Computer-aided classification of melanocytic lesions using dermoscopic images

Laura K. Ferris, MD, PhD, Jan A. Harkes, MS, Benjamin Gilbert, MS, Daniel G. Winger, MS, Kseniya Golubets, MD, MHS, Oleg Akilov, MD, PhD, and Mahadev Satyanarayanan, PhD

Pittsburgh, Pennsylvania

Background: Computer-assisted diagnosis of dermoscopic images of skin lesions has the potential to improve melanoma early detection.

Objective: We sought to evaluate the performance of a novel classifier that uses decision forest classification of dermoscopic images to generate a lesion severity score.

Methods: Severity scores were calculated for 173 dermoscopic images of skin lesions with known histologic diagnosis (39 melanomas, 14 nonmelanoma skin cancers, and 120 benign lesions). A threshold score was used to measure classifier sensitivity and specificity. A reader study was conducted to compare the sensitivity and specificity of the classifier with those of 30 dermatology clinicians.

Results: The classifier sensitivity for melanoma was 97.4%; specificity was 44.2% in a test set of images. In the reader study, the classifier’s sensitivity to melanoma was higher (P < .001) and specificity was lower (P < .001) than that of clinicians.

Limitations: This is a retrospective study using existing images primarily chosen for biopsy by a dermatologist. The size of the test set is small.

Conclusions: Our classifier may aid clinicians in deciding if a skin lesion should be biopsied and can easily be incorporated into a portable tool (that uses no proprietary equipment) that could aid clinicians in noninvasively evaluating cutaneous lesions. (J Am Acad Dermatol 2015;73:769-76.)

Key words: basal cell carcinoma; computer-assisted diagnosis; dermoscopy; information technology; machine learning; melanoma; skin cancer.

In the United States, nearly 10,000 people die of melanoma annually. Early detection of melanoma is vital for reducing melanoma mortality and treatment costs. Dermoscopy can improve accuracy in melanoma detection, although this tool

Abbreviations used:

BCC: basal cell carcinoma
CI: confidence interval
LCB: lower confidence bound

Disclosure: Dr Ferris has served as a consultant to MelaSciences, Castle Biosciences, and DermTech, International. Mr Harkes, Mr Gilbert, Mr Winger, and Drs Golubets, Akilov, and Satyanarayanan have no conflicts to declare.

Accepted for publication July 26, 2015.

Reprint requests: Laura K. Ferris, MD, PhD, Department of Dermatology, University of Pittsburgh, 3601 Fifth Avenue, Fifth Floor, Pittsburgh, PA 15238. E-mail: ferrislk@upmc.edu.

Published online September 16, 2015.

0190-9622/$36.00

© 2015 by the American Academy of Dermatology, Inc.

http://dx.doi.org/10.1016/j.jaad.2015.07.028
is used almost exclusively by dermatologists. The rich images obtained through dermoscopy provide an opportunity to apply principles of machine learning and computer vision to the challenge of interpreting dermoscopic images, potentially making dermoscopy a more useful tool for nondermatologists.

We trained a software classifier on dermoscopic images of benign and malignant skin lesions that uses decision forest classification to generate a severity score for each lesion. We tested the performance of the classifier on a collection of dermoscopic images of biopsy-proven benign and malignant skin lesions. We also asked a cross-section of dermatology providers to participate in a reader study in which they were asked to evaluate a set of dermoscopic images and indicate whether they would biopsy each lesion. The sensitivity and specificity of the classifier on the same set of images were compared with that of the clinicians in the reader study.

METHODS

Ethics approval

This protocol was reviewed and approved by the University of Pittsburgh Institutional Review Board.

Images

Dermoscopic images of skin lesions were collected before biopsy using contact dermoscopy with either isopropyl alcohol or a clear alcohol-based hand sanitizer as an immersion medium using a variety of dermoscopic devices and cameras. All lesions were biopsied based on clinical suspicion of malignancy and images were at least 640 × 640 pixels. Some dermoscopic images were collected prospectively and some were obtained from our collection of existing images. Inclusion criteria included patient 18 years of age or older, biopsy indicated clinically based on suspicion for malignancy, and histologic diagnosis was available in the medical record. All histologic diagnoses were rendered by at least 1 board-certified dermatopathologist and were used as the reference standard for diagnosis. Some dermoscopic images used to train the classifier were obtained from publicly available or purchased image libraries; such images were not included in the reader study or used to test the performance of the classifier. Images collected by the investigators were randomly divided into 2 sets by diagnosis, with half used as candidates for training and half used for testing, with the exception that all high-grade dysplastic nevi were exclusively assigned to the training set to increase the representation of dermoscopic features that could be present in melanoma. All images used in training were curated by 1 dermatologist experienced in dermoscopy (L. K. F.) both for image quality and for clinical representation of the spectrum of melanocytic lesions.

