 0.8 versus ≥ 0.8 of BT were 61% and 75%, respectively. For the latter, ResNetV2, with an area under the ROC curve of 0.76, and EfficientNetB6, of 0.79, outperformed the results obtained by the dermatologist group with 0.70.

Conclusions: EfficientNetB6 recorded the best prediction results, overcoming dermatologists for the comparison of 0.8 mm of BT. DTL could be an ancillary aid to support dermatologists' decision in the near future.

For Peer Review

Introduction

Cutaneous melanoma is responsible for almost 90% of skin cancer deaths, worsening the prognosis when diagnosis is delayed[1]. Breslow thickness (BT) is the main prognostic factor in primary cutaneous melanoma, which measures the microinvasion of the tumour in millimetres (mm) from the granular layer to the deepest of tumour invasion[2]. In addition to ulceration, BT sets out the T classification of 8thAJCC[1].

An accurate diagnosis of early melanoma is one of the major goals of dermoscopy[3], but distinction between in situ (MIS) and invasive melanoma could be challenging even for expert dermatologist[4].

For Peer Review

Deep learning methods are a novel approach for the diagnosis of melanoma, which uses convolutional neural networks (CNNs) to computationally analyse dermoscopic images[5]. One of the drawbacks of deep learning is the need for large amounts of training data to complex patterns within images. To address this issue, deep transfer learning (DTL) is a technique that allows training of CNNs with a lower amount of data, using previous learned model knowledge with minimum training or fine-tuning to perform a new task[6].

To discriminate between MIS and invasive melanoma, some authors created *de novo* CNNs[7–10], but only one pretrained ResNet50 CNN model was previously used for this comparison[10,11]. Human readers outperformed *de novo* CNNs, but not the pretrained CNN[10], so the scope of pretrained CNNs to differentiate MIS versus invasive melanoma should be further analysed.

To our knowledge, no studies compared different DTL approaches using pretrained CNNs. This study aimed to develop, validate, and compare three DTL algorithms to predict whether a melanoma is MIS or invasive and whether the BT is < 0.8 or ≥ 0.8 mm, based on dermoscopic images. Also, we aimed to compare the performance of human readers with that of DTL algorithms.

Material and methods

This retrospective study was performed according to the Declaration of Helsinki and the Standards for Reporting Diagnostic Accuracy (STARD). The Andalusian Review Board and Ethics Committee Virgen Macarena-Virgen del Rocio Hospitals approved the study protocol (ID 0096-N-20). The main dataset was composed of dermoscopic images of histopathologically confirmed melanomas in which BT was measured. To increase clinical relevance, we did not restrict the melanoma subtype or subjects' phototype[12].

Three independent subsets made up our dataset: (i) 1,055 images of 279 cases from the dermoscopic image repository of Virgen del Rocio University Hospital (Seville, Spain) between 2016 and 2022 (Supplemental material); (ii) 193 images of 184 cases from Polesie et al.[13] (iii) 67 images of 67 cases from the ISIC archive[14]. In the (i) and (ii) subsets, some cases had more than one image. These were discarded from the validation and test datasets, but not from the training dataset since it improves training performance. Upsampling was conducted to prevent a higher weight being conferred on cases with more than one image. Thus, we built synthetic copies up to equalling 10 images per case.

Image labelling

The ground truth was established by histopathological diagnosis. Each image was labelled as MIS or invasive melanoma, and < 0.8 or ≥ 0.8 mm of BT. We considered 0.8 mm of BT, as it is the threshold in Europe to perform a sentinel lymph node biopsy, when associating with additional histological risk factors as ulceration[15]. To assess CNNs prediction potential, we performed two comparisons, MIS (283 images) versus invasive melanoma (1,032 images) and < 0.8 (702 images) versus ≥ 0.8 mm (613 images) of BT.

For Peer Review

Image subsets and overfitting prevention

The 80% of the images were used for the training dataset, 10% for validation, and 10% for the test dataset. A cross-validation was not possible since some of our cases comprised more than one image. Instead, we performed the Dropout Regularization[16] and an external test dataset. Also, we conducted 3 trainings with the same test dataset, and the overall means for all outcomes were calculated.

CNN architecture

To train the different CNNs, we resized all images to an appropriate input size, using 299 x 299 pixels for ResNetV2 and InceptionV3, and 512 x 512 pixels for EfficientNetB6. To enhance the external validity to real clinical settings, we did not use any software to modify or curate the dermoscopic images.

As we used different numbers of images per class within each training dataset, the distribution of labelled images in each subset was imbalanced. To solve this, we performed the weight assignment with the *class_weight* function from Keras[17]. To optimize training performance, we established parameters like exponential learning rate scheduler, and callbacks to save model parameters and to perform an early stopping (*monitor='val_loss', patience=10*). To avoid overfitting, we used data augmentation with the following transformations: *rotation_range=40, width_shift_range=0.2, height_shift_range=0.2, horizontal_flip=True*.

Gradient maps

To make CNNs output comprehensible, we created gradient maps with the Gradient-weighted Class Activation Mapping (Grad-CAM)[18]. Grad-CAM is a technique of computer vision that remarks the region of interest of the input images that are relevant for the prediction of a CNN model.

For Peer Review

Statistical analysis

The R software (v.4.1.2) was used to perform all statistical analyses. The capacity to differentiate classes (MIS or invasive, and < or ≥ 0.8 mm of BT) inferred by the model was used to calculate the receiver operating characteristic (ROC) curves. To calculate the ROC curves with 95% confidence intervals, we used the package "nsROC" (v.1.1)[19]. Sensitivity, specificity, positive and negative predictive value, and balanced diagnostic accuracy outcomes were evaluated for each CNN. To examine the performance of the model, we compared the prediction outcomes obtained with those achieved by 5 board-certified dermatologists and 5 dermatology residents. All human readers independently performed predictions for the same test datasets evaluated by all the CNN models. To calculate the interobserver agreement we used the Fleiss' kappa index (k)[20].

