PROF. BYUNG HO OH (Orcid ID : 0000-0001-9575-5665)

Article type : Original Article

Original Article (Revision 1)

Augmented Decision Making for Acral Lentiginous Melanoma Detection Using Deep Convolutional Neural Networks

S Lee, MD[1,2]; YS Chu[3]; SK Yoo[3]; S Choi, MD[4]; SJ Choe, MD[1]; SB Koh, MD, PhD[2]; KY Chung, MD, PhD[4]; L Xing, PhD[5]; B Oh, MD, PhD[4*]; S Yang, PhD[3*]

[1]Department of Dermatology, Yonsei University Wonju College of Medicine, Wonju, Korea
[2]Department of Preventive Medicine, Yonsei University Wonju College of Medicine, Wonju, Korea
[3]Department of Biomedical Engineering, Yonsei University, Wonju, Korea
[4]Department of Dermatology and Cutaneous Biology Research Institute, Severance Hospital, Yonsei University College of Medicine, Seoul, Korea
[5]Department of Radiation Oncology, Stanford University, California, United Status
[*]Corresponding authors

Primary corresponding author: Byungho Oh, MD, PhD
Department of Dermatology, Severance Hospital,
Yonsei University College of Medicine,
50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/JDV.16185

Tel.: +82-2-2228-2090, Fax: +82-2-393-9157, Email: obh505@gmail.com

Keywords:
Deep learning, Machine learning, Artificial Intelligence, Neural network, Acral, Melanoma

Words: 2975
Tables: 1
Figures: 4

Funding/Support: This research was supported in part by the Global Frontier Program, through the Global Frontier Hybrid Interface Materials (GFHIM) of the National Research Foundation of Korea (NRF), funded by the Ministry of Science, ICT & Future Planning (2013M3A6B1078872). This work was supported in part by the Yonsei University Wonju Campus Future-Leading Research Initiative in 2018 (RMS2 2018-62-0058).

Conflict of Interest: The authors have no conflict of interest to declare.

Abstract

Background: Several studies have achieved high-level performance of melanoma detection using convolutional neural networks (CNNs). However, few have described the extent to which the implementation of CNNs improves the diagnostic performance of the physicians.

Objective: This study is aimed at developing a CNN for detecting acral lentiginous melanoma (ALM) and investigating whether its implementation can improve the initial decision for ALM detection made by the physicians.

Methods: A CNN was trained using 1072 dermoscopic images of acral benign nevi, ALM, and intermediate tumours. To investigate whether the implementation of CNN can improve the initial decision for ALM detection, 60 physicians completed a three-stage survey. In Stage I, they were asked for their decisions solely on the basis of dermoscopic images provided to them. In Stage II, they were also provided with clinical information. In Stage III, they were provided with the additional diagnosis and probability predicted by the CNN.

Results: The accuracy of ALM detection in the participants was 74.7% (95% confidence interval [CI], 72.6%–76.8%) in Stage I and 79.0% (95% CI, 76.7%–81.2%) in Stage II. In Stage III, it was 86.9% (95% CI, 85.3%–88.4%), which exceeds the accuracy delivered in Stage I by 12.2%p (95% CI, 10.1%p–14.3%p) and Stage II by 7.9%p (95% CI, 6.0%p–9.9%p). Moreover, the concordance between the participants considerably increased (Fleiss-κ of 0.436 [95% CI, 0.437–0.573] in Stage I, 0.506 [95% CI, 0.621–0.749] in Stage II, and 0.684 [95% CI, 0.621–0.749] in Stage III).

Conclusions: Augmented decision making improved the performance of and concordance between the clinical decisions of a diverse group of experts. This study demonstrates the potential use of CNNs as an adjoining, decision-supporting system for physicians' decisions.

Introduction

Recent advances in computing systems and deep learning algorithms have led to several efforts to develop automated diagnosis systems using deep convolutional neural networks (dCNNs) to interpret medical images; these efforts have included computed tomography,[1] ultrasound,[2] histopathology,[3] and fundoscopy images.[4] They have performed at levels comparable with those of medical experts in the field, especially in the detection of malignant tumours.

Malignant melanoma is a major contributor to the global burden of skin disease,[5] and its incidence and mortality rates have increased over recent decades.[6,7] Among its subtypes, acral lentiginous melanoma (ALM) is rare in Caucasians but has a relatively higher incidence rate in non-Caucasians, specifically Asians.[8,9] ALM is more likely to exhibit poorer prognosis owing to delayed diagnoses due to the difficulty of early detection.[10]

For earlier diagnosis, dermoscopy has been increasingly used; this follows some meta-analyses reporting its superiority in melanoma detection over naked-eye examination[11-13] and its advantage in obtaining detailed plane images of pigmentary lesions. Nonetheless, it is difficult for non-experts and even dermatologists to extract helpful features using dermoscopy, and specialized training is required for proficient diagnoses. Therefore, there is a considerable difference between the performance delivered by experts and non-experts.[14,15]

Beyond the development of human-level automated detection tools, implementing artificial intelligence is aimed at the development of an augmentation method to enhance clinical decision making when detecting such critical conditions in clinical practice.

