Dermatology Practical & Conceptual

Comparison of Convolutional Neural Network Architectures for Robustness Against Common Artefacts in Dermatoscopic Images

Florian Katsch[1], Christoph Rinner[1], Philipp Tschandl[2]

1 Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria
2 Department of Dermatology, Medical University of Vienna, Vienna, Austria

Key words: image classification, object detection, instance segmentation, artefacts, dermatoscopy,

Citation: Katsch F, Rinner C, Tschandl P. Comparison of convolutional neural network architectures for robustness against common artefacts in dermatoscopic images. *Dermatol Pract Concept.* 2022;12(3):e2022126. DOI: https://doi.org/10.5826/dpc.1203a126

Accepted: December 7, 2021; **Published:** July 2022

Funding: None.

Competing interests: PT reports consulting fees from Silverchair, honoraria from FotoFinder, Novartis and Lilly, and grants from MetaOptima Technology and Lilly, outside the submitted work. All other authors declare no conflict of interest.

Authorship: All authors have contributed significantly to this publication.

Corresponding author: Philipp Tschandl, PhD, Department of Dermatology, Medical University of Vienna, Währinger Gürtel 18-20, 1090 Vienna, Austria, E-mail: philipp.tschandl@meduniwien.ac.at

ABSTRACT

Introduction: Classification of dermatoscopic images via neural networks shows comparable performance to clinicians in experimental conditions but can be affected by artefacts like skin markings or rulers. It is unknown whether specialized neural networks are more robust to artefacts.

Objectives: Analyze robustness of 3 neural network architectures, namely ResNet-34, Faster R-CNN and Mask R-CNN.

Methods: We identified common artefacts in the HAM10000, PH2 and the 7-point criteria evaluation datasets, and established a template-based method to superimpose artefacts on dermatoscopic images. The HAM10000-dataset with and without superimposed artefacts was used to train the networks, followed by analyzing their robustness against artefacts in test images. Performance was assessed via area under the precision recall curve and classification results.

Results: ResNet-34 and Faster R-CNN models trained on regular images perform worse than Mask R-CNN on images with superimposed artefacts. Artefacts added to all tested images led to a decrease in area under the precision-recall curve values of 0.030 for ResNet-34 and 0.045 for Faster R-CNN in comparison to only 0.011 for Mask R-CNN. However, changes in model performance only became significant with 40% or more of the images having superimposed artefacts. A loss in performance occurred when the training was biased by selectively superimposing artefacts on images belonging to a certain class.

Conclusions: As Mask R-CNN showed the least decrease in performance when confronted with artefacts, instance segmentation architectures may be helpful to counter the effects of artefacts, warranting further research on related architectures. Our artefact insertion mechanism could be useful for future research.

Introduction

Epidemiological studies show an increasing trend in the incidence rates of melanoma and non-melanoma skin cancer worldwide over the last 30 years [1]. According to the American Joint Committee on Cancer melanoma staging system, stage I malignant skin alterations with a five-year survival rate of more than 90% contrasts with a survival rate of less than 15% for stage IV patients. This indicates a clear need for early, reliable and consistent diagnosis and treatment [2]. The desire for automatic lesion analysis is further intensified by a high dependency between the diagnostic quality and the examiners experience in dermoscopy, as well as a high degree of inter- and intra-variability of diagnoses [3,4].

Methods of automatic skin lesion analysis have been the focus of research for decades, and have gained interest in recent years [5,6]. These methods are intended to support tele-dermatologic settings, improve management decisions or aid in difficult clinical scenarios, but often suffer, among other things, from the presence of artefacts in dermatoscopic images [7-12].

