Toward the discrimination of early melanoma from common and dysplastic nevus using fiber optic diffuse reflectance spectroscopy

Bruce W. Murphy
Rebecca J. Webster
The University of Western Australia
School of Electrical, Electronic, and Computer Engineering
Optical+Biomedical Engineering Laboratory
M018, 35 Stirling Highway
Crawley, Western Australia 6009
Australia

Berwin A. Turlach
The University of Western Australia
School of Mathematics and Statistics
M019, 35 Stirling Highway
Crawley, Western Australia 6009
Australia

Christopher J. Quirk
Christopher D. Clay
Royal Perth Hospital
Department of Dermatology
Perth, Western Australia 6000
Australia

Peter J. Heenan
Cutaneous Pathology
26 Leura Street
Nedlands, Western Australia 6009
Australia

David D. Sampson
The University of Western Australia
School of Electrical, Electronic, and Computer Engineering
Optical+Biomedical Engineering Laboratory
M018, 35 Stirling Highway
Crawley, Western Australia 6009
Australia

Abstract. We describe a study of the discrimination of early melanoma from common and dysplastic nevus using fiber optic diffuse reflectance spectroscopy. Diffuse reflectance spectra in the wavelength range 550 to 1000 nm are obtained using 400-μm core multimode fibers arranged in a six-illumination-around-one-collection geometry with a single fiber-fiber spacing of 470 μm. Spectra are collected at specific locations on 120 pigmented lesions selected by clinicians as possible melanoma, including 64 histopathologically diagnosed as melanoma. These locations are carried through to the histopathological diagnosis, permitting a spatially localized comparison with the corresponding spectrum. The variations in spectra between groups of lesions with different diagnoses are examined and reduced to features suitable for discriminant analysis. A classifier distinguishing between benign and malignant lesions performs with sensitivity/specificity of between 64/69% and 72/78%. Classifiers between pairs of the group common nevus, dysplastic nevus, *in situ* melanoma, and invasive melanoma show better or similar performance than the benign/malignant classifier, and analysis provides evidence that different spectral features are needed for each pair of groups. This indicates that multiple discriminant systems are likely to be required to distinguish between melanoma and similar lesions. [DOI: 10.1117/1.2135799]

Keywords: diffuse reflectance spectroscopy; elastic scattering spectroscopy; white light spectroscopy; tissue optics; melanoma; skin cancer.

Paper 05085R received Mar. 29, 2005; revised manuscript received Jun. 17, 2005; accepted for publication Jun. 17, 2005; published online Nov. 18, 2005.

1 Introduction

Among the wide variety of skin cancers, cutaneous melanoma stands out as the major cause of fatality, and its incidence continues to increase.[1] Early detection and removal is critical, as the prognosis for melanoma worsens with lesion thickness. Data from an American cancer registry show a five-year survival rate for thin melanoma to be close to 100%, whereas ulcerated invasive lesions exceeding 4 mm in thickness, even without recorded metastases, have a five-year survival rate of 45%.[2] Early diagnosis is vital but problematic. Unaided visual examination of pigmented lesions by a physician is the most common technique, but studies report unsatisfactory diagnostic accuracy (compared with histopathology) ranging from 56[3] to 85%[4] and high interphysician variability. A 25-year review of melanoma in the United Kingdom revealed that a correct clinical diagnosis was made in only 41% of cases.[5] It is apparent that positively distinguishing melanoma from benign pigmented lesions is often difficult until the melanoma's invasion is advanced. Physicians overcome this difficulty by excision of any suspicious lesions. The resulting unnecessary excision of many benign lesions has a significant monetary cost and causes patient discomfort and disfigurement. There is scope for substantial improvement. We describe our research into the noninvasive clinical diagnosis of early melanoma based on the spectroscopy of diffuse reflectance of white light

Address all correspondence to Bruce Murphy, Electrical, Electronic and Computer Engineering M018, University of Western Australia, 35 Stirling Highway, Crawley, Western Australia 6009 Australia. Tel: +61 8 6488 7253; Fax: +61 8 6488 1065; E-mail: packrat@ee.uwa.edu.au

1083-3668/2005/10(6)/064020/9/$22.00 © 2005 SPIE

from the lesion using a fiber optic, nonimaging contact probe.

For the purposes of our study, melanoma and melanoma-like pigmented lesions were classified into four groups.[6] Common nevus (N) exhibits a proliferation of melanocytes without severe atypia. These melanocytes are located at the dermo-epidermal junction or in the dermis. Compound nevus are a form of common nevus exhibiting melanocytes in both locations. In a dysplastic nevus (DN), signs of dysplasia, including pleomorphism (changes in the size and shape of cells and their nuclei) are observed among melanocytes. In *in situ* melanoma (ISM), proliferation of severely atypical melanocytes (in a layer or in nests) is evident, but confined to the epidermis. Invasive melanoma (IM) exhibits abnormal melanocytes invading the dermis. In this case, the proliferation of abnormal melanocytes has extended through the basement membrane of the epidermis and may eventually reach the subcutis. Advanced invasive melanomas may exhibit metastasis, the spreading of the melanoma to remote sites via the lymphatic or circulatory systems. Early melanomas are primarily *in situ*, but may have areas of shallow dermal invasion.

Scattering and absorption are known to characterize the biochemical and morphological state of tissues,[7] but the spectral signature of melanoma compared with nevus is not well characterized, particularly in the early stages of melanoma development. The most definite cytological feature of malignancy is nuclear pleomorphism, which is usually accompanied by disorder in the tissue architecture on the scale of tens of cells. Little is known about the differences in the distribution of melanin and melanosomes in benign and malignant lesions, or the timing of the onset of increased blood supply to lesions via angiogenesis, the creation of new blood vessels from pre-existing ones. All of these features may alter the diffuse reflectance signature.