Image processing

The images were captured by different dermoscope/camera combinations that varied in resolution and lighting characteristics, but all were at least 640 × 640 pixels. Each image was segmented manually to delineate lesion from surrounding skin using a World Wide Web–based mark-up tool of our own creation that allows the user to shade the lesion in the image; nonshaded areas (ie, surrounding skin) were excluded from further analysis. Images were rescaled to 640 pixels on the longest axis and 54 features were computed for all segmented lesions. Extracted features included values related to lesion geometry, border gradient, color, and texture. Variations of some features described in Zortea et al were included.

Classifier training

The 54 computed features for lesions in the training set were used to train a decision forest classifier consisting of 1000 decision trees. Lesions used for training were divided into 2 categories: benign and malignant. The malignant training set was composed of 105 melanomas (77 invasive melanomas, 28 melanoma in situ), 29 high-grade dysplastic nevi (included in training because these lesions contain dermoscopic features of melanoma although they are not actual malignancies), 23 basal cell carcinomas (BCC), and 3 squamous cell carcinomas. The benign training set was composed of 93 benign melanocytic lesions (42 benign nevi, 6 blue nevi, 9 lentigines, 33 low-grade dysplastic nevi, and 3 acral nevi) and 20 other benign lesions (4 dermatofibromas, 13 seborrheic keratoses, and 3 angiomas). Features extracted and their relative importance in the decision forest are reported (Fig 1). A decision forest approach was used for creation of the
Through incremental exploratory analysis on labeled images of lesions, 54 features were deemed relevant. These include features such as border irregularity, eccentricity, length of major and minor axes, and color histogram properties. Discovering how these features should be weighted and combined is the problem addressed by the decision forest approach. One thousand binary trees were generated using a randomized procedure in which each interior node of a tree corresponded to a predicate on a feature. The 2 links out of each node corresponded to that predicate being true or false, respectively. A subset of the images in the training set was chosen randomly, and the feature that best divided the subset into malignant and benign groupings was identified. This feature was assigned to the root of a new decision tree. This process was then repeated, generating a new interior node of the tree each time, until we had only images in a single category. The result was 1 decision tree. This process was repeated to construct a forest of 1000 decision trees, with each tree based on a different random subset of the training data. When the classifier was applied to a new image, its features were passed from root to leaf of every tree, thus generating 1000 individual malignant or benign results. The fraction of trees in which the path ends in “malignant” was then used to generate the severity score. A threshold

![Feature importance](image)

**Fig 1.** Relative weight of feature in the random forest classifier in the same order as described in the “Methods” section and detailed in the following chart. The tallest 5 peaks are: number of distinct color histogram values (45), Earth Mover’s Distance (EMD) color symmetry along major axis (22), minimal lesion intensity value (9), SD of border gradient scores (19), and average color histogram value (43). *PvsC*, Periphery of lesion vs center of lesion.
value for this score could be set to measure sensitivity and specificity. The threshold was set to 0.4 in our classifier; a lesion was classified as malignant if its image traced a path to a malignant node in at least 40% of the trees.

Classifier testing

The performance of the classifier was evaluated using a set of images with known histologic diagnoses that were distinct from those images used in training. Lesions used in classifier testing consisted of 39 melanomas (25 invasive and 14 in situ; Breslow depth for invasive melanomas ranged from 0.2-2.98 mm with mean depth of 0.76 mm, median depth of 0.5 mm), 11 BCC, 3 squamous cell carcinomas, and 120 benign lesions (42 benign nevi, 47 low-grade dysplastic nevi, 10 lentigines, 5 blue nevi, 2 Spitz nevi, 11 seborrheic keratoses, 2 angiomatous, and 1 dermatofibroma). Because these lesions had all been selected for biopsy on clinical grounds and thus would be expected to represent the more dermoscopically atypical end of the spectrum of lesions encountered in clinical practice, a separate unbiopsied benign test set was also evaluated. This set consisted of 27 images of lesions that were chosen as not appropriate for biopsy by 2 dermatologists, and thus did not have available histology. The classifier specificity on the unbiopsied benign set was calculated and reported solely to provide additional information, but not used to calculate overall specificity.

