Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images

Michael A. Marchetti, MD,[a] Noel C. F. Codella, PhD,[b] Stephen W. Dusza, DrPH,[a]
David A. Gutman, MD, PhD,[c] Brian Helba,[d] Aadi Kalloo, MHS,[a] Nabin Mishra, PhD,[e]
Cristina Carrera, MD, PhD,[f] M. Emre Celebi, PhD,[g] Jennifer L. DeFazio, MD,[a] Natalia Jaimes, MD,[h,i]
Ashfaq A. Marghoob, MD,[a] Elizabeth Quigley, MD,[a] Alon Scope, MD,[a,j] Oriol Yélamos, MD,[a]
and Allan C. Halpern, MD,[a] for the International Skin Imaging Collaboration
New York, Yorktown Heights, and Clifton Park, New York; Atlanta, Georgia; Rolla, Missouri; Barcelona, Spain; Conway, Arkansas; Medellín, Colombia; Miami, Florida; and Tel Aviv, Israel

Background: Computer vision may aid in melanoma detection.

Objective: We sought to compare melanoma diagnostic accuracy of computer algorithms to dermatologists using dermoscopic images.

Methods: We conducted a cross-sectional study using 100 randomly selected dermoscopic images (50 melanomas, 44 nevi, and 6 lentigines) from an international computer vision melanoma challenge dataset (n = 379), along with individual algorithm results from 25 teams. We used 5 methods (nonlearned and machine learning) to combine individual automated predictions into "fusion" algorithms. In a companion study, 8 dermatologists classified the lesions in the 100 images as either benign or malignant.

Results: The average sensitivity and specificity of dermatologists in classification was 82% and 59%. At 82% sensitivity, dermatologist specificity was similar to the top challenge algorithm (59% vs. 62%, P = .68) but lower than the best-performing fusion algorithm (59% vs. 76%, P = .02). Receiver operating characteristic

From the Dermatology Service,[a] Department of Medicine, Memorial Sloan Kettering Cancer Center, New York; IBM Research Division,[b] Thomas J. Watson Research Center, Yorktown Heights; Departments of Neurology, Psychiatry, and Biomedical Informatics,[c] Emory University School of Medicine, Atlanta; Kitware Inc,[d] Clifton Park; Stoecker & Associates,[e] Rolla; Melanoma Unit,[f] Department of Dermatology, Hospital Clinic, Institut d'Investigacions Biomèdiques August Pi i Sunyer, CIBER de Enfermedades Raras, Instituto de Salud Carlos III, University of Barcelona; Department of Computer Science,[g] University of Central Arkansas, Conway; Dermatology Service,[h] Aurora Centro Especializado en Cáncer de Piel, Medellín; Department of Dermatology and Cutaneous Surgery,[i] University of Miami Miller School of Medicine, Miami; and the Department of Dermatology,[j] Sheba Medical Center, Sackler School of Medicine, Tel Aviv University.

Drs Marchetti and Codella contributed equally to this article.

Supported in part by the National Institutes of Health/National Cancer Institute Cancer Center Support Grant P30 CA008748.

Disclosure: Dr Codella is an employee of IBM and an IBM stockholder. Dr Scope is a consultant to Emerald Inc. Drs Marchetti, Dusza, Halpern, Marghoob, DeFazio, Yélamos, Carrera, Jaimes, Mishra, Kalloo, Quigley, Gutman, Helba, and Celebi have no conflicts of interest to declare.

Preliminary data of this study were presented at the Annual Meeting of the Society for Melanoma Research Congress, November 8, 2016, Boston, Massachusetts.

Accepted for publication August 7, 2017.

Correspondence to: Allan C. Halpern, MD, Dermatology Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, 16 E 60th St, New York, NY 10022. E-mail: halperna@mskcc.org.

Published online September 29, 2017.

0190-9622/$36.00

http://dx.doi.org/10.1016/j.jaad.2017.08.016

area of the top fusion algorithm was greater than the mean receiver operating characteristic area of dermatologists (0.86 vs. 0.71, $P = .001$).

Limitations: The dataset lacked the full spectrum of skin lesions encountered in clinical practice, particularly banal lesions. Readers and algorithms were not provided clinical data (eg, age or lesion history/symptoms). Results obtained using our study design cannot be extrapolated to clinical practice.

Conclusion: Deep learning computer vision systems classified melanoma dermoscopy images with accuracy that exceeded some but not all dermatologists. (J Am Acad Dermatol 2018;78:270-7.)