Results

We compared 3 pretrained CNNs, which test set performance is shown in Table 1. The test set for the comparison between MIS and invasive melanoma (n = 111 images) consisted of 51 MIS and 59 invasive melanomas. The test set for the comparison between melanoma < or ≥ 0.8 BT (n = 86) was constituted by 55 and 31 cases, respectively.

For the MIS versus invasive melanoma prediction model, EfficientNetB6 presented the highest diagnostic accuracy (61%) and sensitivity (72%), whilst it performed the lowest specificity value (31%) of the three pretrained models. The mean area under the ROC curve (AUC) was 0.59, 0.63 and 0.54 for ResNetV2, InceptionV3 and EfficientNetB6, respectively (Figure 1). The readers outperformed the three models for this comparison, showing a diagnostic accuracy of 64% and an AUC of 0.64 (Table 2). Interobserver agreement of dermatologists for this comparison was moderate, k = 0.46.

For Peer Review

In the assessment of invasive melanomas ≥ 0.8 mm of BT, EfficientNetB6 achieved the highest diagnostic accuracy (84%) and specificity (84%). This CNN model recorded a sensitivity of 58%, being the lowest value compared to the 60% and 65% achieved by ResNetV2 and InceptionV3, respectively. The mean AUC was 0.76, 0.75 and 0.69 for ResNetV2, InceptionV3, and EfficientNetB6, respectively (Figure 2). For this comparison, the mean diagnostic accuracy achieved by the dermatologist was 69%, with an AUC of 0.70 (Table 2). Interobserver agreement of dermatologists was fair, k = 0.35. Radar charts illustrated the juxtaposition of the main results to compare the performance of each pretrained CNN model and dermatologists (Figure 3).

Gradient maps

Gradient maps spotlight influential areas of dermoscopic images, where the red colour remarks high attribution area for a specific prediction (Figure 4). Figure 4 remarks gradient map examples of a true positive (TP), true negative (TN), false positive (FP), and false negative (FN) dermoscopic image.

Discussion

This study analysed the performance of three DTL models for the prediction of microinvasion using melanoma dermoscopic images. EfficientNetB6 achieved the highest diagnostic accuracy for the comparison between MIS and invasive melanoma, and < 0.8 versus ≥ 0.8 of BT. For the latter, ResNetV2 and EfficientNetB6 outperformed the dermatologist group.

The state of the art related to the use of deep learning as a decision support system has grown in recent years. Most of the studies that focused on the differentiation between MIS and invasive, developed, and implemented a *de novo* CNN; only Polesie et al.[10] and Chu et al.[11] used a pretrained ResNet50 CNN. The first author tested CNN performance

For Peer Review

on a dataset of 523 dermoscopic images, achieving an AUC of 0.83. This result was not significantly outperformed by 438 international readers for the comparison between MIS and invasive, who reached an AUC of 0.85. The AUC results in this study were superior to ours in the three pretrained CNNs, but it should be noted that our dermatologists did not receive any baseline educational program prior to discrimination of dermoscopic images. This was conducted to avoid recall bias and to match the research context with daily clinical practice.

Chu et al.[11] used ResNet50 to differentiate between MIS and invasive and depth of microinvasion, but in acral lentiginous melanoma. Despite being in a different clinical setting, the CNN performed effectively to distinguish between < 0.8 mm and ≥ 0.8 mm of BT, with and AUC of 0.90 in 57 dermoscopic images. Similarly, we achieved favourable results to discriminate between MIS and invasive and for the level of microinvasion.

Regarding *de novo* CNNs, all studies recorded a fair to moderate accuracy for the distinction between MIS and invasive melanoma like the performance of our pretrained CNNs. *De* novo CNNs were not superior to dermatologists in the classification of MIS or invasive melanoma, whereas only pretrained CNNs outperformed them, as Polesie et al.[10] confirmed. This could help to achieve early detection and accurate stratification at the time of diagnosis. In our results, pretrained CNNs performed better for the prediction of melanomas ≥ 0,8 mm of BT, possibly related to the fact that most of the dermoscopic features between MIS and thin melanomas are overlapped[21].

In addition to similar performance as *de novo* CNNs and outperformance versus dermatologists to predict between MIS and invasive melanoma, further advantages of pretrained CNNs compared to *de novo* have been reported in the scientific literature. DTL uses open code pretrained CNNs that store the information images from other problems

For Peer Review

and utilize for a different target. Additionally, it presents better image feature extraction, lower sample size and shorter time for the learning process, making DTL an optimal tool to obtain quicker and more cost-effective results in healthcare[22].

Clinical and research implications

Our pretrained CNNs appears to be useful ancillary aids in selecting the optimal surgical approach and tumour staging in cutaneous melanoma based on the threshold of 0.8 mm BT. The implementation of CNNs in the dermatology field could help to support the triage and prioritize early cases. A recent study confirmed that most patients were open to CNN use in diagnosis, but always under dermatology supervision[23]. Within the research framework, the implementation of these CNN needs to be prospectively examined in the real-world clinical setting. Also, a call for action is needed for the standardisation of dermoscopic imaging that could lead to robust results of algorithms. We propose attaching the DICOM standard (supplement 221)[24], which suggests adding metadata to images, so the CNN could handle patient data to make the prediction as dermatologists make in their daily practice.

Limitations

A major pitfall of our dermoscopic image repository, found in the background literature, is that people with skin of colour were not involved. Thus, as Butt et al.[25] declared it is imperative to conduct future studies in this population, to avoid racial bias and to evaluate CNN performance. Due to the retrospective design of the study, patient metadata was not available to be included in the dermoscopic image.

Conclusions

This study showed a suitable prediction of three pretrained CNN for discrimination between MIS and invasive melanoma and for the distinction between a BT < 0.8 and $\geq$

For Peer Review

0.8 mm, using 1,315 dermoscopic images. For both models, EfficientNetB6 recorded the best prediction results, overcoming dermatologists for the comparison of 0.8 mm of BT. DTL could be an ancillary aid to support dermatologists' decision in the near future. Nevertheless, a standardisation on dermoscopic images is necessary to achieve the best output of CNN.