Clinical diagnostic instruments should approach or exceed the performance of specialist physicians. Recently quoted average values for sensitivity (percentage of malignant lesions correctly diagnosed), specificity (percentage of benign lesions correctly diagnosed), and diagnostic accuracy (their mean) are 80, 50, and 65%, respectively.[8] In all studies, the gold standard is conventional histopathology. A key factor in the adoption of new instruments is their ease of use; a fiber optic probe is particularly attractive, as a measurement can be conducted in seconds, making whole body scans possible within minutes. Fiber optic probes provide spatially localized information on the scale of the fiber bundle. There has been little research reported on their use in skin cancer diagnosis. Our fiber-based approach employs diffuse reflectance spectroscopy (sometimes known as elastic scattering spectroscopy[9]). It measures the modification of the spectrum of remitted light, i.e., light that has propagated some distance into the skin, been scattered, and recollected at the skin's surface. This is accomplished via a fiber optic probe usually placed in direct contact with the skin. The first published work using nonimaging spectroscopic measurement of lesions was reported by Marchesini et al.[10] in 1992, who studied melanoma and nevus with a 5-mm probe but later abandoned this approach in favor of imaging spectroscopy.[11] Zeng et al. developed a system suitable for studying the autofluorescence or diffuse reflectance of skin *in vivo*,[12] but the system was not used to study melanoma. The first major study of melanoma diagnosis by fiber probe spectroscopy was conducted by Wallace et al.,[13,14]

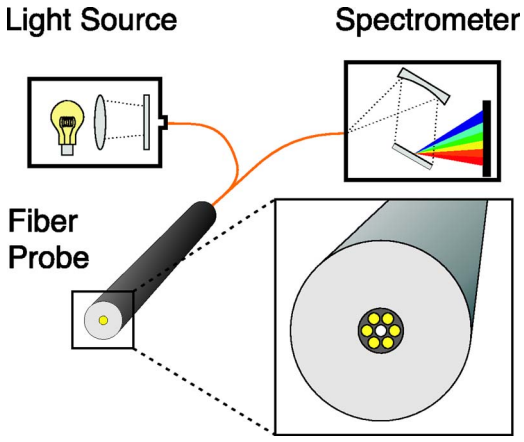

Fig. 1 Schematic diagram of the fiber probe-based skin spectroscopy instrument.

who examined invasive melanomas and nevi. A contact fiber probe was coupled to a spectrometer with a wavelength range of 320 to 1100 nm. McIntosh et al.[15] developed a noncontact fiber probe coupled to a spectrometer with an extended spectral range of 400 to 2500 nm, and examined a range of nonmelanoma skin lesions. Most recently, Garcia-Uribe et al.[16] developed a system employing oblique fiber illumination and a linear array of collection fibers, which they used to examine a number of nonmelanoma pathologies.

We begin in Sec. 2 with a description of the hardware and methods we employed to record the spectra, compare them with histopathological diagnoses, and utilize both in the creation of a classifier. We report in Sec. 3 the performance of a range of classifiers and consider the most important spectral features in each. In Sec. 4, we discuss our results and consider the wider implications, before concluding in Sec. 5.

2 Methods

The primary focus of this study was the collection of multiple distinct white-light diffuse reflectance spectra from pigmented lesions that were suspected of being melanoma. The subjects recruited for the study were patients who had been evaluated by their physician as having one or more clinically suspicious lesions that required excision. This resulted in our clinical dataset containing an unusually high proportion of lesions that were difficult to diagnose clinically. The study was approved by the Ethics Committee of the University of Western Australia.

2.1 *Spectrometer System*

The instrument used to record the spectra is shown schematically in Fig. 1. Diffuse reflectance spectra in the wavelength range 550 to 1000 nm were obtained using a pocket spectrometer (CVI Laser Systems, Albuquerque, NM) with a 2048-pixel linear charge-coupled device detector and spectral resolution of 2 nm. The light source is a lamp utilizing a regulated 2900-K tungsten-halogen bulb. The fiber probe contains seven 400-μm core step-index fibers optimized for low loss in the UV/VIS region in a close-packed six-illumination-around-one-collection arrangement with center-center separations of 470 μm and an outer diameter of 1.3 mm (CVI Laser

Systems). Data were acquired from the spectrometer via a digital acquisition card (National Instruments). The acquisition software was written in LabView and produced data files suitable for further processing and analysis in Matlab.

For calibration, reference spectra from a white diffusing reflectance standard (CVI Laser Systems) and a dark scan were recorded using the same integration time as the measurement. The fiber probe end was held at a fixed angle and distance to the reflectance standard in a blackened, light-tight enclosure designed to eliminate secondary reflections and stray light. This distance and angle were designed to yield a reflectance slightly greater than that of pale skin in contact with the probe. A standard used in a contact geometry with optical properties closer to that of skin would be preferable, but we know of none with sufficiently well characterized properties. The lamp produced stable output after less than five minutes of operation. Warm-up times of at least five minutes were used for all calibration and measurement. Spectra were recorded by taking ten spectral measurements over a period of five seconds, each with an integration time of 30 ms. These measurements were averaged to obtain the final spectrum. Varying probe pressure did affect the spectra, but spectra measured with this system from a small area of normal skin were found to vary by less than 5%. A trained operator was able to achieve repeatability on the order of 2%.

2.2 *Measurement Protocol and Histopathology*

In this study, we sought (for the first time to our knowledge) to more closely couple the process of histopathological diagnosis with the spatial localization provided by the fiber probe. This was achieved in three steps: 1. record a localized spectrum; 2. transfer the location of the probe from the tissue *in vivo* to the microscopic histological section; and 3. record a histopathological diagnosis of a type and scale relevant to the recorded spectrum. This last represents a major departure from routine histopathology practice.

