Digital Imaging Biomarkers Feed Machine Learning For Melanoma Screening

Appendix S1: Supplementary Methods and Discussion

S1 – Study Overview

Figure S1 shows the scheme for our double-blind study, which attempted prediction of the biopsy-based histological diagnosis (melanoma or atypical nevus) using only the pre-biopsy dermoscopy image.

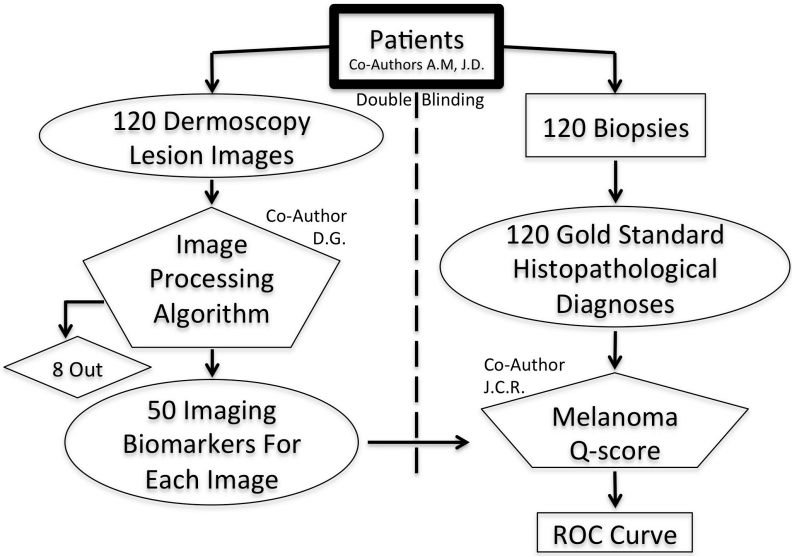

Figure S1. Study Design This double-blinded study retrospectively tested prediction of the histopathological biopsy diagnosis using only the dermoscopy images acquired just prior to biopsy. Co-Author S.Y. assembled and distributed the latter to Co-Author D.G. and the former to Co-Author J.C. Excluding (8 Out) eight images where the algorithm failed to identify the lesion border (see Figure S11), the Melanoma Q-score was produced as a standardized diagnostic MIB and tested by producing the receiver-operator characteristic (ROC) Curve.

The study included dermoscopy images of pigmented lesions on 120 patients who underwent surgical excision or biopsy of a suspicious pigmented lesion, and provided written, informed consent to be part of imaging research as approved by the Institutional Review Board of Memorial Sloan Kettering Cancer Center. Cases with extreme atypia such as those that were ulcerated, nodular/palpable, or did not fit within the field of view of the dermatoscope were excluded. All lesions were pigmented lesions that did not demonstrate a benign pattern (1) under dermoscopy. The study cohort included 60 melanomas and 60 nevi. Alcohol-coupled, non-polarized dermoscopy images were acquired with the EpiFlash™ (Canfield Inc., NJ) dermatoscope attached to a Nikon D80 camera. Each image contained 1-5 megapixels after cropping. After imaging and surgical excision of the imaged lesion, the standard diagnostic method of histopathological evaluation was carried out as part of routine clinical care to yield a diagnosis (melanoma or nevus) for each lesion in the study cohort. The 120 binary diagnoses, along with 120 correlating dermoscopic images comprised the study data. No information about the patient's age, sex, state of sun damage or anatomical location of the lesion was used. Dermoscopy images were randomized and coded to remove all

patient identifiers, then injected into the blind study arm that generated the image-processing algorithm targeting melanoma features by extracting diagnostic melanoma imaging biomarkers (MIBs) without knowledge of the histopathological diagnosis. An example of the extraction of an MIB is shown in Figure S2. The designer of the MIBs received only the dermoscopy images, and the designer of the classification scheme received only the histopathological diagnoses and the MIBs.

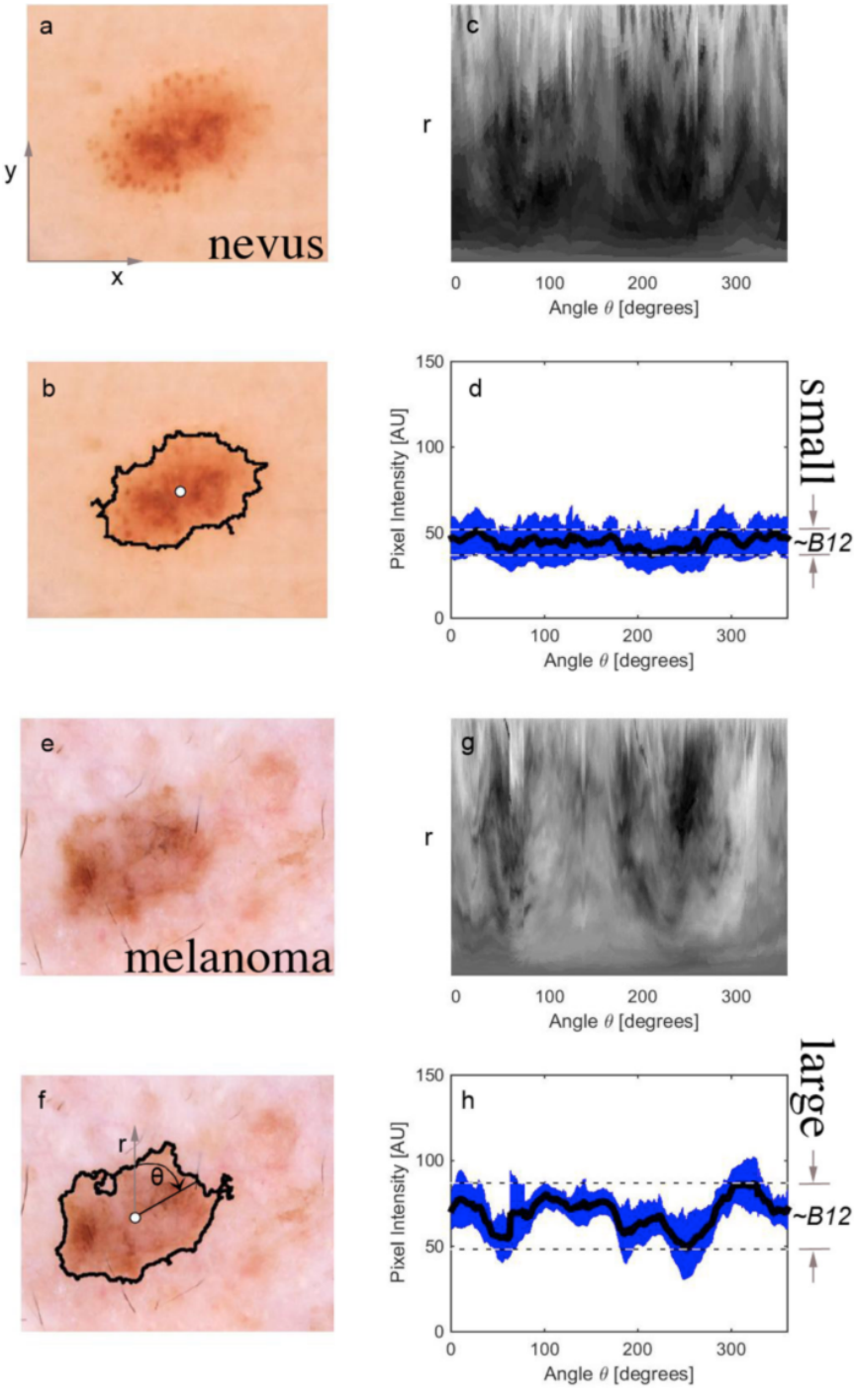

Figure S2. Coordinate Transformation And illustration Of MIB Derivation Using Angular Sweep Analysis In images of a nevus (a) and a melanoma (e), lesion border and center (b,f). (c,g) show the blue channel data under a coordinate transformation from x-y to R-θ such that the bottom row of pixels in (g) is the same pixel in f, namely the center pixel (white circle) and the top row of pixels in (g) traces out the lesion border clockwise. (d,h) analyze the pixel brightness statistics (mean in black and standard deviation in blue) of (c,g) in the vertical direction which is along the radial in (b,f). In (d,h), MIB B12 is derived from the radial variation range, which is the vertical separation of the horizontal dashed lines (d,h).

Ultimately, a machine learning algorithm yielded an ensemble classifier from the set of individual classifiers in Table S2, each designed to combine the MIBs into a diagnostic score, the probability of melanoma, that maximized the partial area under a standard receiver-operator characteristic (ROC) curve (Figure S3). The partial area chosen was the right-hand half of the ROC curve, since high sensitivity is a priority. During

construction of the ROC curve, the threshold that was compared to the Q-score for binary classification (melanoma/nevus) was varied and at each threshold value, the number of true positives, true negatives, false positives and false negatives were tabulated and used to specify the sensitivity and specificity at that particular threshold.

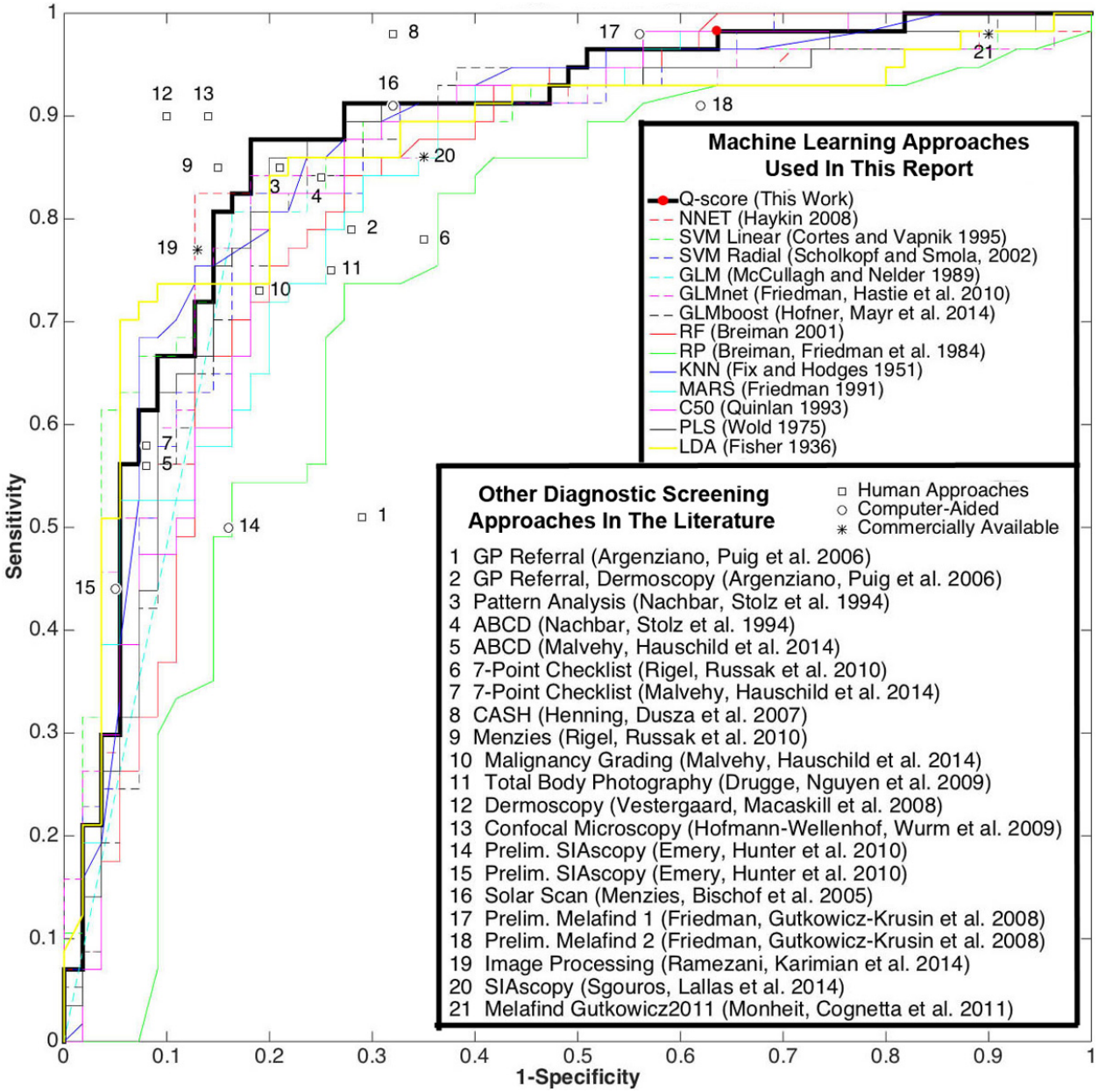

Figure S3. Diagnostic Performance Results Vs. Published Techniques The receiver-operator characteristic (ROC) curves for the individual machine learning approaches (thin colored lines) are outperformed by the compound melanoma Q-score (thick black line) with clinically relevant sensitivity = 98% and specificity = 36% (red circle). Data points of comparative approaches from the literature are marked with symbols that indicate their nature (machine or human-derived) and are numbered by the legend references. The machine learning approaches are abbreviated in the legend and numbered with their respective literature references (see Table S2).

