Review article

Automated melanoma diagnosis: where are we at?

Greg R. Day and Robert H. Barbour

University of Waikato, Hamilton, New Zealand

Background: It has been over a decade since the first mention was made of the computer as a tool for assisting clinicians in diagnosing skin lesions. This review tabulates and summarises the major research papers, and comments on the state of the field after a decade of research.

Conclusions: We conclude that epiluminescent microscopy has become the image-capture technique of choice in this field. However, the reporting of research to date has been less than exemplary, making “reinvention of the wheel” likely. It also appears that although the goal of a clinically useful diagnostic system is closer, the complexity and variation displayed by skin lesions, coupled with the ad hoc direction and reporting of research, may hinder the achievement of this goal for some time to come.

Key words: digital imaging – computer diagnosis – melanoma – skin cancer – dermatoscopy

Melanoma, the most serious type of skin cancer, is an increasing problem worldwide. It is especially a problem in New Zealand, perhaps due to the ingrained ‘culture of the sun’ that New Zealanders appear to subscribe to, coupled with high ultraviolet conditions. Despite some research reports that indicate a leveling of the death rate from this disease (1) the increasing incidence of melanoma indicates that the melanoma problem is not yet solved.

Although melanoma is almost completely curable if detected and treated early enough, there is sufficient variation in skin lesions and conditions to present considerable difficulty to the clinician. A number of researchers have investigated diagnostic accuracy; and although results vary, the underlying message is that general practitioners especially are not very good at diagnosing skin lesions (2, 3). Dermatologist accuracy also varies, but most studies report sensitivity (percent of melanoma classified as melanoma) of around 80% (for example, 3).

The idea of using a computer to assist in diagnosis of skin lesions was first proposed in the literature around 1985. Images of skin lesions are analysed by the computer to capture certain features thought to be characteristic of malignancy. These features (expressed as numeric values) are then used to classify the image and report a diagnosis.

So how close are we to a useful automated diagnostic system for melanoma, given that work has been going for over a decade? The developments in computer-based diagnostic systems for melanoma up to the end of 1994 are reviewed in the next section, followed by the research from the later half of the decade. The research is summarised in Tables 1 and 2, for the pre-1995 and post-1994 periods, respectively. Brief summaries of each of the main components (segmentation, image analysis and classification) are presented. Also presented are the sensitivity (percentage of melanoma classified as melanoma) and specificity (percentage of benign lesions classified as benign lesions) reported. The number of images is also reported in the form xxx/yyyy, where xxx is the total number of lesions, and yyyy is the number of melanomas. Finally, a brief comment on the source of the images is presented. The data is intended only as a guide, and other authors planning to use this data should refer to the original papers.

Pre-1995 period
In this period, the main method of clinical diagnosis was examination with the naked eye by an experienced clinician. Most computer-based methods therefore, also used clinical view (naked-eye images). Hall et al. (4) present an excellent review of the different techniques involved in this research. Other reviews that deal with this period include Stoecker & Moss (5) and Stoecker et al. (6).

From Table 1, we can see a brief summary of the major papers presented. Columns 3 to 5 indicate that a variety of differing techniques were used for seg-
mentation, image analysis and classification. Segmentation techniques (Table 1, column 3) appear to have been developed for the image set the researcher was working on, with no evidence of a standard approach.

This circumstance is certainly understandable, and it highlights one of the difficulties with this research: the lack of a standard set of test images. This lack will be discussed later. It is also relatively rare for the seg-