To determine the tissue volume over which the probe was most sensitive, we assumed the sensitivity to tissue changes to be proportional to the photon fluence.[17] Light transmission and collection through skin was simulated with a layered model (to be described elsewhere) and Monte Carlo modeling based on a modified version of the multi-layered Monte Carlo software package (MCML).[18] We concluded that the area of highest sensitivity was confined to a 400-μm-diam cylinder immediately below the central collection fiber. This cylinder had a wavelength-dependent depth range of 200 to 400 μm, extending into the papillary dermis.

2.2.1 *Record a localized spectrum*

To make multiple spectral measurements from different areas of the same lesion, the skin surrounding the lesion was marked with a 2-mm reference grid using ink that remained visible throughout the excision and subsequent tissue processing steps. These markings were used to guide the placement of the probe via a matched template attached to it. At each location, the probe was placed in light contact with the skin (using index-matching glycerol) and a spectrum was recorded. Figure 2 shows schematic diagrams of the grid, the matching template, the scan locations on pathology tissue blocks, and a photograph of a marked lesion. A spectrum was also taken from the normal tissue adjacent to the lesion.

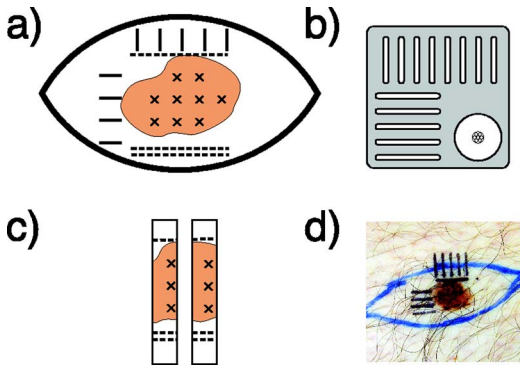

Fig. 2 Measurement procedure: (a) lesion with marked measurement grid, shrinkage references, ellipse, and scan positions; (b) template used to position probe; (c) excised tissue blocks showing offset scan positions; and (d) photograph of a lesion.

2.2.2 *Transfer probe locations to histopathological sections*

Immediately following the recording of the spectra, the lesion was excised following an inscribed ellipse and placed in formalin. An incision was made along one side of the grid as a position reference. When the size of the ellipse permitted, a second double incision was made along the opposite grid edge to permit shrinkage measurement. The ellipse and incisions are shown schematically in Fig. 2(a). The incisions appeared on the sections as notches, as illustrated in Fig. 3. The excised tissue was cut along the grid lines into 2-mm-wide blocks, noting the side adjacent to the probe location [Fig. 2(c)]. Standard histological procedures for cutaneous lesion diagnosis[19] were then followed. The blocks of tissue were dehydrated, clarified, embedded in wax, and sliced with a microtome, initially to cut back the uneven tissue surface until continuous sections were obtained, and then to yield 5-μm-thick sections. These sections were stained using hematoxylin and eosin solution and mounted on a microscope slide.

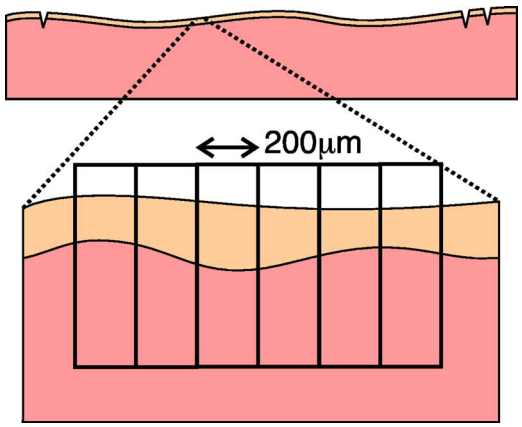

Fig. 3 Example section showing shrinkage and position reference incisions, and a magnified location showing relative size of diagnostic grid.

Table 1 Lesions and localized spectra recorded during this study.

Diagnosis	Code	Number	Total spectra
Common nevus	N	23	75
Dysplastic nevus	DN	33	167
In situ melanoma	ISM	24	114
Invasive melanoma	IM	40	224

Table 2 Local diagnoses (columns) as a percentage for each category of overall histopathological diagnosis (rows).

	Localized diagnosis			
Whole lesion diagnosis	IM	ISM	DN	N
Dysplastic nevus	0	0	83	17
In situ melanoma	0	82	3	15
Invasive melanoma	81	9	1	9

The initial cutting back of the tissue with the microtome was determined to remove approximately 0.5 mm of tissue, but this varied between samples, leading to some uncertainty in position. The measurement grid was offset by 0.5 mm from the marked grid lines to compensate for this. Tissue shrinkage also caused uncertainty in position. It may occur during the excision, fixation, and histopathology processing steps. Thin sections may exhibit nonuniform shrinkage, resulting in uneven curvature of the skin's surface, which impedes the accurate measurement of length. Lateral shrinkage was determined from a comparison of the measured distance between the reference marks on the mounted section and the known grid spacing. Measured values were in the range 0 to 25%. These values were used to adjust the locations for localized histopathology.

Shrinkage could only be measured when the excision area was large enough to permit both reference incisions to be made. Cosmetic considerations frequently prevented this, resulting in shrinkage data being available for fewer than half of the recorded lesions.

2.2.3 *Record localized histopathology*

Based on the results of our modeling, we produced a template overlay for the histopathology diagnosis that divided up the slide at the probe location into six 200-μm-wide vertical sections, as shown in Fig. 3. Each of the locations was divided into an epidermis and a dermis portion. As vertical shrinkage was not quantified, this easily identified division was used in preference to a specific depth. The distribution of atypical melanocytes was evaluated for each section. Within the epidermis, these melanocytes could be absent, present in a layer along the dermal-epidermal junction, or present in nests in the lower epidermis, or through the whole depth of the epidermis. Invasions into the dermis were recorded with notes on the depth, extent, and volume proportion of melanocytes in the invaded tissue. The presence of regression was also recorded.

