Four-Class Classification of Skin Lesions With Task Decomposition Strategy

Kouhei Shimizu*, Hitoshi Iyatomi, M. Emre Celebi, Kerri-Ann Norton, and Masaru Tanaka

***Abstract*—This paper proposes a new computer-aided method for the skin lesion classification applicable to both melanocytic skin lesions (MSLs) and nonmelanocytic skin lesions (NoMSLs). The computer-aided skin lesion classification has drawn attention as an aid for detection of skin cancers. Several researchers have developed methods to distinguish between melanoma and nevus, which are both categorized as MSL. However, most of these studies did not focus on NoMSLs such as basal cell carcinoma (BCC), the most common skin cancer and seborrheic keratosis (SK) despite their high incidence rates. It is preferable to deal with these NoMSLs as well as MSLs especially for the potential users who are not enough capable of diagnosing pigmented skin lesions on their own such as dermatologists in training and physicians with different expertise. We developed a new method to distinguish among melanomas, nevi, BCCs, and SKs. Our method calculates 828 candidate features grouped into three categories: color, subregion, and texture. We introduced two types of classification models: a layered model that uses a task decomposition strategy and flat models to serve as performance baselines. We tested our methods on 964 dermoscopy images: 105 melanomas, 692 nevi, 69 BCCs, and 98 SKs. The layered model outperformed the flat models, achieving detection rates of 90.48%, 82.51%, 82.61%, and 80.61% for melanomas, nevi, BCCs, and SKs, respectively. We also identified specific features effective for the classification task including irregularity of color distribution. The results show promise for enhancing the capability of the computer-aided skin lesion classification.**

***Index Terms*—Basal cell carcinoma (BCC), dermoscopy, image processing, melanoma, skin lesion classification.**

I. Introduction

INCIDENCE of skin cancer has been increasing over the decades and early treatment is becoming more and more important [1]–[3]. The five year survival rate of melanoma, the most fatal skin cancer is only 9–15% [4] at stage IV, while this rate increases to 85–99% if detected early at stage II. Basal cell carcinoma (BCC), the most common skin cancer is rarely fatal, but it destroys surrounding tissue if left untreated [5]. Thus, early detection and appropriate treatment are essential [6].

Detection of skin cancers is difficult due to the confusing appearance of wide variety of skin lesions. Melanomas and nevi are especially difficult to differentiate. Even with dermoscopy, which uses a magnifying glass with a polarization filter and a uniform light source, the accuracy of melanoma diagnosis by expert dermatologists remains at 75–84% [7]. Biopsy provides a definitive diagnosis, however, it can cause metastasis, and therefore, is only allowed based on the premise of following surgical operation within a month. In addition, these are invasive operations and make unpleasant experiences to the patient [8].

To avoid unnecessary biopsy, several researchers investigated noninvasive computer-aided methods to distinguish melanomas from nevi using dermoscopy images [9]–[16]. These methods usually consist of three steps: 1) border detection of skin tumor; 2) feature extraction; and 3) classification. The border detection process finds the border of the tumor in the dermoscopy image, which is essential for an accurate skin lesion classification. Several methods have been proposed such as the dermatologist-like method [17], SRM [18], hybrid thresholding [19], threshold fusion [20], and so on. The feature extraction process obtains discriminating image features that facilitate classification such as general color statistics, contour shape, and texture information. Wavelet coefficients that capture color and shape information have also been investigated [9]. The classification process determines the type of skin lesions from the extracted image features. General pattern classifiers such as linear discriminant classifier [10], k-NN [11], artificial neural networks [12], [13], and support vector machines (SVMs) [14] are often used. Based on the aforementioned three steps, researchers have improved the automated classification methods. Although there are several limitations, these studies reported superior classification performance compared to experts. For instance, Celebi *et al.* [14] achieved 93.3% SE (sensitivity: correct classification rate for melanomas) and 92.3% SP (specificity: correct classification rate for benign tumors). Besides, we developed an internet-based melanoma screening system [13] (current URL is https://dermoscopy.k.hosei.ac.jp), which we continually update to improve the accuracy and reliability. Anyone who has a dermoscopy image can use our online system.

The aforementioned conventional studies have several problems: 1) only limited types of skin lesions are acceptable for the classification; 2) the systems do not explain the reasons for the classification results; and 3) the systems were developed and evaluated with only ideal condition images and did not consider the condition of test images. In this paper, we focus on the first issue, i.e., the limitation of applicable skin lesion types. That

Manuscript received March 25, 2014; revised July 13, 2014; accepted August 3, 2014. Date of publication August 15, 2014; date of current version December 18, 2014. This work was supported in part by Japan Science and Technology A-STEP Program (AS251Z01435P, 2013), NIH, and ACS. *Asterisk indicates corresponding author.*

*K. Shimizu is with the Department of Applied Informatics, Hosei University, Tokyo 102-8160, Japan (e-mail: kohei.shimizu.66@adm.hosei.ac.jp).

H. Iyatomi is with the Department of Applied Informatics, Hosei University, Tokyo 102-8160, Japan (e-mail: iyatomi@hosei.ac.jp).

M. E. Celebi is with the Department of Computer Science, Louisiana State University, Shreveport, LA 71115 USA (e-mail: emrecelebi1980@gmail.com).

K. A. Norton is with the Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: kerri.norton@gmail.com).

M. Tanaka is with the Department of Dermatology, Medical Center East, Tokyo Women's Medical University, Tokyo 162-8666, Japan (e-mail: masarutanaka@1984.jukuin.keio.ac.jp).

Digital Object Identifier 10.1109/TBME.2014.2348323

is, most of the conventional works handled only melanocytic skin lesions (MSLs) such as melanomas and nevi, which originate from melanocytes, whereas nonmelanocytic skin lesions, (NoMSLs) indicating all the other pigmented skin lesions except MSLs such as BCCs and seborrheic keratoses (SKs) have been relatively neglected [21]. This is because melanoma is the most fatal skin cancer and especially difficult to differentiate from nevus even by expert dermatologists. However, BCC is also harmful and accounts for 80% of the skin cancer incidences [22]. SKs are observed in most Caucasian people above the age of 50 [23] and are sometimes confused with melanomas [24]. Although classification of NoMSLs is considered to be easier than that of MSLs for expert dermatologists, it is not always easy for inexperienced dermatologists or physicians with different expertise. Therefore, if we open the system also for those potential users, as was the case with the aforementioned internet-based system [13], it is necessary to handle both MSLs and NoMSLs.

Actually, some researchers handled both MSLs and NoMSLs for automated classifications [25]–[29]. Stoecker *et al.* [25] proposed a method to detect BCCs from MSLs and NoMSLs with the luminance-related features that describe the semitranslucent area seen in BCCs. Cheng *et al.* [26] used features that characterize the topologies of blood vessels of BCCs. However, these studies were not concerned with detecting melanomas. Ballerini *et al.* [27] presented a method to distinguish among nevi, BCCs, SKs, actinic keratoses, and squamous cell carcinomas (SCC). Although they handled these common types of skin lesions, once again melanomas were not included. Sigurdsson *et al.* [28] used Raman spectra, which is obtained from lesion exposure to laser beams and reported a classification rate of 80% for melanomas and over 90% for nevi, BCCs, SKs, and normal skins. This indicates that such nonimage information is also useful for the classification. Nevertheless, dermoscopy imaging has its own merit of being widely available in dermatology and is covered by most health insurance.