<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Segmentation technique</th>
<th>Classifier</th>
<th>Number of algorithms</th>
<th>Sens.</th>
<th>Spec.</th>
<th>Number of images (total/melanoma)</th>
<th>Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dhawan (21)</td>
<td>1988</td>
<td>Automatic: partially described</td>
<td>Expert system</td>
<td>18: partially described</td>
<td>Work in progress</td>
<td>Work in progress</td>
<td>Not available</td>
<td>'Nevoscope' images</td>
</tr>
<tr>
<td>Green et al. (22)</td>
<td>1991</td>
<td>Automatic: partial description analysis</td>
<td>Discriminant analysis</td>
<td>11: described</td>
<td>80%</td>
<td>72%</td>
<td>70/5</td>
<td>Digitised video images</td>
</tr>
<tr>
<td>Cascinelli et al. (23)</td>
<td>1992</td>
<td>Automatic: not described</td>
<td>Simple thresholding and feature summation</td>
<td>8: partially described</td>
<td>83%</td>
<td>60%</td>
<td>169/45 train, 44/10 test</td>
<td>TV camera mag: ×3</td>
</tr>
<tr>
<td>Schindewolf et al. (24)</td>
<td>1992</td>
<td>Automatic: described</td>
<td>Decision trees (CART)</td>
<td>Unknown. Described in general Shape only. Described in (26)</td>
<td>94%</td>
<td>88%</td>
<td>353/215</td>
<td>Digitised slides: mag ×0.5</td>
</tr>
<tr>
<td>Bostock et al. (25)</td>
<td>1993</td>
<td>See (26)</td>
<td>Artificial neural network (ANN)</td>
<td>92%</td>
<td>68%</td>
<td>124/68</td>
<td>See (26)</td>
<td></td>
</tr>
<tr>
<td>Ercal et al. (19)</td>
<td>1994B</td>
<td>Hand drawn</td>
<td>ANN’s and Fuzzy inference</td>
<td>17: described</td>
<td>96%</td>
<td>62%</td>
<td>399/135</td>
<td>Digitised slides</td>
</tr>
<tr>
<td>Andreassi et al. (28)</td>
<td>1995</td>
<td>Automatic: not described</td>
<td>Not applicable</td>
<td>8: described</td>
<td>Unknown</td>
<td>Unknown</td>
<td>430/50</td>
<td>Digitised video images</td>
</tr>
<tr>
<td>Hintz-Madsen et al. (29)</td>
<td>1996</td>
<td>Unknown</td>
<td>Artificial neural network</td>
<td>21: colour and texture not described Unknown: not described, commercial</td>
<td>59% of the test set was classified correctly</td>
<td>180/60 train, 60 test (unknown distribution)</td>
<td>170/75</td>
<td>Digitised photographs Digitised dermaphot</td>
</tr>
<tr>
<td>Menzies et al. (17)</td>
<td>1997</td>
<td>Automatic: not described</td>
<td>Statistical: logistic regression Statistical: linear classifier</td>
<td>93%</td>
<td>67%</td>
<td>104/30</td>
<td>Digitised ELM slides – Only MM and AMN</td>
<td></td>
</tr>
<tr>
<td>Gutkowicz-Krusin et al. (18)</td>
<td>1997</td>
<td>Automatic: described</td>
<td>Statistical: discriminant analysis Decision tree (RPART)</td>
<td>93%</td>
<td>95%</td>
<td>917/65 used in study 90/31 compared to human diagnosis</td>
<td>Video microscope. Mag: ×10–50 ELM video camera</td>
<td></td>
</tr>
<tr>
<td>Seidenari et al. (30)</td>
<td>1998</td>
<td>Automatic: not described</td>
<td>Statistical: discriminant analysis</td>
<td>21: described</td>
<td>89–100%†</td>
<td>80–84%</td>
<td>221/45</td>
<td></td>
</tr>
<tr>
<td>Bischof et al. (16)</td>
<td>1998</td>
<td>Automatic: partially described</td>
<td>Decision tree (RPART)</td>
<td>Unknown: not described,commercial</td>
<td>89–100%†</td>
<td>80–84%</td>
<td>221/45</td>
<td></td>
</tr>
</tbody>
</table>

* Experiments 1b and 2b did not use dysplastic naevi. Experiment 2 used fewer attributes to allow for different film types.

** Of the three experiments, only Experiment 3 used a separate training set.

† Resubstitution gave 100%, whilst cross-validation gave 89%. They state the likely value is somewhere between the two.
mentation technique to be fully detailed, making reproduction of the research difficult.

The image analysis algorithms (Table 1, column 5) used in these papers tend to be based on easily reproducible human guidelines for use at the clinical view. A popular example of such a guideline is the ABCD Asymmetry, Border irregularity, Colour variegation, Diameter) criteria proposed by Friedman et al. (7). In most cases, these algorithms are described well, although several papers present incomplete descriptions, making it impossible to replicate the research.

Several of these clinical view-based systems produced excellent results on fairly large datasets (Table 1, columns 6–8). To our knowledge however, there are no examples of clinical-view based diagnostic systems in clinical use today. The trend of using clinical view images continued until around the end of 1994.