A common neural network used for classification is ResNet, two well-known neural network architectures in computer vision are Faster R-CNN and Mask R-CNN (Figure 1) [13,14]. The first is performing "object detection", a process where one or multiple objects in an image can be detected and located with a rectangular "bounding box". The latter is performing "instance segmentation" where one or more objects in an image can be found and their respective area (i.e. pixels) in the image outlined ("segmented"), and can be regarded as a CNN-based multi-instance generalisation of computer-vision based techniques of lesion segmentation [15,16]. Object detection has been used in the field of automated skin cancer detection on clinical images [17], but training of instance segmentation neural networks in dermatoscopy has not yet been reported on successfully, most probably because of missing ground-truth data.

Objectives

Our hypothesis is that in contrast to ResNet, the other network architectures intrinsically have to "concentrate" on regions of the classified object in an image and hence may offer robustness against artefacts surrounding the lesions. Robustness in this case describes the consistency of the obtained diagnoses under the influence of artefacts in the input image data. These networks could potentially be used as off-the-shelf methods with little customization-effort needed and could enable us to focus less on tedious image pre-processing such as removal of bubbles or hairs [18].

Methods

Image datasets

The primary source of dermatoscopic images was the HAM10000 dataset [19]. This dataset also includes publicly available lesion segmentation masks for every image, as described previously, which are necessary for training the Faster R-CNN and Mask R-CNN architectures [8]. It contains 10,015 images, each with 600x450 pixels and 3-8 bit color channels. Each image is assigned one of seven diagnostic classes: actinic keratosis / intraepithelial carcinoma

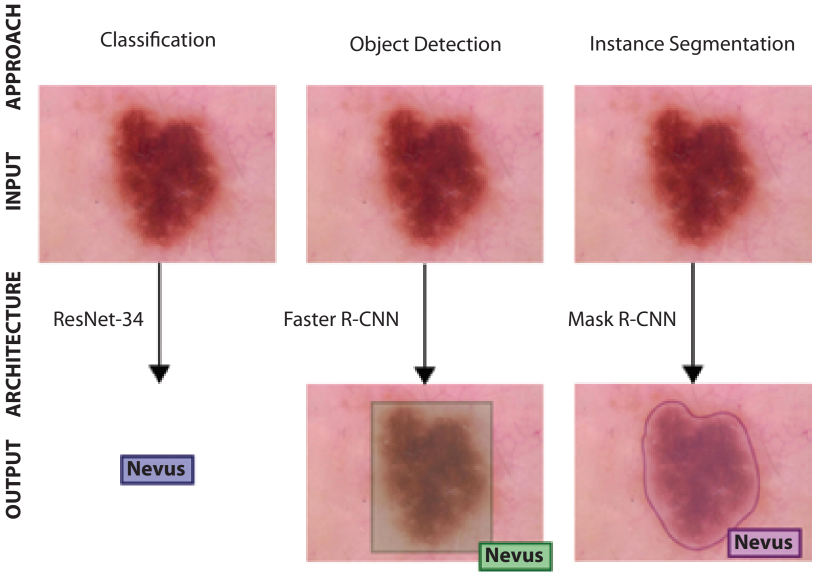

Figure 1. A visual representation of the outputs of the three approaches. Image classification (ie ResNet-34) classifies the image as a whole, object detection (ie Faster R-CNN) finds objects and their approximate position in the image and instance segmentation (ie Mask R-CNN) finds objects and their exact spatial delimitation.

(akiec), basal cell carcinoma (bcc), benign keratotic lesion (bkl), dermatofibroma (df), nevus (nv), melanoma (mel), or vascular lesion (vasc). Also, the PH2 and the 7-point criteria evaluation dataset were reviewed and several images were utilized to extract artefacts from [20,21]. Images from those datasets were not used for other purposes within this study. We used the ISIC2018 test-set as the test-set to keep variation as low as possible, as it sources from the same origin as the HAM10000 dataset and includes the same classes.