Acknowledgments

On behalf of all authors, we would like to thank all dermatology who voluntarily participated as readers of the dermoscopic images.

Data availability

The data underlying this article will be shared on reasonable request to the corresponding author.

References:

1. Garbe C, Amaral T, Peris K, Hauschild A, Arenberger P, Basset-Seguin N, et al. European consensus-based interdisciplinary guideline for melanoma. Part 1: Diagnostics: Update 2022. Eur J Cancer. 2022;170:236–55.

2. Breslow A. Thickness, Cross-Sectional Areas and Depth of Invasion in the Prognosis of Cutaneous Melanoma. Ann Surg. 1970;172(5):902–8.

3. Ring C, Cox N, Lee JB. Dermatoscopy. Clin Dermatol. 2021;39(4):635–42.

4. Polesie S, Jergéus E, Gillstedt M, Ceder H, Dahlén Gyllencreutz J, Fougelberg J, et al. Can Dermoscopy Be Used to Predict if a Melanoma Is In Situ or Invasive? Dermatol Pract Concept. 2021;2021079.

5. Aractingi S, Pellacani G. Computational neural network in melanocytic lesions diagnosis: artificial intelligence to improve diagnosis in dermatology? Eur J Dermatol. 2019;29(S1):4–7.

6. Sarker IH. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN Comput Sci. 2021;2(6):420.

For Peer Review

7. Polesie S, Gillstedt M, Ahlgren G, Ceder H, Dahlén Gyllencreutz J, Fougelberg J, et al. Discrimination Between Invasive and In Situ Melanomas Using Clinical Close-Up Images and a De Novo Convolutional Neural Network. Front Med. 2021;8.

8. 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(3):647–9.

9. Gillstedt M, Mannius L, Paoli J, Dahlén Gyllencreutz J, Fougelberg J, Johansson Backman E, et al. Evaluation of Melanoma Thickness with Clinical Close-up and Dermoscopic Images Using a Convolutional Neural Network. Acta Derm Venereol. 2022;102:adv00790.

10. Polesie S, Gillstedt M, Kittler H, Rinner C, Tschandl P, Paoli J. Assessment of melanoma thickness based on dermoscopy images: an open, web-based, international, diagnostic study. J Eur Acad Dermatology Venereol. 2022;36(11):2002–7.

11. Chu YS, Lee S, Lee SG, Chung KY, Roh MR, Yang S, et al. Deep Learning Algorithms for Predicting Breslow Thickness from Dermoscopic Images of Acral Lentiginous Melanomas. J Invest Dermatol. 2022;142(8):2268-2271.e2.

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

13. Polesie S, Jergéus E, Gillstedt M, Ceder H, Dahlén Gyllencreutz J, Fougelberg J, et al. Can Dermoscopy Be Used to Predict if a Melanoma Is In Situ or Invasive? Dermatol Pract Concept. 2021;e2021079.

14. International Skin Imaging Collaboration [Internet]. Available from: https://www.isic-archive.com/

15. Garbe C, Amaral T, Peris K, Hauschild A, Arenberger P, Basset-Seguin N, et al. European consensus-based interdisciplinary guideline for melanoma. Part 2: Treatment - Update 2022. Eur J Cancer. 2022;170:256–84.

16. Xiaobo LiangLijun WuJuntao LiYue WangQi MengTao QinWei ChenMin

For Peer Review

ZhangTie-Yan Liu. R-Drop: Regularized Dropout for Neural Networks. ArXiv. 2021;abs/2106:14448.

17. Chollet F. Imbalanced classification: credit card fraud detection [Internet]. Keras. 2020. Available from: https://keras.io/examples/structured_data/imbalanced_classification/

18. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Int J Comput Vis. 2020;128(2):336–59.

19. Pérez-Fernández S, Martínez-Camblor P, Filzmoser P, Corral N. nsROC: An R package for Non-Standard ROC Curve Analysis. R J [Internet]. 2019;10(2):55. Available from: https://journal.r-project.org/archive/2018/RJ-2018-043/index.html

20. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378–82.

21. Lallas A, Longo C, Manfredini M, Benati E, Babino G, Chinazzo C, et al. Accuracy of Dermoscopic Criteria for the Diagnosis of Melanoma In Situ. JAMA Dermatology. 2018;154(4):414.

22. Pan SJ, Yang Q. A Survey on Transfer Learning. IEEE Trans Knowl Data Eng. 2010;22(10):1345–59.

23. Lim K, Neal-Smith G, Mitchell C, Xerri J, Chuanromanee P. Perceptions of the use of artificial intelligence in the diagnosis of skin cancer: an outpatient survey. Clin Exp Dermatol. 2022;47(3):542–6.

24. Caffery LJ, Clunie D, Curiel-Lewandrowski C, Malvehy J, Soyer HP, Halpern AC. Transforming Dermatologic Imaging for the Digital Era: Metadata and Standards. J Digit Imaging. 2018;31(4):568–77.

25. Butt S, Butt H, Gnanappiragasam D. Unintentional consequences of artificial intelligence in dermatology for patients with skin of colour. Clin Exp Dermatol. 2021;46(7):1333–4.

For Peer Review

Supporting information:

The CNNs code can be found at https://github.com/juaherrod/UHVR_DERMOSCOPY_BRESLOW. The dataset of dermoscopic images from Virgen del Rocio University Hospital is available as online supplementary material.

Figure legends:

Figure 1. ROC curves for the prediction model between in situ and invasive melanoma for (a) ResNetV2, (b) InceptionV3 and (c) EfficientNetB6.

Figure 2. ROC curves for the prediction model between melanoma < 0.8 or ≥ 0.8 mm of Breslow thickness for (a) ResNetV2, (b) InceptionV3 and (c) EfficientNetB6.

Figure 3. Radar Chart for the comparison of the performance of the three pretrained CNN for (a) MIS vs invasive melanoma and (b) melanoma < 0.8 mm or ≥ 0.8 mm of BT.