1995 to now
In this period, the major new development in automated diagnosis of melanoma involved the use of epiluminescent microscopy. Epiluminescent microscopy, or ELM (also termed dermatoscopy, skin surface microscopy, incident light microscopy, Auflichtmikroskopie, and dermoscopy) started to gain in popularity as a technique for diagnosing melanoma in the early 1990s. The history of skin microscopy is long, as described by Stolz et al. (8), but has only recently become a primary tool for diagnosis. A large amount of literature details the method, for example (8–11).

It has been reported that ELM allows trained specialists to achieve a higher diagnostic accuracy rate than is possible by simply using the clinical view (11–14).

Prior to 1995, as reported above, almost all work in automated systems for melanoma diagnosis used clinical-view images. From 1995 on, however, reports about ELM research in this area became more common, and research into clinical-view automated systems becomes very apparent (See Table 2). A brief review dealing with some of the related research is presented by Kopf et al. (15).

Table 2 shows the techniques used in these systems. Segmentation of images is only reported in one of the papers. Image analysis algorithms are fully described in two papers. Commercial confidentiality was the reason for the lack of detail in the other two papers [Bischof et al. (16) and Menzies et al. (17)], whilst the partial description from Gutkowicz et al. (18) was not considered adequate to enable the replication of all of their algorithms.

The ELM-based papers all report excellent results. All surpass that reported for dermatologists in previous surveys (for example, 3). However, again, none of these systems are used in a clinical situation. There does exist a system called “MoleMax II” promoted as a clinical tool by DermaInstruments, Austria. This system, although primarily a method of image capture, reference and storage, also includes an “Expertiser” function that reportedly attempts to classify lesions. The performance of this function is not known.

Summary
The above review, and Tables 1 and 2, illustrate quite clearly the progression of ideas in this field. Prior to 1995, most of the research was based on clinical-view images. The images were analysed with algorithms derived from criteria provided by experts. The most popular example (mainly because of the ease with which it is reproduced algorithmically, rather than its clinical usefulness) is the well-known ABCD criteria proposed by Friedman et al. (7). In these systems, very good results were reported, for example (19, 20).

After 1994, however, with the expanding popularity of the dermatoscope, epiluminescent microscopy images were favored as input. ELM reportedly allows better differentiation between marginal lesions, as well as indicating malignancy earlier than observation of clinical view images. However, Binder et al. (14) report that this technique may actually decrease diagnostic accuracy in the hands of untrained clinicians. It is also unclear whether or not the algorithmic techniques used to analyse ELM images allow better differentiation between lesions, although the results presented in Table 2 suggest that this might be the case.

Discussion
What may not be completely apparent from Tables 1 and 2 is that segmentation techniques, image analysis algorithms and even classification techniques tend to be different from paper to paper. This lack of coordinated efforts and standard techniques exhibited throughout the reviewed papers may be indicative of what Hall et al. (4) term “directing computer technology at this problem in an ad hoc fashion...”. One of the circumstances contributing to this lack of coordinated effort is the scarcity of good quality clinical-view image sets. Without such an image set, standard techniques will not be agreed upon, as different techniques may be required for different image sets. Provision for such an image set should be one of the primary goals of this field. An image set that provides both standard clinical view- and ELM-images may be the most flexible for researchers in this field.

What is apparent from the tables is a lack of rigor-
ous reporting in this field. In many cases, the benchmark of all research reporting, that the research is replicable, is not met. In some cases, this failure is because of commercial interests. In others, the reason is not clear. It is very apparent that this lack of rigor can frustrate the efforts of researchers, and can lead to “reinventing the wheel”, an intensely undesirable situation.

**Future**

The future for automated diagnosis of skin lesions is far from clear. It is likely that ELM-based systems will continue to dominate research efforts as an understanding of the technique continues to be refined. Algorithms will evolve, and may produce better results over a larger range of images. What is required is an easily available test set of images, containing both clinical and ELM-views, which will enable different techniques from different researchers to be tested and compared. Finally, however, it may be the case that the large variability exhibited by skin lesions will hinder the development of a clinically useful system for some time to come.

**References**

28. Andreassi L, Perrotti R, Burroni M, Dell’eva G, Biagioli M.


Address:
Greg Day
c/o CSMTER
University of Waikato
Private Bag 3105
Hamilton
New Zealand

Tel: 64 7 856 2889 extn 8151
Fax: 64 7 838 4272
e-mail: grd@waikato.ac.nz