Artefact generation

As with every real-world picture, dermatoscopic images can contain content considered as "artefacts". Examples are hairs, dark corners, vignettes, medical devices, different sorts of rulers, ink markings in different shapes, styles and colors, air bubbles or reflections. This work focuses on three of them: "bubbles" that originate from trapped air in the liquid between skin and the dermatoscope, "rulers" used to show the spatial dimension of a lesion, and ink "markings" on the patient's skin used to highlight the lesion for excision or review.

In order to generate artefact-modified cases, we selected 60 images from the HAM10000, PH2 and 7-point criteria dataset which contain either a bubble, a ruler or a marking artefact. From those images we extracted the artefacts by manually repairing the images areas with Adobe® Photoshop's® (version CC 2018 (19.1.9), Adobe Inc.) content aware image repair mechanism and using the difference, per RGB channel, to the untouched image as a template (Figure 2). The insertion of those templates was done in a way that the position of artefacts varies according to observed patterns, using the provided segmentation mask of the target image. In Figure 3, a dermatoscopic image with automatically superimposed artefacts is shown. The source code will be made available upon publication of this work at https://github.com/thisismexp/artefact_insertion.

Using the artefact insertion mechanism, several dataset mutations of the original HAM10000 dataset were created, where artefacts were superimposed on either none or all of the images and on every image belonging to a certain diagnosis. The test portion of the HAM10000 dataset, corresponding to the ISIC2018 challenge Task 3 test-set with 1,511 images, was altered in the same way. Additionally, artefacts were inserted in a certain percentage of images in 20% step increments.

Neural Network Training

As representatives for image classification, object detection and instance segmentation we trained a ResNet-34, a Faster

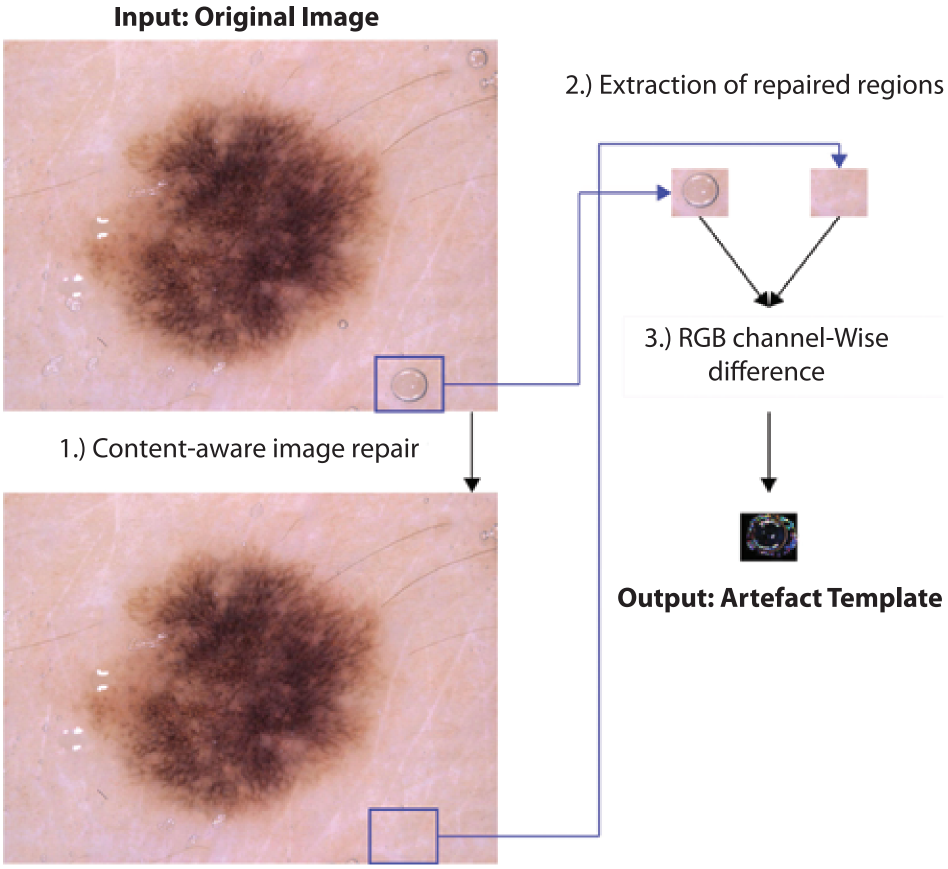
Input: Original Image
2.) Extraction of repaired regions
3.) RGB channel-Wise difference
1.) Content-aware image repair
Output: Artefact Template