As detailed in section S3.2, the study set was randomly partitioned into training (80%) and testing sets (20%) for tuning the classification algorithms. This training-testing scheme was replicated 1000 times for the evaluation of classifiers robustness. The Q-score was the median melanoma probability from the ensemble of methods since this approach proved to outperform the accuracy of individual classifiers. The ROC curve in Figure S3 allows comparison of the Q-score to other dermoscopic diagnostic algorithms and computer-automated techniques reported in the literature, besides our approach.

Thus, our analytical approach, drawing on the image processing and machine learning fields, included a two-step process: (1) derivation of discrete MIBs; and (2) combination of MIBs into a function that produced an optimized diagnostic score. These two steps are detailed in sections S2 and S3, respectively, with explicit formulae for all MIBs in section S5. The computational time to extract MIBs from each image ranged from 1 to 10 minutes, depending on the size of the lesion. The training of the Q-score classifier took 15 hours to run the machine learning algorithms.

S2 – Image Processing And Resulting MIBs

Images were cropped to exclude artificial pen markings and scale bars commonly used in the clinic and for lesions that were small, to exclude vast swaths of surrounding normal skin such that the area of lesion and normal skin were comparable. The algorithm then operated independently on each color layer of the 3-color Red/Green/Blue (RGB) image, and performed a segmentation to determine the border of the lesion. Though sophisticated segmentation methods may be applied such as active contours, (2) we chose the commonly used Otsu's method (3) for simplicity.

A series of 50 computer-generated MIBs, developed to target diagnostic features in pigmented lesions, of two types were extracted from each image: A set of 43 single-color-channel (SC) MIBs was extracted from single color channels, and a set of 7 Multi-Color (MC) MIBs was extracted from the entire RGB color channel triad directly. While MC MIBs were derived from variation between the gray-scale values in the separate color channel images and also derived from variations in the MIBs derived from the various color channel data, three versions of each SC MIB (one from each color channel) were extracted for each lesion. For example, the SC mIB $B12$ illustrated in Figure 1 and Figure S2 was extracted from each of the 3 RGB color channels of each of the 120 dermoscopy images. The rationale for extracting each mIB from each color channel was that it was immediately noted that the MIBs had different values when extracted from the various color channels. An example of how this phenomena lead to diagnostically preferential extraction from particular channels is shown in Figure S4. Overall, 27 MIBs extracted from single color channels were statistically significant ($p<0.05$) for melanomas versus nevi. Four additional statistically significant MIBs used information from all color channels. Mathematical formulae for each of the statistically significant MIBs are given in section S5.

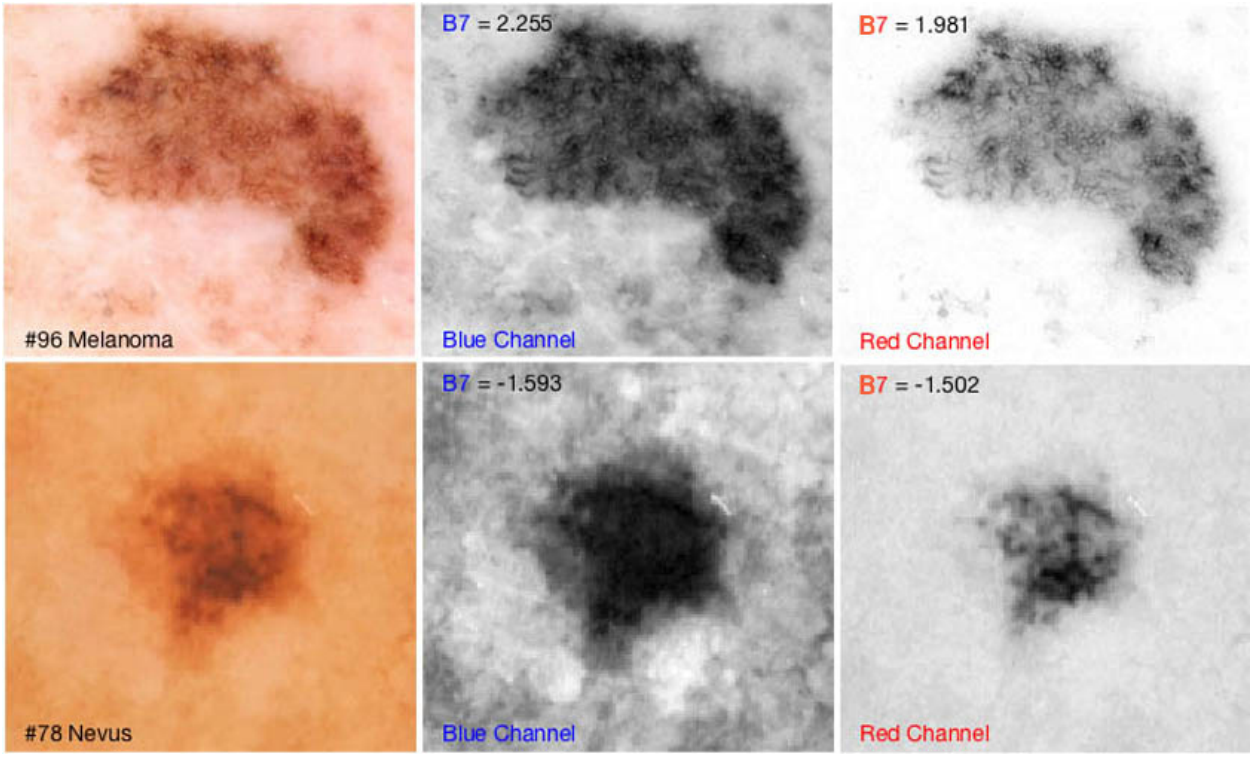

Figure S4. Spectral Diagnostic Content Variance
MIB B7 was chosen as evaluated on the blue channel over its red color-channel version B7 because in the blue channel, MIB B7 had a greater differential between nevus and melanoma. MIB B7, adapted from Equation 13 and visualized in Figure S5, quantified deviations in pixel brightness from a smooth mathematical model near the edge of the lesion.

Some of the MIBs directly targeted standard dermoscopy analysis features, e.g., number of colors present, diameter, characteristics of pigment network branches (4), and features of symmetry and border. Other MIBs were indirectly impacted by standard dermoscopic melanoma features such as: atypical pigment networks, atypical globules, off center blotches, peripheral tan structure-less areas and regression structures.

A large set of novel MIBs were created based on our novel angular sweep analysis. We quantified brightness variation on an angular sweeping arm (Figure S2f) that connected the geometric center of the lesion and a point on the border tracing that border clockwise. From the center, radial arms projected to the lesion border and rotating clockwise were used as regions of interest to quantify image characteristics along the arc of rotation. The series of arcs created by radial sweep around the center covering the entire 360-degree view of the lesion, was analogous to the sweep of hands around an analog clock. The MIB-producing mathematical operations (given in section S5) either produced direct transformations of the actual data (i.e. Figure S2) or quantified differences between the data and mathematical models used to estimate the data's deviation from smoothly transitioning features (i.e. Figure S5).

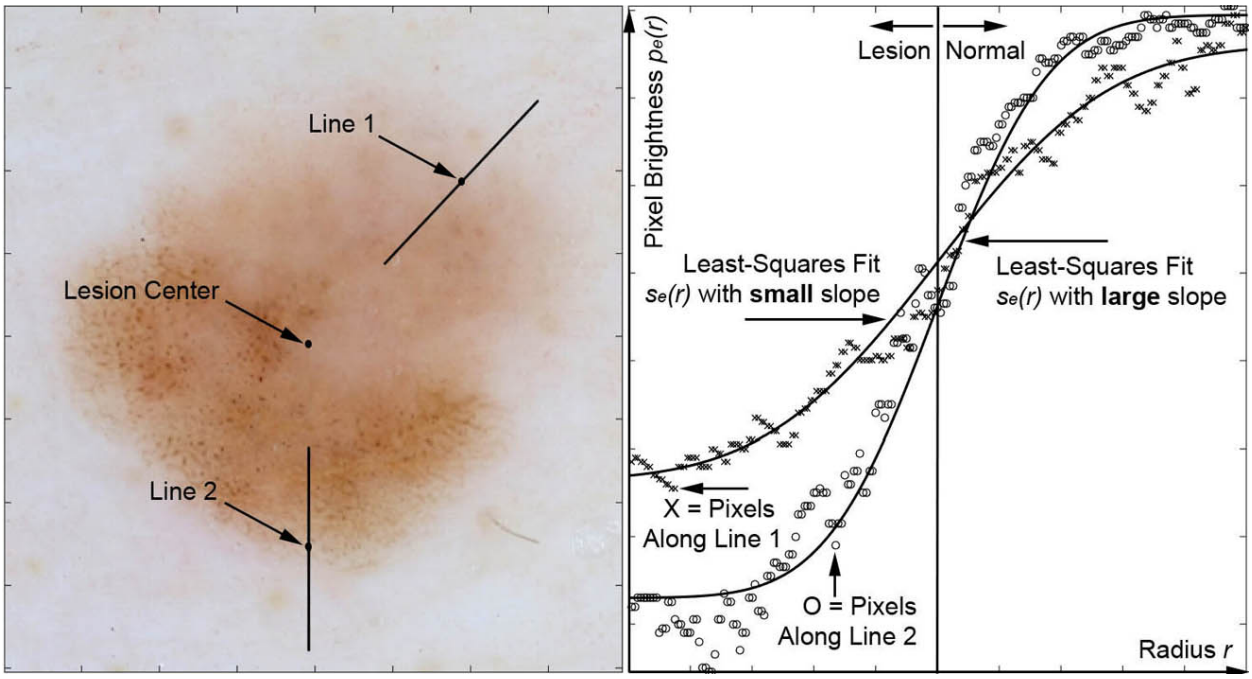

Figure S5. Fitting for Edge Demarcation Edge demarcation was quantified as the slope of the transitioning from dark pixels inside the lesion to bright pixels outside the lesion. Increased slope of the fitting mathematical function resulted from increased lesion border demarcation. The two radial lines (Line 1, Line 2) drawn on the lesion include the lesion border from inside the lesion where the pixels are dark inside the lesion to outside the lesion where the pixels are bright in normal skin indicate illustrate two locations where the demarcation is gradual (Line 1) and sharp (Line 2). The pixel brightness extracted along these two lines (x for Line 1 and o for Line 2), $p_e(r)$ was fit to a mathematical model), $s_e(r)$ to yield the fitting parameters, which were used to produce MIBs B3, B4, B9, B13, B14, R1, R5 and R10 marked "Border" in Figure 2. This includes the edge demarcation slope, which is the slope of the solid line at the lesion border between normal skin and lesion and the error in the fit, which is the sum of the squared differences between the data points, $p_e(r)$ and the error function fit (solid line) , $s_e(r)$. Melanomas had a sharper border, a higher degree in variability of border sharpness and a greater fitting error.