Key words: computer algorithm; computer vision; dermatologist; International Skin Imaging Collaboration; International Symposium on Biomedical Imaging; machine learning; melanoma; reader study; skin cancer.

CAPSULE SUMMARY

- Computer vision has shown promise in medical diagnosis.
- A machine learning fusion algorithm using predictions from 16 algorithms exceeded the performance of most dermatologists in the classification of 100 dermoscopic images of melanomas and nevi.
- These results should not be extrapolated to clinical practice until validation in prospective studies.

The early diagnosis of melanoma remains challenging.[1] Estimates of the sensitivity of dermatologists for melanoma in reader studies were 70% for the Nevisense trial[2] and 78% for the MelaFind trial.[3] In addition, because nonphysicians detect the majority of melanomas[4] and because population-based melanoma screening by clinicians is not recommended in the United States,[5] there is not only interest in the development of automated image analysis algorithms to help dermatologists classify dermoscopic images, but also to aid laypersons or nondermatology physicians in melanoma detection.[6-13] To date, the lack of a large, public dataset of skin images has limited the ability to directly compare the diagnostic performance of competing automated image analysis approaches against clinicians.

To address this limitation, the International Skin Imaging Collaboration (ISIC) Melanoma Project created an open-access archive of dermoscopic images of skin lesions for education and research.[14] We describe the melanoma classification results from a challenge conducted by the ISIC Archive[15] at the 2016 International Symposium on Biomedical Imaging (ISBI) involving 25 competing teams.[16] We further performed a companion reader study with 8 experienced dermatologists on a subset of images; these results served as a reference comparator to the automated algorithm approaches.

MATERIALS AND METHODS

Institutional review board approval

Institutional review board approval was obtained at Memorial Sloan Kettering and the study was conducted in accordance with the Helsinki Declaration.

ISBI 2016 melanoma detection challenge dataset

Details of the challenge tasks, evaluation criteria, timeline, and participation have been previously described.[15,17,18] In December 2015, 1552 lesions were chosen from ~12,000 dermoscopic images in the ISIC Archive; after excluding 273 for inadequate image quality, 1279 lesions (248 [19.3%] melanomas and 1031 [80.7%] nevi or lentigines) were included. Images were excluded because of poor focus or if they included multiple lesions or lesions that encompassed the entire field of view. The dataset was randomly divided into training (n = 900 [19.2% melanomas]) and test (n = 379 [19.8% melanomas]) sets. All melanomas and a majority of the nevi/lentigines (n = 869, 84%) had been histopathologically examined. Nonhistopathologically examined nevi (n = 162) originated from a longitudinal study of children; selection from this dataset was biased to include lesions with the largest diameters, and all images were reviewed by ≥2 dermatologists to confirm their benign nature.[19] Images used in this challenge were obtained with multiple camera/dermatoscope combinations and originated from >12 dermatology clinics around the world.

Twenty-five teams participated in the challenge, all of which used deep learning, a form of machine learning that uses multiple processing layers to automatically identify increasingly abstract concepts present in data. Computer algorithms were ranked using average precision, which corresponds to the

Abbreviations used:

ISBI:	International Symposium on Biomedical Imaging
ISIC:	International Skin Imaging Collaboration
ROC:	receiver operating characteristic
SVM:	support vector machine

integral under a precision-recall curve (which depicts positive predictive value [ie, the proportion of positive results that are true positives] and sensitivity [ie, the proportion of positive results that are correctly identified]), and the final results and rankings are publically available.[15,18]

Reader study

A reader study was performed on 50 randomly selected melanomas (31 invasive, 19 in situ) and 50 benign neoplasms (44 nevi, 6 lentigines) from the 379 test images. Nonhistopathologically examined benign lesions were excluded from this image set. The median (range) Breslow depth for the invasive melanomas was 0.70 mm (0.10-2.06 mm). Eight experienced dermatologists from 4 countries were invited on May 13, 2016, and all agreed to participate. The mean (range) number of years of (1) postresidency clinical experience and (2) use of dermoscopy among readers was 13 years (range, 3-31 years) and 13.5 years (range, 6-27 years), respectively, and all had a primary clinical focus on skin cancer. For each dermoscopic image, readers: (1) classified the lesion (benign vs. malignant) and (2) indicated management (obtaining a biopsy specimen or observation/reassurance). Readers were blinded to diagnosis and clinical images/metadata. There were no time restrictions and participants could complete evaluations over multiple sittings.