We have been working on the development of classification methods for both MSLs and NoMSLs. First, we developed a general border detection algorithm [21] for MSLs and NoMSLs. Finding the border of NoMSLs was a challenging task because they often have unclear borders. With this sophisticated algorithm, we found that a linear classifier with only two image features ("skewness of bright region on the major axis" and "difference in blue intensity between the peripheral and the normal skin") discriminated MSLs from NoMSLs with performance of 98.0% SE and 86.6% SP [30]. We further developed a system to detect melanomas from other MSLs (nevi) and NoMSLs. Using 548 MSL and 110 NoMSL images, the system achieved 97.6% SE and 87.7% SP (89.5% SP for nevi and 79.1% SP for NoMSLs) [31]. However, this study focused only on discriminating melanomas from all other lesions, thus clumping BCCs with benign skin tumors.

In this paper, we propose a method to distinguish among four types of skin lesions: melanoma, nevus, BCC, and SK, using a significantly larger dataset. Melanoma and BCC account for 80% of all skin cancer incidences. Accurate identification of nevus and SK are clinically important since they are sometimes confused with melanomas. This paper is organized as follows: Section II describes the image datasets, Section III explains the proposed method, Sections IV and V evaluate the classification performance, and Section VI discusses the significance and concludes this paper.

II. Dataset

In this study, we used 968 digital dermoscopy images categorized into four types: melanoma, nevus, BCC, and SK. The details are given as follows.

1) *Melanoma:* 105 images (30 from Keio University Hospital and 75 from the University of Naples and Graz), a malignant melanocytic tumor (MSL), the most fatal skin cancer.
2) *Nevus:* 692 images (448 from Keio University Hospital and 244 from the University of Naples and Graz), a benign melanocytic tumor (MSL), often difficult to differentiate from melanomas.
3) *BCC:* 69 images (20 from Keio University Hospital and 49 from Tokyo Women's Medical University), a malignant nonmelanocytic tumor (NoMSL), the most common skin cancer.
4) *SK:* 98 images (42 from Keio University Hospital and 56 from Tokyo Women's Medical University), a benign nonmelanocytic tumor (NoMSL), which commonly occurs in the elderly and is sometimes confused with melanomas.

These images have different resolutions ranging from 512×384 to 3641×2732. The diagnosis of the skin lesions was determined by histopathological examination or clinical agreement by several expert dermatologists.

III. Method

A. Border Detection

From each skin lesion image, we extracted the border between the tumor and the surrounding normal skin area. Accurate border detection usually results in better classification performance. Conventional automated methods of border detection mostly focused on only melanocytic skin lesions (MSLs). In our previous study, we developed a general border detection algorithm [21] for both MSLs and NoMSLs. The core of the algorithm is color thresholding, removal of artifacts such as microscope border and hair, and inclusion of bright area seen specifically in NoMSLs. The algorithm outperformed other state-of-the-art methods (dermatologist-like method [17], SRM [18], hybrid thresholding [19], *k*-means++[32], and JSEG [33]) for NoMSLs and showed equivalent or better performance for MSLs. For more details, please refer to our recent paper [21].

B. Feature Extraction

After determining the border of the tumor, we segmented the skin lesion image into four regions as illustrated in Fig. 1: normal skin, peripheral, central tumor, and whole tumor. The whole tumor consists of all pixels within the extracted border. In contrast, the normal skin is all pixels on the outside of the border. The peripheral is the first 30% of the whole tumor area,

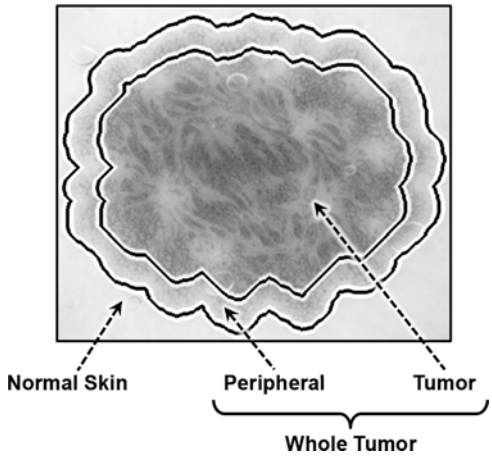

Fig. 1. Four regions in the skin lesion image.

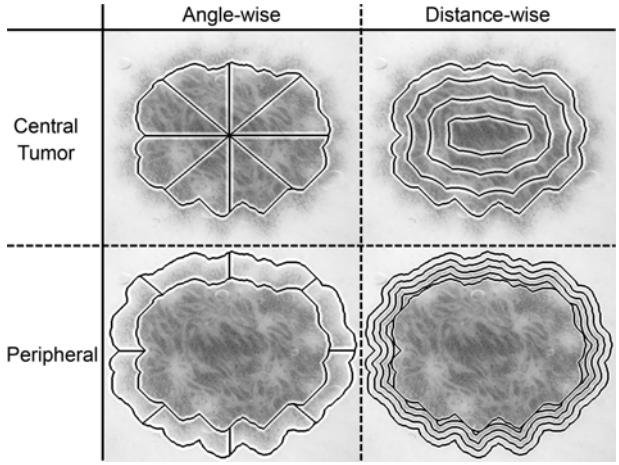

Fig. 2. Subregions of central tumor and peripheral.

obtained by going inward from the border as in our previous studies [13]. Finally, the central tumor is obtained by removing the peripheral from the whole tumor.

For preprocessing, we rotated the images to make the major axis of the whole tumor parallel to its horizontal axis (X-axis). We also resized the images such that the major axis of the whole tumor was 512 pixels in length due to the disparate image resolutions and to reduce the computation time.

After preprocessing, we calculated 828 candidate image features, which are mostly variants of the 428 image features from our previous studies [13]. The reason for introducing new features is that the previous 428 features were designed purely for detecting melanomas, while in this study, we distinguish among four types of skin lesions.

The 828 features are grouped into the three categories: color (300), subregion (144), and texture (384). The numbers in the parentheses denote those of the features in the corresponding categories. Next, we explain the details of each category.

1) Color-Related Features: As for color-related features, we calculated ten statistics (min, max, standard deviation, skewness, entropy, 5%-tile, 25%-tile, 50%-tile, 75%-tile, and 95%-tile) of the intensity of six color channels (R: red, G: green, B: blue, H: hue, S: saturation, and V: luminance) for each of the three tumor regions (peripheral, central tumor, and whole tumor shown in Fig. 1). This yielded 180 parameters (10 statistics $\times$ 6 channels $\times$ 3 regions). We also calculated the difference in the same ten statistics on the six color channels between central tumor and peripheral and those between peripheral and normal skin area, which yielded 120 parameters (10 statistics $\times$ 6 channels $\times$ 2 pairs-of-regions). We expect these difference-oriented features to be robust over variations of dermoscopy images caused by different photographic conditions. In total, there are 300 color-related features ($180 + 120$).

The main change made from our previous studies is the adoption of percentile statistics. The reason for using percentile is that they are expected to be robust over artifacts such as black hairs and shiny bubbles compared to min, mean, or max.