Nevertheless, to date, there have been few studies describing the way implementation of artificial intelligence can affect the clinical decisions of the physicians and the extent to which their diagnostic performance is improved. We hypothesized that augmented decision making using a dCNN not only improved the diagnostic performance of the physicians also help in minimizing the difference between the performances delivered by experts and non-experts, especially if applied to diagnostic modalities that not all physicians are competent in. Therefore, the aim of this study is to develop a dCNN for classifying ALM from acral pigmentary lesions by interpreting dermoscopy images and to investigate whether its implementation can improve the ALM detection performed by the physician.

Materials and Methods

Dataset

This retrospective analysis included dermoscopic images of acral pigmentary lesions collected from 2014 to 2019 at the Department of Dermatology, Severance Hospital, Seoul, Korea. As a means of ensuring quality control, from the dermoscopic image repository of our institution, we constructed a balanced dataset of 500 acral benign nevi (BN) images without atypical or dysplastic features and 500 ALM images. An additional 72 images of acral nevi with atypical or dysplastic features in their histopathological findings were separately collected and labelled as intermediate tumours.[16]

All dermoscopic images were captured using a DermLite Cam® (3 Gen, California) in a cross-polarized or non-polarized mode, and with or without the use of contact media. Although dCNNs generally require extensive amounts of training data, the relatively higher homogeneity of dermoscopic images allowed us to considerably reduce the amount of data required to train the network.[17]

Because the data were acquired from a population-based clinical setting, many of the BN images were provided alongside their respective clinical diagnoses from follow-up examinations. To ensure accurate diagnoses, these were reviewed by three board-certified dermatologists during the dataset-construction stages. The other samples were obtained following the complete excision of acral lesions by us (K.Y.C. and B.O.), enabling histopathologic assessment. The haematoxylin and eosin, and immunohistochemical (if available) stains were reviewed by board-certified pathologists and dermatologists, and the results served as the diagnostic gold standard throughout the study.

In total, 800 (80%) randomly selected BN and ALM images, and all 72 intermediate tumour images, were used to train the dCNN. The remaining 200 (20%) were used for validation (test-set-200, n=200). Two datasets were reviewed to ensure lesion-stratified partitioning.

Clinical information was obtained by reviewing patients' electronic medical charts (Table S1). This study was approved by the institutional review board of Yonsei University College of

Medicine (Approval No. 4-2019-0282) and was carried out in accordance with the Declaration of Helsinki. A waiver of written informed consent was granted for the de-identified data used.

Deep Convolutional Neural Network

We utilized ResNet with 50 residual layers (Fig. S1)[18]. Given the characteristics of the acral site images, in which unnecessary background and colour oversaturation were more likely, we pre-processed images and implemented a modified version of the method used in our previous study, for training and inference (Fig. S2).[19,20] To investigate an optimal method of training for ALM detection of intermediate tumours, we trained four dCNNs (Fig. S3). To identify influential pixels for tumour detection, class activation maps were generated using global average pooling (Fig. S4).[21] The model with the best ALM detection performance was adopted as the final model, and is referred to as ALMnet. All procedures were repeated 5 times to ensure reproducibility and comparability. No overlap between the training and test data was allowed. Details on the training and inferential methods of the dCNN, including image pre-processing and architecture configuration, are described in Supplementary Method in Supplementary Materials.

Decision Study

We recruited a total of 60 physicians (20 board-certified dermatologists, 20 dermatology residents, and 20 general physicians) to investigate how well physicians could make an initial diagnosis of ALM by interpreting dermoscopic images, and how much their performance could be improved with the aid of ALMnet. For ease of response, 100 images from test-set-200 were extracted for the decision study (human-set-100) (Fig. 1). Using Google Survey, an anonymous online-based diagnostic questionnaire was conducted in three stages over 2-week intervals (Fig. S5). In Stage I, participants were shown 100 dermoscopic images only, and asked for their decision in each case (either 'It is more likely a benign nevus. Follow-up is needed.' or 'It is more likely a malignant melanoma. Immediate biopsy or intervention is required.'). In Stage II, they were provided with additional clinical information for the same cases and were asked to answer in the same manner. In most previous studies, human evaluators were asked for their diagnoses without being given any clinical information (equivalent to Stage I). However, this does not occur in clinical settings,

which is why we prepared this intermediate stage. In Stage III, they were further provided with the ALMnet diagnosis (either BN or ALM) and the confidence score for each case, and then asked for their answer. The suggested diagnosis was determined by a dichotomous classification (BN versus ALM), produced by interpreting the prediction output with a fixed cut-off (> 0.5 favours ALM). The performance of participants was assessed by comparing their predictions and associated decisions to the diagnostic gold standard. At each stage, the sequences of cases were shuffled to ensure unbiased responses at each stage. The reference diagnosis for each case and participants' scores were not disclosed until the end of the study.