Automated predictions

We report the performance of the 5 top-ranked individual algorithms of the ISBI 2016 Challenge on the reader set of 100 dermoscopic images. In addition, we implemented 5 methods of fusing all automated predictions from the 25 participating teams in the ISBI challenge into a single prediction. These methods included 2 nonlearned approaches (prediction score averaging and voting) and 3 machine learning methods: greedy ensemble fusion,[20] linear binary support vector machine (SVM), and nonlinear binary SVM (histogram intersection kernel) (Supplemental materials; available at http://www.jaad.org).[21] Test set images that were not involved in the reader study (n = 279) were used to train fusion methods; fusion algorithms were ranked by average precision on the reader set of 100 images.

Statistical analysis

The primary outcomes and measures were sensitivity, specificity, and area under the receiver operating characteristic (ROC) curves. Sensitivity in classification was defined as the percentage of melanomas that were correctly scored as malignant. Sensitivity in management decision was defined as the percentage of melanomas for which obtaining a biopsy specimen was correctly indicated. Specificity in classification was defined as the percentage of benign lesions that were correctly scored as benign. Specificity in management decision was defined as the percentage of benign lesions that were correctly indicated for observation/reassurance.

Computers submitted predictions between 0.0 and 1.0, with 0.5 used as a dichotomous threshold in the ISBI Challenge: values $\leq$0.5 were benign and values $>$0.5 to 1.0 were malignant. For analyses here, we considered scores closer to 0 to indicate a higher probability of a benign diagnosis and scores closer to 1 to indicate a higher probability of malignancy.

As ground truth data provided to participants in the ISBI 2016 challenge was restricted to classification (benign vs. malignant) and did not include management data (obtaining a biopsy specimen vs. observation/reassurance), we chose classification performance as the primary outcome. To inform clinical practice, however, we also compared management decisions of dermatologists to computer classification performance as an exploratory outcome; another rationale for reporting management decision performance was that most studies comparing human readers to computers have used management decision, and not classification, as the primary outcome. Our primary comparison between readers and computers was specificity at average dermatologist sensitivity; the secondary comparison between readers and computers was ROC area of the algorithm and mean ROC area of the dermatologists.

Descriptive statistics, such as relative frequencies, means, and standard deviations were used to describe the dermatologist lesion classifications and management decisions for each evaluation. Overall percent agreement, kappa, and intraclass correlation were used to evaluate reader responses for lesion classification and management. Levels of interrater agreement were evaluated as percent agreement and multirater kappa. In addition, patterns of agreement for the dermatologist assessments were evaluated on the lesion level, where lesions were classified as having unanimous agreement between readers or not.

Table I. Reader results

	Classification			Management		
Reader no.	**Sensitivity**	**Specificity**	**ROC area**	**Sensitivity**	**Specificity**	**ROC area**
1	68%	72%	0.70	72%	68%	0.70
2	68%	66%	0.67	86%	40%	0.63
3	98%	54%	0.76	100%	38%	0.69
4	86%	62%	0.74	88%	56%	0.72
5	88%	34%	0.61	90%	30%	0.60
6	74%	68%	0.71	76%	66%	0.71
7	82%	58%	0.70	100%	32%	0.66
8	92%	60%	0.76	96%	48%	0.72
Average	82%	59%	0.71	89%	47%	0.68

ROC, Receiver operating characteristic.

Table II. Results of the International Symposium on Biomedical Imaging Challenge top 5 individual algorithms and fusion algorithms on the reader study dataset on 100 images evaluated by dermatologists

Algorithm	Sensitivity	Specificity	Specificity at 82% sensitivity	Specificity at 89% sensitivity	ROC1 at 82% sensitivity*	ROC1 at 89% sensitivity*	ROC2†	Average precision‡
Rank 1	52%	92%	0.62	0.38	0.72	0.64	0.79	0.84
Rank 2	60%	80%	0.56	0.40	0.69	0.65	0.80	0.83
Rank 3	36%	96%	0.48	0.34	0.65	0.62	0.79	0.81
Rank 4	68%	84%	0.50	0.38	0.66	0.64	0.80	0.83
Rank 5	26%	100%	0.60	0.52	0.71	0.71	0.81	0.84
Average fusion	46%	92%	0.78	0.66	0.80	0.78	0.86	0.86
Voting fusion	56%	90%	0.82	0.60	0.82	0.75	0.86	0.86
Greedy fusion	58%	92%	0.76	0.64	0.79	0.77	0.86	0.87
Linear SVM fusion	66%	86%	0.68	0.42	0.75	0.66	0.82	0.85
Nonlinear SVM fusion	70%	88%	0.68	0.34	0.75	0.62	0.82	0.86