2) Subregion-Related Features: Subregion-related features describe geometrical distribution of the color. First, we divided the central tumor and the peripheral into smaller even subregions as illustrated in Fig. 2. We used two types of subdivisions: angle wise and distance wise. The angle wise is based on the angle from the center of gravity of the central tumor to the edge of the region. The distance wise is based on the Euclidean distance from the outer border. We used three numbers of subregions: 4, 8, and 16 for the angle-wise manner and 2, 4, and 8 for the distance-wise manner.

For each subregion, we calculated three statistics (mean, standard deviation, and skewness) on four color channels (R, G, B, and S). Here, we left out H and V because these two channels did not contribute to the classification performance in our preliminary experiments. Finally, we calculated the standard deviation of these statistics within all subregions. This yielded 144 subregion features (2 target regions $\times$ 2 types of subdivisions $\times$ 3 numbers of subregions $\times$ 4 color channels $\times$ 3 statistics for each subregion).

In our previous studies, we had the asymmetry features to describe geometrical distribution of the color. However, in our preliminary experiments, we determined that the subregion features are more effective for the four-class skin lesion classification.

3) Texture-Related Features: As for texture-related features, we adopted the gray level cooccurrence matrix (GLCM) [34]. We obtained the GLCMs with the following settings: two target regions (central tumor and whole tumor), three quantization levels ($N = 16$, 32, and 64), four distances ($\delta = 1$, 2, 4, and 8 pixels), and four directions ($\theta = 0^\circ$, 45°, 90°, and 135° from the major axis). From each GLCM, we extracted four GLCM statistics (energy, correlation, entropy, and homogeneity).

To make the directional settings (θ) more meaningful, we extracted min, mean, max, and difference (i.e., max$-$min) of the aforementioned GLCM-statistics in four main directions (θ) as was also recommended in the original literature of the GLCM [34]. This is an extension of our previous studies [13]. In total, there are 384 texture features (2 regions $\times$ 3 quantization levels $\times$ 4 distances $\times$ 4 directions (e.g., max) $\times$ 4 GLCM statistics).

For postprocessing, we normalized all of the 828 features so that they have mean of 0 and variance of 1 over all images in

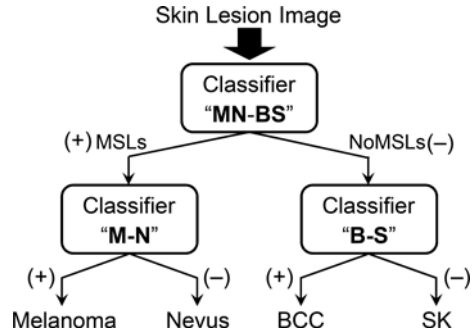

Fig. 3. Overview of the layered model.

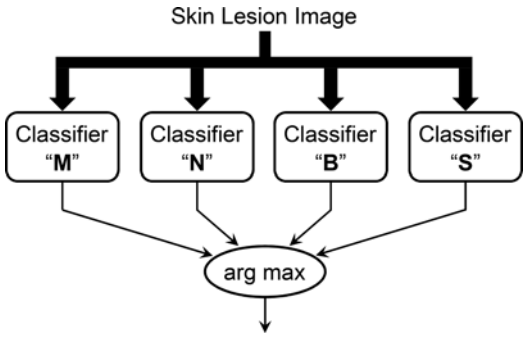

Fig. 4. Overview of the flat model.

the datasets. Note that only a small number of features were selected from the 828 for the classifier development as will be described later.

C. Classification

In this section, we introduce the proposed layered model as the primary classification model and the flat model as a performance baseline shown in Figs. 3 and 4, respectively. The letters M, N, B, and S in the figures denote melanoma, nevus, BCC, and SK, respectively. We used linear classifiers over nonlinear ones in order to gain a clear understanding of the relationship between the inputs and the outputs of the models and to facilitate a comparison of the classification performance.

1) Layered Model (Proposed): The first-step classifier "MN-BS" identifies the input skin lesion as MSL if the output value is greater than the classifier's threshold value or as NoMSL otherwise. These are shown by (+) and (−) in Fig. 3. If the result is an MSL, the second-step classifier "M-N" distinguishes melanoma from nevus in the same manner by comparing its output value with the threshold value. If the result from the first-step classifier is a NoMSL, the second-step classifier "B-S" distinguishes BCC from SK.

The idea of the layered model is to decompose the whole classification task to 1) the broad classification of MSL and NoMSL by the "MN-BS" and 2) the detailed classification of "melanoma and nevus" and that of "BCC and SK" by the "M-N" and "B-S," respectively. It may be inferred that the first-step classifier "MN-BS" must have high accuracy because misclassifications at this phase are fatal. We designed this model based on the results of our past studies that distinguishing MSLs from NoMSLs is relatively easy [30].

One of the most important steps for the classifier development is feature selection. It is well known that too many features or irrelevant features lead to poor performance and the overfitting problem. Therefore, it is necessary to select an appropriate subset of features for each of the classifiers "MN-BS," "M-N," and "B-S."

We adopted Wilks' Lambda stepwise feature selection method [35] as in our previous studies. This algorithm begins with no selected features and repeats the step of adding or removing a feature one by one iteratively. The feature added is the one which gives the highest increase in linear regression fitness under the F-test ($p < 0.05$). A feature is removed when it no longer contributes to the linear regression fitness ($p > 0.10$). This iterative process of adding and removing features continues until no features pass the test for addition or removal.

After selecting the input features, we trained the linear classifiers. The assigned supervisory outputs were either $+1$ or -1 as specified by (+) and (−) in Fig. 3. After the training step, we adjusted the threshold values of the three linear classifiers by full search to optimize classification performance, e.g., detection rate of melanoma. This is described in detail in Section IV.

In our preliminary experiment, we also tested a different layered model, which distinguishes cancer (melanoma and BCC) from no-cancer (nevus and SK) at the top level. However, the performance was not satisfactory mainly because of the difficulty in the classification between cancer and no-cancer.

2) Flat Models (Performance Baseline): We introduce two types of flat models, namely the 'flat model I" and the 'flat model II" as the performance baseline. Each of the flat models has four linear classifiers: "M," "N," "B," and "S" whose output values estimate the presence/absence of the corresponding classes: melanoma, nevus, BCC, and SK, respectively. This kind of classification model is typically used for the multiclass classification [36].

To compare the outputs of the four classifiers, the following score F_i is calculated for each classifier i (e.g., the "M").

$$F_i = \alpha_i \times (O_i - \xi_i). \tag{1}$$

Here, O_i is the normalized output value of the classifier i whose standard deviation is 1. The ξ_i and the α_i are the threshold value and the scaling factor, respectively. The classification result is given by $\arg\max_i(F_i)$. Note that scaling factors used in the flat models are not necessary for the layered model.

The flat model I and the flat model II are different in how the classifiers possess the features. In the flat model I, all classifiers share the same features. We select the features with the Wilks' Lambda stepwise method with the strategy that it improves overall classification performance. In the flat model II, each classifier possesses its own features. We select the features specifically effective for each classifier by the Wilks' Lambda stepwise method as well as the layered model.