Figure 4. Gradient maps. (a) True positive (TP): the algorithm identified a blue-white veil area as an important region of a melanoma ≥ 0.8 BT. (b) True negative (TN): a regular network area was identified by the algorithm to predict a melanoma < 0.8 mm BT. (c) False positive (FP): the algorithm focused on the healthy surrounding skin of a tiny nodular melanoma ≥ 0.8 mm. (d) False negative (FN): the algorithm was unable to focus on a specific dermoscopic structure to correctly perform the prediction in a melanoma $<$ 0.8 mm BT.

For Peer Review

Table 1. Comparison of the Performance of Convolutional Neural Networks for prediction of melanoma Breslow thickness

Comparison	Training dataset	Se	Sp	PPV	NPV	Accuracy	F1-score	AUC
In situ vs Invasive	ResNetV2	0.61 ± 0.15	0.50 ± 0.14	0.66 ± 0.15	0.45 ± 0.19	0.54 ± 0.02	0.61 ± 0.01	0.59 ± 0.03
	InceptionV3	0.63 ± 0.13	0.51 ± 0.12	0.68 ± 0.14	0.45 ± 0.14	0.56 ± 0.03	0.64 ± 0,04	0.63 ± 0.02
	EfficientNetB6	0.72 ± 0.03	0.39 ± 0.04	0.70 ± 0.04	0.41 ± 0.09	0.61 ± 0.02	0.71 ± 0.01	0.54 ± 0.06
< 0.8 mm vs ≥ 0.8 mm of BT	ResNetV2	0.60 ± 0.17	0.76 ± 0.08	0.57 ± 0.07	0.79 ± 0.11	0.70 ± 0.04	0.57 ± 0.04	0.76 ± 0.06
	InceptionV3	0.65 ± 0.22	0.70 ± 0.14	0.53 ± 0.17	0.79 ± 0.17	0.65 ± 0.04	0.54 ± 0.04	0.75 ± 0.01
	EfficientNetB6	0.58 ± 0.20	0.84 ± 0.06	0.61 ± 0.09	0.81 ± 0.10	0.75 ± 0.03	0.58 ± 0.09	0.69 ± 0.06

AUC, area under the ROC curve; NPV, negative predictive value; PPV, positive predictive value; Se, sensitivity; Sp, specificity

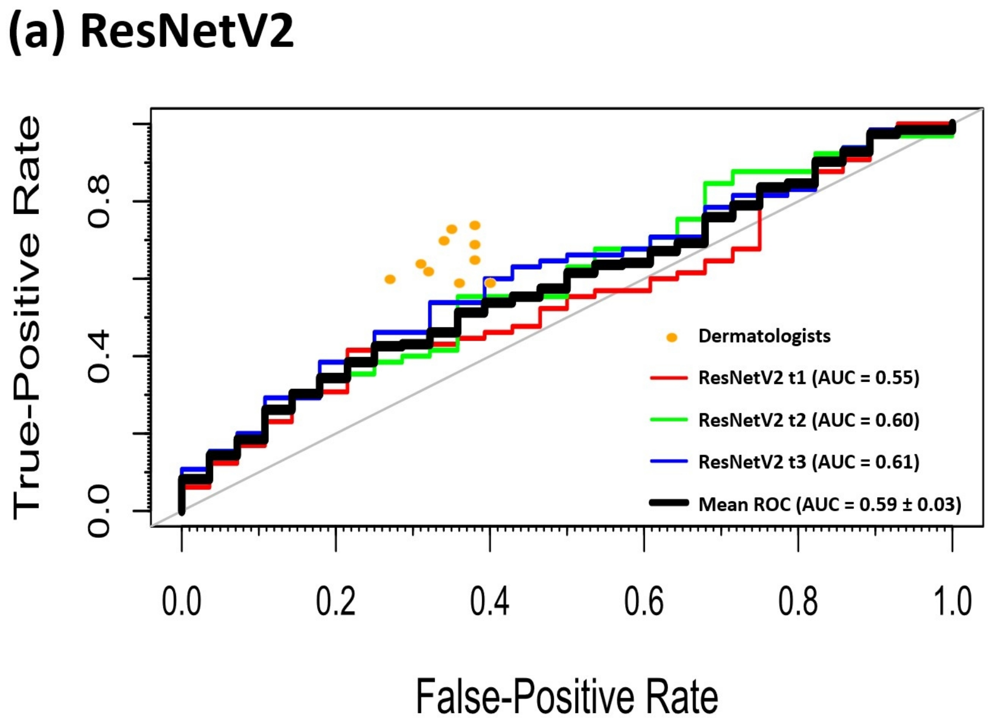
For Peer Review

Table 2. Dermatologist performance on dermoscopic images

Readers	AUC	95%CI	Accuracy	Se	Sp
Melanoma in situ vs invasive					
Readers	AUC	95%CI	Accuracy	Se	Sp
1	0.69	0.60 to 0.77	0.69	0.73	0.65
2	0.68	0.59 to 0.77	0.67	0.74	0.62
3	0.63	0.54 to 0.72	0.64	0.65	0.62
4	0.68	0.59 to 0.77	0.68	0.70	0.66
5	0.60	0.53 to 0.68	0.63	0.60	0.73
6	0.69	0.60 to 0.78	0.65	0.69	0.62
7	0.58	0.49 to 0.70	0.59	0.59	0.60
8	0.58	0.50 to 0.66	0.60	0.59	0.64
9	0.64	0.56 to 0.73	0.65	0.64	0.69
10	0.62	0.54 to 0.71	0.64	0.62	0.68
Mean	0.64	0.61 to 0.67	0.64	0.66	0.65
Melanoma < 0.8 vs ≥ 0.8 millimetres Breslow thickness					
Readers	AUC	95%CI	Accuracy	Se	Sp
1	0.84	0.76 to 0.93	0.87	0.88	0.87
2	0.70	0.59 to 0.80	0.72	0.61	0.78
3	0.71	0.64 to 0.82	0.68	0.54	0.89
4	0.48	0.37 to 0.59	0.52	0.33	0.63
5	0.72	0.65 to 0.82	0.70	0.55	0.90
6	0.75	0.66 to 0.86	0.77	0.66	0.84
7	0.75	0.68 to 0.86	0.77	0.65	0.86
8	0.69	0.62 to 0.76	0.60	0.47	1.00
9	0.67	0.58 to 0.77	0.63	0.49	0.85
10	0.69	0.60 to 0.80	0.67	0.53	0.85
mean	0.70	0.64 to 0.76	0.69	0.57	0.84