ROC, Receiver operating characteristic; *SVM*, support vector machine.
*Based on dichotomizing data for response.
†Based on using continuous probability generated from algorithm.
‡Integral under a precision-recall curve (or positive predictive value—sensitivity curve).

Levels of classification and management accuracy were calculated for each individual reader and for the readers as a group. To describe the study sample, and to provide comparisons between measures of diagnostic performance between readers and algorithms, 2-sample tests for proportions along with chi-square tests were used. In addition, regression models for binary outcomes using a general estimating equations approach with a log link and an exchangeable correlation structure were used. In these models, readers were considered a covariate, allowing for between-reader comparisons of accuracy and stratified analyses. The exchangeable correlation structure was used to adjust the standard error estimates for the potential of clustered observations within readers. In addition, ROC curves were estimated for the individual readers and for the readers as a group. Comparisons of area under the ROC curves were performed to assess differences in reader performance and to make comparisons between the reader and algorithm performance. For dichotomous predictions, area under ROC curves is equivalent to the average of sensitivity and specificity. Alpha level was set at 5% and all presented *P* values are 2-sided. All analyses were performed with Stata SE software (version 14.1; Stata Corp, College Station, TX).

RESULTS

Diagnostic accuracy of dermatologists for melanoma

The average (min-max) sensitivity and specificity of the 8 readers for lesion classification (ie, benign vs. malignant) was 82% (68-98%) and 59% (34-72%), respectively (Table I). This corresponded to an average (min-max) ROC area of 0.71 (0.61-0.76). The average (min-max) sensitivity for melanoma in situ and invasive melanoma was 68.4% (53-95%) and 89.1% (75-100%), respectively. Data describing levels of agreement among dermatologists are included in the Supplementary materials.

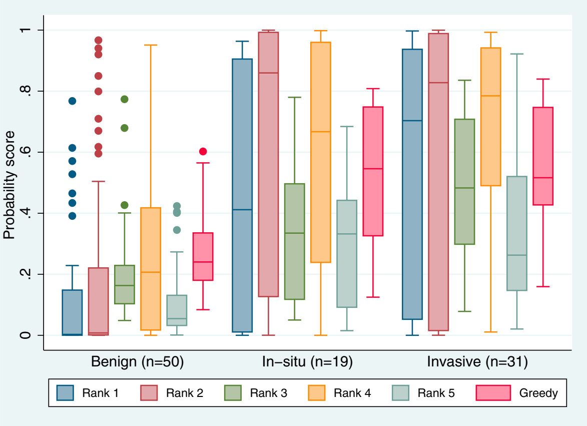

Fig 1. Algorithm probability scores. Mean probability score for the top 5 algorithms and best fusion algorithm (Greedy) by lesion diagnosis (ie, benign nevi or lentigines, melanoma in situ, and invasive melanoma). Probability scores from computer algorithms were in the range 0 to 1, with scores closer to 0 indicating a greater probability of a benign diagnosis and scores closer to 1 indicating a greater probability of a malignant diagnosis. The upper and lower bounds of the boxed area represent the 25th and 75th percentiles, the line transecting the box is the median value, and whiskers indicate the 5% and 95% percentiles. Dots that fall outside of the whiskers indicate extreme or outlier values.

Diagnostic accuracy of computer algorithms

Performance of automated systems on the 100 images evaluated by the dermatologists is shown in Table II. Ranked on average precision, greedy fusion was the top performing fusion algorithm (selected 16 algorithms for fusion from the 25 total). While nonlearning methods performed similarly, more complex SVM models demonstrated a slight reduction in performance. The relearned probabilistic SVM thresholds increased sensitivity of the corresponding systems considerably. Fig 1 shows the mean probability score for the top 5 algorithms and the best performing fusion algorithm by diagnosis.