For the flat model I, it is necessary to take all four classifiers into consideration when selecting a feature to add or remove in the stepwise method. This issue was addressed in [37], which

TABLE I
TOP THREE SELECTED FEATURES FOR EACH LINEAR CLASSIFIER

	Classifier	First Feature	Second Feature	Third Feature
Layered Model	"MN-BS" (MSL versus NoMSL)	Col: 95 percentile (S:p-n) †	Tex: homogen(ct) [θdif, N64, δ4] ⋆	Col: 25 percentile (H:wt)
	"M-N" (melanoma versus nevus)	Col: σ(V:p)	Sub: $\sigma(\mu$(R:ct)) [angle-16]	Col: 25 percentile (G:p-n)
	"B-S" (BCC versus SK)	Sub: $\sigma(\kappa$(S:ct)) [angle-4] ⋄	Tex: entropy(ct) [θdif, N16, δ8]	Sub: $\sigma(\sigma$(R:ct)) [distance-4]
Flat Model I	shared by all classifiers	Tex: homogen(wt) [θdif, N64, δ4]	Col: 75 percentile (S:p-n)	Col: σ(V:p)
Flat Model II	"M" (melanoma versus else)	Col: σ(V:p)	Sub: $\sigma(\mu$(R:ct)) [angle-16]	Col: κ(G:p-n)
	"N" (nevus versus else)	Tex: homogen(wt) [θdif, N64, δ4]	Col: 5 percentile (S:t-p)	Sub: $\sigma(\mu$(R:ct)) [angle-16]
	"B" (BCC versus else)	Col: 25 percentile (H:wt)	Sub: $\sigma(\sigma$(S:ct)) [angle-16]	Col: 5 percentile (B:p-n)
	"S" (SK versus else)	Col: 95 percentile (S:p-n)	Sub: $\sigma(\mu$(B:ct)) [distance-8]	Tex: entropy(wt) [θmin, N64, δ8]

Feature categories: Col: color, Sub: subregion, Tex: texture.
Target regions: ct: central tumor, wt: whole tumor, p: peripheral, n: normal skin area, ct-p: difference between central tumor and peripheral.
Statistics: μ: mean, σ: standard deviation, κ: skewness.
†: Difference in 95 percentile of saturation (S) between peripheral (p) and normal skin area (n).
⋆: Difference (i.e., max−min) of homogeneity of the GLCM (quantization level = 64, distance = 4 pixels) among the four directions 0°, 45°, 90°, and 135° in central tumor (ct).⋄: Skewness (κ) of saturation (S) is calculated for each of the four angle-wise subregions in central tumor (ct). Standard deviation (σ) of these skewness values is used as the feature.

TABLE II
CLASSIFICATION PERFORMANCE OF THE THREE MODELS UNDER THE CONDITION OF %M > 90%

Model	#Features ⋆	AUC †	%M	%N	%B	%S	min (%N, %B, %S)
Layered Model	(9, 6, 5)	**0.824**	**90.48**	**75.58**	**86.96**	**76.53**	**75.58**
Flat Model I	20	0.775	90.48	69.94	72.46	71.43	69.94
Flat Model II	(4, 4, 9, 3)	0.750	90.48	69.08	69.57	68.37	68.37
Layered Model	(9, 10, 6)	**0.856**	**90.48**	**82.51**	**82.61**	**80.61**	**80.61**
Flat Model I	25	0.802	90.48	74.57	75.36	74.49	74.49
Flat model II	(6, 3, 8, 8)	0.795	90.48	74.42	75.36	74.49	74.42
Layered Model	(11, 12, 7)	**0.864**	**90.48**	**82.66**	**84.06**	**82.65**	**82.65**
Flat Model I	30	0.821	90.48	75.58	76.81	75.51	75.51
Flat Model II	(8, 5, 9, 8)	0.802	90.48	76.59	76.81	78.57	76.59

†: AUC denotes the area of the receiver-operating characteristic (ROC) curve between %M and min(%N, %B, %S).
⋆: The numbers in the parenthesis denote those of the features assigned to each classifier in the order of the "MN-BS," "M-N," and "B-S" for the layered model and the "M," "N," "B," and "S" for the flat model II, respectively. The flat model I has all classifiers share the same 20, 25, or 30 features.

examined multiclass classification. It suggests two methods: either to optimize "average" or "maximum" of the four error reduction amounts associated with the four outputs. We chose the "average" method because it showed better performance in our preliminary experiments.

After the feature selection step, we trained the four classifiers for each of the two flat models. The supervisory outputs were $+1$ for the target type (melanoma for the classifier "M") and -1 for the rest (nevus, BCC, and SK for the "M"). Finally, we adjusted the threshold ξ_i and the scaling factor α_i shown in (1) by means of full search to optimize the classification performance. The search scope of α_i was empirically defined as $[2^{-0.5}, 2^{1.5}]$.

IV. RESULTS

Table I shows the top three selected features for each classifier of the layered model and the two flat models. The features are written in the "category : detail" format. For example, the first feature "Col: σ(V: p)" of the classifier "M-N" is a color-related feature that computes standard deviation of luminance (V) in the peripheral.

Table II summarizes the result of the classification performance under the tenfold cross-validation test. We divided all skin lesion images into ten sets, where the number of images belonging to each skin lesion type was set equal for all ten sets in order to prevent possible biases.

In the #features column, the left number denotes the sum of the assigned features for all classifiers. The right numbers in the parenthesis show the number assigned to each classifier in the order of the "MN-BS," "M-N," and "B-S" for the layered model and the "M," "N," "B," and "S" for the flat model II, respectively. We adjusted these numbers in the following manner: 1) set the sum to 20, 25, or 30 as specified in the table and 2) adjust these numbers to maximize the area under the curve (AUC), a statistic of classification performance, which will be detailed later. For example, the layered model with total 20 features showed the highest AUC (0.824) when 9, 6, and 5 features were assigned to the "MN-BS," "M-N," and "B-S," respectively.

The AUC is the area under the receiver-operating characteristic (ROC) curve. Fig. 5 shows the ROC curves drawn from the classification results by the layered model and the two flat models with 25 features. The horizontal axis is the detection rate of melanoma (%M), which we define as the ratio of the correctly classified melanoma images over all the melanoma images in the dataset (105 as described in Section II). The vertical axis is the minimum of the detection rates of nevus, BCC, and SK [min(%N, %B, %S)]. The reason for using the minimum is to measure the detection rate applicable to all three nonmelanoma

types of skin lesions. We made the curves by optimizing the thresholds and the scaling factors of the linear classifiers to maximize min(%N, %B, %S) under the condition imposed on %M ranging to 100% from 0%. Larger AUC indicates a better performance.

Seeing the ROC curves, we notice that min(%N, %B, %S) does not reach 100% even when %M is set to zero unlike the typical ROC curves seen in the studies of binary classification between melanoma and the rest. This is because min(%N, %B, %S) would reach 100% only if all the nevi, BCCs, and SKs could be perfectly classified as such, whereas in the binary classification, the detection rate of nonmelanoma, i.e., SP (specificity) reaches 100% simply by increasing the threshold enough to dismiss all the melanomas as nonmelanoma. This is the reason why the AUC in Table II seems comparatively lower than those reported in other conventional works, e.g., 0.937 in [9].

Finally, the %M, %N, %B, and %S in the table show the result under the condition that the detection rate of melanoma (%M) should be at least 90%.