AUC, area under the ROC curve; Se, sensitivity; Sp, specificity, 95% CI, 95% confidence interval

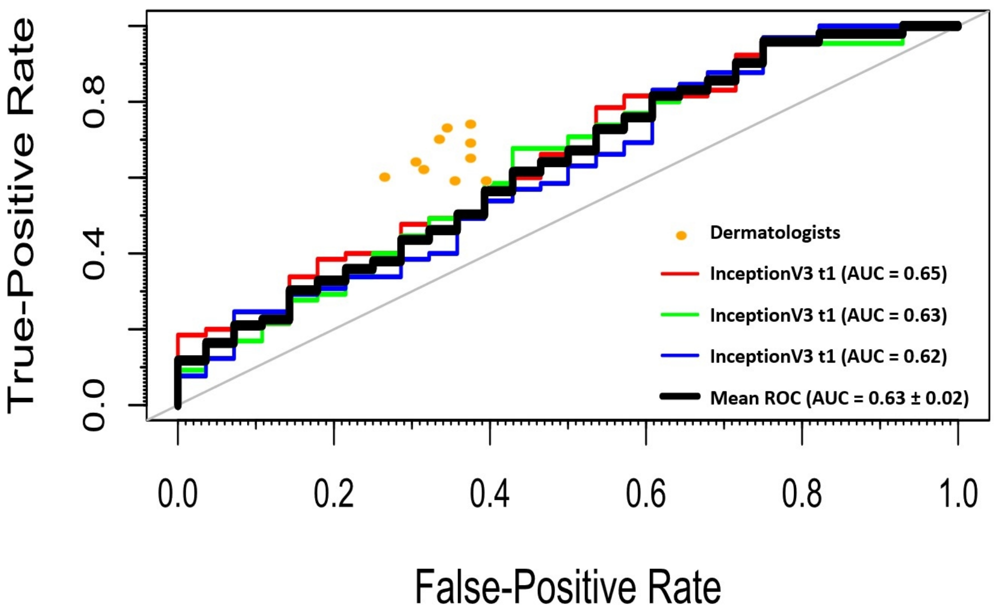
For Peer Review



ROC curves for the prediction model between in situ and invasive melanoma for (a) ResNetV2, (b) InceptionV3 and (c) EfficientNetB6.

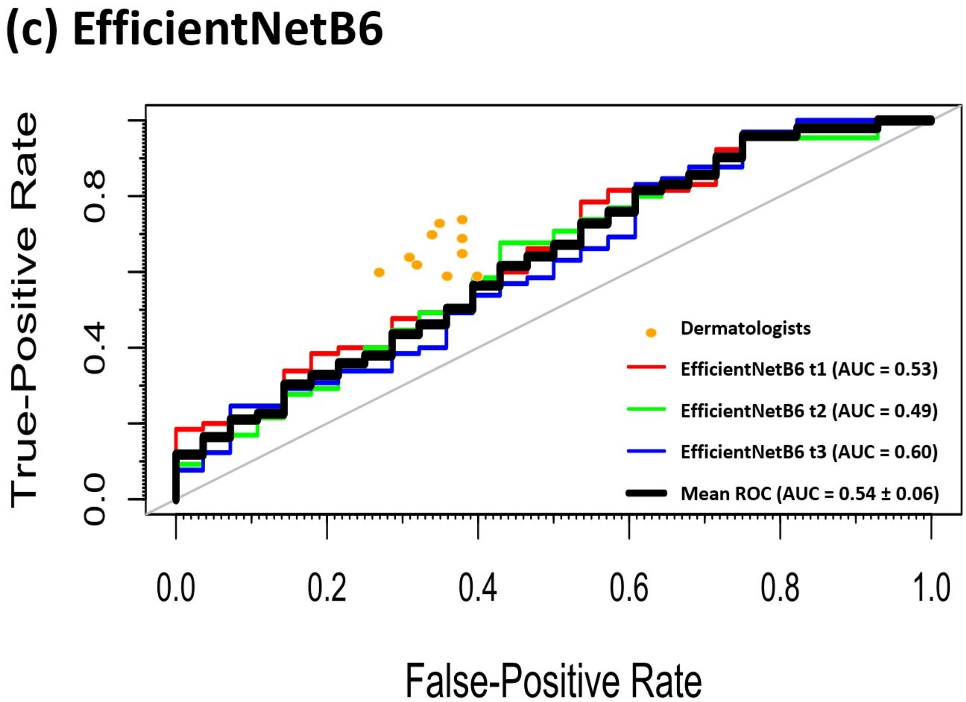
232x167mm (144 x 144 DPI)