Comparison of diagnostic accuracy of dermatologists to computers

The ROC area of the best fusion computer algorithm (greedy fusion) was 0.86, which was significantly greater than the mean ROC area of 0.71 of the 8 readers in classification (*P* = .001). Using the dermatologist mean sensitivity value for classification (82%) as the operating point on the computer algorithm ROC curves (Figs 2, *A* and *B*), the top fusion algorithm specificity was 76%, which was higher than the average dermatologist specificity of 59% (*P* = .02) and the top-ranked individual algorithm specificity of 62% (*P* = .13). Using the dermatologist mean sensitivity value for management (89%) as the operating point on the computer algorithm ROC curves, the fusion algorithm specificity was 64%, which was higher than the average dermatologist specificity of 47% (*P* = .02) and the top-ranked individual algorithm specificity of 38% (*P* = .009). At this cut-off threshold, there was no difference between the average dermatologist specificity and the top-ranked individual algorithm (47% vs. 38%, *P* = .22).

DISCUSSION

We compared the melanoma diagnostic performance of computer algorithms from an international challenge to the average performance of 8 experienced dermatologists using 100 dermoscopic images of pigmented lesions. We found that individual computer algorithms have comparable diagnostic accuracy to dermatologists; at 82% sensitivity, average reader specificity was similar to the top computer algorithm. Fusion techniques significantly improved computer performance; at 82% sensitivity, the top-ranked fusion algorithm had higher average specificity than dermatologists. In our exploratory analysis using arguably the most clinically relevant sensitivity value, the dermatologists' mean sensitivity in management decision (89%), dermatologists had specificity similar to the top algorithm, but lower than the top fusion algorithm approach. It is worth noting that some dermatologists had higher diagnostic performance than all individual and fusion algorithms in classification or management.

There has been considerable interest in developing computer vision systems for melanoma diagnosis, but few groups have directly compared computer algorithms to human performance. In 2017, Esteva et al[22] trained a deep learning convolution neural network on 129,450 images of 2032 different diseases and reported dermatologist-level classification of skin cancer. In the corresponding reader studies using clinical images (33 melanomas and 97 nevi) and dermoscopy images (71 melanomas, 40 benign), the convolution neural network had a ROC area of 0.94 and 0.91, respectively, which was superior to dermatologists.[22] In 2015, Ferris et al[8] compared the diagnostic accuracy of a computer classifier to 30 dermatology health care providers on a dataset of 65 lesions (25 melanomas, 32 nevi, 4 lentigines, and 4 seborrheic keratoses); the computer algorithm had a sensitivity of 96% and specificity of 42.5%, and the human readers—including dermatologists, dermatology residents, and physician assistants—had a mean sensitivity of 70.8% and specificity of 58.7%. In 2005, Menzies et al[11] reported on the performance of SolarScan and included a reader study of 78 lesions (13 melanomas, 63 nevi, and 2 lentigines). The computer classifier had a sensitivity