Outcome Measurement and Statistical Analysis

The area under curve (AUC) of the receiver operating characteristics (ROC) curve was taken as the primary measurement for comparing the performances of multiple dCNNs. ROC curves were depicted using the prediction output for dichotomous classification of BN and ALM. Two-tailed independent t-tests were used to compare the AUC values, which measured 5 repeats for each of the 4 models. For the model with the highest performance, a repeat with the median AUC was adopted as ALMnet.

The primary outcome of the decision study was the difference in accuracy across all human evaluators between Stages I, II, and III. The secondary outcomes include the differences in sensitivity and specificity. For human-set-100, participant accuracies, sensitivities, and specificities were calculated for each stage. Sensitivity was calculated as the proportion of correctly classified ALM images, and specificity was calculated as the proportion of correctly classified BN images. Repeated measure analysis of variance was used to compare the performance of participants across the physician groups and study stages.

The exact Clopper-Pearson method[22] was used for calculating the 95% confidence intervals (CIs) of the dCNN performance parameters. Fleiss-κ statistics[23] were calculated to assess concordance among participants' responses. A heatmap was used to visualize inter-participant agreement rate, with clustering via a hierarchical agglomerative method. All statistical analyses were conducted in R version 3.5.0 (R Foundation for Statistical Computing) at a significance level of 5%. Four R packages were used: ROCit for plotting ROC curves,[24] IRR[25] and raters[26] for calculating κ-statistics, and pheatmap[27] to produce the heatmap.

Results

Performance of Deep Convolutional Neural Networks

The performances of four dCNNs for test-set-200 in ALM detection are summarized in Table S2 and Fig. S6. The ensembled model (Model 4) showed the best performance, with a highest AUC value of 0.976 (95% CI, 0.974–0.978), significantly higher than those of the other models. Model 2 in which intermediate tumours were trained as BN showed higher specificity compared to Model 4. However, a loss in the sensitivity resulted in a sub-optimal AUC and accuracy. On the contrary, Model 3 where these intermediate tumours were trained as ALM showed a higher sensitivity compared to Model 4, but the overall AUC and accuracy was sub-optimal owing to the loss of specificity. From the 5 repeats for Model 4, the median performance (AUC of 0.976) was adopted as ALMnet. With a cut-off value of 0.5 for predictive output of ALM classification, ALMnet demonstrated an accuracy of 92.5% (95% CI, 87.9%–95.7%), a sensitivity of 90.0% (95% CI, 82.4%–95.1%), and a specificity of 95.0% (95% CI, 88.7%–98.4%). For human-set-100, it demonstrated an accuracy of 94.0% (95% CI, 87.4%–97.8%), a sensitivity of 92.0% (95% CI, 80.8%–97.8%), and a specificity of 96.0% (95% CI, 86.3%–99.5%).

Performance of Human Evaluators in Stages I and II

A total of 60 participants were provided with 100 dermoscopic images (human-set-100) and made initial decisions for each case. The performances of the participants and ALMnet are summarized in Table 1 and Figs 2 and 3. The interrater agreement rates between participant responses, ALMnet predictions, and the diagnostic gold standard are presented in Fig. 4.

In Stage I, participants showed an accuracy of 74.7% (95% CI, 72.6%–76.8%), a sensitivity of 79.9% (95% CI, 76.2%–83.5%), and a specificity of 69.5% (95% CI, 65.1%–73.8%) (Table 1I). Despite some exceptions, most participants were outperformed by ALMnet (Fig. 2). Moreover, a prominent disagreement (Fig. 4A, Fleiss-κ of 0.436 [95% CI, 0.370–0.503]) and performance gap (Fig. S7A) were observed across participants.

In Stage II, participants were given dermoscopic images along with each patient's age, sex, and site of lesion. The accuracy, sensitivity, and specificity were 79.0% (95% CI, 76.7%–81.2%),

81.5% (95% CI, 77.7%–85.2%), and 76.4% (95% CI, 72.5%–80.4%), respectively. The differences compared to Stage I were 4.3%p (95% CI, 2.5%p–6.1%p), 1.6%p (95% CI, -1.0%p to 4.2%p), and 7.0%p (95% CI, 3.2%p–12.4%p), respectively. Although the availability of clinical information increased participants' overall performances, many were still outperformed by ALMnet. Regarding interrater agreement, we see that despite a slightly improved concordance between participants, a considerable disagreement was still observed (Fig. S7B and Fig. 4B, Fleiss-κ of 0.506 [95% CI, 0.437–0.573]).