V. Discussion

A. Classification of Melanoma

We see from Table II that the layered model significantly outperformed the two flat models. All three models improved as the #features increased. However, it is preferable to achieve good performance with smaller numbers of features to avoid the overfitting problem. We think that the layered model with 25 features has a good balance between #features and classification performance, i.e. it achieved over 80% detection rates of nevus, BCC, and SK (%N, %B, %S) while keeping 90% detection rate of melanoma (%M).

One might point out that the flat model II might be disadvantaged having relatively smaller number of features for each individual classifier. We also evaluated the flat model II with larger, the total of 40, features. The results were AUC of 0.826 and min(%N, %B, %S) of 78.26%, which is still inferior to the layered model with total 25 features.

As shown by the #features of the layered model, the two classifiers "M-N" and "MN-BS" required more features than the "B-S." We consider the reason is that the "M-N" must differentiate melanomas from nevi, which is generally a difficult task and the "MN-BS" must have high accuracy because misclassifications here cannot be compensated by the following classifiers.

Looking at the ROC curves in Fig. 5, it seems that the flat model II outperformed the other models when %M is set nearly to 100%. However, the min(%N, %B, %S) under this condition remained below 75% and the overall AUC of this model was the lowest.

For more insight into the classification performance, Tables III–V show the confusion matrices of the classification results by the three models with 25 features using tenfold cross-validation test, respectively.

For all three models, mistaking nevus as melanoma was the most common misclassification. Also, melanoma was more likely to be mistaken as nevus than BCC or SK.

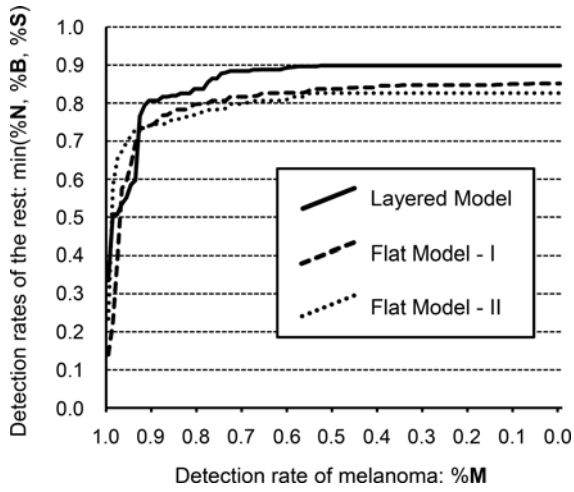

Fig. 5. ROC curves by the three models with 25 features.

TABLE III
CONFUSION MATRIX (%) BY THE LAYERED MODEL WITH 25 FEATURES

Result dataset	Melanoma	Nevus	BCC	SK
Melanoma	**90.48**	6.67	0.95	1.90
Nevus	15.17	**82.51**	0.72	1.59
BCC	5.80	1.45	**82.61**	10.14
SK	5.10	7.14	7.14	**80.61**

TABLE IV
CONFUSION MATRIX (%) BY THE FLAT MODEL I WITH 25 FEATURES

Result dataset	Melanoma	Nevus	BCC	SK
Melanoma	**90.48**	6.67	2.86	0.00
Nevus	16.18	**74.57**	2.89	6.36
BCC	13.04	2.90	**75.36**	8.70
SK	12.24	4.08	9.18	**74.49**

TABLE V
CONFUSION MATRIX (%) BY THE FLAT MODEL II WITH 25 FEATURES

Result dataset	Melanoma	Nevus	BCC	SK
Melanoma	**90.48**	9.52	0.00	0.00
Nevus	16.91	**74.42**	2.31	6.36
BCC	8.70	0.00	**75.36**	15.94
SK	10.20	0.00	15.31	**74.49**

B. Classification of Cancers

Next, we will consider the case where the detection rates should be high for both melanomas and BCCs (%M, %B). The %M is definitely important since melanoma is the most fatal skin cancer. However, as explained earlier, BCC is also harmful and has the highest incidence rate among all the skin cancers.

Therefore, we now impose additional conditions on detection rate of either 90% or 85% for BCCs (%B) while keeping that of

TABLE VI
CLASSIFICATION PERFORMANCE UNDER CONDITIONS ON BOTH %M AND %B

Condition		Layered Model		Flat Model I		Flat Model II	
%M	%B	%N	%S	%N	%S	%N	%S
90	85	82.51	79.59	53.47	57.14	59.97	60.20
90	90	78.90	75.51	unattainable ⋆		38.58	23.47

⋆: No threshold value met the condition.

90% for melanomas (%M). Under such conditions, we measured the classification performance of the three models; Table VI summarizes the results. The number of the features was set to 25 for the two flat models as well as the layered model shown in Table II. The threshold values and the scaling factors of the linear classifiers were adjusted in the following manner: 1) keep %M and %B greater than the predefined values, i.e., shown in Condition columns in Table VI and 2) maximize the minimum of the two other detection rates [min(%N, %S)]. Note that "unattainable" in the table indicates that no threshold value and scaling factor for the flat model I meets these conditions.

We see that the layered model showed much higher performance than the two flat models. Compared to the test without any condition on %B (#features = 25 in Table II), the detection rates of nevus and SK (%N, %S) decreased. This is an inevitable tradeoff for achieving high detection rates for both melanomas (%M) and BCCs (%B).

We think that the appropriate adjustment is to give the first priority to the detection rate of melanoma (%M) then the second priority to that of BCC (%B) while keeping those of nevus and SK (%N, %S) within an acceptable range.

Besides, we also tested a cascade model based on the notion that SK looks more similar to MSLs than BCC. This model first distinguishes "MSLs and SK" from BCC, then MSLs from SK, and finally, melanoma from nevus. However, it was even inferior to both of the flat models.

C. Feature Interpretation

For more insight into the developed model, we will examine how the selected features contributed to the classification. Fig. 6 shows examples of images classified correctly by the layered model with 25 features. The left column shows MSL images: (I) and (II) are melanomas, and (III) is a nevus. The right column shows NoMSL images: (IV) and (V) are BCCs, and (VI) is a SK.

Fig. 7 shows the scatter plot of the first feature "Col: σ(V: p)" and the second feature "Sub: $\sigma(\mu$(R: ct)) [angle-16]" of the classifier "M-N" (see Table I). The circles and dots represent the images of melanomas and those of nevi, respectively. The plots corresponding to the images (I), (II), and (III) in Fig. 6 are specified by the arrows. The dashed line shows the classification boundary for the 90% melanoma detection rate. This is a rough criterion to distinguish between the two types of skin lesions. Although some images are still misclassified due to the overlapped area, we used more features to improve the classification performance as specified by #features in Table II.

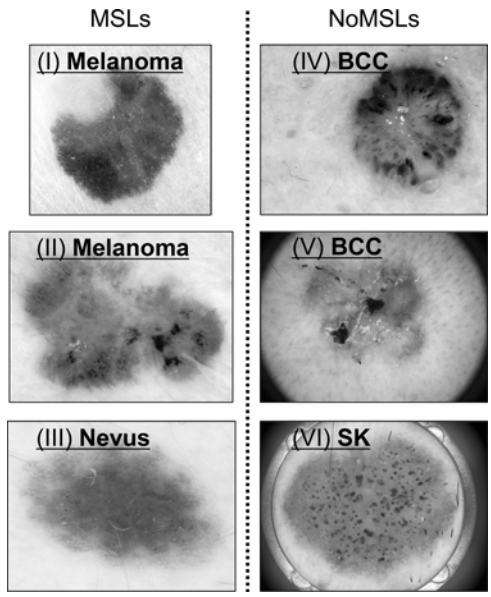

Fig. 6. Examples of correctly classified images.