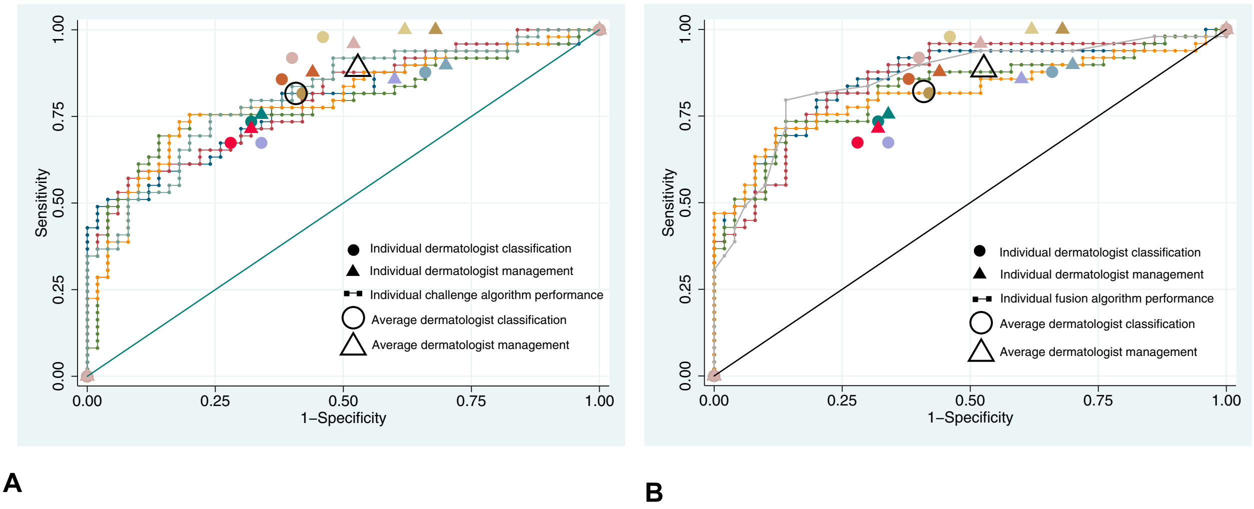

Fig 2. Diagnostic accuracy of algorithms and dermatologists for melanoma on the 100-image dataset. Receiver operating characteristic curves demonstrating sensitivity and specificity for melanoma of (**A**) top 5 ranked individual algorithms and (**B**) 5 fusion algorithms, with melanoma classification and management performance of 8 dermatologists indicated by small colored solid circles and triangles, respectively. Small colored solid circles and triangles of the same color indicate the performance of an individual dermatologist. The large transparent circle and triangle with black outline indicate the average diagnostic performance of dermatologists in classification and management, respectively.

of 85% and specificity of 65%; this compared to a mean sensitivity and specificity of 79.5% and 50.8% for the 13 human readers. Differences in study design make comparisons of our computer algorithm results to these data challenging, highlighting the importance of creating open datasets like ours.

Compared to previous investigations, there are novel aspects to our study: (1) we compared multiple computer classifiers from around the world and an aggregated model of their performance to dermatologists, increasing the likelihood that the computer-vision results reflect the current state-of-the-art; (2) our dataset originated from >12 dermatology clinics, possibly increasing the generalizability of our findings; (3) the readers originated from 4 countries, which may have improved the generalizability of our dermatologists' results; and (4) our dataset is public, permitting external and independent analysis and use as a future reference dataset by developers of diagnostic tools.

Our results should be interpreted with caution. A significant limitation is that our dataset did not sufficiently include: (1) the complete spectrum of skin lesions encountered in clinical practice that can mimic melanoma, including pigmented seborrheic keratoses; (2) less common presentations of melanoma, such as amelanotic, nodular, or desmoplastic types that are challenging to identify; (3) lesions from all anatomic sites, skin types, genetic backgrounds, and age ranges; and (4) a sufficiently representative group of benign lesions from which biopsy specimens would not typically be obtained or histopathologic examination conducted. It can reasonably be inferred that >99.9% of all benign skin lesions are routinely correctly classified by dermatologists (eg, nevi, angiomas, seborrheic keratoses, and lentigines), and therefore results obtained using our study design cannot be extrapolated to clinical practice. Our study setting was artificial because computer algorithms and readers did not have access to clinical data that might have improved diagnostic performance (eg, age, lesion history/symptoms, etc).[23] It has also been shown that the real-world performance of a computer-based system for melanoma in the hands of nonexperts is lower than that expected from experimental data; diagnostic accuracy depends on the ability of users to identify appropriate lesions for analysis.[24] Finally, participants of the ISBI 2016 melanoma challenge were instructed that computer algorithms would be ranked using average precision, a metric that does not target a clinically relevant sensitivity or specificity threshold; therefore, the algorithms were not optimized for comparison to dermatologist diagnostic performance.

Our results underscore the value of public challenges conducted in open-access image resources like the ISIC archive. This platform permits comparison of the performance of individual algorithms, as well as the type of fusion experiments presented here. Larger and more diverse collections of public, clinically validated images are needed to advance the field of computerized lesion classification, education, and clinical decision support. A bigger 2017 ISIC challenge represents a further step in this direction.[25] In addition to providing a larger and more diversified set of images, including seborrheic keratoses, the current challenge tailors the performance metrics for comparison to the current state of clinical diagnosis.

In conclusion, in this artificial study setting without integration of clinical history, state-of-the-art computer vision systems are comparable to dermatologist diagnostic accuracy for melanoma dermoscopy images and, when using fusion algorithms, can exceed dermatologist performance in classification of some but not all dermatologists. Although these results are preliminary and should be viewed with caution, development and comparison of deep learning methods on larger, more varied datasets is likely to accelerate the potential use and adoption of computer vision for melanoma detection. Strategies for including common skin lesions from which biopsy specimens are not typically obtained in these datasets are critical for optimizing the generalizability of computer vision algorithms.