2.3 *Description of Clinical Data*

Table 1 lists the number of lesions and spectra collected during this study in each of the four pathology groups. The 120 lesions were obtained from 115 patients. The number of spectra for each lesion varied with the number of grid intersections that lay within the lesion. Lesion sizes recorded in our study varied with type: dysplastic nevi ranged in size from 2 to 15 mm in diameter and melanomas from 2 to 20 mm. The average number of spectra taken for each lesion type ranged from 3.2 in nevus, to 5.6 in invasive melanoma and 5.0 in dysplastic nevus.

A new feature of this study is the collection of multiple spectra from each lesion. Our measurement grid ensures that these spectra are from nonoverlapping areas spaced over the whole lesion. Indications of the variability of localized histological diagnosis within our measured lesions are provided in Table 2. Each of the abnormal pathology subgroups have a localized diagnosis of common nevus in around 10% or more of instances, and invasive melanoma has a localized diagnosis of *in situ* melanoma in around 10% of instances. The whole-lesion histopathological diagnosis (the gold standard) is as severe as the worst local diagnosis. For example, an *in situ* melanoma may be associated with areas of common and dysplastic nevus, but will not contain any regions of invasive melanoma.

Average spectra recorded by the system for normal skin, dysplastic nevus, *in situ* melanoma, and invasive melanoma are shown in Fig. 4. The shaded areas represent the regions within one standard deviation above and below the average curves shown. The figures illustrate the variability between and within the groups of spectra, and provide an indication of the difficulty of the classification problem.

2.4 *Data Analysis*

Little information currently exists on the relationship between specific types of pigmented lesions and their reflectance spectra. In the absence of such information, we approach the problem of lesion classification with statistical techniques and seek classifiers that can separate lesions into pathology groups. Next, we describe the extraction of numeric features from the spectra, the construction and evaluation of classifiers, and the evaluation of the importance of features.

2.4.1 *Feature definition*

The spectral range and resolution limit spectra to 225 potentially independent data points, which is an intractably large number for our statistical approach. We sought to reduce this to a manageable number of data points or features using two approaches. The first follows Wallace et al.[13] and contains intensity and slope features placed at points of spectral difference between pathology groups. (The use of slopes in addition to levels for classifying spectra is common within the field of chemometrics.) To these, we have added features specific to chromaphores within skin for a total of 29 features. The second approach, in common with McIntosh et al.,[15] uses 50

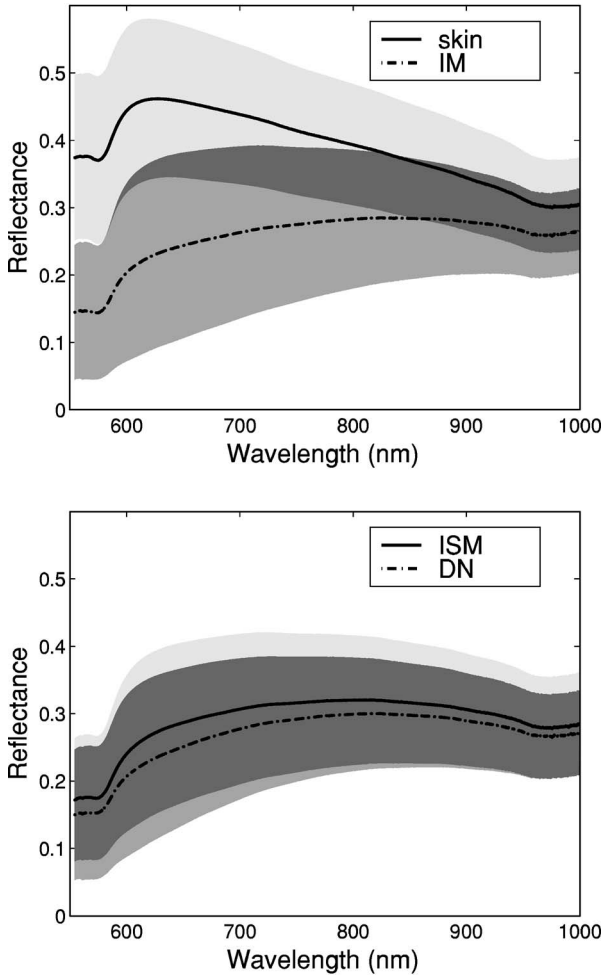

Fig. 4 Measured spectra showing average curves and ±1 SD boundaries for skin and invasive melanoma (upper) and dysplastic nevus and *in situ* melanoma (lower).

features constructed from a uniform sampling of the intensity across the collected spectrum. Some previous studies have made use of normalized spectra, i.e., the spectra from the lesion are divided by the spectrum of normal skin adjacent to the lesion. We collected normal skin spectra, but found that classification using normalized spectra was less accurate. We do not further describe these results.

2.4.2 *Classification system*

Several different approaches may be used to classify lesion spectra. Neural network classifiers are more commonly employed when a very large number of input values are used. They are prone to overfitting, the failure to generalize to data outside the training set,[20] and are difficult to use in quantifying feature importance.[21] For these reasons, we restrict our attention to discriminant analysis. Linear and quadratic discriminant classification systems generate a numerical value by combining input features; binary classification is performed by applying a threshold to this value. Quadratic discriminant analysis is capable of distinguishing sets that are inseparable by linear methods but significantly increases the number of features that must be considered[20] and may be inaccurate when used on non-normal data. The features used in our analysis were tested for normality, and significant kurtosis and truncation were observed. Furthermore, tests on an initial subset of the data showed quadratic discriminants performed worse than linear discriminants. For these two reasons, we restricted our subsequent analysis to linear classifiers.