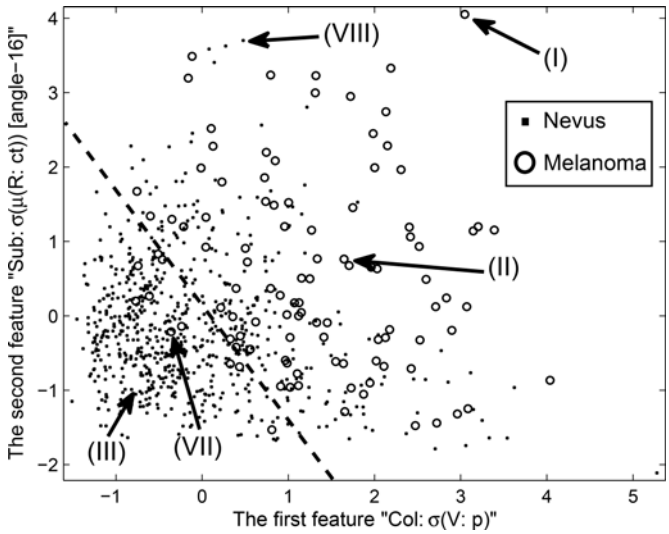

Fig. 7. Plots of the two selected features of the "M-N".

The first feature "Col: σ(V: p)" is the standard deviation of luminance in the peripheral. This feature tends to be higher for melanomas than nevi. For example, the nevus image (III) in Fig. 6 shows a gradual increase of luminance going outward from the center while the melanoma image (I) shows an irregular decrease of luminance at the peripheral, causing a relatively high contrast against the surrounding normal skin area. The second feature "Sub: $\sigma(\mu$(R: ct)) [angle-16]" is the difference of the red channel between the angle-wise subregions shown in Fig. 2. This feature was also larger for melanomas than nevi possibly because melanomas tend to have an uneven or irregular color distribution as established by the ABCD-rule [38] and the seven-point check list [39], two common references for melanoma diagnosis.

Regarding NoMSLs, Fig. 8 shows the scatter plot of the first and the second features of the classifier "B-S." The plots

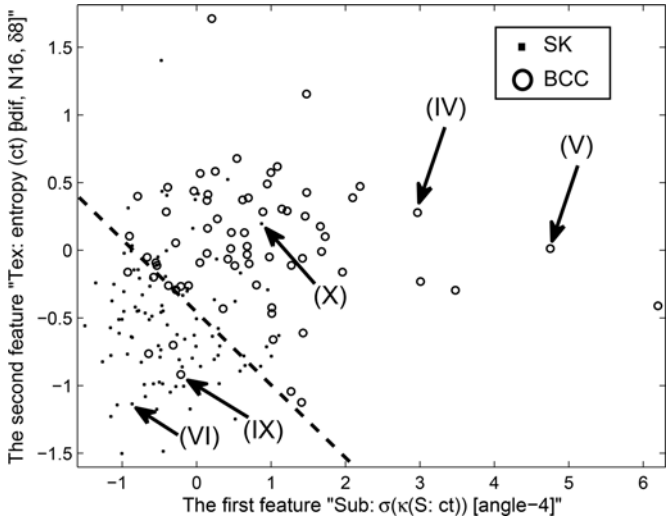

Fig. 8. Plots of the two selected features of the "B-S."

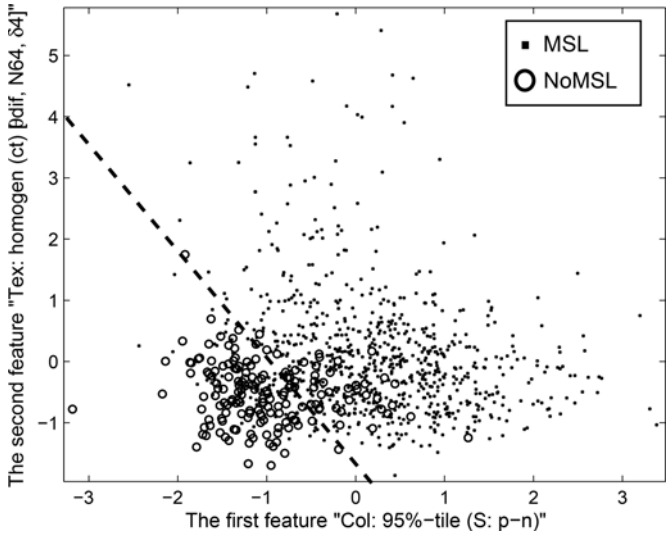

Fig. 9. Plots of the two selected features of the "MN-BS."

corresponding to the images (IV), (V), and (VI) in Fig. 6 are also shown. The dashed line shows the boundary for the BCC detection rate of 90% based on the two features.

The first feature "Sub: $\sigma(\kappa$(S: ct)) [angle-4]" is the difference in skewness on saturation channel between the angle-wise subregions. This feature was higher for BCCs than SKs on average mainly due to the presence of different local objects seen specifically in BCCs [40] such as dark pigments (IV) and blood vessels (V). The second feature "Tex: entropy (ct) [θdif, N16, δ8]" is the directionality of the coarseness in the central tumor. Some of the SKs in our datasets showed a lot of holes with no preference of direction or location as seen in (VI), making this feature especially low.

Fig. 9 shows the scatter plot of the first and the second features of the classifier "MN-BS." The dashed line is the boundary for 90% classification rate of MSLs. Looking at the distribution, this classification is more accurate than that between melanoma and nevus shown in Fig. 7. Actually, the AUC computed on this scatter plot was 0.911, which is higher than 0.888 on Fig. 7 and 0.891 on Fig. 8.

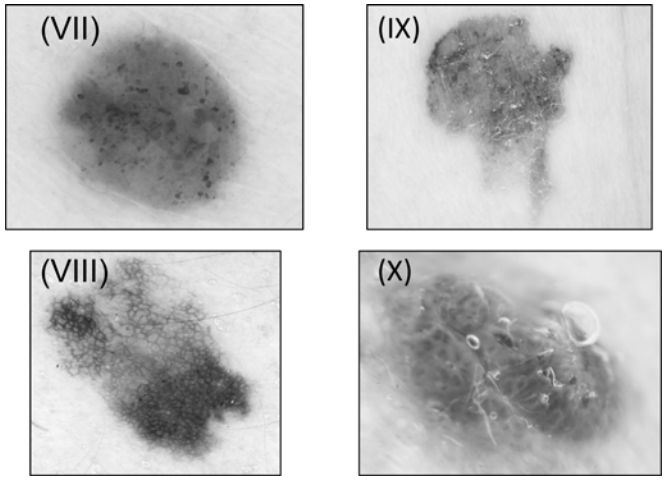

Fig. 10. Examples of misclassified images. (VII) melanoma misclassified as nevus. (VIII) nevus misclassified as melanoma. (IX) BCC misclassified as SK. (X) SK misclassified as BCC.