Augmented Decision Making of Human Evaluators in Stage III

In Stage III, participants were further provided with the diagnosis and its probability, as predicted by ALMnet for each case. Compared to Stage II, 811/6000 (13.5%) of participant responses were changed as a result of augmented decision making, and 643/811 (79.3%) of these resulted in increased accuracy (Fig. S8). The accuracy, sensitivity, and specificity were 86.9% (95% CI, 85.3%–88.4%), 88.7% (95% CI, 86.0%–91.5%), and 85.0% (95% CI, 82.7%–87.3%), respectively. The differences compared to Stage I were 12.2%p (95% CI, 10.1%p–14.3%p), 8.9%p (95% CI, 6.1%p–11.7%p), and 15.5%p (95% CI, 11.4%p–19.7%p), respectively. Similarly, those compared to Stage II were 7.9%p (95% CI, 6.0%p–9.9%p), 7.3%p (95% CI, 4.5%p–10.0%p), and 8.6%p (95% CI, 5.3%p–11.8%p), respectively, demonstrating a significant improvement in participants' performances. This improvement was particularly emphasized in the relatively inexperienced groups of dermatology residents and general physicians. Moreover, the performance gap across participants was considerably diminished, and the overall agreement across participants greatly improved (Fig. S7C and Fig. 4C, Fleiss-κ of 0.684 [95% CI, 0.621–0.749]).

Discussion

Main Findings

This study demonstrated that augmented decision making with the dCNN can improve the quality of clinical decisions and minimize discrepancies between human experts. The accuracy of the initial decision making of 60 participants was increased significantly by 7.9%p (95% CI, 6.0%p–

9.9%p) and the interrater reliability across participants was considerably improved with the predictions derived from the dCNN. In terms of optimizing the algorithms, the dCNN showed the best performance in ALM detection when it took the form of an ensembled output of two models, in which intermediate tumours were considered as BN and ALM, and trained, respectively.

Deep Learning Approaches to Melanoma Detection

The detection of malignant melanoma is one of the tasks in which the dCNN achieves the highest performance. A few studies have reported upon deep learning approaches to detect malignant melanoma and have compared their approach performances to those of human experts. Han et al.[28] reported a dCNN with an AUC of 0.96 when using clinical photographs, which is a similar performance to that of a dermatologist. Esteva et al.[29] reported AUCs of 0.94 and 0.91 when analysing clinical photographs and dermoscopy images, respectively. Similarly, Haenssle et al.[30] reported a dCNN with an AUC of 0.95 for dermoscopy images, outperforming dermatologists, who achieved an AUC of 0.86. In the aforementioned studies, dCNNs consistently showed a superior performance over human experts in interpreting dermoscopy images, and our results support these findings. This may be because, even for dermatologists, dermoscopy images are unfamiliar. Therefore, applying deep learning approaches to the extraction of diagnostic features from dermoscopy images would be of greater benefit than applying it to clinical photographs.

Our model demonstrated fair performance despite relatively limited data. We hypothesize that the following factors contributed to the model's performance. i) Only images taken with a single dermoscopy model and magnification were used. It is important to develop algorithms that can consistently handle various heterogeneous inputs for generalizability, however the benefits of data homogenization would be greater for rare conditions such as ALM. ii) We optimized the training and inferential method by introducing an image pre-processing and augmentation technique, tailored to the characteristics of the images taken at the acral sites. iii) Intermediate tumours, such as atypical or dysplastic nevi, were trained in an ensemble manner. These two models showed a lower performance compared to other models, suggesting that they cannot be entirely morphologically interpreted as either BN or ALM, and have mixed and ambivalent features. Additional studies using more data would be required to determine if they should be trained as a distinctive category for optimal performance.

Clinical Implementation as a Decision-support System

Several studies have reported the high diagnostic accuracy of deep learning approaches compared to those of human experts. Despite the expert-level performance of the deep learning models, there are still many issues to be solved before they can be implemented as a decision-supporting system, including uncertainty and unactionable problems.[31-34] Moreover, over-reliance on automated diagnosis can lead to a critical and fatal outcome for false negative findings.

Because of the limitations of the current models, there would be a considerable disparity between the predictions of the models and the final decisions of physicians. Therefore, it is necessary to investigate how deep learning algorithms can change patterns of physician decision making, and how much these algorithms can improve their eventual performances. However, there have been only few studies describing how these approaches eventually affect the clinical decisions of physicians.[35,36] Cho et al. reported a significant increase in the accuracy of less experienced physicians (non-dermatologists) in detection of malignant lip diseases with the aid of algorithms.[35] Similarly, Bien et al. reported similar results in which clinical experts achieved a better detection of diverse knee abnormalities and injuries, by analysing magnetic resonance images with algorithms.[36] The findings of our study also indicate the potential of current deep learning-based models as adjoining support systems for physicians' own decisions. With augmented decision making, not only the overall performance of physicians, but also the concordance in clinical decisions between diverse groups of experts, can be increased, minimizing the inter-professional practice gap in detection of acral melanocytic lesions.

Strengths and Limitations

The main strength of this study is the implementation of augmented decision making via a dCNN to improve the overall diagnosis performance of a relatively large group of 60 physicians, including both dermatologists and general physicians, this has been scantly reported in previous studies. Moreover, although this study included less data than studies analysing more common diseases, our data was acquired from a tertiary referral centre that, as a single institution, manages the largest number of patients with ALM.