The full set of 50 MIBs was filtered by excluding 17 non-significant ($p>0.05$) MIBs using univariate two-tailed unpaired t-tests (for continuous variables) and Fisher Exact Test (for categorical variables). Eighty-six percent of the 33 significant MIBs had larger values for melanomas than for nevi. Significant SC MIBs were found in the blue channel (B1-B15), in the red channel (R1-13), and in the green channel (G1).

Among the most diagnostically significant SC mIBs were edge demarcation features (e.g. metric $B1$, Equation 10, Figure S6), where a sharper edge (higher slope) and more edge sharpness variation were present in melanomas. A program to identify length, darkness, and end points of each branch in a pigment network is illustrated in Figure 1 and this information was used for MIBs targeting pigmented patterns such as the variation in pigmented network branch length. The complete significant MIB set is shown as a heat map in Figure S6.

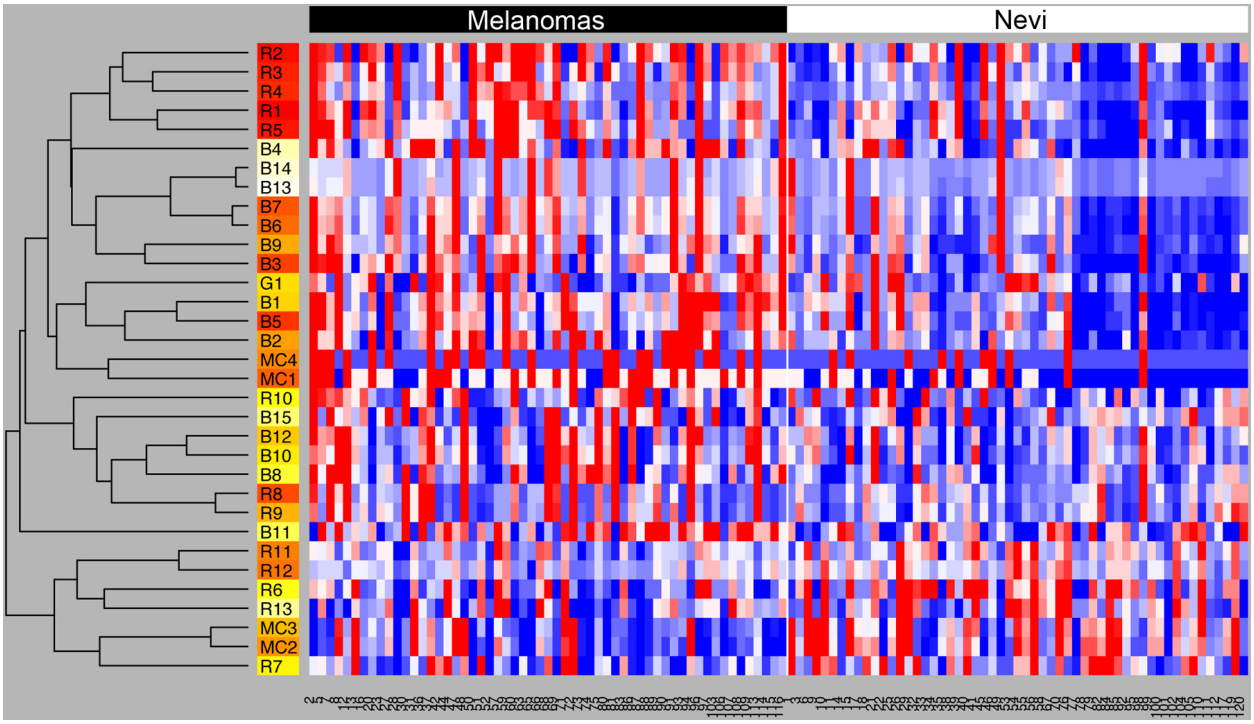

Figure S6. Imaging Biomarkers Heat Map Correlation structure among imaging biomarkers and their discriminative ability are represented in a heat map built with Hierarchical clustering. Distances among MIBs were evaluated with Euclidean distance and Mcquitty algorithm for the standardized set of measurements (z-scores). In the heat map, the binary partition at the top of the hierarchy isolates a subset of MIBs in which z-scores evaluated for nevi are, on average, higher than the ones in melanomas. Those MIBs are: R6, R7, R11, R12, R13, MC2 and MC3.

There was a pattern of mutual exclusivity of SC MIB importance in the red and blue channels. The data in Figure 2 show that most MIBs were best computed in particular color channels, with highest diagnostic value in blue or red channel. Figure S7 illustrates the most significant 6 metrics with examples of lesions for which they yielded high raw values versus low raw values and figure S4 illustrates visually the importance of picking the color channel where each MIB is most significant. In this case, the MIBs are expressed as coefficients of variation compared to the average MIB value for the entire data set. The MIB B7, which is the degree to which the radial lesion edge profile data (see Figure S5) deviate from a smooth mathematical edge function, has a greater difference between the melanoma and the nevus in the blue channel because in the blue channel, the lesion appears homogeneously dark. The melanoma evidences its deviation as the bright and dark structures are retained to large extent when isolating the blue channel.

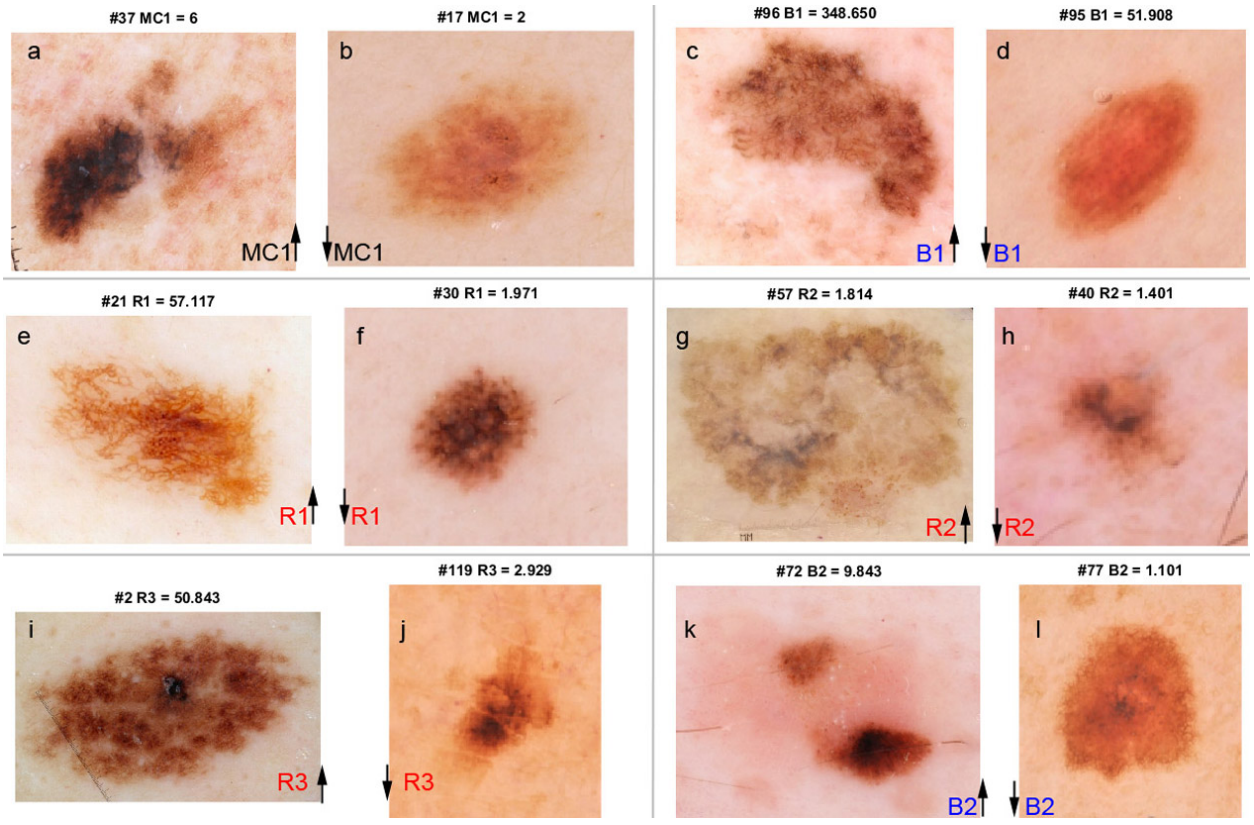

Figure S7. Good Metrics The six most diagnostic MIBs (MC1, B1, R1, R2, R3, B2) yielded high values on (a, c, e, g, i, k) and low values on (b, d, f, h, j, l)
a-b Lesion #037 and #017 contain 6 and two colors, respectively. Lesion 037 has an atypical network with central white area with peppering (i.e. atypical dots) suggestive of regression.
c-d Lesion #096 and #095 contain different degrees of instantaneous brightness shifting over the angular sweep. Lesion #096 also contains an atypical reticular pigment network, which is shown in detail in Figure 1.
e-f Lesion #021 and #030 contain different levels of variation of the demarcation edge slope as illustrated in Figure S5. Lesion #021 contains dark brown and light brown atypical reticular network patterns with dots, asymmetry and uneven border. The reticular pigmented network has dots overlying the pigment network and areas of missing pigmented network which could be interpreted as areas of regression.
g-h Lesion #057 has a higher fractal dimension than #040. Lesion #040 contains polygonal lines suggestive of an atypical network, although they may be distributed by hair follicle and off centered/ atypical dots.
i-j Lesion #002 and #119 contain different ratios of branch points to end points in pigment network analysis. Lesion 002 contains a patchy reticular pattern with the melanoma specific feature off centered blotch. The dark region surrounded by lighter pigment with even lighter areas of potential regression caused variation in the angular analysis and a high Q-score. Lesion 119 contains light brown homogenous lesion with dark brown off centered blotch.
k-l Lesion #096 and #095 contain different levels of variation over the angular sweep of variation along the sweeping arm.

S3 – Statistical Methods: MIB Combination To Form The Melanoma Q-score

In the second computational step where the predictive model was built, the 33 MIBs achieving diagnostic statistical significance ($p<0.05$), 4 MC and 29 SC (using the most significant RGB color channel version), were input into the 12 statistical/machine learning algorithms as predictive informatics programs listed in Table S2. Our strategy for MIB selection was based on choosing the SC MIB from the best color channel when there was any color channel for an MIB that achieved statistical significance, including that SC MIB, and also including any significant MC MIB.

The statistical classification methods for machine learning were chosen to represent the broad universe of base classifiers (5). Figure S8 shows an example classification method, the C5.0 decision tree, that prioritized a subset of the MIBs and conducted a series of comparisons between these MIBs and a set of thresholds to ultimately lead to a classification of each lesion. Each method output the melanoma likelihood for each lesion.