We briefly showed how the selected features contributed to successful classification. However, misclassifications still occurred due to the varied appearances of the skin lesions. Fig. 10 shows examples of images misclassified by the layered model. These images are also specified by the arrows in Figs. 7 and 8. The melanoma (VII) was misclassified as nevus due to the lack of certain characteristics of melanoma, e.g., the difference in color between the angle-wise subregions. Nevertheless, the presence of the irregular dark globules might be seen as a sign of melanoma. Dealing with such detailed patterns will be the goal of future work.

Despite these difficulties, the layered model with 25 features achieved a detection rate of 90% for melanomas and over 80% for nevi, BCCs, and SKs. These results might seem inferior to those reported in other studies of the binary classification between melanoma and nevus such as 93.3% SE and 92.3% SP [14]. However, this seems inevitable since our methods distinguished among four types of skin lesions instead of two. Also, our analysis used a comparatively large number of images from different data sources, which makes the analysis more realistic. Note that the performance of the method will possibly improve by using nonlinear classifiers, e.g., SVM.

In this study, we dealt with four types of skin lesions, while we did not include SCC the second most common skin cancer due to unavailability of datasets. We are planning to deal with those skin lesions in near future.

VI. Conclusion

In this paper, we proposed a method to distinguish among melanomas, nevi, BCCs, and SKs. For the classification model, we introduced a layered model for task decomposition and two flat models to serve as the baseline.

We evaluated the models with 964 dermoscopy images and showed that the layered model outperformed the two flat models. The layered model with 25 features achieved a detection rate of 90% for melanomas and over 80% for each of the three other types of skin lesions. The result of this study shows promise for broadening the range of users for classification and enhancing the capability of the computer-aided skin lesion classification.

REFERENCES

[1] E. de Vries, L. V. van de Poll-Franse, W. J. Louwman, F. R. de Gruijl, and J. W. Coebergh, "Predictions of skin cancer incidence in the Netherlands up to 2015," *Brit. J. Dermatol.*, vol. 152, no. 3, pp. 481–488, Mar. 2005.

[2] A. Jemal, M. Saraiya, P. Patel, S. S. Cherala, J. Barnholtz-Sloan, J. Kim, C. L. Wiggins, and P. A. Wingo, "Recent trends in cutaneous melanoma incidence and death rates in the United States, 1992–2006," *J. Amer. Acad. Dermatol.* vol. 65, no. 5, pp. S17.e1–S17.e11, Nov. 2011.

[3] F. Bath-Hextall, J. Leonardi-Bee, C. Smith, A. Meal, and R. Hubbard, "Trends in incidence of skin basal cell carcinoma. Additional evidence from a UK primary care database study," *Int. J. Cancer*, vol. 129, no. 9, pp. 2105–2108, Nov. 2007. C. M. Balch, A. C. Buzaid, S. J. Soong, M. B. Atkins, N. Cascinelli, D. G. Coit, I. D. Fleming, J. E. Gershenwald, A. Jr. Houghton, J. M. Kirkwood, K. M. McMasters, M. F. Mihm, D. L. Morton, D. S. Reintgen, M. I. Ross, A. Sober, J. A. Thompson, and J. F. Thompson

[4] C. M. Balch, A. C. Buzaid, S. J. Soong, M. B. Atkins, N. Cascinelli, D. G. Coit, I. D. Fleming, J. E. Gershenwald, A. Jr. Houghton, J. M. Kirkwood, K. M. McMasters, M. F. Mihm, D. L. Morton, D. S. Reintgen, M. I. Ross, A. Sober, J. A. Thompson, and J. F. Thompson, "Final version of the American Joint Committee on Cancer staging system for cutaneous melanoma," *J. Clin. Oncol.*, vol. 19, no. 16, pp. 3635–3648, Aug. 2001.

[5] R. C. Brooke, "Basal Cell Carcinoma," *Clinical Med.*, vol. 5, no. 6, pp. 551–554, Nov. 2005.

[6] R. Marks, "An overview of skin cancers: Incidence and causation," *Cancer Suppl.*, vol. 75, no. S2, pp. 607–612, Jan. 1995.

[7] G. Argenziano, H. P. Soyer, S. Chimenti, R. Talamini, R. Corona, F. Sera, M. Binder, L. Cerroni, G. De Rosa, G. Ferrara, R. Hofmann-Wellenhof, M. Landthaler, S. W. Menzies, H. Pehamberger, D. Piccolo, H. S. Rabinovitz, R. Schiffner, S. Staibano, W. Stolz, I. Bartenjev, A. Blum, R. Braun, H. Cabo, P. Carli, V. De Giorgi, M. G. Fleming, J. M. Grichnik, C. M. Grin, A. C. Halpern, R. Johr, B. Katz, R. O. Kenet, H. Kittler, J. Kreusch, J. Malvehy, G. Mazzocchetti, M. Oliviero, F. Ozdemir, K. Peris, R. Perotti, A. Perusquia, M. A. Pizzichetta, S. Puig, B. Rao, P. Rubegni, T. Saida, M. Scalvenzi, S. Seidenari, I. Stanganelli, M. Tanaka, K. Westerhoff, I. H. Wolf, O. Braun-Falco, H. Kerl, T. Nishikawa, K. Wolff, and A. W. Kopf "Dermoscopy of pigmented skin lesions: results of a consensus meeting via the Internet," *J. Amer. Acad. Dermatol.*, vol. 48, no. 5, pp. 679–693, May 2003.

[8] J. de Leeuw, N. van der Beek, W. D. Neugebauer, P. Bjerring, and H. A. Neumann, "Fluorescence detection and diagnosis of non-melanoma skin cancer at an early stage," *Lasers Surg. Med.*, vol. 41, no. 2, pp. 96–103, Feb. 2009.

[9] R. Garnavi, M. Aldeen, and J. Bailey, "Computer-aided diagnosis of melanoma using border- and wavelet-based texture analysis," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 6, pp. 1239–1252, Nov. 2012.

[10] S. Seidenari, G. Pellacani, and C. Grana, "Pigment distribution in melanocytic lesion images: A digital parameter to be employed for computer-aided diagnosis," *Skin Res. Technol.*, vol. 11, no. 4, pp. 236–241, Nov. 2005.

[11] H. Ganster, A. Pinz, R. Rohrer, E. Wildling, M. Binder, and H. Kittler, "Automated melanoma recognition," *IEEE Trans. Med. Imag.*, vol. 20, no. 3, pp. 233–239, Mar. 2001.

[12] P. Rubegni, G. Cevenini, M. Burroni, R. Perotti, G. Dell'Eva, P. Sbano, C. Miracco, P. Luzi, P. Tosi, P. Barbini, and L. Andreassi, "Automated diagnosis of pigmented skin lesions," *Int. J. Cancer*, vol. 101, no. 6, pp. 576-580, Oct. 2002.

[13] H. Iyatomi, H. Oka, M. E. Celebi, M. Hashimoto, M. Hagiwara, M. Tanaka, and K. Ogawa, "An improved Internet-based melanoma screening system with dermatologist-like tumor area extraction algorithm," *Comput. Med. Imag. Graph.*, vol. 32, no. 7, pp. 566–579, Oct. 2008.