2.4.3 *Generating and ranking classifiers*

We have performed our feature selection using an all-subsets approach.[22] This approach may be susceptible to overfitting and coincidental matches if the classifier is generated using purely internal analysis, i.e., if the whole dataset is used for both training and testing. We avoided this problem by using k-fold cross validation[20] to robustly evaluate the classifier performance. K-fold cross validation is known to underestimate the classification accuracy, because its test classifiers are constructed on subsets of the data. In general, classifiers improve their performance with dataset size, but this performance is asymptotic to the maximum classification accuracy. It is difficult to determine the required size of dataset for a particular classifier, but empirical rules of thumb exist; Huberty[22] suggests that for two-group classifiers, the group dataset sizes should exceed 3p, for a p-feature classifier. We do not expect multiple spectra from the same lesion to be wholly independent. If we assume no independence (a conservative view), some of our pathology groups are just sufficient in size to test a five-feature classifier.

The optimal number of features to include in a classifier is closely tied to the size of the dataset, but also depends on the changes in classifier performance as the number of features is increased. In general, classifier performance on a training set improves asymptotically to a maximum with the number of features, but the ability to generalize to new data is reduced at the same time. Hastie, Tibshirani, and Friedman[20] suggest adopting the smallest sized set of features that approaches the asymptotic limit. We have considered three-, four-, and five-feature classifiers. The small performance increase between four- and five-feature classifiers indicates that we are approaching the asymptotic limit for a classifier set by our dataset size.

In ranking the classifiers, a single five-fold partition was used. When the mean five-fold cross-validated accuracy from ten random partitions was calculated for the classifiers, the ranking of classifiers was almost unchanged. The results presented next use classifiers ranked by a single five-fold cross-validation.

3 Results

We present indicative results based on sensitivity and specificity at a single threshold, as well as an illustrative set of receiver operator characteristic (ROC) curves, which characterize the trade-off of sensitivity against specificity as the classifier threshold is adjusted. In this and following sections, feature set A is the set of features based on spectral characteristics, and feature set B is the set of uniformly sampled points across the spectra.

3.1 *Classifiers*

Three- and five-feature classifiers were constructed using feature sets A and B for a range of groupings reported with

Table 3 Five-fold cross-validated classifier performance for three- and five-feature classifiers expressed as percentage sensitivity/specificity for the feature sets A and B.

	Set A		Set B	
Classifier	Three features	Five features	Three features	Five features
N+DN versus ISM+IM	60/61	60/65	60/61	64/69
N versus IM	67/65	72/70	63/60	73/73
N versus DN	70/64	72/69	76/68	77/69
DN versus ISM	55/63	59/66	56/66	56/71
ISM versus IM	63/63	64/60	63/54	69/58

sensitivity/specificity in Table 3. Classifiers were constructed to distinguish between benign (N and DN) and malignant (ISM and IM) pigmented lesions. The best classifier had sensitivity/specificity of 64/69% and used five features from the feature set B. Under internal analysis, the same classifier had sensitivity/specificity of 72/78%. For comparison, classifiers were constructed between N and IM. Although not clinically useful for early melanoma, this classification facilitates the comparison with other studies in Sec. 4. The best classifier had sensitivity/specificity of 73/73% and used five features from feature set B. Under internal analysis, the same classifier had sensitivity/specificity of 74/83%.

Given the modest performance of these classifiers, three additional classifier types were generated to test the hypothesis that the different stages of melanoma development produce distinct changes to the recorded spectra that are not easily distinguished by a single benign (B)/malignant (M) classifier. The classifiers were set up between N and DN, DN and ISM, and ISM and IM, and constructed from five-feature classifiers from feature set B. Figure 5 displays representative ROC curves for each classifier as well as for the B/M classifier. Ideally, the B/M classifier should exhibit significantly better performance than the others, as it is based on a larger dataset. In fact, the performance of the N/DN classifier exceeds that of the B/M classifier, and the DN/ISM and ISM/IM classifiers display better performance over much of the threshold range. These results suggest that an approach utilizing multiple classifiers may perform better. This hypothesis is explored further in Sec. 4. We cannot test multiple classifier performance directly, as the subdivided datasets are too small for such an analysis. We instead describe a method for evaluating and comparing the significant spectral features in each classifier.

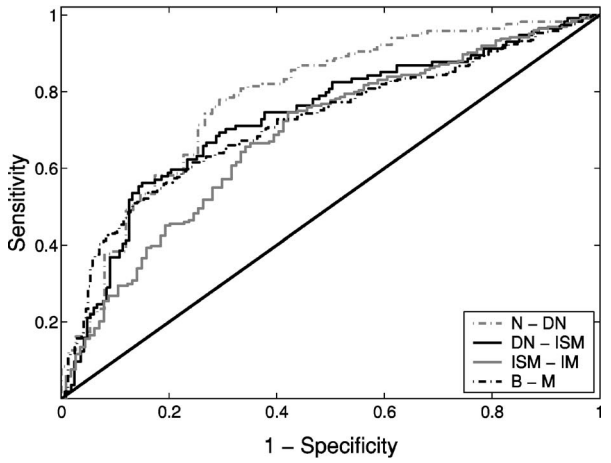

Fig. 5 Receiver-operator characteristic curves for selected diagnosis-based groups of melanoma and nevus. The straight curve represents random classification.

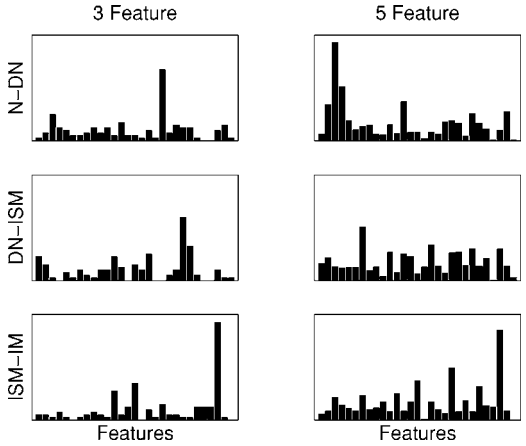

Fig. 6 Feature significance counts for feature set A for each classification group for three- and five-feature classifiers. The *x* axis enumerates features.