This study reports the efforts of the International Skin Imaging Collaboration (ISIC), the members of which can be found at http://isdis.net/isic-project/working-group-members/. We thank the leadership of the ISIC collaboration: H. Peter Soyer, MD, Dermatology Research Centre, The University of Queensland, Brisbane, Australia (Technique Working Group Co-Leader); Clara Curiel-Lewandrowski, MD, University of Arizona Cancer Center, Tucson, AZ, USA (Technique Working Group Co-Leader); Harald Kittler, MD, Department of Dermatology, Medical University of Vienna, Austria (Terminology Working Group Leader); Liam Caffery, PhD, The University of Queensland, Brisbane, Australia (Metadata Working Group Leader); Josep Malvehy, MD; Hospital Clinic of Barcelona, Spain (Technology Working Group Leader); and Rainer Hofmann Wellenhof, MD, Medical University of Graz, Austria (Archive Group Leader).

We thank the organizing committee of The International Symposium on Biomedical Imaging (ISBI), the chairs of the 2016 ISBI Grand Challenges (Bram van Ginneken, Radboud University Medical Center, Netherlands; Adriënne Mendrik, Utrecht University, Netherlands; and Stephen Aylward, Kitware Inc), and all the participants of the 2016 ISBI Challenge "Skin Lesion Analysis Towards Melanoma Detection."

REFERENCES

1. Marghoob AA, Scope A. The complexity of diagnosing melanoma. *J Invest Dermatol*. 2009;129:11-13.
2. Malvehy J, Hauschild A, Curiel-Lewandrowski C, et al. Clinical performance of the Nevisense system in cutaneous melanoma detection: an international, multicentre, prospective and blinded clinical trial on efficacy and safety. *Br J Dermatol*. 2014;171:1099-1107.
3. Monheit G, Cognetta AB, Ferris L, et al. The performance of MelaFind: a prospective multicenter study. *Arch Dermatol*. 2011;147:188-194.
4. Brady MS, Oliveria SA, Christos PJ, et al. Patterns of detection in patients with cutaneous melanoma. *Cancer*. 2000;89:342-347.
5. Bibbins-Domingo K, Grossman DC, Curry SJ, et al. Screening for skin cancer: US Preventive Services Task Force recommendation statement. *JAMA*. 2016;316:429-435.
6. Celebi ME, Kingravi HA, Uddin B, et al. A methodological approach to the classification of dermoscopy images. *Comput Med Imaging Graph*. 2007;31:362-373.
7. Iyatomi H, Oka H, Celebi ME, et al. An improved Internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm. *Comput Med Imaging Graph*. 2008;32:566-579.
8. Ferris LK, Harkes JA, Gilbert B, et al. Computer-aided classification of melanocytic lesions using dermoscopic images. *J Am Acad Dermatol*. 2015;73:769-776.
9. Zortea M, Schopf TR, Thon K, et al. Performance of a dermoscopy-based computer vision system for the diagnosis of pigmented skin lesions compared with visual evaluation by experienced dermatologists. *Artif Intell Med*. 2014;60:13-26.
10. Blum A, Luedtke H, Ellwanger U, Schwabe R, Rassner G, Garbe C. Digital image analysis for diagnosis of cutaneous melanoma. Development of a highly effective computer algorithm based on analysis of 837 melanocytic lesions. *Br J Dermatol*. 2004;151:1029-1038.
11. Menzies SW, Bischof L, Talbot H, et al. The performance of SolarScan: an automated dermoscopy image analysis instrument for the diagnosis of primary melanoma. *Arch Dermatol*. 2005;141:1388-1396.
12. Rajpara SM, Botello AP, Townend J, Ormerod AD. Systematic review of dermoscopy and digital dermoscopy/artificial intelligence for the diagnosis of melanoma. *Br J Dermatol*. 2009;161:591-604.
13. Rubegni P, Cevenini G, Sbano P, et al. Evaluation of cutaneous melanoma thickness by digital dermoscopy analysis: a retrospective study. *Melanoma Res*. 2010;20:212-217.
14. International Skin Imaging Collaboration archive. Available at: https://isic-archive.com/. Accessed September 2, 2016.
15. International Symposium on Biomedical Imaging website. ISBI 2016: skin lesion analysis towards melanoma detection. Available at: https://challenge.kitware.com/#challenge/n/ISBI_2016%3A_Skin_Lesion_Analysis_Towards_Melanoma_Detection. Accessed September 2, 2016.
16. International Symposium on Biomedical Imaging website. ISBI 2016 challenges: International Symposium on Biomedical Imaging: from nano to macro. Available at: http://biomedicalimaging.org/2016/?page_id=416. Accessed September 2, 2016.
17. Codella N, Nguyen QB, Pankanti S, et al. Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM J Res Develop*. 2017;61. Available at: https://arxiv.org/ftp/arxiv/papers/1610/1610.04662.pdf. Accessed August 17, 2017.