There are some limitations of this study. The dermoscopic images used in this study were taken with only a single model camera. To achieve consistent performance for images taken in different environments, an additional training step with a larger dataset would be required. In addition, the performance of human participants may have been underestimated in this experimental setting. Because dermoscopy is an interactive diagnostic tool, in a clinical setting, physicians would have been possible to extract more diagnostic features using various windows and media.

In conclusion, this study demonstrated that the dCNN can detect ALM by interpreting dermoscopy images and can achieve a comparable or even superior performance than most human experts. In addition, this study showed the benefit of augmented decision making using the dCNN. A considerable improvement in the performance and concordance of clinical decisions across a large group of 60 physicians was shown, suggesting its potential use as a decision supporting system. Moreover, these results would suggest its considerable benefit not only in detection of ALM but also in detection of various diseases, using other imaging sources. A prospective clinical trial would be required to investigate the efficacy of its usage in real clinical settings, and to see if it eventually improves patients' outcomes.

Acknowledgments

We would like to thank all the physicians who participated in the online survey, without whom this study would not have been possible.

References

1. Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. *Lancet* (London, England). 2018;392(10162): 2388-2396.
2. Li X, Zhang S, Zhang Q, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *Lancet Oncol*. 2019;20(2): 193-201.
3. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women with Breast Cancer. *JAMA*. 2017;318(22): 2199-2210.
4. Ting DSW, Cheung CY, Lim G, et al. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multi-ethnic Populations with Diabetes. *JAMA*. 2017;318(22): 2211-2223.
5. Karimkhani C, Dellavalle RP, Coffeng LE, et al. Global Skin Disease Morbidity and Mortality: An Update from the Global Burden of Disease Study 2013. *JAMA Dermatol*. 2017;153(5): 406-412.
6. Jemal A, Saraiya M, Patel P, et al. Recent trends in cutaneous melanoma incidence and death rates in the United States, 1992-2006. *J Am Acad Dermatol*. 2011;65(5 Suppl 1): S17-25.e11-13.
7. Nikolaou V, Stratigos AJ. Emerging trends in the epidemiology of melanoma. *Br J Dermatol*. 2014;170(1): 11-19.
8. Bradford PT, Goldstein AM, McMaster ML, Tucker MA. Acral lentiginous melanoma: incidence and survival patterns in the United States, 1986-2005. *Archives of Dermatology*. 2009;145(4): 427-434.
9. Chang JW. Acral melanoma: a unique disease in Asia. *JAMA Dermatol*. 2013;149(11): 1272-1273.
10. Oh BH, Lee SH, Nam KA, Lee HB, Chung KY. Comparison of negative pressure wound therapy and secondary intention healing after excision of acral lentiginous melanoma on the foot. *Br J Dermatol*. 2013;168(2): 333-338.
11. Bafounta ML, Beauchet A, Aegerter P, Saiag P. Is dermoscopy (epiluminescence microscopy) useful for the diagnosis of melanoma? Results of a meta-analysis using

techniques adapted to the evaluation of diagnostic tests. *Archives of Dermatology*. 2001;137(10): 1343-1350.

12. 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(3): 669-676.
13. Salerni G, Teran T, Puig S, et al. Meta-analysis of digital dermoscopy follow-up of melanocytic skin lesions: a study on behalf of the International Dermoscopy Society. *J Eur Acad Dermatol Venereol*: JEADV. 2013;27(7): 805-814.
14. Kittler H, Pehamberger H, Wolff K, Binder M. Diagnostic accuracy of dermoscopy. *The Lancet Oncol*. 2002;3(3): 159-165.
15. Marino ML, Carrera C, Marchetti MA, Marghoob AA. Practice Gaps in Dermatology: Melanocytic Lesions and Melanoma. *Dermatol Clin*. 2016;34(3): 353-362.
16. J. Eduardo Calonje TB, Alexander J Lazer, Steven D Bbillings. *McKee's Pathology of the Skin*. 5th edition ed. Amsterdam, Netherland: Elsevier; 2019.
17. Varnousfaderani ES, Yousefi S, Belghith A, Goldbaum MH. Luminosity and contrast normalization in color retinal images based on standard reference image. Paper presented at: Medical Imaging 2016: Image Processing 2016.
18. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016.
19. Yu C, Yang S, Kim W, et al. Acral melanoma detection using a convolutional neural network for dermoscopy images. *PLoS One*. 2018;13(3): e0193321.
20. Perez F, Vasconcelos C, Avila S, Valle E. Data augmentation for skin lesion analysis. In: *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Springer; 2018: 303-311.
21. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. Paper presented at: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016.
22. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934;26(4): 404-413.
23. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychological bulletin*. 1971;76(5): 378.