Each lesion in the test set was given a severity score of 0 to 1, corresponding to the fraction of decision trees that classify a lesion as malignant. The sensitivity and specificity of the classifier was calculated using a threshold severity score of 0.4 (ie, at least 40% of decision trees classifying a lesion as malignant).

Reader study

A reader study was performed that included a subset of a total of 65 lesions: 25 melanomas (15 invasive and 10 in situ) and 40 benign lesions (16 low-grade dysplastic nevi, 14 benign nevi, 2 blue nevi, 4 lentigines, and 4 seborrheic keratoses). The Breslow depths for invasive melanomas in the reader study set ranged from 0.2 to 2.98 mm (mean depth 0.93 mm, median depth 0.74 mm). All images used for the reader study were first evaluated for image quality by 1 investigator who was blinded to their diagnosis. Although the 1 melanoma that was misclassified as benign by the classifier was intentionally included in the reader study, the remaining images were randomly selected among those determined to be of suitable image quality for display on a computer screen.

Thirty of 35 invited participants, all of whom self-reported some training and experience with the use of dermoscopy, completed the reader study. Participants were assigned to 1 of 3 subgroups based on their self-reported level of training in dermatology: board-certified dermatologist, dermatology resident, or dermatology physician assistant. Participants were asked how often they use dermoscopy when examining pigmented skin lesions and also to evaluate the 65 dermoscopic images detailed above and to indicate if they would biopsy the lesion.

Statistical analysis

Receiver operating characteristic curves were generated and areas under the curve calculated to evaluate the performance of the classifier on both the reader study and extended test sets. The sensitivity and specificity of the classifier, with 95% confidence intervals (CI) and 95% lower confidence bounds (LCB), were computed. The average sensitivity and specificity of the readers were calculated based on biopsy decisions on the reader study test set and compared with the sensitivity and specificity of the classifier on the same set of images using a 2-sided, 1-sample t test. The Fleiss kappa statistic was calculated to quantify degree of agreement between readers.

RESULTS

Performance of classifier

To determine the optimal threshold severity score to use to maximize classifier sensitivity and specificity, a receiver operating characteristic curve was generated showing the performance of the classifier on the 39 melanomas and 120 benign lesions in our test set (area under the curve 0.818) (Fig 2). Based on these findings a threshold severity score of 0.4 was established and used to calculate classifier sensitivity and specificity (Table I). Among lesions with known histologic diagnosis and for which greater than 5 examples were present in our data sets, the median severity score of lesions was highest for invasive melanomas (0.772) and lowest for benign nevi (0.337). Median severity scores by diagnosis are shown (Fig 3).

Of the 39 melanomas evaluated, 38 were correctly classified as malignant, yielding a sensitivity to melanoma of 97.4% (95% CI 86.5%-99.9%, and 95% LCB 88.4%). Eight of the 11 BCC and all of the 3 squamous cell carcinomas were correctly classified as malignant, yielding a sensitivity to nonmelanoma skin cancer of 78.6%. Among the 120 benign lesions
evaluated, 53 were correctly classified as benign, yielding a specificity of 44.2% (95% CI 35.1%-53.5%, and 95% LCB 36.5%). In addition, among the 27 unbiopsied but presumed benign lesions evaluated, 20 were classified as benign, yielding a specificity of 74.1% (95% CI 53.7%-88.9% and 95% LCB 58.8%) (Table I).

### Performance of dermatology practitioners versus classifier

A total of 30 dermatology practitioners (12 board-certified dermatologists, 10 dermatology residents, and 8 physician assistants currently practicing dermatology) completed the reader study (response rate of 85.7%). Readers were asked how frequently they use dermoscopy when evaluating pigmented lesions in their clinical practice. All reported using dermoscopy at least some of the time. Among board-certified dermatologists, 67% reported using dermoscopy “always/almost always” or “very frequently.” This degree of dermoscopy use was reported by 90% of the dermatology residents and 75% of the physician assistants who participated in the reader study. Performance of each individual reader is shown (Fig 2). Overall, clinician sensitivity to melanoma was 70.8%, and specificity was 58.7%. On the same set of images, the sensitivity of the classifier to melanoma was 96.0%, and the specificity of the classifier was 42.5% (Table II). There was a fair level of agreement among readers in aggregate, with a kappa score of 0.346. In total, 2 melanomas (8.0%) were chosen to be biopsied by all readers. The melanoma in situ that was incorrectly classified as benign by the classifier was chosen for biopsy by 23.3% of the readers. Examples of lesions from the reader study test set and 1 lesion from the unbiopsied benign test set are shown (Fig 4).