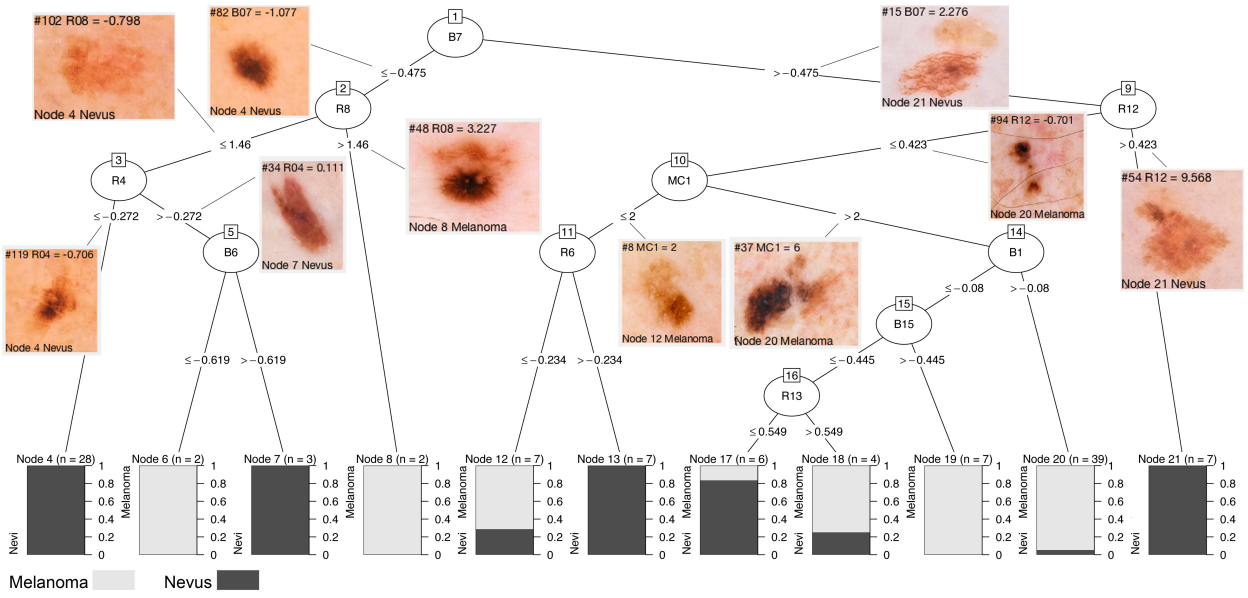

Figure S8. Decision Tree Built With The C5.0 Algorithm The algorithm was applied to predict lesions type (melanoma vs. nevus) with the full data set that included 112 lesions and 33 MIBs. The final decision tree has 10 decision nodes (#1, #2, #3, #5, #9, #10, #11, #14, #15, #16) and 11 terminal nodes (#4, #6, #7, #8, #12, #13, #17, #18, #19, #20 and #21). The algorithm selected decision nodes based on four mIBs from the blue channel (B1, B6, B7, and B15), five mIBs from the red channel (R4, R6, R8, R12, and R13) and one multicolor mIB (MC1). At the terminal nodes the proportion of melanomas (light gray) and nevi (dark gray) are shown with stacked bar plots. The final classification has yielded 7 pure terminal nodes (#4, #6, #7, #8, #13, #19 and #21) where melanoma or nevi have 100% prevalence. The nodes #4 and #20 together have 59.8% of the lesions and they perfectly discriminate nevi and melanoma, respectively.

The likelihoods produced by all methods were combined into the overall endpoint of the analysis, the ultimate best estimate of melanoma probability between zero and one, the melanoma Q-score. Thus, we built a predictive model that combined the MIBs into a risk score for probability of melanoma. Figure S9 shows how the Q-scores of the entire image set were optimized to fill the dynamic range from 0 to 1.

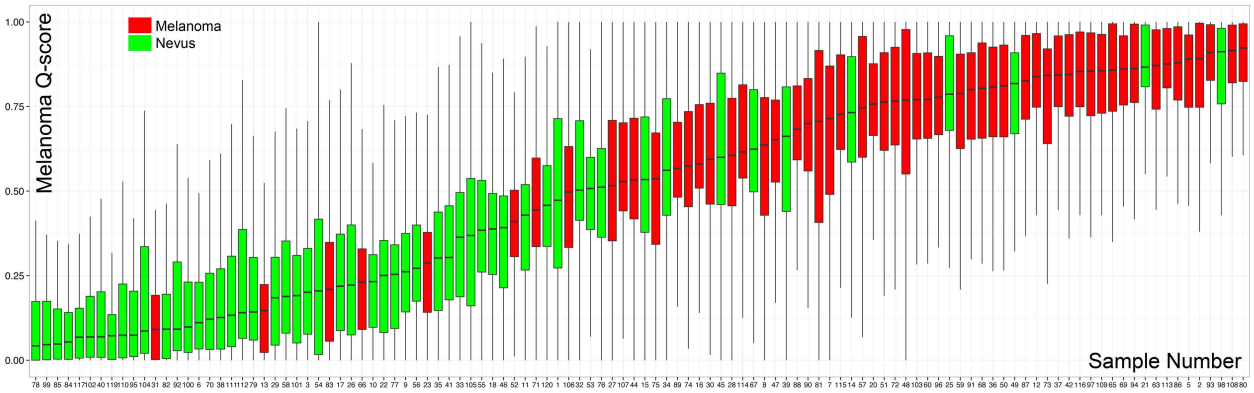

Figure S9. Distribution Of The Melanoma Q-Score Over The Entire Data Set
The melanoma Q-Scores for each lesion (x-axis) were obtained by the median melanoma probabilities according to the classification methods in Table S2. These median Q-scores are shown as horizontal black lines. The box-plot diagram shows, for each lesion, the variability in Q-Score from the Monte Carlo experiment described in the Statistical Methods (see supplementary material). A total of 1000 training and test sets with 75% and 25% of the lesions, respectively were randomly selected and every time a particular lesion was included in the test set, the Q-Score was evaluated. The number of evaluations for each lesion was approximately 250, the expected number of times a lesion would be selected as part of the test sample. In this diagram, the 25th, 50th (median) and 75th percentiles for the Q-Score, over different test sets, are delimiters for the boxes and the whiskers are obtained from the minimum and maximum Q-Score. The lesions are ranked on the x-axis from the lowest median Q-Score to the highest.

Figure S10 illustrates several examples of nevi, melanomas and their assigned Q-scores to provide a frame of reference for the overall classification method. Shown are examples of lesions that were well classified (i.e. true positives and true negatives) versus images that were not (i.e. false positives and false negatives). Since sensitivity is more important than specificity, the characteristics of misclassified melanomas were particularly important. Two of these melanomas are largely “featureless,” and two others appear to have significant regression, which is not directly scored in the current method.

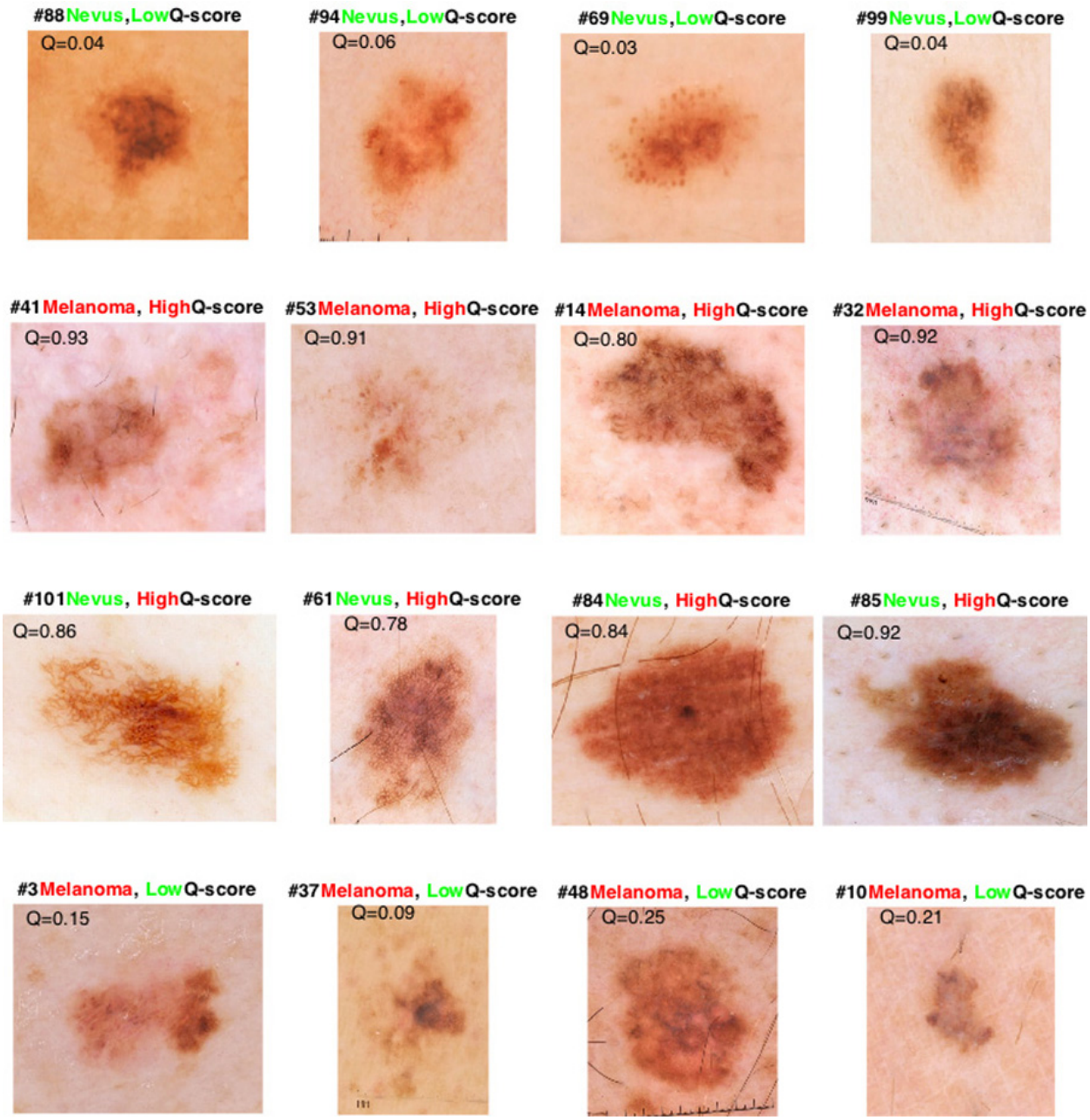

Figure S10. Well vs. Poorly-Classified Images Ensembles of well-classified (top two rows) and poorly-classified (bottom two rows) lesions.

Our framework was set first to identify the most discriminative MIBs upon which the predictive model will be built. These MIBs and their discriminative ability are illustrated in the heat map in Figure S6. To this end, we first evaluate the differences between melanoma and nevus for each one of the seven multicolor MIBs and also for the RGB channel-specific MIBs (41 MIBs for each Red, Green and Blue channels). For this univariate assessments, two-sided unpaired t-tests, Wilcoxon-Mann-Whitney and chi-square tests were used for continuous (e.g. MIB B1), ordinal (e.g. MIB MC1) and

categorical (e.g. MIB MC4) MIBs, respectively. Of the total 130 MIBs, 33 (4 multicolor, 13 Red, 1 Green and 15 Blue MIBs) were selected as the most significant discriminators ($p<0.05$) between melanoma and nevi (See Figure 2) to continue to the multivariate discrimination stage. When significant differences for a given MIB were found in more than one channel the most discriminative channel regarding its p-value was selected. Among the selected 33 most discriminative MIBs were Blue(15), and Red(12)-channel MIBs, followed by four multicolor MIBs and only one MIB from the Green channel. This set of 33 discriminative MIBs measured from all the 112 lesions were used as inputs for our predictive model.

S3.1 – Predictive model

We used a collection of 12 classification methods that range from simple to sophisticated and altogether cover different data structures. The collection of classification algorithms is listed in Table S2 and includes the K-nearest neighbors (KNN) (6), a simple and efficient nonparametric distance-based method that has been successfully applied for more than 60 years. Artificial neural networks (7) and support vector machines (SVM) (8, 9) were included as to represent high-dimensional complex nonlinear functions. To accommodate complex interactions between predictors we incorporated four methods following the decision tree/recursive partitioning paradigm: CART (10), C5.0 (11), Multiple Adaptive Regression Splines (MARS) (12) and Random Forests (13). Logistic regression (14) and Linear Discriminant Analysis (15) are based on solid statistical foundations. The former permits inference upon parameters by a probabilistic approach and the latter is one of the oldest techniques for dimensionality reduction. Partial Least Squares (PLS) regression (16) (17) is of more recent development and simultaneously performs regression/classification and dimensionality reduction.