24. Khan MRAA. ROCit-An R Package for Performance Assessment of Binary Classifier with Visualization. 2019.
25. Gamer M, Lemon J, Gamer MM, Robinson A, Kendall's W. Package 'irr'. *Various coefficients of interrater reliability and agreement*. 2012.
26. Ripamonti PQaE. raters: A Modification of Fleiss' Kappa in Case of Nominal and Ordinal Variables. https://CRAN.R-project.org/package=raters. Published 2014. Accessed May, 20, 2018.
27. Kolde R, Kolde MR. Package 'pheatmap'. *R Package*. 2015;1(7).
28. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. *J Invest Dermatol*. 2018;138(7): 1529-1538.
29. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639): 115-118.
30. Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol*. 2018;29(8): 1836-1842.
31. Begoli E, Bhattacharya T, Kusnezov D. The need for uncertainty quantification in machine-assisted medical decision making. *Nat Mach Intell*. 2019;1(1): 20.
32. Shortliffe EH, Sepulveda MJ. Clinical Decision Support in the Era of Artificial Intelligence. *JAMA*. 2018;320(21): 2199-2200.
33. Maddox TM, Rumsfeld JS, Payne PRO. Questions for Artificial Intelligence in Health Care. *JAMA*. 2019;321(1): 31-32.
34. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1): 44-56.
35. Cho SI, Sun S, Mun J, et al. Dermatologist-level classification of malignant lip diseases using a deep convolutional neural network. *Br J Dermatol*. 2019. 2019 Aug 26. doi: 10.1111/bjd.18459. [Epub ahead of print]
36. Bien N, Rajpurkar P, Ball RL, et al. Deep learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med*. 2018;15(11): e1002699.

Figure Legends

Figure 1. Schematic Illustration of the Study Design and Flow

A decision study was performed to investigate the performance of physicians making their initial diagnosis of ALM by interpreting dermoscopic images (Stage I), images with accompanying clinical information (Stage II), and images with the aid of a deep neural network (Stage III). Human-set-100 was generated by extracting 100 dermoscopic images from test-set-200 and was distributed to the human participants through Google Survey (Fig. S5). This online-based survey was conducted over 2-week intervals, and the sequences of cases were shuffled for participants' independent responses. The diagnostic gold standard (histopathological findings) for each case and participants' scores were not disclosed until the end of the study.

Abbreviations: BN, benign nevus; ALM, acral lentiginous melanoma.

Figure 2. Receiver Operating Characteristics Curve Depicting the Performances of ALMnet and the 60 Human Evaluators in the Decision Study

The receiver operating characteristics curve represents the performance of ALMnet for dichotomous classification of BN and ALM images for human-set-100. The dots (points) represent sensitivity and (1- specificity) of the human participants in each stage (grey dots in Stage I, black dots in Stage II, and coloured dots in Stage III). In Stages I and II, most participants were outperformed by ALMnet. The crosses represent the averaged sensitivity and (1 - specificity) of human participants in each stage. The differences between those in Stage III and those in Stage II were considerably larger than the differences between those in Stage II and those in Stage I.

(a) All human participants.

(b) Board-certified dermatologists.

(c) Dermatology residents.

(d) General physicians.

Abbreviations: BN, benign nevus; ALM, acral lentiginous melanoma.

Figure 3. Improvement in Performance of Human Evaluators with Augmented Decision Making using the Deep Neural Network

The differences in the performances of human participants in Stages II and III, in comparison to those in Stage I (baseline). The improvement was more prominent in the relatively inexperienced groups (dermatology residents and general physicians) than in the board-certified dermatologist group.