Figure 2. Workflow for extracting artefact templates. Manually selected original images (Input) were repaired manually (1), and corresponding image areas extracted (2). The channel-wise difference (3) was stored as a template for the corresponding artefact type.

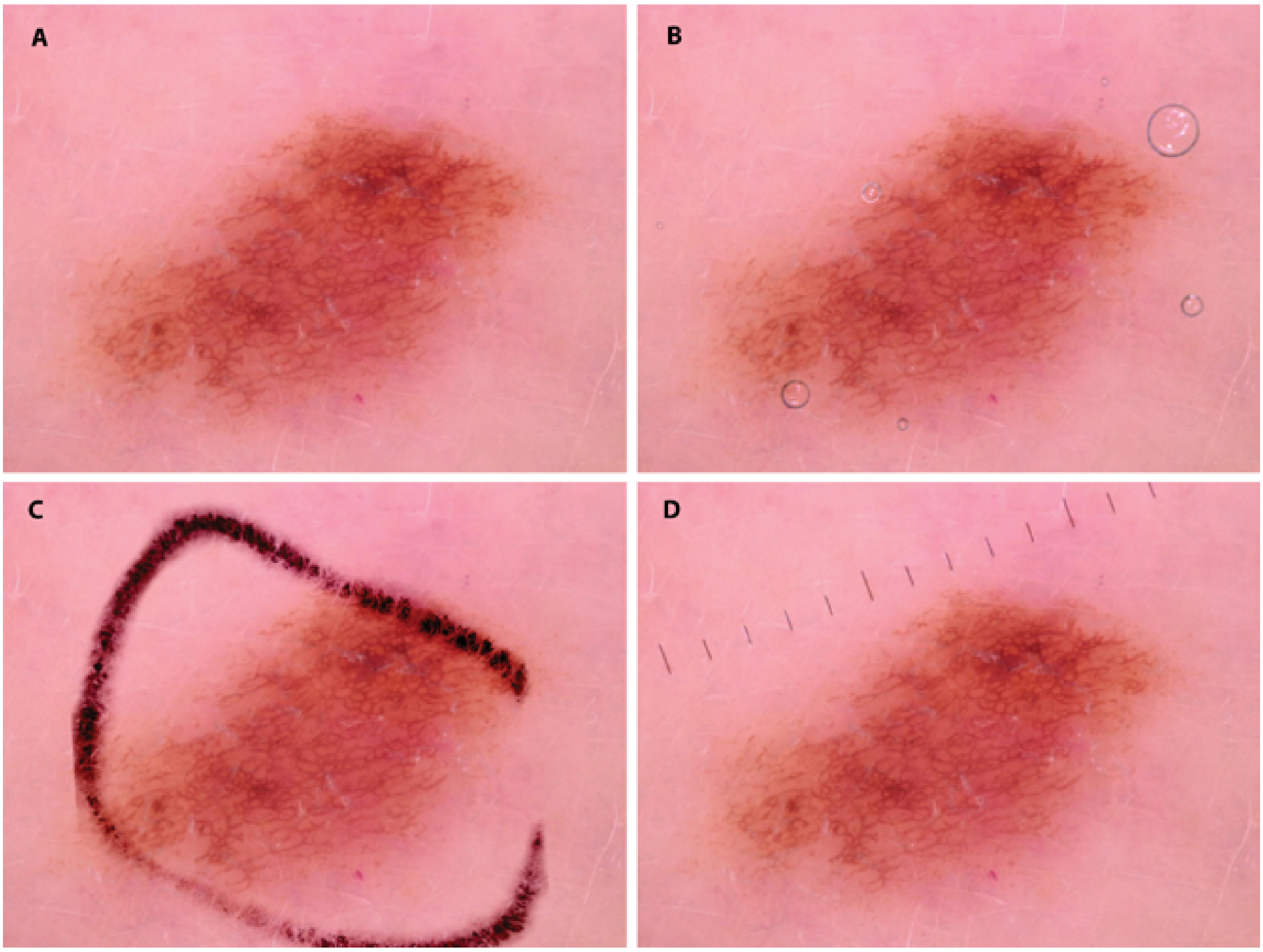

Figure 3. Example of automatically superimposed artefacts on a dermatoscopic image. (A) In the top left the original image without artefact is shown. The other 3 images show the lesion with the superimposed artefacts bubbles (B), ink markings (C) and a ruler (D).

R-CNN (with ResNet-34 backbone) and a Mask R-CNN (also with a ResNet-34 backbone) model as provided by the Torchvision package of the open source machine learning framework PyTorch [22]. All models were trained on all of the 9 generated datasets in a 5-fold cross validation fashion. Transfer-learning and data augmentation including random crops, resize, rotations, mirroring operations as well as color jitter operations were used.

Statistics

To evaluate diagnostic accuracy, all trained network models are tested against the 13 test datasets and performance was reported in terms of area under the precision recall curve (PR-AUC), precision, recall, false positive (FPR) and false negative rates (FNR) and differences thereof (calculated using scikit-learn version 0.24.1) [23]. To visualize spatial activations, Gradient based Class Activation Map (Grad-CAM) visualizations were used. A two-sided p-value of 0.05 was regarded as statistically significant, and all calculations were performed using statsmodels version 0.12.2 [24].

Results

Baseline performance in terms of PR-AUC of our models trained and tested with no additional artefacts was 0.8 for ResNet-34 and 0.72 for Faster R-CNN as well as Mask R-CNN. Introduction of artefacts in only the test dataset led to a reduction in performance for all three architectures (Figure 4) increasing with the proportion of artefacts present in the test dataset, and more severe