3.2 *Quantifying Feature Importance*

An obvious approach for determining the most effective classifier features is to declare those features in the best classifier to be the most important. This approach is flawed, as the single best classifier may have arisen through a coincidental fit and may not perform well on new data. Our feature importance metric is the number of occurrences of each feature within the top 1% of classifier feature sets, which is similar to the approach suggested by Huberty.[20] We are able to confidently report feature significance, because we evaluate the performance of all classifiers for each of the feature sets. Figure 6 displays histograms of the frequency of occurrence of these features in the top-ranked three- and five-feature classifiers based on feature set A for the N/DN, DN/ISM, and ISM/IM classifiers. Figure 7 shows the same feature frequency histograms for the top classifiers obtained using feature set B. These results are discussed in the next section.

4 Discussion

We first discuss some of the interesting aspects of our results in Sec. 4.1, and then consider the wider implications in Sec. 4.2.

4.1 *Issues Arising from this Study*

4.1.1 *Feature importance*

The histograms presented in Figs. 6 and 7 show that only a small number of features appear frequently within the top 1%

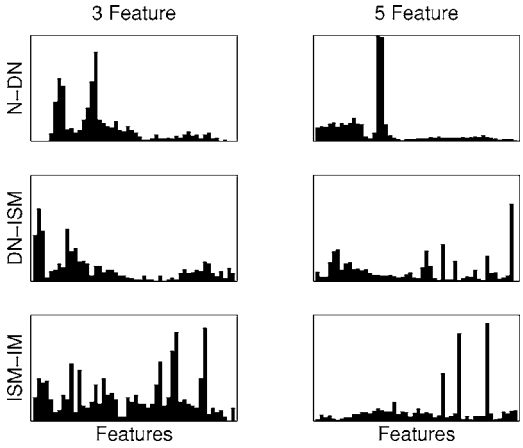

Fig. 7 Feature significance counts for feature set B for each classification group for three- and five-feature classifiers. The *x* axis represents the spectral sampling point.

of classifiers. If we construct a new classifier based on the most frequently appearing features, then the high performance of this classifier demonstrates that the selection of important features by this metric is not spurious. Indeed, in each case, this most-frequent-feature classifier was within the top 1% of classifiers.

The histograms also illustrate that the classifiers for each of the three paired pathologies have a different set of most-important features. The classifiers for the two pairs farthest from the B/M threshold, N/DN and ISM/IM, share almost no features. This result is consistent with the histopathology of early melanoma. The predominant change exhibited in early dysplasia, nuclear pleomorphism in the lower epidermis, is quite distinct from the proliferation of epidermal nests and dermal invasion observed during later *in situ* and invasive melanoma.[6] As a consequence, these two stages can be expected to have distinct spectral signatures, and this is supported by these results. For the pathology pair astride the B/M threshold, DN/ISM, the feature frequency histograms show that the best classifier shares some features with both the N/DN and ISM/IM classifiers. This is consistent with *in situ* melanoma being characterized by features of both early stage dysplasia and later stage proliferation.

We note that these observations are consistent for feature sets A and B and for both three- and five-feature classifiers. This eliminates the possibility that they have arisen from coincidental features of a single feature set or classifier type.

4.1.2 *Pathology-based subclassification*

Our classifier performance results demonstrated similar or slightly increased classification ability when our benign and malignant lesion classes were subdivided into specific pathology groups. As this subdivision typically decreases the dataset size by a factor of 2, a decrease in classification accuracy would be expected, unless the division simplifies the classification task. These results suggest that a classification system for melanocytic lesions should be designed to take advantage of the distinct signatures of the pathology stages. Our results on feature importance lend weight to this suggestion. Several classification techniques exist for constructing multigroup classifiers, but they require more data than is available in this study.

4.1.3 *Spatially localized histopathology*

Our choice of a 200-μm grid for localizing the modified histopathological diagnosis took into consideration the probed volume, the accuracy of the localization, and the workload of the diagnosing histopathologist. This grid was large enough to encompass architectural histopathological features that appear on the scale of tens of cells. Table 2 illustrates the heterogeneous nature of many lesions on this scale. We expect that the inclusion of measurements with conflicting diagnoses within their volume of sensitivity would reduce the performance of any empirically derived classifier and should ideally be removed from the training set. During this study, as described, accurately localized histopathology was not available for all lesions. Our analysis required inclusion of the whole dataset, and so could not take full advantage of the localized histopathology. A significant benefit still accrued from our measurement procedure, which ensured that multiple measurements from a lesion were evenly spaced, ensuring maximum independence and lesion coverage. The development of a multigroup classifier with a larger dataset would certainly require localized histopathology to ensure that the finer grained pathology groups were not compromised by scans from differently structured areas of the lesion.

4.2 *Wider Implications*

Ours is the only reported study based on a single illumination-collection fiber spacing. Wallace et al.'s probe had 18 illumination and 12 collection fibers arranged symmetrically with several different spacings and a total diameter of 1.5 mm,[13] similar to our own (1.3 mm). McIntosh et al., in contrast, used a much larger probe with a 7 mm diameter held slightly above the surface of the skin, causing the returned signal to be averaged over the lesion.[15] Neither study quantified the volume of sensitivity of the probe. Garcia-Uribe employed a probe with varying fiber spacings across the lesion and recorded the spectra returned from each fiber.[16] No correction was reported for the different sampling depths toward one side of the lesion, the effect of the probe's asymmetry, or the contribution of lesion extent to these readings. While these studies provide a useful starting point, the use of varying fiber spacings and unquantified volumes of probe sensitivity prevent the studies from being extended to make use of localized histopathology. Furthermore, sampling of a whole lesion could prevent the identification of small morphological changes that are important in the histological detection of early melanoma.