18. Gutman D, Codella NC, Celebi E, et al. Skin lesion analysis toward melanoma detection: a challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). Available at: https://arxiv.org/abs/1605.01397. Accessed August 17, 2017.
19. Scope A, Marchetti MA, Marghoob AA, et al. The study of nevi in children: principles learned and implications for melanoma diagnosis. *J Am Acad Dermatol*. 2016;756:813-823.
20. Yan R, Fleury MO, Merler M, Natsev A, Smith JR. Large-scale multimedia semantic concept modeling using robust subspace bagging and MapReduce. In: *Proceedings of the First ACM Workshop on Large-scale Multimedia Retrieval and Mining*. Beijing, China: LS-MMRM; 2009.
21. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol*. 2011;2:1-27.
22. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542:115-118.
23. Binder M, Kittler H, Dreiseitl S, Ganster H, Wolff K, Pehamberger H. Computer-aided epiluminescence microscopy of pigmented skin lesions: the value of clinical data for the classification process. *Melanoma Res*. 2000;10:556-561.
24. Dreiseitl S, Binder M, Hable K, Kittler H. Computer versus human diagnosis of melanoma: evaluation of the feasibility of an automated diagnostic system in a prospective clinical trial. *Melanoma Res*. 2009;19:180-184.
25. International Skin Imaging Collaboration 2017: skin lesion analysis towards melanoma detection. Available at: https://challenge.kitware.com/#challenge/n/ISIC_2017%3A_Skin_Lesion_Analysis_Towards_Melanoma_Detection. Accessed December 21, 2016.

SUPPLEMENTARY MATERIALS

Methods

Greedy ensemble fusion is a supervised machine learning method that can calculate weighted or unweighted averages of multiple predictors to maximize a defined metric of performance. A binary SVM is a supervised machine learning technique that calculates a "hyperplane" (a multi-dimensional boundary) to separate a dataset according to supplied binary labels. The dataset is represented in a multi-dimensional feature space. The output of the SVM is a classification score, which is a signed distance from the hyperplane designating class membership. In comparison to greedy ensemble selection, an SVM is a more complex model, providing opportunity to better fit intricate patterns in data, but at risk to potentially fitting noise in the data.

For score averaging, all algorithm prediction scores on individual images are averaged into a single prediction for that image, with no prior filtering or selection of models. For voting, all algorithm predictions are first dichotomized to values of 0 or 1, using 0.5 as a threshold, and then subsequently averaged for each image. For greedy ensemble fusion, a selection process is employed: algorithm predictions are sorted according to performance, in terms of average precision. An iterative process ensures, whereby for each iteration n, the top n performing algorithm predictions are averaged, and the performance of the average is recorded. The iteration that yields the best overall performance determines which algorithm predictions are selected to be averaged into a single prediction score. For SVMs, feature vectors were created using all participant predictions (sigmoid normalized). A C value of 1 was employed, and thresholds were re-learned according to a probabilistic approach.[17]

Results

Agreement among dermatologists. The overall kappa for classification and management was 0.53 and 0.47, respectively. Of the 100 lesions, readers were 100% concordant (8/8 agreement) in disease classification of 44 lesions; 26 (59%) were true-positives, 13 (30%) were true-negatives, and 5 (11%) were false-positives. Readers were discordant in diagnosis of 56 lesions, of which 32 were benign and 24 malignant. Regarding the proportion of readers whose disease classification agreed with the reference-standard diagnosis (i.e., histopathology), 8/8 (100%) agreed with the reference-standard diagnosis on 39 lesions, 7/8 (87.5%) on 15 lesions, 6/8 (75%) on 10 lesions, 5/8 (62.5%) on 3 lesions, 4/8 (50%) on 8 lesions, 3/8 (37.5%) on 8 lesions, 2/8 (25%) on 5 lesions, 1/8 (12.5%) on 7 lesions, and 0/8 (0%) on 5 lesions.