S3.2 – Model Estimation/Training

To estimate each of the classifiers' parameters and to evaluate the distribution of the prediction error empirically, we created a Monte-Carlo experiment. During each of 500 iterations, the set of 112 lesions was randomly partitioned into training (75%) and test (25%) sets. For each classifier, model parameters were estimated by maximizing a partial area under the ROC curve obtained by limiting the specificity to be within the range 0-40% and tuning parameters were estimated by 10-fold Cross-Validation. The best configuration for each classifier was used to predict the 25% hold-out lesions in the test set. The performance of each classifier on the test set (across the 500 iterations) is shown in Figure S3 and clearly indicates that the CART(RP) algorithm is outperformed by all others. Notably, as measured by the AUC curve, the classifiers performance become more similar in the specificity range 0-0.4 where the partial AUC was maximized. Despite the fact that the Random Forests algorithm performed slightly better than the Q-Score (see Figure S3) in the region of high sensitivity, the Q-Score was more robust over the entire range of specificities, and it reached the highest value for the Total Area Under the Curve. Moreover, considering the bootstrapped 95%

confidence interval estimate for the Partial Area Under the Curve (specificity between 0 and 0.4) both Random Forests and Q-Score were equivalent.

Recent developments in machine learning and pattern recognition have been shown that an ensemble of predictive algorithms is likely to generate more accurate predictions than a single algorithm (18), (19). Following this paradigm, we developed the melanoma Q-Score, a diagnostic for melanoma discrimination obtained by evaluating the median probability across K available classifiers.

$$Q_{Score} = median\{Prob_i(Melanoma|\mathbf{M})\}; i = 1, 2, \ldots, K$$

Where $\mathrm{Prob_i} \in \{0,1\}$ is the probability of the lesion being a melanoma, as predicted by the i-th classifier based on a set of MIBs $\mathbf{M}$. The Monte Carlo experiment setting allows us to obtain the empirical distribution of the Q-Score for each lesion, which is represented in Figure S9. The Q-score distribution shows that the number of false-positives (melanomas classified as nevi) is lower than the false-negatives; indicating that our classification strategy is more sensitive than specific. If we set up a threshold of 0.5 for the Q-Score, only in 5 melanomas (lesions D031, D13, D83, D66 and D23) the 75th percentile of the Q-Score distribution is below 0.5.

Classification performance was assessed through a set of standard approaches including: sensitivity, specificity, positive and negative predictive values, as presented in Supplementary Table S4. Supplementary Table S4 shows that the strategy (Best Channel+Multicolor) outperforms the classification obtained within any individual color channel, resulting in a classification 8% more sensitive than the most discriminative channel (Blue).

S4 – Discussion

S4.1 – Clinical Impact and Translation

Our results show that discrete, quantitative MIBs achieved diagnostic significance. Combining MIBs with machine learning to stratify risk during screening may augment current techniques that guide biopsy.

Because false negative screening must be avoided, we evaluated our technology in the high sensitivity range of the ROC (Figure S3). The Q-score diagnostic precision (sensitivity/specificity=98%/36%) is exceeded by the technique (sensitivity/specificity =98%/44%) (20) it was initially modeled after and the preliminary study (sensitivity/specificity=98%/44%) (21) of the commercial technology that achieved se/sp=98%/10% in subsequent studies (22). Because no other reports in the literature specify a sensitivity of 98% or greater, we believe our result to be the most promising in the high sensitivity range that is required for clinical utility.

Though we report the median melanoma likelihood produced by the various machine learning approaches as the Q-score, one approach (the C5.0 decision tree approach) outperformed the Q-score at 98% sensitivity, yielding sensitivity/specificity = 98%/44%.

This result was produced using branching logic. Figure S8 shows an illustration of this branch choice approach, which may be the most promising approach as instructive to visual examination. This analysis may be able to "teach back" to dermatologists both new visual dermoscopic features and new ways to combine MIB evaluations sequentially. Our full diagnostic, which involves calculating a melanoma probability for each machine learning approach and taking the median of those probabilities as the Q-score can likely be reduced for translation to visual screening. Thus, automated, unconstrained digital engineering of diagnostically optimized decision trees revealed intriguing analytical pathways, though the feasibility of translation for human use without computer vision remains to be shown. Also, it may be desirable from the point of view of decreasing computation time and/or complexity during evaluation to reduce the number of statistical classification approaches from the 13 used here to 1 or 2 (e.g. the C5.0 tree in figure S8) if such a small subset continues to outperform the Q-score ensemble approach that uses the median result of all the classifiers.

The goal of algorithms performing automatic differentiation, diagnosis, and/or screening is to avoid subjective judgment. However, there are many potential dermoscopic MIBs and it is unclear if a "one size fits all" approach will be optimal, since dermatologists evaluate features that may be particular to their patients' skin type and other factors. Furthermore, MIBs and our statistical approach may be applicable to screening and differential clinical identification of other skin diseases and conditions besides melanoma such as non-melanoma skin cancer and seborrheic keratosis. To remain objective in this study, we removed the least diagnostic (below the $p = 0.05$ threshold, see Figure 2) from the MIB library. However, the significance of various MIBs will change when this analysis is applied to different image databases, from different dermoscopists, on different patient populations. Future investigation must quantify the variability of the level of statistical significance of each MIB among image databases. The reproducibility of each MIB may become a second-degree quality measurement to the p-values used in this work.

S4.2 – Limitations

Though promising based on the ROC in Figure S3, our algorithm for melanoma detection had several limitations. The limitations of the algorithm were:

1) In the image-processing phase, eight images (Figure S11) were excluded due to the algorithm's inability to automatically identify the lesion border. The inability of the algorithm to converge on a lesion border in 6.7 % of the cases was due to either excessive hair or lesion segments that extended beyond the image border, particularly in lesions that had extremely subtle border in which the whole field of view could be considered lesion. In 11 single color channels of 10 lesions, the algorithm identified two lesion segments separated by normal skin. In these cases the largest lesion segment was chosen as the lesion segment and smaller segments were ignored. In 9 out of 10 of those cases, the algorithm completed the border analysis but in one the algorithm failed. That failure along with 7 other failures are the 8 cases out of 120 shown in Figure S11 where the algorithm failed to produce a result due to the inability to identify the lesion border.

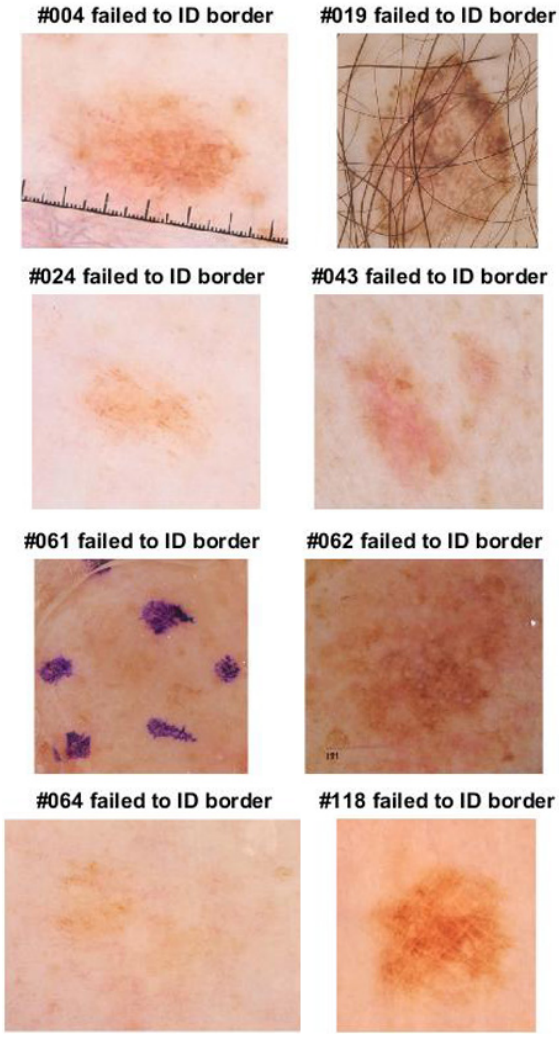

Figure S11. Failed Border Detection In the eight cases for which the algorithm failed to determine the lesion border, the lesions that did not result in a defined border were either covered with hair or markers or had borders that extended beyond the field of view in one or more color channels. Lesions 019 and 062 were melanomas, the rest were nevi.
Lesion 019 was covered too heavily with hair
Lesion 043 had two lesion segments in both the blue and green channels
Lesion 061 had pen markings that were too difficult to remove
Lesion 062 had a poorly defined border that extended beyond the field of view

2) Our algorithm was only semi-automated. Non-automated tasks were the cropping of the image to include a spatial range roughly equal to twice the lesion diameter and choice of sample dermoscopic colors in the data set. Though the training exercise of choosing dermoscopic colors led to good overall performance in mimicking what trained dermoscopists identify as particular colors, at a rate of about 10%, the trained algorithm identified colors that may not be supported by trained dermoscopists.

3) The limitation to generalization of the Q-score diagnostic thus far is that it considered only one group of clinicians with one dermatoscope and camera combination. A logical expansion to include multiple devices and users is needed to quantify inter-observer agreement, although the approach may be optimized with a single imaging platform. We must determine the degree to which the Q-score needs training for each imaging platform and user. Future work will analyze multiple data sets from different users, identifying methods to overcome variations in such data towards a generalizable endpoint diagnostic that may be presented as a mobile medicine solution.

4) The study cohort of 120 patients represents a small sample size. By comparison, the study size that enabled the Melafind system to achieve FDA approval was approximately ten times larger. We intentionally chose a small data set on which to develop our algorithm with the intention to expand to multiple data sets in future studies. Our choice of a small data set for this work facilitated the evaluation of feature recognition efficacy, visually, on the entire data set. This made rapid development possible. The ROC curve we present in Figure S3 with the point showing 98% sensitivity and 36% specificity is an average result that was obtained after evaluating probabilities of melanoma and nevus in randomly selected test sets. The data was

balanced regarding the prevalence of positive cases and the evaluation in this Monte Carlo experiment was performed in an attempt to overcome the small sample size.

5) Only a small fraction of the established dermoscopy nomenclature has been automated in this report and the MIB library needs expansion. One important example is that the presence of regression and negative networks is not specifically targeted by any MIB. These features impact the MIBs indirectly but are worth some future efforts to specifically target.