Wallace et al.[13] reported a good classification performance of 100/84% with a set of 15 melanoma and 32 compound nevi (i.e., common nevi containing melanocytes in the epidermis and dermis). An extension to this study[14] reported a reduced performance of 83/88% using a neural network classifier on a larger dataset with 26 melanoma and 49 common nevi. This classification task differed from the one we report by the exclusion of the intermediate cases of dysplastic nevus and *in situ* melanoma. McIntosh and Garcia-Uribe both reported high classification accuracies (97 and 100%, respec-

tively) in distinguishing nevus from dysplastic nevus. Our N/DN classifier had the highest performance, suggesting it is the most tractable of the clinically useful classifiers.

Most other studies of the diagnosis of pigmented lesions have employed a camera-based system. This has been coupled with image processing to replicate the procedures followed by physicians, or has included images taken at multiple wavelengths and their analysis. Camera-based systems are further developed than other technologies for lesion classification and have been the subject of larger clinical trials. The B/M classification performance (sensitivity/specificity) of these systems include the Spectrophotometric Intracutaneous Analysis system SIAscope's 80.1/82.7% on a set of 384 pigmented lesions with 52 melanomas with six *in situ*;[23] Electro-Optical Sciences (Irvington, NY) MelaFind's 100/85% on a set of 246 lesions with 63 melanomas comprising 33 invasive and 30 *in situ*;[24] and the diagnostic and neural analysis of skin cancer (DANAOS) study that yielded 82/85% with a dataset containing 2218 lesions with 187 melanoma.[25] Several of these studies relied on leave-one-out analysis to verify their classification accuracies. This technique has greater variance than five-fold cross-validation and may yield overestimates for individual classifiers.[20]

The presence and handling of the intermediate lesion pathologies (DN and ISM) in these studies varied considerably. While the SIAscope and MelaFind studies did distinguish between *in situ* and invasive melanoma, the classifier was constructed to distinguish the whole melanoma set. The small proportion of *in situ* melanoma in the SIAscope study is not representative of early melanoma. The DANAOS study did not distinguish between melanoma development stages; instead, melanoma were grouped by pathology type with the majority being superficial spreading melanoma. The proportion of dysplastic nevus also determines the relevance of the study to early melanoma. The DANAOS and SIAscope studies presented 11 and 3% of their total nevus as dysplastic. The MelaFind study reported 60%, which was similar to our clinically observed population. Our own results (Table 3) suggest that populations lacking the intermediate cases may yield artificially high classification performances. In our view, most lesion classification studies reported to date should be considered as preliminary, and require larger and more refined clinical trials to further evaluate and develop their potential.

5 Conclusion

This study achieves modest performance in the classification of pigmented skin lesions, from 139 patients selected from a prescreened population presenting clinically suspicious lesions to physicians. Our results (Table 2) demonstrate the heterogeneity of lesions and the need for localization of the measurement volume. We described a probe design and protocol for the generation of a training set that takes this into account, although this was only partially utilized due to the lack of data. The improved or similar classification performances on groups of lesions that were divided by specific lesion pathology, in combination with the importance of different features for different groups, suggest that distinct spectral signatures distinguish each lesion type. These results support the development of classifiers for individual lesion development stages, rather than simple classifiers that seek to establish a benign or malignant case.

Our results highlight a need for a new clinical dataset, large enough that the individual groups are adequate to accurately train classifiers and assess their performance. In this case, each of the groups will need to contain at least 100 lesions. As statistical errors can be introduced by a training set that is of significantly different composition to the screened population, an effort will have to be made to ensure that malignant pathologies do not dominate. If such a device were to be designed for screening by nonspecialists, this would require the inclusion of a large number of benign lesions. A further obstacle to such a study is the ethical consideration that prevents excision and histopathology of examined benign lesions. Without excision, gold-standard evaluation of lesion microstructure is impossible. An extension of this work is the development of classifiers that are not restricted to threshold-based binary outcomes. This would permit the degree of malignancy and invasion to be better assessed, which would likely be of the most interest to specialists. Despite the modest performance of our classifiers, our study suggests avenues for further development of the next generation of more accurate fiber-probe-based lesion classification systems.

Acknowledgments

We wish to gratefully acknowledge the various contributions to the studies reported here made by our colleagues: Paul Bond, Ha Dang, Philippe Lauper, Martin Punke, Ian Walton, Harry Whyte, and Kirrily Wong. We also wish to thank the doctors and specialists who assisted with the clinical data collection.