for the ResNet-34 and Faster R-CNN model. With a maximum relative reduction of 0.05 PR-AUC the Faster R-CNN model was affected the most, ResNet-34 (-0.03) the second most, and Mask R-CNN was the most robust (-0.01). For ResNet-34 and Faster R-CNN, changes in predictive performance compared to baseline was significant at and above 40% of introduced artefacts in the test set ($P < 0.01$; tested using McNemar test with Edwards correction on binarized predictions). For Mask R-CNN we did not detect a significant difference in predictions in all used test sets (all P values > 0.17).

Introducing artefacts in the training data led to biased results in all three examined architectures. Artefacts introduced into all images of the melanocytic nevi class during training decreased recall values on average by 0.218 for ResNet-34, 0.129 for Faster R-CNN and by 0.155 for Mask R-CNN in comparison to the respective unbiased models. Reduction in recall values indicate that those are indeed biased by artefacts for specific classes. This effect was more apparent the bigger the proportion of biased samples in the dataset is. Considering the FPR and FNR for specific classes, a selective bias towards classes that were corrupted by artefacts during training could be observed for all three architectures. The increase in FPR for the class with inserted artefacts during training, and a simultaneous increase in FNR for all others, in fact showed a shift in classifications towards the biased class. This effect could not be observed if artefacts were inserted into none or all of the images.

When inspecting heat map representations of the Grad-CAM we observed that training with artefacts shifted the attention of the object detection and instance segmentation network away from the artefact itself towards areas of the lesion (Figure 5). These mappings indicate an increase in robustness against these very artefacts for Faster and Mask R-CNN models, if trained with inserted artefacts in the dataset.