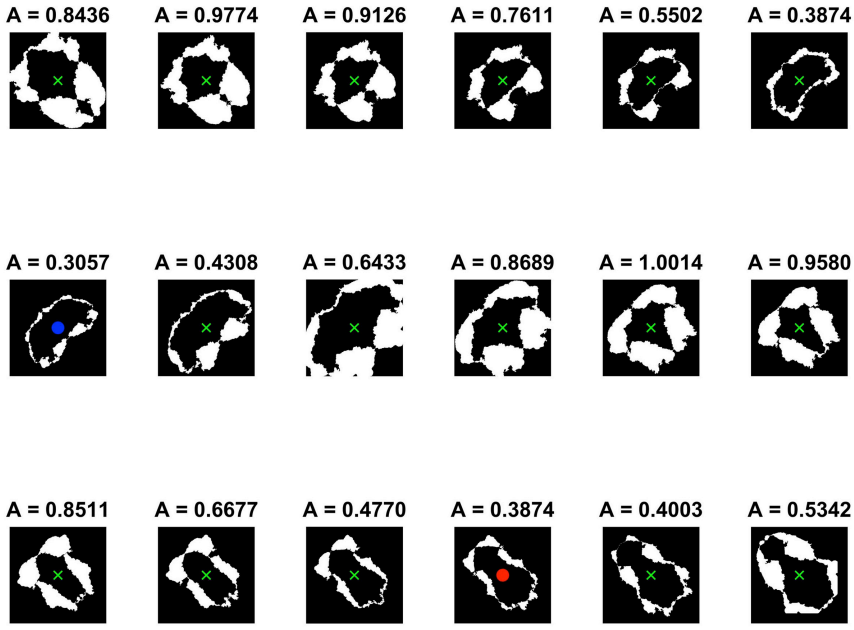

Figure S12. Determination of Border Asymmetry The border asymmetry was analyzed perpendicular to the axis of most symmetry as dermoscopy texts train practitioners to do. Each image in this figure represents analysis at a particular angle, in increments of 10 degrees, over a 180-degree range. The image whose center is marked with a blue dot indicates the angle where the lowest variable value was found and the image whose center is marked with a red dot indicates the perpendicular angle where the MIB was scored. The MIB R13 measuring border asymmetry as determined in the clinically conventional way, the degree of symmetry of the silhouette of the lesion, proved to be less effective at differentiating melanoma from nevi thank MIB R11, which, instead of evaluating A at the figure marked by the blue dot, simply took the maximum value of A and divided it by the elipsity of the lesion.

S4.3 – Competing Technology

Though there are a number of mobile medicine services that will transfer an image of a pigmented lesion to a certified dermatologist, there are relatively few that offer any automated diagnostic capabilities. The FDA-approved device and service to automate interpretation of pigmented lesions used “wavelet” transforms of features that largely distinguish normal, benign nevi from melanomas in a proprietary analysis algorithm. As commercialized, this approach had 98% sensitivity and 10% specificity for melanoma classification (22) which is far inferior to accuracy of the best manual methods. (20) (23) These mathematical transformations of visual characteristics produce diagnostic data that are not strongly and intuitively linked to visual or structural features of pigmented lesions. Early App. technologies may miss-classify 30% or more of melanomas as un-concerning. (24) Some have high sensitivity such as SkinVision and others are more specific such as Skin Analytics but none yet deliver significantly more than 10% specificity at the requisite 98% sensitivity that is clinically useful. Thus, in this work, we

set goals of having high sensitivity and at least as good specificity as the median dermatologist in the USA, with the added benefit of potentially instructing dermatologists about new image characteristics that may be helpful in overall lesion evaluations, and with the potential to eventually expand non-expert, standardized approaches into primary care settings.

The emergence of Watson Health from International Business Machines, Inc. represents an upcoming trend that will likely set the pace for advancing computational diagnostics. Watson is originally a project that aimed to use language processing and machine learning to reveal insights from large amounts of data. Watson analyzes this data and through trial and error improves its performance, becoming smarter and smarter over time. (25) While the first efforts were made towards language recognition, Watson has lately been focusing on using its databases for medical purposes. Notably, the group working closely with the Memorial Sloan Kettering Cancer Center to train its software to recognize many different kinds of cancers, including melanoma. (26)

This tool could be used for diagnosis on personal devices (such as phone apps), but currently Watson is focused on creating a support system for physicians specifically. The system would prompt the physician for characteristics and questions regarding the patients, and the software would then present several hypotheses with various degrees of confidence, as well as providing publications and guidelines used to form those.

While Watson seems to be working towards automated machine diagnosis, there have been no major efforts to translate this technology into accessible medical software for non-physician users. (27) Though non-peer reviewed internet publications claim that Watson is 95% accurate in identifying melanoma (28), we could find no rigorous tests of Watson's melanoma diagnostic potential in the literature. In the context of early melanoma detection based on computational image processing, there is a limitation to Watson Health in that although images can be fed directly into machine learning approaches with gold standard histological diagnoses as training material, the dermoscopic diagnostic criteria are not yet digitized. Thus, unconstrained machine learning does not yet leverage dermoscopy pattern recognition that has been developed to date. The digitization of dermoscopic features achieved by our MIBs is one of the components that this work cannot be achieved by machine learning alone. MIBs provide constraints that will likely outperform unconstrained machine learning by leveraging visual pattern recognition. Our second machine-learning step (after the MIBs that digitize the dermoscopy features are created) can be substituted with the Watson Health analysis.

S4.4 – The Light-tissue Interaction

Some logical clinical MIBs correlated well with melanoma such as the presence of multiple dermoscopic colors (marked "Color" in Figure 2). Other intuitive MIBs, such as the asymmetry of the lesion border ("Asymmetry" in Figure 2, illustrated in Figure S10) had lower values for melanomas than for nevi and were of marginal diagnostic significance.

Our results supported the general principle that melanomas have more spectral and structural irregularity than benign nevi and that red, blue, and green light-based images show differing tissue characteristics. MIBs naturally segregated as either diagnostic in the blue channel or the red channel, but not both.

A hypothesized mechanism for the spectral dependence of diagnostic importance is deeper penetration by longer wavelengths (eg. red), and thus ability to differentially visualize differing 3-dimensional tissue or chromophore characteristics of melanoma invading the dermis as well as superficial epidermal imaging by the shorter (eg. blue) wavelengths of basal layer atypia or junctional nests of melanocytes associated with melanoma. The green channel may have included some information from both but mixed the information content becoming nonspecific to either. This study raises the question whether ultraviolet and infrared imaging may enable even more specific surveillance of the superficial and deep atypia, respectively. The resulting spectral dependence of diagnostic importance (Figure 2) suggests that RGB MIBs may be missing spectral information to be probed using MIBs in conjunction with hyperspectral imaging.

Though the melanoma Q-score combined information from the red green and blue channels, analysis at isolated single color channels showed the blue channel outperformed the red channel and the red channel outperformed the green channel (Table S4).

The diagnostic criteria we developed still need to be related to underlying tissue structure, including proliferative and invasion patterns of melanoma cells, and molecular pathways that could produce differing pigmentation. Beyond these features, one also needs to consider the contribution of hemoglobin to lesion pigmentation. Optically testing hemoglobin for saturation/desaturation may effectively detect metabolically active areas of growing melanoma cells or active areas of immune-response to tumors. If improved with additional analysis of multiple discrete wavelengths in the green range, hemoglobin spectroscopy may enable oxymetric measurement of metabolic activity as well as surveillance of polymorphic vasculature associated with melanoma and other skin cancers such as basal cell and squamous cell carcinoma. Within a hyperspectral image of a pigmented lesion, one can include measures of hemoglobin saturation and desaturation, which may help to identify metabolically active regions within lesions. The basis for the “steeper edge slope” (MIB $R5$) in melanomas does not yet have a cellular basis, but we speculate that it might represent growth of melanocytes in nests at the dermal-epidermal junction at the edge of a melanoma, whereas atypical nevi tend to have only individual junctional melanocytes (nevus cells) that are decreasing in number at the edge of this kind of lesion, whereas deeper nests of melanocytes/nevus cells are organized in the central or “body” region of an atypical nevus. (29)

S5 – Mathematical formulae for MIBs

S5.1 – Multi-Color MIBs

The process for dermoscopic color identification was as follows: Sample dermoscopic colors were chosen within the data set blind to the gold standard diagnoses and visually inspected the entire data set. At least three examples of each dermoscopic color were chosen for pixel extraction and generation of the threshold values in Table S3. This simplified colorimetric approach then classified each lesion pixel in each image as potentially one of 6 dermoscopic colors [light brown, dark brown, black, red, blue-gray and white].

For each potential color, if the pixel ratio of red to blue (R/B) was within one standard deviation of the mean for that color, and the same was true for R/G and B/G, then the pixel was assigned that color. For each pixel, a sequential check was made for the presence of colors in the order [light brown, dark brown, black, red, blue-gray and white]. In this manner, the two most common colors [light brown, dark brown], were first identified as the least suspicious group. Next, red black and red were identified as more suspicious. Finally blue-gray and white were identified as most suspicious. The algorithm checked each pixel for each color, leaving the assignment of the last checked (most suspicious) color as the designated color for that pixel.

A color list *CL* was produced for each lesion indicating the presence or absence of each color. For instance *CL* = [1 1 1 0 0] would result from a dermoscopic image where the lesion contained light brown, dark brown and red but no black or blue-gray/white. *MC1* is then the number of dermoscopic colors identified in the lesion.

$$MC1 = \sum_{i=1}^{5} CL(i) \tag{1}$$

Let *L(y,x)* denote an image mask of the lesion segment with value 1 inside the lesion and value 0 outside the lesion. Let $\mathrm{L_{red}(y,x)}$, $\mathrm{L_{green}(y,x)}$ and $\mathrm{L_{blue}(y,x)}$ be masks derived from the red, green and blue channels of the color image, respectively. MC2 is then the normalized difference in lesion size between the red and blue color channels

$$MC2 = \frac{\sum_{x=1}^{Nx}\sum_{y=1}^{Ny} L_{red}(y,x) - \sum_{x=1}^{Nx}\sum_{y=1}^{Ny} L_{blue}(y,x)}{\sum_{x=1}^{Nx}\sum_{y=1}^{Ny} L_{red}(y,x)} \tag{2}$$

Let $\mathrm{R(\theta)}$ be the length of the radial between the geometric center of the lesion and a point on the lesion border that sweeps over the angle θ from $\theta = 0$ to $\theta = 2\pi$ radians. Let $\mathrm{R_R(\theta), R_G(\theta)}$ and $\mathrm{R_B(\theta)}$ be three versions where the geometric centers and the borders are those extracted from $\mathrm{L_{red}(y,x)}$, $\mathrm{L_{green}(y,x)}$ and $\mathrm{L_{blue}(y,x)}$, respectively.

$$R_{var}(\theta) = \frac{\sigma(R_R(\theta), R_G(\theta), R_B(\theta))}{\langle R_R(\theta), R_G(\theta), R_B(\theta)\rangle} \tag{3}$$

MC3 is then the mean coefficient of variation of lesion radii among the color channels, where ⟨⟩ denotes the expectation value or mean operator.

$$MC3 = \langle R_{var}(\theta) \rangle|_{\theta=0}^{\theta=2\pi} \tag{4}$$

where, as an illustration of the definition of the mean value, for a set x that contains n elements:

$$\langle x \rangle = \frac{\sum_{i=1}^{n} x_i}{n} \tag{5}$$

MC4 is the binary presence of blue-gray or white in the image.

$$MC4 = CL(5) \tag{6}$$

S5.2 – MIBs With Blue Channel Diagnostic Significance

Let $\mathrm{p}(\mathrm{r}_1)$ be the pixel brightness along a radial line r_1 connecting the center point of the lesion and a point on the peripheral edge of the lesion. Let $\mathrm{R_m}(\theta)$ be the mean pixel brightness $\langle \mathrm{p}(\mathrm{r}_1) \rangle$ along a set of lines that vary as specified by the angle θ. As θ varies in increments of $\mathrm{d}\theta$ one full rotation from zero to 2π radians (360 degrees), the set of lines r_1 sweep the lesion like a clock arm sweeping an analog clock.

$$R_m(\theta) = \langle p(r_1) \rangle|_{\theta=0}^{\theta=2\pi} \tag{7}$$

$$R_{std}(\theta) = \sigma(p(r_1))|_{\theta=0}^{\theta=2\pi} \tag{8}$$

where, an illustration of the definition of the standard deviation, for a set x that contains n elements:

$$\sigma(x) = \left(\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \langle x \rangle)^2 \right)^{\frac{1}{2}} \tag{9}$$

B1 is then the average of the absolute value of the derivative of $\mathrm{R_m}(\theta)$ over the angular sweep is the mean instantaneous brightness shift from one angular analysis position to the next over the entire 360-degree angular range.