(a) Accuracy

(b) Sensitivity

(c) Specificity

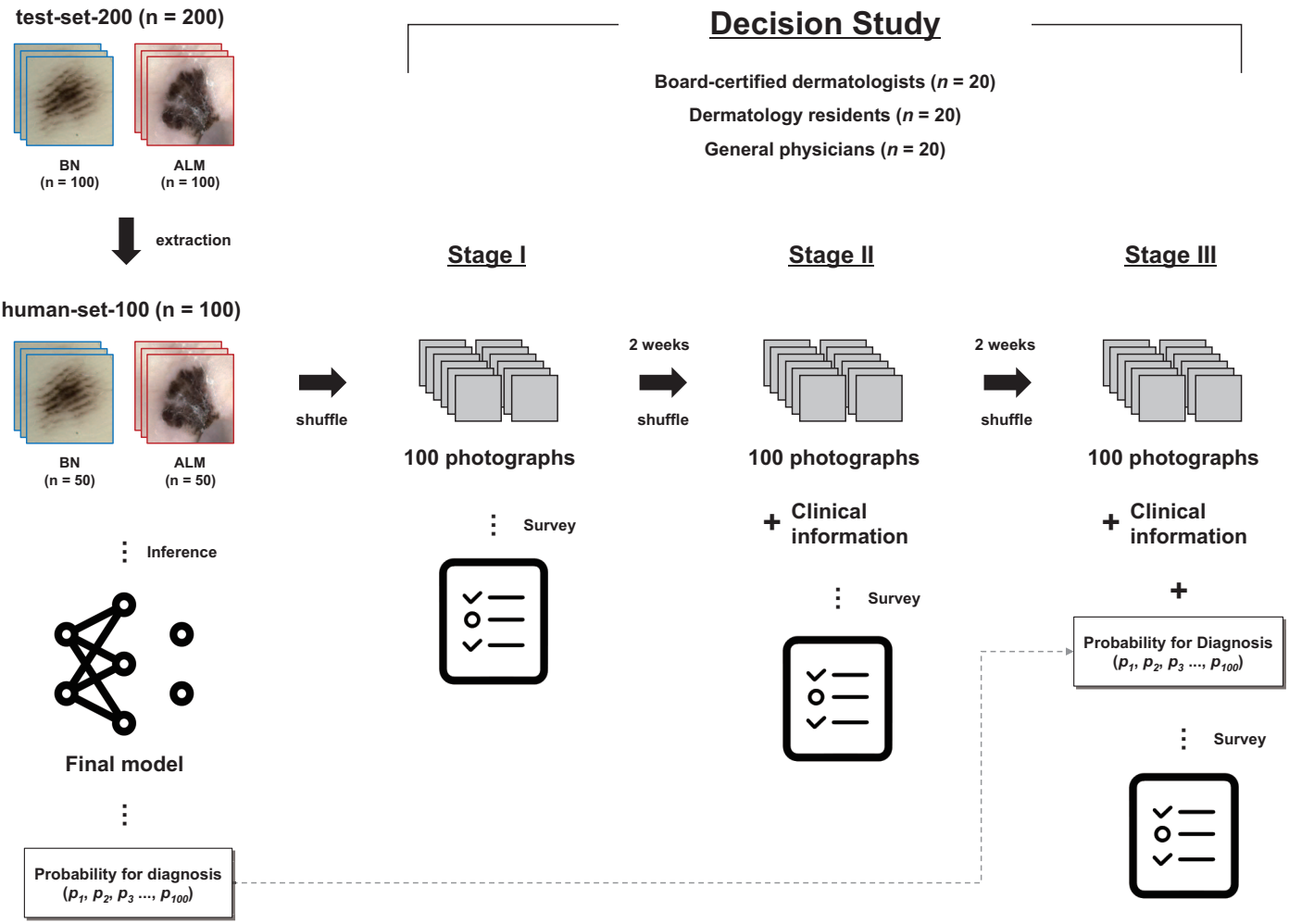
Figure 4. Interrater Agreement and Reliability Among Human Evaluators

The heatmaps here visualize the interrater agreement rate across the responses of 60 human participants, the prediction of ALMnet, and the histopathological diagnosis. Comparing (a) Stage I and (b) Stage II, a considerable improvement can be seen in the concordance across the evaluators in (c) Stage III. The Fleiss-κ statistics and their 95% confidence intervals for all participants, and within each physician group, are represented along with the heatmap to quantify the interrater reliability

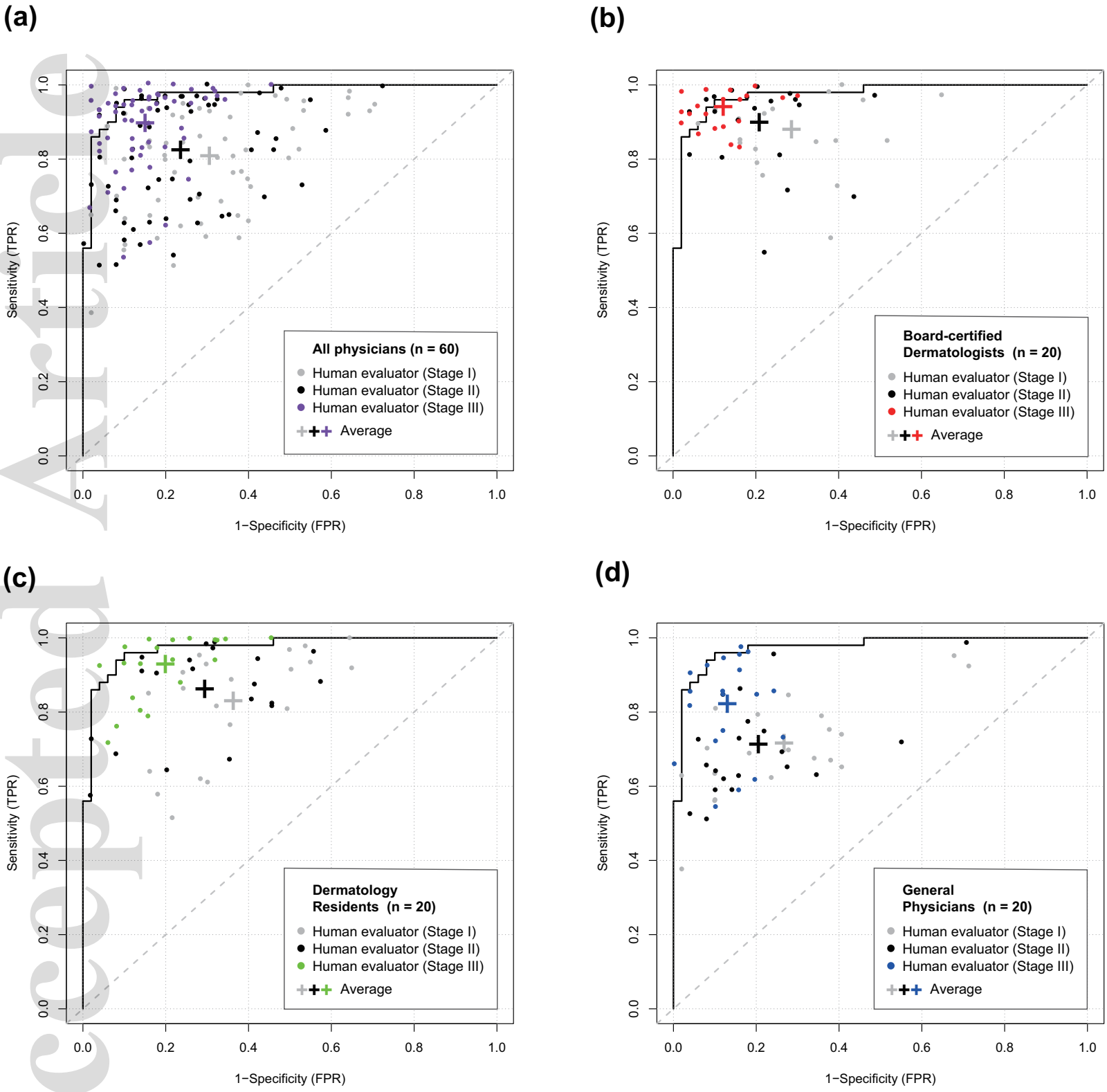
Table 1. Performances of Human Evaluators