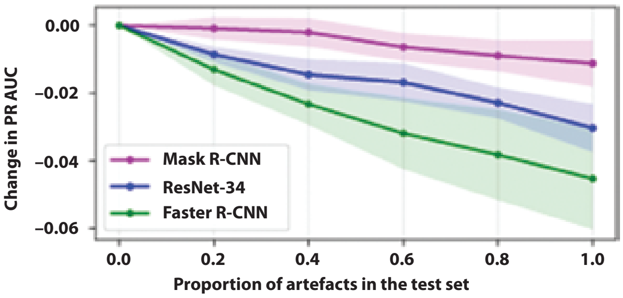

Figure 4. Neural networks show different robustness to inserted artefacts on the test set. Precision recall curve (PR-AUC) as achieved by training without additional artefacts in the train and test set was used as the baseline (0%). With increasing proportion of inserted artefacts, PR-AUC decreases for ResNet-34 (blue) and Faster R-CNN (green), but almost not for Mask R-CNN (purple). Shaded areas denote 95%-confidence intervals.

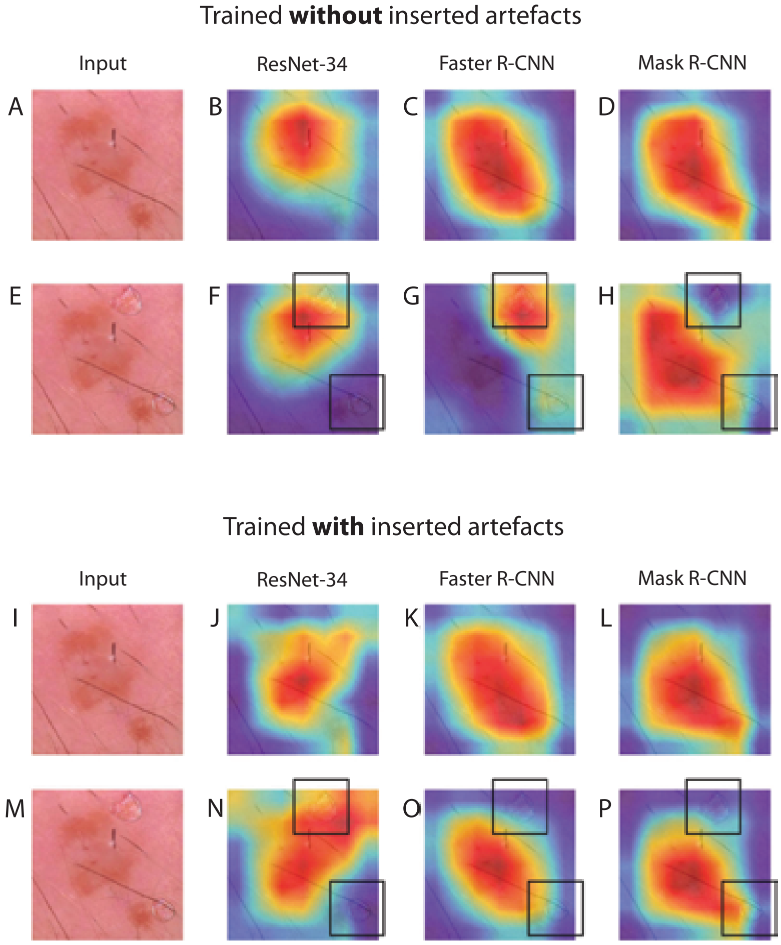

Figure 5. Grad-CAM for used network architectures. The first column shows the input image for the corresponding row, in its original form (top) and with bubble artefacts inserted (bottom). Grad-CAM heatmaps show the ResNet-34 increases attention towards the bubble-area after training with artefacts (N), where the Faster R-CNN network loses its initial attention towards the artefact (G) afterwards (O). The Mask R-CNN architecture seems to ignore the artefact throughout (H and P). Black boxes denote positions of inserted bubble artefacts.