(b) InceptionV3



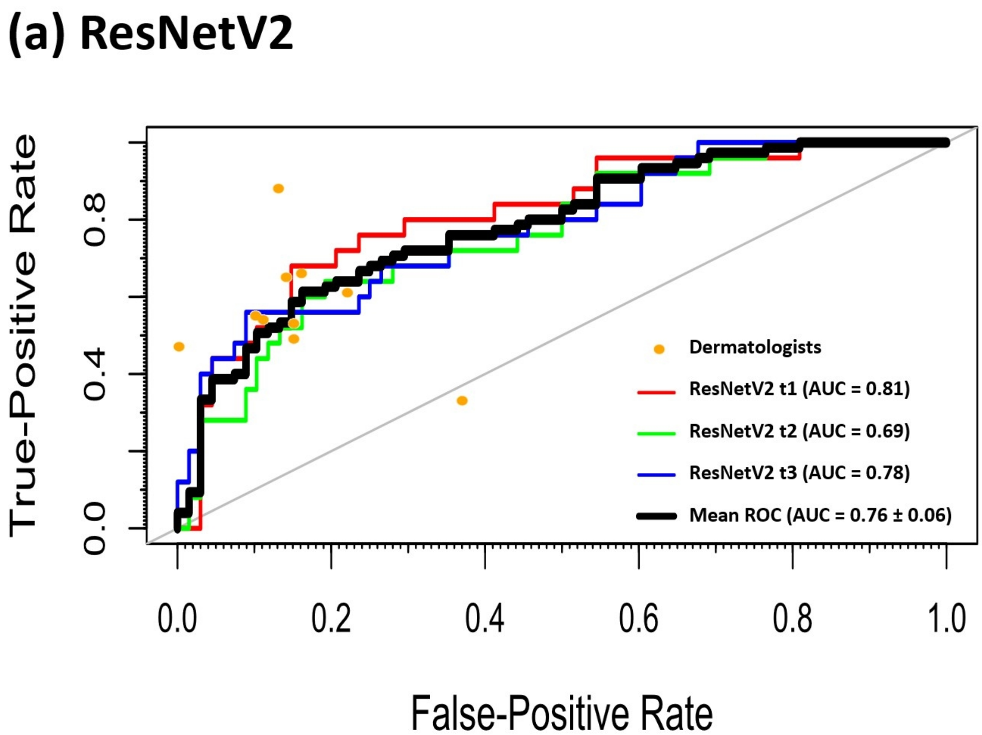
ROC curves for the prediction model between in situ and invasive melanoma for (a) ResNetV2, (b) InceptionV3 and (c) EfficientNetB6.

236x169mm (144 x 144 DPI)



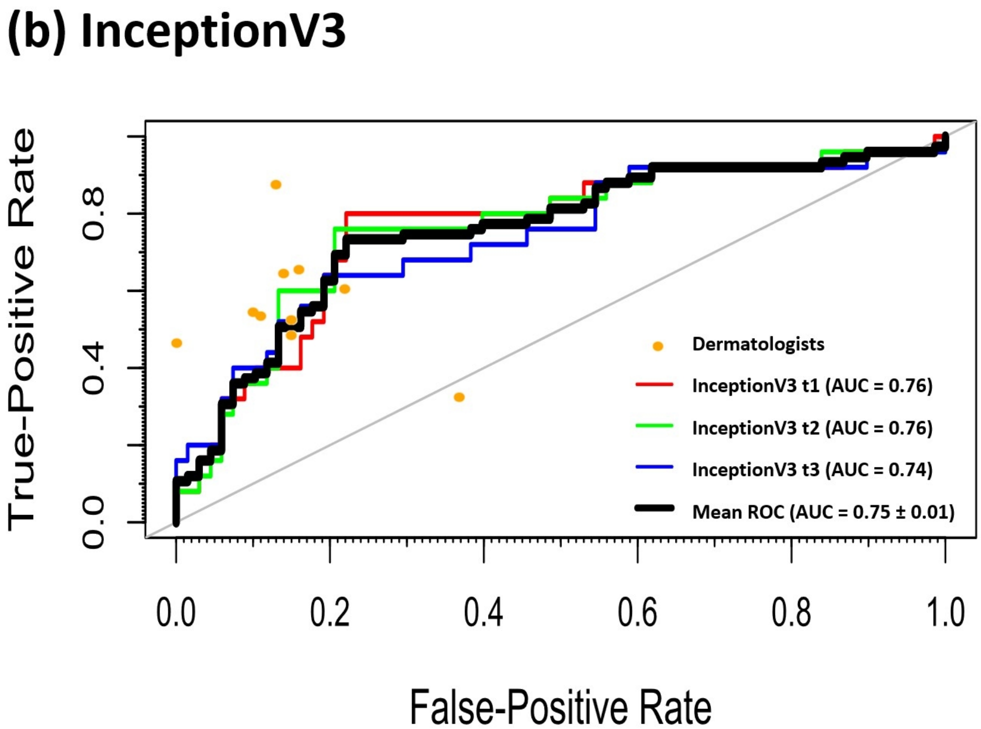
ROC curves for the prediction model between in situ and invasive melanoma for (a) ResNetV2, (b) InceptionV3 and (c) EfficientNetB6.

236x166mm (144 x 144 DPI)



ROC curves for the prediction model between melanoma < 0.8 or ≥ 0.8 mm of Breslow thickness for (a) ResNetV2, (b) InceptionV3 and (c) EfficientNetB6.

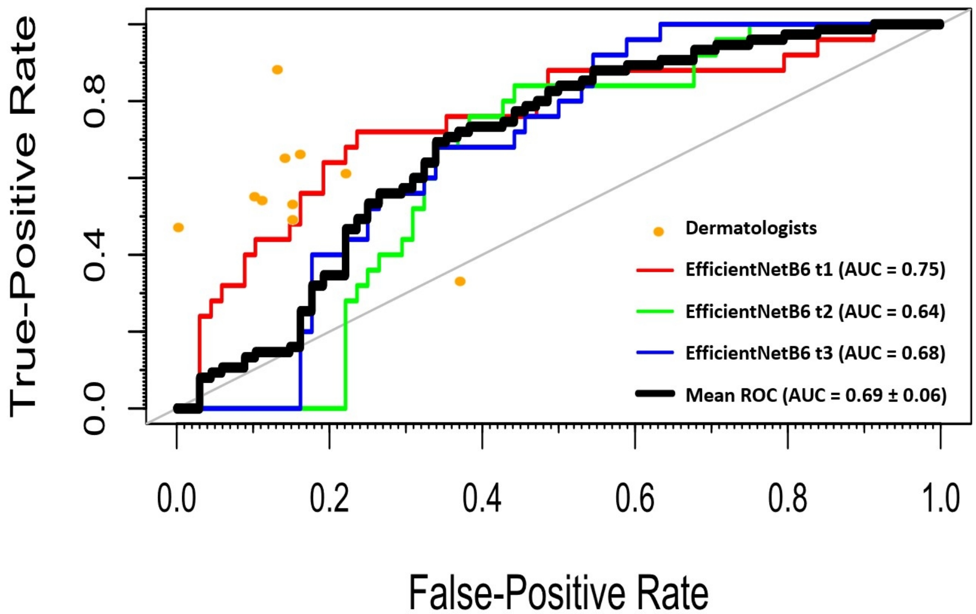
233x171mm (144 x 144 DPI)



ROC curves for the prediction model between melanoma < 0.8 or ≥ 0.8 mm of Breslow thickness for (a) ResNetV2, (b) InceptionV3 and (c) EfficientNetB6.