[14] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, and R. H. Moss, "A methodological approach to the classification of dermoscopy images," *Comput. Med. Imag. Graph.*, vol. 31, no. 6, pp. 362–371, Sep. 2007.

[15] Z. She and P. S. Excell, "Skin pattern analysis for lesion classification using local isotropy," *Skin Res. Technol.*, vol. 17, no. 2, pp. 206–212, May 2011.

[16] M. A. Sheha, M. A. Mabrouk, and A. Sharawy, "Automatic detection of melanoma skin cancer using texture analysis," *Int. J. Comput. Appl.*, vol. 42, no. 20, pp. 22–26, Mar. 2012.

[17] H. Iyatomi, H. Oka, M. Saito, A. Miyake, M. Kimoto, J. Yamagami, S. Kobayashi, A. Tanikawa, M. Hagiwara, K. Ogawa, G. Argenziano, H. P. Soyer, and M. Tanaka, "Quantitative assessment of tumour extraction from dermoscopy images and evaluation of computer-based extraction methods for an automatic melanoma diagnostic system," *Melanoma Res.*, vol. 16, no 2, pp. 183–190, Apr. 2006.

[18] M. E. Celebi, H. A. Kingravi, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, R. H. Moss, J. M. Malters, J. M. Grichnik, A. A. Marghoob, H. S. Rabinovitz, and S. W. Menzies, "Border detection in dermoscopy images using statistical region merging," *Skin Res. Technol.*, vol. 14, no. 3, pp. 347–353, Aug. 2008.

[19] R. Garnavi, M. Aldeen, M. E. Celebi, G. Varigos, and S. Finch, "Border detection in dermoscopy images using hybrid thresholding on optimized color channels," *Comput. Med. Imag. Graph.*, vol. 35, no. 2, pp. 105–115, Mar. 2011.

[20] M. E. Celebi, Q. Wen, S. Hwang, H. Iyatomi, and G. Schaefer, "Lesion border detection in dermoscopy images using ensembles of thresholding methods," *Skin Res. Technol.*, vol. 19, no. 1, pp. e252–258, Feb. 2013.

[21] K. A. Norton, H. Iyatomi, M. E. Celebi, S. Ishizaki, M. Sawada, R. Suzaki, K. Kobayashi, M. Tanaka, and K. Ogawa, "Three-phase general border detection method for dermoscopy images using non-uniform illumination correction," *Skin Res. Technol.*, vol. 18, no. 3, pp. 290–300, Aug. 2012.

[22] K. Nakayama, "Growth and progression of melanoma and non-melanoma skin cancers regulated by ubiquitination," *Pigment Cell Melanoma Res.*, vol. 23, no. 3, pp. 338–351, Jun. 2010.

[23] J. M. Yeatman, M. Kilkenny, and R. Marks, "The prevalence of seborrhoeic keratoses in an Australian population: does exposure to sunlight play a part in their frequency?" *Brit. J. Dermatol.*, vol. 137, no. 3, pp. 411–414, Sep. 1997.

[24] S. V. Deshabhoina, S. E. Umbaugh, W. V. Stoecker, R. H. Moss, and S. K. Srinivasan, "Melanoma and seborrheic keratosis differentiation using texture features," *Skin Res. Technol.*, vol. 9, no. 4, pp. 348–356, Nov. 2003.

[25] W. V. Stoecker, K. Gupta, B. Shrestha, M. Wronkiewiecz, R. Chowdhury, and R. J. Stanley, "Detection of basal cell carcinoma using color and histogram measures of semitranslucent areas," *Skin Res. Technol.*, vol. 15, no. 3, pp. 283–287, Aug. 2009.

[26] B. Cheng, D. Erdos, R. J. Stanley, W. V. Stoecker, D. A. Calcara, and D. D. Gómez, "Automatic detection of basal cell carcinoma using telangiectasia analysis in dermoscopy skin lesion images," *Skin Res. Technol.*, vol. 17, no. 3, pp. 278–287, Mar. 2011.

[27] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees, "Color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions," in *Color Medical Image Analysis*, M. E. Celebi, and G. Schaefer, Eds. New York, NY, USA: Springer, 2012, pp. 63–86.

[28] S. Sigurdsson, P. A. Philipsen, L. K. Hansen, J. Larsen, M. Gniadecka, and H. C. Wulf, "Detection of skin cancer by classification of Raman spectra," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 10, pp. 1784–1793, Oct. 2004.

[29] B. Cheng, R. Joe Stanley, W. V. Stoecker, S. M. Stricklin, K. A. Hinton, T. K. Nguyen, R. K. Rader, H. S. Rabinovitz, M. Oliviero, and R. H. Moss, "Analysis of clinical and dermoscopic features for basal cell carcinoma neural network classification," *Skin Res. Technol.*, vol. 19, no. 1, pp. e217–e222, Feb. 2013.

[30] H. Iyatomi, K. A. Norton, M. E. Celebi, G. Schaefer, M. Tanaka, and K. Ogawa, "Classification of melanocytic skin lesions from non-melanocytic lesions," in *Proc. IEEE Eng. Med. Biol. Soc.*, 2010, pp. 5407–5410.

[31] K. Shimizu K, H. Iyatomi, K. A. Norton, and M. E. Celebi, "Extension of automated melanoma screening for non-melanocytic skin lesions," in *Proc. Int. Conf. Mechatronics Mach. Vis. Practice*, 2012, pp. 16–19.

[32] H. Zhou, M. Chen, R. Gass, L. Ferris, L. Drogowski, and J. M. Rehg, "Spatially constrained segmentation of dermoscopy images," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2008, pp. 800–803.

[33] M. E. Celebi, Y. A. Aslandogan, W. V. Stoecker, H. Iyatomi, H. Oka, and X. Chen, "Unsupervised border detection in dermoscopy images," *Skin Res. Technol.*, vol. 13, no. 4, pp. 454–462, Nov. 2007.

[34] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. 3, no. 6, pp. 610–621, Nov. 1973.
[35] B. S. Everitt and G. Dunn, *Applied Multivariate Data Analysis*. London, U.K.: Edward Arnold, 1991, pp. 219–220.
[36] T. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *J. Artif. Intell. Res.*, vol. 2, no. 1, pp. 263–286, Aug. 1994.
[37] X. Chen, X. Zeng, and D. V. Alphen, "Multi-class feature selection for texture classification," *Pattern Recog. Lett.*, vol. 27, no. 14, pp. 1685–1691, Oct. 2006.
[38] W. Stolz, A. Riemann, A. B. Cognetta, L. Pillet, W. Abmayr, and D. Holzel, "ABCD rule of dermatoscopy: a new practical method for early recognition of malignant melanoma," *Eur. J. Dermatol.*, vol. 4, no. 7, pp. 521–527, 1994.
[39] G. Argenziano, G. Fabbrocini, P. Carli, V. de Giorgi, E. Sammarco, and M. Delfino, "Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions. Comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis," *Arch. Dermatol.*, vol. 134, no. 12, pp. 1563–1570, Dec. 1998.
[40] A. N. Crowson, "Basal cell carcinoma: Biology, morphology and clinical implications," *Mod. Pathol.*, vol. 19, pp. s127–s147, Sep. 2006.

Authors' photographs and biographies not available at the time of publication.