References

1. A. Jemal, S. S. Devesa, T. R. Fears, and P. Hartge, "Cancer surveillance series: Changing patterns of cutaneous malignant melanoma mortality rates among whites in the United States," *J. Natl. Cancer Inst.* **92**, 811–818 (2000).
2. C. M. Balch, A. C. Buzaid, S. J. Soong, M. B. Atkins, N. Cascinelli, D. G. Coit, I. D. Fleming, J. E. Gershenwald, A. Houghton, Jr., 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.* **19**(16), 3635–3648 (2001).
3. C. A. Morton and R. M. MacKie, "Clinical accuracy of the diagnosis of cutaneous malignant melanoma," *Br. J. Dermatol.* **138**, 283–287 (1998).
4. C. M. Grin, A. W. Kopf, B. Welkovich, R. S. Bart, and M. J. Levenstein, "Accuracy in the clinical diagnosis of malignant melanoma," *Arch. Dermatol.* **126**, 763–766 (1990).
5. R. A. G. Parslew and L. E. Rhodes, "Accuracy of diagnosis of benign skin lesions in hospital practice: A comparison of clinical and histological findings," *J. Eur. Acad. Dermatol. Venereol* **9**, 137–141 (1997).
6. H. Kamino and A. B. Ackerman, "Malignant melanoma in situ: The evolution of malignant melanoma within the epidermis," in *Pathology of Malignant Melanoma*, A. B. Ackerman, Ed., Masson Publishing, New York (1981).
7. V. V. Tuchin, *Tissue Optics*, SPIE Press, Bellingham, WA (2000).
8. A. A. Marghoob, L. D. Swindle, C. Z. M. Moricz, F. A. Sanchez Negron, B. Slue, A. C. Halpern, and A. W. Kopf, "Instruments and new technologies for the *in vivo* diagnosis of melanoma," *J. Am. Acad. Dermatol.* **49**, 777–797 (2003).
9. I. J. Bigio and J. R. Mourant, "Ultraviolet and visible spectroscopies for tissue diagnostics: Fluorescence spectroscopy and elastic-scattering spectroscopy," *Phys. Med. Biol.* **42**, 803–814 (1997).

10. R. Marchesini, N. Cascinelli, M. Brambilla, C. Clemente, L. Mascheroni, E. Pignoli, A. Testori, and D. R. Venturoli, "In vivo spectrophotometric evaluation of neoplastic skin pigmented lesions II: Discriminant analysis between nevus and melanoma," *Photochem. Photobiol.* **55**, 515–522 (1992).
11. R. Marchesini, M. Ballerini, C. Bartoli, E. Pignoli, A. E. Sichirollo, S. Tomatis, S. Zurrida, and N. Cascinelli, "Telespectrophotometry of human diseases by means of a CCD camera," *Proc. SPIE* **2081**, 168–173 (1993).
12. H. Zeng, C. MacAulay, B. Palcic, and D. I. McLean, "A computerised autofluorescence and diffuse reflectance spectroanalyser system for *in vivo* studies," *Phys. Med. Biol.* **38**, 231–240 (1993).
13. V. P. Wallace, D. C. Crawford, P. S. Mortimer, R. J. Ott, and J. C. Bamber, "Spectrophotometric assessment of pigmented skin lesions: Methods and feature selection for evaluation of diagnostic performance," *Phys. Med. Biol.* **45**, 735–751 (2000).
14. V. P. Wallace, J. C. Bamber, D. C. Crawford, R. J. Ott, and P. S. Mortimer, "Classification of reflectance spectra from pigmented skin lesions, a comparison of multivariate discriminant analysis and artificial neural networks," *Phys. Med. Biol.* **45**, 2859–2871 (2000).
15. L. M. McIntosh, R. Summers, M. Jackson, H. H. Mantsch, J. R. Mansfield, M. Howlett, A. N. Crowson, and J. W. P. Toole, "Towards non-invasive screening of skin lesions by near-infrared spectroscopy," *J. Invest. Dermatol.* **116**, 175–181 (2001).
16. A. Garcia-Uribe, N. Kehtarnavaz, G. Marquez, V. Prieto, M. Duvic, and L. V. Wang, "Skin cancer detection by spectroscopic oblique-incidence reflectometry: Classification and physiological origins," *Appl. Opt.* **43**, 2643–2650 (2004).
17. M. Hiraoka, M. Firbank, M. Essenpreis, M. Cope, S. R. Arridge, P. van der Zee, and D. T. Delpy, "A Monte Carlo investigation of optical pathlength in inhomogenous tissue and its application to near-infrared spectroscopy," *Phys. Med. Biol.* **38**, 1859–1876 (1993).
18. L. Wang, S. L. Jacques, and L. Zheng, "MCML—Monte Carlo modeling of light transport in multi-layered tissues," *Comput. Methods Programs Biomed.* **47**(2), 131–146 (1995).
19. E. Calonje, "ACP best practice No. 162. The histological reporting of melanoma—Association of Clinical Pathologists," *J. Clin. Pathol.* **53**(8), 587–590 (2000).
20. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, New York (2001).
21. M. H. Hassoun, *Fundamentals of Artificial Neural Networks*, MIT Press, Cambridge, MA (1995).
22. C. J. Huberty, *Applied Discriminant Analysis*, Wiley-Interscience, New York (1994).
23. M. Moncrieff, S. Cotton, E. Claridge, and P. Hall, "Spectrophotometric intracutaneous analysis: A new technique for imaging pigmented skin lesions," *Br. J. Dermatol.* **146**, 448–457 (2002).
24. M. Elbaum, A. W. Kopf, H. S. Rabinovitz, R. G. B. Langley, H. Kamino, M. C. Mihm, A. J. Sober, G. L. Peck, A. Bogdan, D. Gutkowicz-Krusin, M. Greenebaum, S. Keem, M. Oliviero, and S. Wang, "Automatic differentiation of melanoma from melanocytic nevi with multispectral digital dermoscopy: A feasibility study," *J. Am. Acad. Dermatol.* **44**(2), 207–218 (2001).
25. K. Hoffman, T. Gambichler, A. Rick, M. Kreutz, M. Anscheutz, T. Grunendick, A. Orlikov, S. Gehlen, R. Perotti, L. Andreassi, J. Newton Bishop, J. P. Cesarini, T. Fischer, P. J. Frosch, R. Lindskov, R. Mackie, D. Nashan, A. Sommer, M. Neumann, J. P. Ortonne, P. Bahadoran, P. F. Penas, U. Zora, and P. Altmeyer, "Diagnostic and neural analysis of skin cancer (DANAOS). A multicentre study for collection and computer-aided analysis of data from pigmented skin lesions using digital dermoscopy," *Br. J. Dermatol.* **149**, 801–809 (2003).