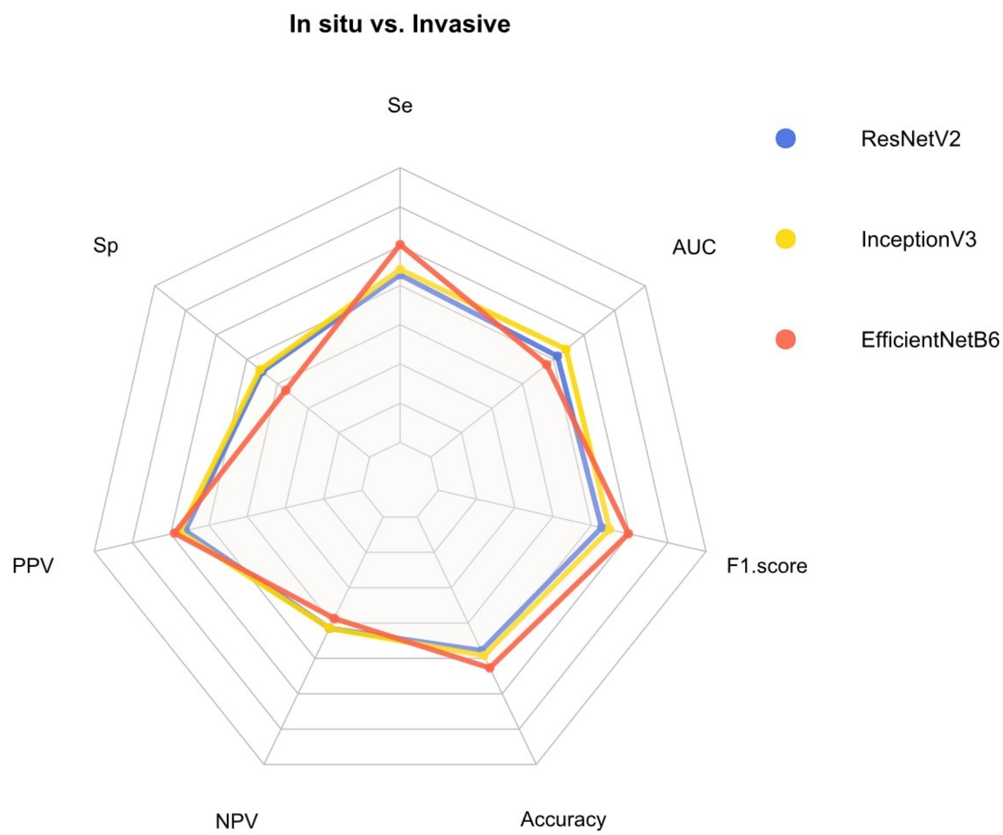
233x170mm (144 x 144 DPI)

(c) EfficientNetB6



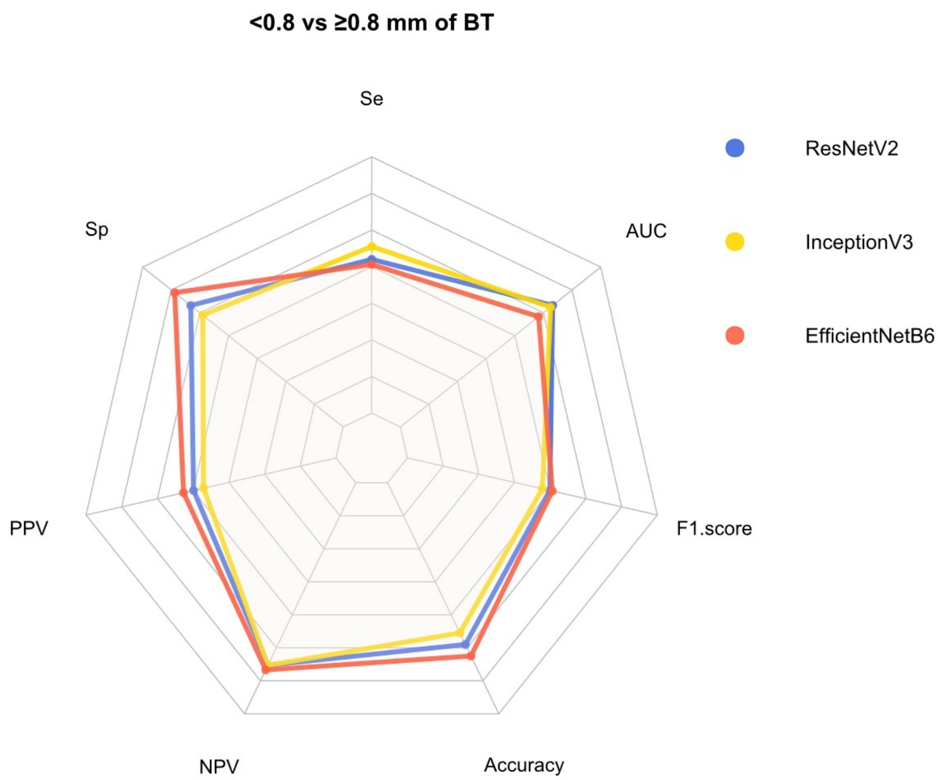
ROC curves for the prediction model between melanoma < 0.8 or ≥ 0.8 mm of Breslow thickness for (a) ResNetV2, (b) InceptionV3 and (c) EfficientNetB6.