$$B1 = \langle (|R_m(\theta_n) - R_m(\theta_{n+1})|) \rangle|_{\theta=0}^{\theta=2\pi} \tag{10}$$

B2 is the variance over the angular sweep of the variance in pixel brightness over the radial sampling arm. This variable is increased when there are some angles at which the lesion contains even pigmentation but others that contain variable pigmentation such as in reticular or globular patterns of bright and dark areas.

$$B2 = \sigma(R_{std}(\theta))|_{\theta=0}^{\theta=2\pi} \tag{11}$$

Let $p_e(r_2)$ be the pixel brightness along a second radial line r_2 of the same length as r_2 and at the same angular sweep angle θ but extending from half-to-1.5 times the lesions radius $R(\theta)$ instead of 0-to-1 such as to be centered on the border between lesion and normal skin. $p_e(r)$ has the characteristic that half of its pixels (within the lesion) are darker than the other half of its pixels (outside the lesion). Let $s_e(r)$ be a mathematical model error function across the lesion border with three fitting parameters: Min, Max and $Slope$ that are iteratively adjusted to minimize the least squares difference between $p_e(r)$, the data and $s_e(r)$ (Figure S5). $erf(x)$ is defined as twice the integral of the Gaussian distribution with 0 mean and variance of 1/2, as shown below with the dummy variable t. Considering r_b as the lesion border pixel with approximately the mean pixel brightness in $p_e(r)$ and exactly the mean brightness of $s_e(r)$, $s_e(r)$ is defined as:

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt \tag{12a}$$

$$f_e(r) = \frac{erf(\frac{r-r_b}{Slope})}{2} \tag{12b}$$

$$s_e(r) = Min + (f_e(r) - min(f_e(r))) \times Max \tag{12c}$$

$B3$ is then the mean error between the model $s_e(r)$ and the data $p_e(r)$ evaluated over a range equal to the distance between the center and the lesion border but centered on the edge of the lesion. This error measurement is high if the lesion brightness does smoothly transition between dark inside the lesion and bright outside the lesion. The fitting algorithm, *fminsearch()* in Matlab (Mathworks Inc., Natick MA) , was limited to 200 fitting iterations. If convergence was reached before the 200-iteration limit, the result was flagged as one type while fits that were cut off at the 200-iteration limit were flagged as a second type. $B3$ included only results of the second type, that did not converge by the time the iteration limit was reached.

$$B3 = \left\langle \sum_{r=0.5D}^{r=1.5D} (p_e(r) - s_e(r))^2 \right\rangle |_{\theta=0}^{\theta=2\pi} \tag{13}$$

$B4$ Is the mode error, calculated the same as B3 but with the mode() operator instead of the mean⟨⟩ operator, calculated for only the data that exceeded the number (200) of fitting iterations allowed.

$$B4 = mode \left(\sum_{r=0.5D}^{r=1.5D} (P_e(r) - erf(r))^2 \right) |_{\theta=0}^{\theta=2\pi} \tag{14}$$

B5 is the standard deviation of the set of derivative values of the mean brightness. The variance of the derivative of brightness describes how much variability in the instantaneous change in brightness there is over the angular sweep. If some angular ranges are flat (low intra-range brightness derivative) and some ranges vary wildly, this variable will have a high value.

$$B5 = \sigma\left(\frac{dR_m}{d\theta}\right) = \sigma(|R_m(\theta_n) - R_m(\theta_{n+1})|)|_{\theta=0}^{\theta=2\pi} \quad (15)$$

B6 was calculated like B3 except that it used all data and was not restricted to the data requiring more fitting iterations than Matlab was allowed to execute. Similarly, B7 used only the fits that did not require more iterations than (200) the maximum number of fitting iterations allowed.

A watershed analysis was developed to identify pigmented network branches. First, gray-scale images extracted from individual channels were passed through a rank filter which reset the gray-scale value of each pixel to the rank in brightness of the pixel under consideration with its group of neighboring pixels. This step was needed prior to the watershed analysis to act as a high-pass spatial filter and eliminate overall brightness variations in the lesion, leaving the local variations such as those caused by pigmented networks to be identified by the watershed analysis. Branches, which were skeletonized to a single pixel width down their spine, were characterized by three features: their length, their mean brightness and their angle with respect to the lesion centroid. The MR clock sweep scored the mean pixel intensity of the branches $I_{branch}(\theta)$, the standard deviation of intra-branch pixel intensity variation σ_{branch}, the mean length of the branches $L_{branch}(\theta)$ and the total number of branches $N_{branch}(\theta)$ within a differential angle element that traced with the clock MR clock sweep. B8 is then the normalized inter-branch pixel intensity variation.

$$B8 = \frac{\sigma(I_{branch}(\theta)|_{\theta=0}^{\theta=2\pi})}{\langle I_{branch}(\theta)|_{\theta=0}^{\theta=2\pi}\rangle} \quad (16)$$

B9 Is the standard deviation of the error measurement like in B3, except that the standard deviation operator σ is used instead of the mean <> operator. B9 was evaluated only for fits requiring more fitting iterations than the 200 iterations allowed.

$$B9 = \sigma\left(\sum_{r=0.5D}^{r=1.5D}(P_e(r) - erf(r))^2\right)|_{\theta=0}^{\theta=2\pi} \quad (17)$$

B10 is the normalized angular coefficient of brightness variation.

$$B10 = \frac{\sigma(R_m(\theta))}{\langle R_m(\theta)\rangle} \quad (18)$$

B11 The standardized variance of branch lengths.

$$B11 = \frac{\sigma(L_{branch})|_{\theta=0}^{\theta=2\pi})}{\langle L_{branch}|_{\theta=0}^{\theta=2\pi}\rangle} \tag{19}$$

B12 is the normalized range of angular brightness depicted in Figure 1 and Figure S2.

$$B12 = \frac{max(R_m(\theta)) - min(R_m(\theta))}{\langle R_m(\theta)\rangle} \tag{20}$$

B13 is calculated B6 except the standard deviation operator σ is used instead of the mean <> operator. Like B6, B13 used all the data.

$$B13 = \sigma\left(\sum_{r=0.5D}^{r=1.5D}(P_e(r) - erf(r))^2\right)|_{\theta=0}^{\theta=2\pi} \tag{21}$$

B14 Is the standard deviation σ() of the error measurement as in B13 except that B14 was evaluated only for the fits that completed within the allowed number (200) of fitting iterations.

Let $\mathrm{Perim_G}$ be the length of the perimeter of the lesion segment in the green channel $\mathrm{L_{green}}$. G1 is the length of the lesion segment border normalized by the square root of the area of the lesion segment.

$$G1 = \frac{Perim_G}{\sqrt{\sum_{x=1}^{Nx}\sum_{y=1}^{Ny} L_{green}}} - \frac{2\pi}{\sqrt{\pi}} \tag{22}$$

B15 Is the mean intra-branch coefficient of variation.

$$B15 = \left\langle \frac{\sigma(I_{branch}(\theta))}{\langle I_{branch}(\theta)\rangle}|_{\theta=0}^{\theta=2\pi} \right\rangle \tag{23}$$

S5.2 – MIBs With Red Channel Diagnostic Significance

The fitting algorithm depicted in Figure S5 yielded a slope S for the sigmoidal edge fit. R1 was the standard deviation of the slope fit values

$$R1 = \sigma(S)|_{\theta=0}^{\theta=2\pi} \tag{24}$$

R2 is the fractal dimension of the lesion segment binary image as defined by (30)

$$R2 = D(L_{red}(y,x)) \tag{25}$$

Each branch segment in terminated on two ends in either a branch point or an end point. R3 is the connectedness of the pigmented network, defined as the ratio of the number of branch points $\mathrm{N_{BP}}$ to the number of endpoints $\mathrm{N_{EP}}$.

$$R3 = \frac{N_{BP}}{N_{EP}} \tag{26}$$

$\mathrm{R4}$ is the size of the lesion segment $\mathrm{L_{red}}$, which is the sum of the binary mask valued at one inside the lesion segment and zero outside the lesion segment.

$$R4 = \sum_{x=1}^{Nx} \sum_{y=1}^{Ny} L_{red} \tag{27}$$

$\mathrm{R5}$ is the mean slope (S) for the edge fit function $\mathrm{s_e(r)}$ (as used in Eq. 13) evaluated only for the fits that did not require more iterations of the *fminsearch()* operator than the 200 allowed.

$$R5 = \langle S|_{\theta=0}^{\theta=2\pi} \rangle \tag{28}$$

Let the instantaneous radius of the lesion, as in Eq. 3, be denoted by $\mathrm{R_R(\theta)}$ over the angular sweep of θ. $\mathrm{R6}$ is then the coefficient of variation in the lesion radius over the angular sweep

$$R6 = \frac{\sigma(R_{red}(\theta)|_{\theta=0}^{\theta=2\pi})}{\langle R_{red}(\theta)|_{\theta=0}^{\theta=2\pi} \rangle} \tag{29}$$

Let $\mathrm{N_b(\theta, d\theta)}$ be the number of pigmented network branches identified in a differentiual angle emelment $\mathrm{d\theta}$ as a function of angle θ over the angular sweep. $\mathrm{R7}$ is then the range in number of branches detected as a function of angle.

$$R7 = max(N_{branch}(\theta, d\theta)) - min(N_{branch}(\theta, d\theta)) \tag{30}$$

$\mathrm{R8}$ is the range in the standard deviation of pixel brightnes on the angular sweep arm over the angular sweep.

$$R8 = max(R_{std}(\theta)|_{\theta=0}^{\theta=2\pi}) - min(R_{std}(\theta)|_{\theta=0}^{\theta=2\pi}) \tag{31}$$

Pixels with the lesion segment were scored as a set $\mathrm{P_{lesion}}$. The coefficient of variation for pixels within the lesion segment was calculated by dividing the standard deviation of intra-lesional pixel brightness by the mean lesional pixel brightness. $\mathrm{R9}$ is then the coefficient of variation in pixel brightness within the lesion.

$$R9 = \frac{\sigma(P_{lesion})}{\langle P_{lesion} \rangle} \tag{32}$$

$\mathrm{R10}$ is the mode error, calculated the same as $\mathrm{B4}$ but evaluated only for the fits that did not exceed the number of fitting iterations (200) allowed.

$$R10 = mode\left(\sum_{r=0.5D}^{r=1.5D}(P_e(r) - erf(r))^2\right)|_{\theta=0}^{\theta=2\pi} \quad (33)$$

The maximum asymmetry of the lesion (Figure S12) was normalized by the eccentricity of the lesion E as calculated using the *stats.Ecentricity* function in Matlab. This normalization enabled de-emphasis of uniform ovals as asymmetric. $R11$ is then the maximum asymetry of the lesion silhouette

$$R11 = max\left(\frac{A}{E}\right) \quad (34)$$

$R12$ is the sum of the normalized derivitive in lesion radius D over the angular sweep

$$R12 = \sum_{\theta=0}^{\theta=2\pi} abs(R_{red}(\theta, d\theta) - R_{red}(\theta - 1, d\theta)) \quad (35)$$

$R13$ is the asymetry of the lesion sihlouette evaluated in the standard technique (Figure S10)

$$R13 = A|\theta_{sym} - \frac{\pi}{2} \quad (36)$$

Supplementary Tables

Table S1. Diagnostic sensitivity and specificity for melanoma detection by human pattern recognition (*) and by machine-augmented pattern recognition (**). The final two listed other techniques (***) represent the current commercially available clinical machine-vision systems. † = dermatologist referral by general practitioner non-expert dermoscopist ‡ = averaged over in situ and stage I melanoma.