### Table I. Classifier performance on the test set

<table>
<thead>
<tr>
<th>Lesion</th>
<th>Total</th>
<th>TP</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Melanoma</td>
<td>39</td>
<td>38</td>
<td>97.4%</td>
</tr>
<tr>
<td>Invasive</td>
<td>25</td>
<td>25</td>
<td>100%</td>
</tr>
<tr>
<td>In situ</td>
<td>14</td>
<td>13</td>
<td>92.9%</td>
</tr>
<tr>
<td>Nonmelanoma skin cancer</td>
<td>14</td>
<td>11</td>
<td>78.6%</td>
</tr>
<tr>
<td>BCC</td>
<td>11</td>
<td>8</td>
<td>72.7%</td>
</tr>
<tr>
<td>SCC</td>
<td>3</td>
<td>3</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lesion</th>
<th>Total</th>
<th>TN</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>All biopsied benign</td>
<td>120</td>
<td>53</td>
<td>44.2%</td>
</tr>
<tr>
<td>Benign nevi</td>
<td>42</td>
<td>22</td>
<td>52.4%</td>
</tr>
<tr>
<td>LGDN</td>
<td>47</td>
<td>20</td>
<td>42.6%</td>
</tr>
<tr>
<td>Lentigo</td>
<td>10</td>
<td>2</td>
<td>20.0%</td>
</tr>
<tr>
<td>SK</td>
<td>11</td>
<td>4</td>
<td>36.4%</td>
</tr>
<tr>
<td>Other*</td>
<td>10</td>
<td>5</td>
<td>50.0%</td>
</tr>
<tr>
<td>Unbiopsied benign</td>
<td>27</td>
<td>20</td>
<td>74.1%</td>
</tr>
</tbody>
</table>

*BCC, Basal cell carcinoma; LGDN, low-grade dysplastic nevi; SCC, squamous cell carcinoma; SK, seborrheic keratoses; TP, true positive; TN, true negative.

*Includes blue nevi (5), typical Spitz nevus (2), angioma (2), dermatofibroma (1).
studies. Melanoma than physicians in accompanying reader.

harm than failing to biopsy a melanoma, and thus
unnecessary biopsy of a benign lesion causes less
specificity. Although a high specificity is ideal,
sensitivity to melanoma generally also had the lowest
our study, those practitioners who had the highest
just in devices and tests, but in clinicians as well. In

DISCUSSION
We designed a classifier that uses decision tree
analysis to score dermoscopic images. Using a fixed
cutoff, the sensitivity of our classifier for melanoma
was greater than that of the 30 clinicians who
participated in our reader study, however overall
clinicians’ specificity was higher.

Other investigators have used computer vision to
aid in diagnosing melanoma from dermoscopic
images. Zortea et al used features similar to some of
those used in our classifier and reported that their
classifier had a sensitivity of 86% and specificity of
52% in a small study. One imaging device, MelaFind
(MELA Sciences, Irvington, NY), is Food and Drug
Administration approved for the early detection of
melanoma; it performs multispectral imaging and
provides a lesion score and recommendation of
whether or not a biopsy is indicated. The labeling
for this device restricts its use to dermatologists.

Although the sensitivity of MelaFind is very high
(98.3%), concerns have been raised about the low
specificity (9.9%) of the device. Another device that
aids in noninvasively classifying a lesion as benign or
malignant is Nevisense (Scibase, Stockholm,
Sweden), which uses electrical impedance spectro-
cy, with a sensitivity of 96.6%, and a specificity of
34.4%. Both devices show superior sensitivity to
melanoma than physicians in accompanying reader
studies.

The trade-off of a higher sensitivity at the cost of
specificity is, to some degree, inevitable and seen not
just in devices and tests, but in clinicians as well. In
our study, those practitioners who had the highest
sensitivity to melanoma generally also had the lowest
specificity. Although a high specificity is ideal,
unnecessary biopsy of a benign lesion causes less
harm than failing to biopsy a melanoma, and thus
specificity that significantly sacrifices sensitivity is
unacceptable. Because our classifier generates a
severity score, the threshold at which a lesion is
considered malignant is easily tunable. Our choice of
score was chosen based on our receiver operating
characteristic curves with a goal of maximizing
sensitivity. However, the severity score itself could
likely provide the clinician with guidance on making
a more informed biopsy decision.