Conclusions

We compared representatives of three neural network architectures to classify lesions in dermatoscopic images in regard to their robustness against artefacts. Although as a limitation the baseline performance of the examined models were not the same, we found differences in their vulnerability to performance changes under the influence of artefacts. Mask R-CNN tends to be the most robust. The influence on classification results by artefacts in test images can be reduced by augmenting training data with artificially superimposed artefacts for all three architectures. This is in line with findings by Maron et al, who reduced - but not eliminated - brittleness of their system through data augmentation [25]. We anticipate that automated superimposition of artefacts as presented here as a further evolution of data augmentation, that together with integrating more diverse variants, will enhance robustness of automated classifiers and decision support systems further [26,27]. The initial data, in our view, warrants more in-depth follow up research on this topic, to understand which approaches are the most effective and efficient.

However, this work failed to find evidence for a clinically relevant robustness against artefacts of instance segmentation for several reasons. On the one hand we used a shallow backbone network architecture for our experiments, even though current research and commercial products commonly use deeper models, and an increase in robustness against image distortions has been demonstrated by others with increased backbone capacity [28]. We also used a new template-based approach to superimpose artefacts on images. This approach leaves room for improvement with regard to the number of images the artefacts are extracted from, and a detailed analysis on how different artefact types affect the classification performance. Alternatively, lesions with existing artefacts could be used after manual or automated annotations.

References

1. Apalla Z, Lallas A, Sotiriou E, Lazaridou E, Ioannides D. Epidemiological trends in skin cancer. *Dermatol Pract Concept*. 2017;7(2):1-6. DOI: 10.5826/dpc.0702a01. PMID: 28515985. PMCID: PMC5424654.
2. Balch CM, Soong S-J, Atkins MB, et al. An evidence-based staging system for cutaneous melanoma. *CA Cancer J Clin*. 2004;54(3):131-149; quiz 182-184. DOI: 10.3322/canjclin.54.3.131. PMID: 15195788.
3. Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. *Lancet Oncol*. 2002;3(3):159-165. DOI: 10.1016/s1470-2045(02)00679-4. PMID: 11902502.
4. Korotkov K, Garcia R. Computerized analysis of pigmented skin lesions: a review. *Artif Intell Med*. 2012;56(2):69-90. DOI: 10.1016/j.artmed.2012.08.002. PMID: 23063256.
5. Rubegni P, Burroni M, Cevenini G, et al. Digital dermoscopy analysis and artificial neural network for the differentiation of clinically atypical pigmented skin lesions: a retrospective study. *J Invest Dermatol*. 2002;119(2):471-474. DOI: 10.1046/j.1523-1747.2002.01835.x. PMID: 12190872.
6. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118. DOI: 10.1038/nature21056. PMID: 28117445. PMCID: PMC8382232.
7. Muñoz-López C, Ramírez-Cornejo C, Marchetti MA, et al. Performance of a deep neural network in teledermatology: a single-centre prospective diagnostic study. *J Eur Acad Dermatol Venereol*. 2021;35(2):546-553. DOI: 10.1111/jdv.16979. PMID: 33037709. PMCID: PMC8274350.
8. Tschandl P, Rinner C, Apalla Z, et al. Human–computer collaboration for skin cancer recognition. *Nat Med*. 2020;26(8):1229-1234. DOI: 10.1038/s41591-020-0942-0. PMID: 32572267.
9. Fink C, Blum A, Buhl T, et al. Diagnostic performance of a deep learning convolutional neural network in the differentiation of combined naevi and melanomas. *J Eur Acad Dermatol Venereol*. 2020;34(6):1355-1361. DOI: 10.1111/jdv.16165. PMID: 31856342.
10. Okuboyejo DA, Olugbara OO. A review of prevalent methods for automatic skin lesion diagnosis. *Open Dermatol J*. 2018;12(1):14-53. DOI: 10.2174/187437220181201014
11. Winkler JK, Fink C, Toberer F, 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(10):1135-1141. DOI: 10.1001/jamadermatol.2019.1735. PMID: 31411641. PMCID: PMC6694463.
12. Navarrete-Dechent C, Dusza SW, Liopyris K, Marghoob AA, Halpern AC, Marchetti MA. Automated Dermatological Diagnosis: Hype or Reality? *J Invest Dermatol*. 2018;138(10):2277-2279. DOI: 10.1016/j.jid.2018.04.040. PMID: 29864435. PMCID: PMC7701995.
13. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2017;39(6):1137-1149. . DOI: 10.1109/TPAMI.2016.2577031. PMID: 27295650.
14. He K, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017:2980-2988. DOI: 10.1109/ICCV.2017.322.
15. Kaur R, LeAnder R, Mishra NK, et al. Thresholding methods for lesion segmentation of basal cell carcinoma in dermoscopy images. *Skin Res Technol*. 2017;23(3):416-428. DOI: 10.1111/srt.12352. PMID: 27892649.
16. Mishra NK, Kaur R, Kasmi R, et al. Automatic lesion border selection in dermoscopy images using morphology and color features. *Skin Res Technol*. 2019;25(4):544-552. DOI: 10.1111/srt.12685. PMID: 30868667. PMCID: PMC7173402.
17. Han SS, Moon IJ, Lim W, et al. Keratinocytic Skin Cancer Detection on the Face Using Region-Based Convolutional Neural Network. *JAMA Dermatol*. 2020;156(1):29-37. DOI: 10.1001/jamadermatol.2019.3807. PMID: 31799995. PMCID: PMC6902187.
18. Lee T, Ng V, Gallagher R, Coldman A, McLean D. Dullrazor: A software approach to hair removal from images.