Parameter	Participant type	Performance, % (95% CI)			Difference, %p (95% CI)		
		Stage I	Stage II	Stage III	Stage (II – I)	Stage (III – I)	Stage (III – II)
Primary outcome							
Accuracy, % (95% CI)	**All physicians (*n* = 60)**	74.7 (72.6 to 76.8)	79.0 (76.7 to 81.2)	86.9 (85.3 to 88.4)	4.3 (2.5 to 6.1)	12.2 (10.1 to 14.3)	7.9 (6.0 to 9.9)
	Board-certified dermatologists (*n* = 20)	79.2 (75.1 to 83.3)	84.1 (80.2 to 87.9)	90.5 (88.7 to 92.3)	4.9 (2.6 to 7.1)	11.3 (7.6 to 15.0)	6.5 (3.4 to 9.5)
	Dermatology residents (*n* = 20)	72.9 (70.0 to 75.7)	77.9 (77.4 to 81.4)	86.0 (83.9 to 88.1)	5.1 (1.5 to 8.6)	13.2 (10.1 to 16.2)	8.1 (5.3 to 10.9)
	General physicians (*n* = 20)	72.0 (69.0 to 74.9)	74.9 (71.7 to 78.1)	84.1 (81.0 to 87.2)	3.0 (-0.3 to 6.2)	12.2 (8.0 to 16.3)	9.2 (5.3 to 13.1)
Secondary outcomes							
Sensitivity, % (95% CI)	**All physicians**	79.9 (76.2 to 83.5)	81.5 (77.7 to 85.2)	88.7 (86.0 to 91.5)	1.6 (-1.0 to 4.2)	8.9 (6.1 to 11.7)	7.3 (4.5 to 10.0)
	Board-certified dermatologists	87.0 (82.6 to 91.4)	88.9 (83.7 to 94.1)	93.1 (91.1 to 95.1)	1.9 (-0.9 to 4.7)	6.1 (2.8 to 9.4)	4.2 (0.2 to 8.2)
	Dermatology residents	82.0 (75.6 to 88.4)	85.2 (79.8 to 90.6)	91.9 (88.1 to 95.7)	3.2 (-0.6 to 7.0)	9.9 (6.3 to 13.5)	6.7 (3.0 to 10.4)
	General physicians	70.6 (64.9 to 76.3)	70.3 (64.7 to 75.9)	81.2 (75.6 to 86.8)	-0.3 (-6.3 to 5.7)	10.6 (3.9 to 17.3)	10.9 (4.9 to 16.9)

Specificity, % (95% CI)	**All physicians**	69.5 (65.1 to 73.8)	76.4 (72.5 to 80.4)	85.0 (82.7 to 87.3)	7.0 (3.2 to 10.7)	15.5 (11.4 to 19.7)	8.6 (5.3 to 11.8)
	Board-certified dermatologists	71.4 (64.7 to 78.1)	79.2 (74.1 to 84.3)	87.9 (84.6 to 91.2)	7.8 (3.2 to 12.4)	16.5 (10.0 to 23.0)	8.7 (4.2 to 13.2)
	Dermatology residents	63.7 (57.0 to 70.4)	70.6 (63.4 to 77.8)	80.1 (75.4 to 84.8)	6.9 (-0.7 to 7.0)	16.4 (9.9 to 22.9)	9.5 (4.7 to 14.3)



jdv_16185_f1.eps



jdv_16185_f2.eps