We are encouraged that in our study the magni-
itude of the increase in sensitivity to melanoma of our
classifier compared with that of readers is greater
than the magnitude of the loss of specificity. Further
development and testing of our classifier is needed to
fully determine its clinical use and our study has
limitations. Although in this initial proof-of-concept
study only 1 melanoma in situ among 39 melanomas
was given a severity score that fell under our
threshold, further validation on additional lesions is
needed as our sample size is lower than those in the
pivotal studies of devices that have received or are
seeking regulatory approval (ie, Nevisense and
MelaFind). Also, our reader study was relatively
small, both in the number of readers and number
of lesions included, and data should be interpreted
with this in mind as the ratio of benign to malignant
lesions in a reader study cannot mirror that seen in
clinical practice. In addition, different groups of
clinicians are likely to vary in the choices they
make about how to manage pigmented skin lesions,
even when using similar technologies. Similarly to
studies of other diagnostic aids, the lesions in our
study were primarily selected for biopsy because
they were determined to be suspicious by a derma-
tologist. Encouragingly, however, the specificity of
our classifier was highest on our set of clinically
benign unbiopsied lesions and these lesions had the
lowest severity score of any group of images we
tested, suggesting that our classifier will not recom-
mend biopsy of benign lesions at an unacceptably
high rate. Performance on a similar set of lesions has
not been reported for Nevisense or MelaFind and we
report these data as exploratory because we do not
have histologic confirmation of the diagnosis of
these lesions. Another limitation of our classifier is
that although our sensitivity to melanoma is high, our
sensitivity to BCC in this small series was lower at
72.7%. For comparison, Nevisense has a sensitivity of
100% for nonmelanoma skin cancer; sensitivity of
MelaFind for nonmelanoma skin cancer has not been
reported. Although these tumors are rarely fatal
and early detection is not associated with improved
survival, we would like to optimize our classifier’s
performance on these common tumors. In our
reader study, we only provided dermoscopic images

Table II. Performance of classifier and dermatology practitioners who participated in the reader study on the reader study test set

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>P value</th>
<th>Specificity</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifier</td>
<td>96.0%</td>
<td>-</td>
<td>42.5%</td>
<td>-</td>
</tr>
<tr>
<td>All readers (n = 30)</td>
<td>70.8%</td>
<td>&lt;.001</td>
<td>58.7%</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Board-certified</td>
<td>64.7%</td>
<td>&lt;.001</td>
<td>65.4%</td>
<td>.002</td>
</tr>
<tr>
<td>dermatologists</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(n = 12)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dermatology</td>
<td>70.4%</td>
<td>&lt;.001</td>
<td>59.0%</td>
<td>.002</td>
</tr>
<tr>
<td>residents (n = 10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physician assistants</td>
<td>80.5%</td>
<td>.015</td>
<td>48.1%</td>
<td>.423</td>
</tr>
<tr>
<td>(n = 8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p values for comparison of classifier vs reader sensitivity or specificity.
and not clinical images. This is not representative of the true patient encounter and may have decreased the accuracy of the biopsy decisions made by participants. However, our reader study results are similar to those reported by others and suggest that the lesions used to test our classifier were not clinically obvious cases but represent the type of skin lesions that are diagnostically challenging even for dermatologists.

Many patients with a lesion suspicious for melanoma will not have easy access to a dermatologist. In addition, population-based screening performed primarily by nondermatologists can lower melanoma mortality. This provides an opportunity to aid clinicians, particularly nondermatologists, by providing them an accessible tool to improve early detection of melanoma by aiding them in choosing lesions that are appropriate for biopsy. In developing our classifier, we considered how it may ultimately and practically be incorporated into a tool that can be widely and inexpensively used by a variety of health care practitioners. The classifier was trained on images taken by several different camera and dermoscope combinations so that its use would not be restricted to a single piece of equipment. It can be run using modest computing resources, using a dermoscope attached to a smartphone or tablet computer, reducing its cost of operation. Because a tool that uses our classifier is likely to be of greater use to nondermatologists, who may select different lesions for biopsy than dermatologists, we also plan to validate the performance of our classifier on lesions that were referred to our teledermatology service by primary care providers. Providing useful information to aid in the evaluation and management of skin lesions has the potential to improve melanoma detection, particularly among nondermatologists and in patients with limited access to dermatologists.

We are indebted to the participants in the reader study.
REFERENCES