234x170mm (144 x 144 DPI)



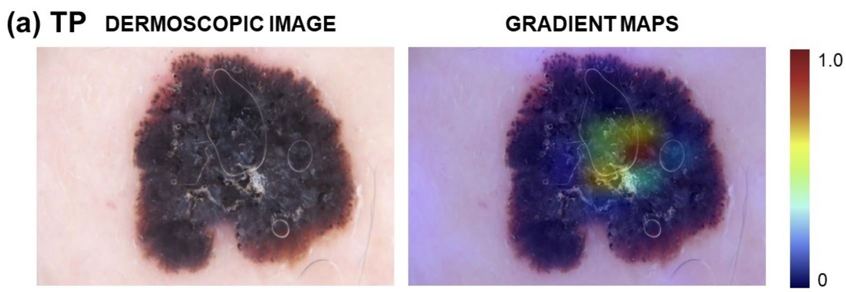
Radar Chart for the comparison of the performance of the three pretrained CNN for (a) MIS vs invasive melanoma and (b) melanoma < 0.8 mm or ≥ 0.8 mm of BT

202x171mm (144 x 144 DPI)



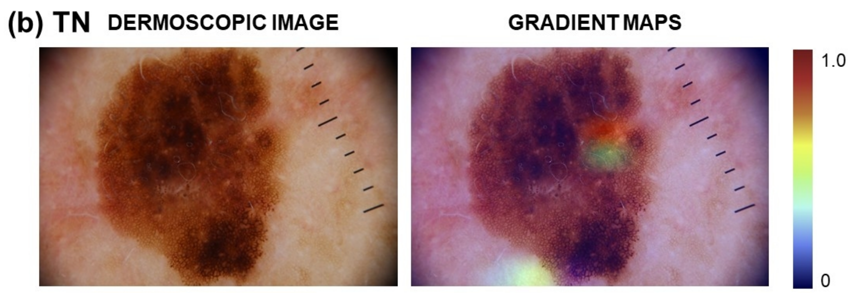
Radar Chart for the comparison of the performance of the three pretrained CNN for (a) MIS vs invasive melanoma and (b) melanoma < 0.8 mm or ≥ 0.8 mm of BT

217x173mm (144 x 144 DPI)



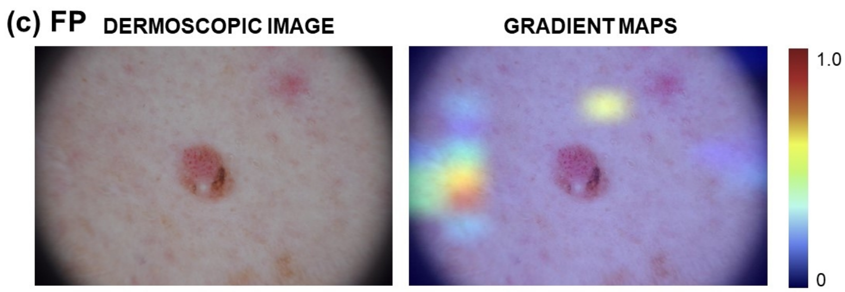
Gradient maps. (a) True positive (TP): the algorithm identified a blue-white veil area as an important region of a melanoma ≥ 0.8 BT. (b) True negative (TN): a regular network area was identified by the algorithm to predict a melanoma < 0.8 mm BT. (c) False positive (FP): the algorithm focused on the healthy surrounding skin of a tiny nodular melanoma ≥ 0.8 mm. (d) False negative (FN): the algorithm was unable to focus on a specific dermoscopic structure to correctly perform the prediction in a melanoma < 0.8 mm BT.

338x190mm (96 x 96 DPI)



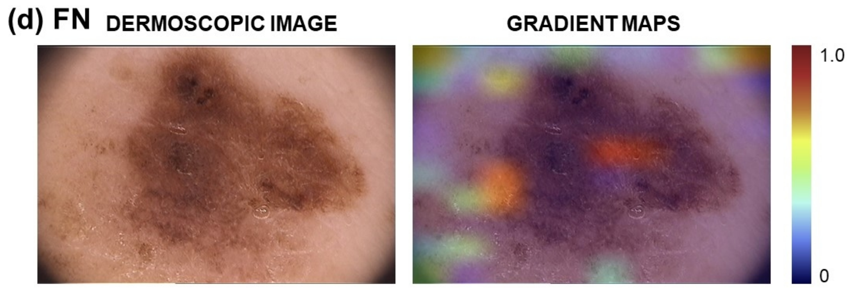
Gradient maps. (a) True positive (TP): the algorithm identified a blue-white veil area as an important region of a melanoma ≥ 0.8 BT. (b) True negative (TN): a regular network area was identified by the algorithm to predict a melanoma < 0.8 mm BT. (c) False positive (FP): the algorithm focused on the healthy surrounding skin of a tiny nodular melanoma ≥ 0.8 mm. (d) False negative (FN): the algorithm was unable to focus on a specific dermoscopic structure to correctly perform the prediction in a melanoma < 0.8 mm BT.

338x190mm (96 x 96 DPI)



Gradient maps. (a) True positive (TP): the algorithm identified a blue-white veil area as an important region of a melanoma ≥ 0.8 BT. (b) True negative (TN): a regular network area was identified by the algorithm to predict a melanoma < 0.8 mm BT. (c) False positive (FP): the algorithm focused on the healthy surrounding skin of a tiny nodular melanoma ≥ 0.8 mm. (d) False negative (FN): the algorithm was unable to focus on a specific dermoscopic structure to correctly perform the prediction in a melanoma < 0.8 mm BT.

338x190mm (96 x 96 DPI)



Gradient maps. (a) True positive (TP): the algorithm identified a blue-white veil area as an important region of a melanoma ≥ 0.8 BT. (b) True negative (TN): a regular network area was identified by the algorithm to predict a melanoma < 0.8 mm BT. (c) False positive (FP): the algorithm focused on the healthy surrounding skin of a tiny nodular melanoma ≥ 0.8 mm. (d) False negative (FN): the algorithm was unable to focus on a specific dermoscopic structure to correctly perform the prediction in a melanoma < 0.8 mm BT.

338x190mm (96 x 96 DPI)