Comput Biol Med. 1997;27(6):533-543. DOI: 10.1016/s0010-4825(97)00020-6. PMID: 9437554.

19. Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data*. 2018;5:180161. DOI: 10.1038/sdata.2018.161. PMID: 30106392. PMCID: PMC6091241.
20. Mendonça T, Ferreira PM, Marques JS, Marcal ARS, Rozeira J. PH2 - A dermoscopic image database for research and benchmarking. In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*.; 2013:5437-5440. DOI: 10.1109/EMBC.2013.6610779. PMID: 24110966.
21. Kawahara J, Daneshvar S, Argenziano G, Hamarneh G. Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets. *IEEE Journal of Biomedical and Health Informatics*. 2019;23(2):538-546. DOI: 10.1109/JBHI.2018.2824327. PMID: 29993994.
22. Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, d' Alché-Buc F, Fox E, Garnett R, eds. *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc.; 2019:8026-8037. Available from: https://papers.nips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html
23. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*. 2011;12:2825-2830. Available from: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf
24. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. In: *Proceedings of the 9th Python in Science Conference*. Vol 57. Austin, TX; 2010:61. Available from: https://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf
25. Maron RC, Haggenmüller S, von Kalle C, et al. Robustness of convolutional neural networks in recognition of pigmented skin lesions. *Eur J Cancer*. 2021;145:81-91. DOI: 10.1016/j.ejca.2020.11.020. PMID: 33423009.
26. Aggarwal SLP. Data augmentation in dermatology image recognition using machine learning. *Skin Res Technol*. 2019;25(6):815-820. DOI: 10.1111/srt.12726. PMID: 31140653.
27. Winkler JK, Sies K, Fink C, et al. Association between different scale bars in dermoscopic images and diagnostic performance of a market-approved deep learning convolutional neural network for melanoma recognition. *Eur J Cancer*. 2021;145:146-154. DOI: 10.1016/j.ejca.2020.12.010. PMID: 33465706.
28. Michaelis C, Mitzkus B, Geirhos R, et al. Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming. *arXiv [csCV]*. Published online July 17, 2019. Available from: http://arxiv.org/abs/1907.07484