Method	Sensitivity	Specificity
*GP Referral † (Argenziano, Puig et al. 2006)	51%	71%
*GP Referral, Dermoscopy † (Argenziano, Puig et al. 2006)	79%	72%
*Pattern Analysis (Nachbar, Stolz et al. 1994)	85%	79%
*ABCD (Nachbar, Stolz et al. 1994)	84%	75%
*ABCD † (Malvehy, Hauschild et al. 2014)	56%	92%
*7-Point Checklist (Rigel, Russak et al. 2010)	78%	65%
*7-Point Checklist ‡ (Malvehy, Hauschild et al. 2014)	58%	92%
*CASH (Henning, Dusza et al. 2007)	98%	68%
*Menzies (Rigel, Russak et al. 2010)	85%	85%
*Malignancy Grading ‡ (Malvehy, Hauschild et al. 2014)	73%	81%
*Total Body Photography (Drugge, Nguyen et al. 2009)	75%	74%
*Dermoscopy (Vestergaard, Macaskill et al. 2008)	90%	90%
*Confocal Microscopy (Hofmann-Wellenhof, Wurm et al. 2009)	90%	86%
**Prelim. SIAscopy (Emery, Hunter et al. 2010)	50%	84%
**Prelim. SIAscopy (Emery, Hunter et al. 2010)	44%	95%
**Solar Scan (Menzies, Bischof et al. 2005)	91%	68%
**Prelim. Melafind 1 (Friedman, Gutkowicz-Krusin et al. 2008)	98%	44%
**Prelim. Melafind 2 (Friedman, Gutkowicz-Krusin et al. 2008)	91%	38%
**Image Processing (Ramezani, Karimian et al. 2014)	77%	87%
***SIAscopy (Sgouros, Lallas et al. 2014)	86%	65%
***Melafind (Monheit, Cognetta et al. 2011)	98%	10%
Q-score (This Work)	98%	36%

Table S1 References

Argenziano, G., et al. (2006). "Dermoscopy improves accuracy of primary care physicians to triage lesions suggestive of skin cancer." J Clin Oncol **24**(12): 1877-1882.

Drugge, R. J., et al. (2009). "Melanoma screening with serial whole body photographic change detection using Melanoscan technology." Dermatol Online J **15**(6): 1.

Emery, J. D., et al. (2010). "Accuracy of SIAscopy for pigmented skin lesions encountered in primary care: development and validation of a new diagnostic algorithm." BMC Dermatol **10**: 9.

Friedman, R. J., et al. (2008). "The diagnostic performance of expert dermoscopists vs a computer-vision system on small-diameter melanomas." Arch Dermatol **144**(4): 476-482.

Henning, J. S., et al. (2007). "The CASH (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy." J Am Acad Dermatol **56**(1): 45-52.

Hofmann-Wellenhof, R., et al. (2009). "Reflectance confocal microscopy--state-of-art and research overview." Semin Cutan Med Surg **28**(3): 172-179.

Malvehy, J., et al. (2014). "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 **171**(5): 1099-1107.

Menzies, S. W., et al. (2005). "The performance of SolarScan: an automated dermoscopy image analysis instrument for the diagnosis of primary melanoma." Arch Dermatol **141**(11): 1388-1396.

Monheit, G., et al. (2011). "The performance of MelaFind: a prospective multicenter study." Arch Dermatol **147**(2): 188-194.

Nachbar, F., et al. (1994). "The ABCD rule of dermatoscopy. High prospective value in the diagnosis of doubtful melanocytic skin lesions." J Am Acad Dermatol **30**(4): 551-559.

Ramezani, M., et al. (2014). "Automatic Detection of Malignant Melanoma using Macroscopic Images." J Med Signals Sens **4**(4): 281-290.

Rigel, D. S., et al. (2010). "The evolution of melanoma diagnosis: 25 years beyond the ABCDs." CA Cancer J Clin **60**(5): 301-316.

Sgouros, D., et al. (2014). "Assessment of SIAscopy in the triage of suspicious skin tumours." Skin Res Technol.

Vestergaard, M. E., et al. (2008). "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 **159**(3): 669-676.

Table S2. Classification Algorithms applied to Melanoma Discrimination

Method	Description
NNET	Feed-forward neural networks with a single hidden layer
SVM (linear and	Support Vector Machines (8, 9)
GLM	Logistic regression within the framework of Generalized Linear Models (31)
GLMnet	Logistic regression with the elastic-net penalization (14)
GLMboost	Logistic regression with gradient boosting (32)
RF	Random Forests (13)
RP	CART (Classification and Regression Trees) algorithm for classification problems (10)
KNN	K-nearest neighbors algorithm developed for
MARS	Multiple Adaptive Regression Splines (12)
C50	C5.0 decision tree algorithm for classification problems
PLS	Partial Least Squares (16)
LDA	Linear Discriminant Analysis (15)

Table S3. Mean ratios for intensities between color channel pairs. These values are the empirical color channel ratios for manually selected regions in the test set blind to the gold standard diagnosis. These were the input values for MIB MC1, which evaluated pixels for the presence of colors.

color	R/B	R/G	B/G
dark brown	1.90±0.43	2.02±0.40	1.07±0.22
light brown	1.72±0.46	1.62±0.26	0.94±0.14
black	0.46±0.55	0.75±0.64	1.74±1.18
red	1.81±0.69	1.98±0.61	1.10±0.28
blue-gray	1.26±0.34	1.32±0.25	1.05±0.13
white	0.92±0.10	1.2008±0.09	1.30±0.08

Table S4. Median Estimates for the Performance of MIBs According to Different Channels Across All Methods. The Sensitivity, specificity, Predictive Negative Value and Predictive Positive Value were evaluated for each one of the 500 test sets according to the described 12 methods. The overall median (across methods and test sets) was used to summarize the classification Performance.

Wavelength	Sensitivity	Specificity	PPV	PNV
Red	0.66	0.64	0.65	0.67
Green	0.64	0.69	0.67	0.64
Blue	0.71	0.69	0.72	0.69
Best, Multicolor	0.79	0.77	0.77	0.75

References:

1. Marghoob Ashfaq A., Malvehy Josep, Braun Ralph P., . Memorial Sloan-Kettering Cancer Center. Atlas of dermoscopy. 2012.

2. Mete M., Sirakov N. M. Dermoscopic diagnosis of melanoma in a 4D space constructed by active contour extracted features. Comput Med Imaging Graph. 2012;36(7):572-9.

3. Otsu N. A threshold selection method from gray-level histogram. IEEE Transactions on System Man Cybernetics. 1979;SMC-9(1):62-6.

4. Leachman S. A., Cassidy P. B., Chen S. C., Curiel C., Geller A., Gareau D., Pellacani G., Grichnik J. M., Malvehy J., North J., Jacques S. L., Petrie T., Puig S., Swetter S. M., Tofte S., Weinstock M. A. Methods of Melanoma Detection. Cancer Treat Res. 2016;167:51-105. doi: 10.1007/978-3-319-22539-5_3. PubMed PMID: 26601859.

5. Kuncheva L. Combining Pattern Classifiers. Methods and Algorithms. 2 ed: Wiley; 2014.

6. Fix E., Hodges J.L. Discriminatory Analysis, nonparametric discrimination: Consistency properties. 1951.

7. Haykin Simon. Neural Networks and Learning Machines. 3 ed. New York: Prentice Hall; 2008 November.

8. Cortes Corinna, Vapnik Vladimir. Support-vector networks. Machine Learning. 1995;20(3):273-97.

9. Scholkopf B. , Smola A. . Learning with Kernels. MIT Press. 2002.

10. Breiman Leo, Friedman Jerome, Stone Charles J., Olshen R. A. Classification and Regression Trees (Wadsworth Statistics/Probability). 1 ed: Chapman and Hall/CRC; 1984 January.

11. Quinlan J.R.: Morgan Kaufmann Publishers; 1993.

12. Friedman Jerome H. Multivariate Adaptive Regression Splines. The Annals of Statistics. 1991;19(1):1-67.

13. Breiman Leo. Random Forests: Kluwer Academic Publishers; 2001. 5-32 p.

14. Friedman J., Hastie T., Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010;33.

15. Fisher R.A. The use of multiple measurements in taxonomic problems. Annals of Eugenics. 1936;7:179-88.

16. Wold H. Soft Modeling by Latent Variables; the Nonlinear Iterative Partial Least Squares Approach. Perspectives in Probability and Statistics Papers in Honour of M S Bartlett. 1975.

17. Rosipal Roman, Krämer Nicole. Overview and Recent Advances in Partial Least Squares. In: Saunders C, Grobelnik M, Gunn S, Shawe-Taylor J, editors. Subspace, Latent Structure and Feature Selection: Springer Berlin Heidelberg; 2006. p. 34-51.

18. Polikar R. Ensemble based system in decision making. IEEE Circuits and Systems Magazine. 2006;6:21-45.

19. Rokach Lior. Ensemble-based classifiers. Artificial Intelligence Review. 2010;33:1-39.

20. Henning J. S., Dusza S. W., Wang S. Q., Marghoob A. A., Rabinovitz H. S., Polsky D., Kopf A. W. The CASH (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy. J Am Acad Dermatol. 2007;56(1):45-52.

21. Friedman R. J., Gutkowicz-Krusin D., Farber M. J., Warycha M., Schneider-Kels L., Papastathis N., Mihm M. C., Jr., Googe P., King R., Prieto V. G., Kopf A. W., Polsky D., Rabinovitz H., Oliviero M., Cognetta A., Rigel D. S., Marghoob A., Rivers J., Johr R., Grant-Kels J. M., Tsao H. The diagnostic performance of expert dermoscopists vs a computer-vision system on small-diameter melanomas. Arch Dermatol. 2008;144(4):476-82.

22. Monheit G., Cognetta A. B., Ferris L., Rabinovitz H., Gross K., Martini M., Grichnik J. M., Mihm M., Prieto V. G., Googe P., King R., Toledano A., Kabelev N., Wojton M., Gutkowicz-Krusin D. The performance of MelaFind: a prospective multicenter study. Arch Dermatol. 2011;147(2):188-94.

23. Rigel D. S., Russak J., Friedman R. The evolution of melanoma diagnosis: 25 years beyond the ABCDs. CA Cancer J Clin. 2010;60(5):301-16.

24. Wolf J. A., Ferris L. K. Diagnostic inaccuracy of smartphone applications for melanoma detection--reply. JAMA Dermatol. 2013;149(7):885. doi: 10.1001/jamadermatol.2013.4337. PubMed PMID: 23864095.

25. Doyle-Lindrud S. Watson will see you now: a supercomputer to help clinicians make informed treatment decisions. Clin J Oncol Nurs. 2015;19(1):31-2. doi: 10.1188/15.CJON.31-32. PubMed PMID: 25689646.

26. Lynch M., Carroll F., Kavanagh A., Honari B., Collins P. Comparison of a semiautomated hand-held device to test minimal erythema dose before narrowband ultraviolet B phototherapy with the conventional method using matched doses. J Eur Acad Dermatol Venereol. 2014;28(12):1696-700. doi: 10.1111/jdv.12371. PubMed PMID: 24456040.

27. Billings Steven D., Cotton Jenny. Inflammatory dermatopathology a pathologist's survival guide. New York, NY: Springer,; 2011. Available from: http://dx.doi.org/10.1007/978-1-60327-838-6.

28. Levis W. R., Holzer A. M., Leonard L. K. Topical diphenylcyclopropenone as a measure of immune competence in HIV-seropositive subjects. J Drugs Dermatol. 2006;5(9):853-8. PubMed PMID: 17039650.

29. Elder D. E. Dysplastic naevi: an update. Histopathology. 2010;56(1):112-20.

30. Costa Alceu Ferraz. Hausdorff (Box-Counting) Fractal Dimension. 2013.

31. McCullagh P., Nelder John A. Generalized Linear Models. Second ed: Chapman and Hall/CRC; 1989 August.

32. Hofner Benjamin, Mayr Andreas, Robinzonov Nikolay, Schmid Matthias. Model-based boosting in R: a hands-on tutorial using the R package mboost. Computational Statistics. 2014;29(1-2):3-35.