Clinical-view versus ELM: An investigation into image types in the context of skin lesion screening

A thesis

submitted in partial fulfilment

of the requirements for the degree

of

Doctor of Philosophy

at the

University of Waikato

by

Greg R. Day

Department of Computer Science

The University of Waikato

Te Whare Wānanga o Waikato

Hamilton, New Zealand

January, 2000

Abstract

Melanoma, the most serious form of skin cancer, is increasing in incidence in countries with predominantly white skinned populations. Automated tools have been proposed to help detect this most visible of cancers. Current automated systems for detecting melanoma analyse images of skin lesions for relevant image features, and classify the images based on those features. There are two types of image available to be used in such systems, Clinical-view or Epiluminescent microscopy (ELM) images. ELM images reportedly allow more accurate assessment of skin lesions in the clinical setting, but this finding has not been proven in the context of an automated system.

This research has evaluated the question of Clinical-view versus ELM images in an automated screening system. Two methods of implementing a screening system were considered in this research. Firstly, the 'diagnosis system', which is based on previous work in this field, and secondly, a 'dermatologist assessment system', which is an original method of implementing an automated screening system. The Clinical-view versus ELM question was considered for both of these systems.

Specifically, two automated systems were developed. The first analysed Clinical-view images, while the second processed ELM images. From the analysis, each system attempted to classify lesion images into two groups. For the diagnosis problem, the lesion was either 'melanoma' or 'benign'. For the 'dermatologist assessment' problem, the groups were 'excised' or 'not excised'. The results raise doubts over the current emphasis on ELM images in the automated diagnosis case. Similarly, it appears that Clinical-view images are of more use for reproducing 'dermatologists assessment'. We have also shown that the 'dermatologist assessment' approach to screening skin lesions is a viable and potentially useful alternative to the current emphasis on the diagnosis approach.

Acknowledgements

So many people to thank, so little space! Where to start?

Dr. Amanda Oakley, my second supervisor who introduced me to the realms of dermatology. A dedicated professional and extremely busy person, without whom this project would never have started. A very great deal of responsibility for the final project must be credited to Amanda. Thank you.

The staff of Waikato Hospital Dermatology Department, in particular, Dr. Ian Coutts and Dr. Anthony Yung, who were nice to an error-prone photographer. Thank you for all of your assistance and patience. Dr Marius Rademaker and Dr Mark Duffill, for giving up a lot of valuable time to assess lesions, and never once refusing to help a struggling student. Dermatology is in safe hands in New Zealand. Also to all of the other helpful people at the hospital, including the staff of Visual Communications, especially Kathy who was very patient.

Dr. Scott Menzies of the Sydney Melanoma Unit, for providing the Sydney Image Set used in this research, and for going out of his way to assist.

To all of CSMTER for an entertaining place to work. I would especially like to thank the Director, Dr. Alister Jones, who never complained (too loudly!) about my internet bills. Raewyn, of course. Also the rest of the staff at CSMTER, for putting up with a computer geek out of his depth in the realms of education.

Professor Mark Apperley for his assistance with the final thesis. Andrew Malcolm for his diversions, and programming help early on in this project. Dr. Len Trigg for his help with WEKA, and for writing an excellent thesis. Eibe Frank, WEKA guru who took over answering my emails once Len left. Suz Killinger, who obtained numerous papers for me once I got to Sydney, and is a pretty good flatmate to boot.

To my readers, firstly, Dave Dravitzki, who read (almost) every last word, and for the games of tennis and chess. I am still sure all in all you still owe me, but who's keeping track? And Kerry Earl (bsw.), who read large portions of this thesis, and distracted me with many "wondrous strange" conversations. Maybe one day we will have both our lives on track...maybe not! You have both improved this thesis far

beyond my meagre efforts could ever have managed.

To Angie, who taught me at least as much as the Ph.D. I'm sorry I was such a slow learner! I hope... well, you know.

Finally, Dr. Bob Barbour. Chief supervisor, and part-time slave driver, who directed this project from start to finish. Although it is social death to be friends with a crusty academic, I hope this is what we are. I think we have both learned from this experience. Almost certainly, he has seen how an open-door policy can be abused! Strangely however, the door never closed. Some people are slow learners I guess....

I hope the future will bring you all the very best. And a minimum of annoyances like me!

Thank you all.

Greg

Contents

List of Figures

List of Tables

Glossary

Atypical naevus A naevus showing some form of clinical atypia. Also referred to as dysplastic naevus. See also Dysplastic Naevus.

Basal layer A layer of cells between the epidermis and the dermis. This layer is 3-5 cells thick and is the deepest sublayer of the epidermis. Cell mitosis, or cell reproduction, occurs mostly here.

Benign The opposite of malignant. A benign lesion has no possibility of causing death (at that point in time). See also Malignant.

Carcinoma A malignant tumour. In the context of skin cancer, two types of carcinoma exist, basal cell and squamous cell carcinomas.

Clinical-view The view (of a lesion) that the clinician sees. This term means what may be viewed without external apparatus. See also Epiluminescent Microscopy.

DPI Abbreviation for Dots Per Inch. See Dots per Inch

Dependent variable A dependent variable is explained or affected by an independent variable. See also Independent variable.

Dermatologist A medical practitioner specialising in skin disease. The dermatologist is normally the medical practitioner who excises suspicious skin lesions. The tissue is then sent to the pathologist for histopathological examination and diagnosis. See also Pathologist.

Dermis The inner layer of the skin. The dermis contains sweat glands, sensory receptors, lymph glands, blood vessels, nerves, and hair follicles.

Diagnosis system An automated system intended to detect melanoma that analyses either Clinical-view or ELM images of skin lesions. This system attempts to reproduce the results of pathologists, and has several methodological shortcomings identified in this work. See also Screening system.

Dots per Inch (DPI) A measurement of resolution. The higher the DPI, the higher the resolution. See also Resolution.

Dysplastic naevus Friedman et al. (1991) states that dysplastic naevus "have one or more of the clinical features of malignant melanoma-i.e., asymmetry, border irregularity, colour variegation, and a diameter greater than 6mm". In general, these are benign moles that are 'unusual'. Many different clinical and histological definitions of the term 'dysplastic naevus' have been proposed, and the term 'atypical naevus' is generally preferred.

Epidermis The outer layer of the skin. The epidermis is arranged in layers of cells, including the basal layer, the prickle cell layer and the stratum corneum. The major cell type in the epidermis is the keratinocyte, which is constantly produced in the basal layer. These cells ascend towards the surface and end up at the stratum corneum, which is a layer of dead keratinocytes. The entire epidermis renews itself every 52-75 days. See also Dermis, Skin.

Epiluminescent Microscopy (ELM) Epiluminescent microscopy refers to the use of a low powered microscope to examine skin lesions. The lesion is covered in oil, and the microscope is placed directly against the lesion. This process removes the scattering of light due to the stratum corneum, and allows light to pass through the epidermis. Examination of the structures in the epidermis and dermoepidermal junction is thus permitted.

False negatives Melanomas incorrectly identified as benign lesions. Some melanomas are inevitably missed in the screening process. See also True negatives, False positives, False negatives.

False positives Benign lesions incorrectly recognised as melanoma. See also True positives, True negatives, False negatives.

Histogram See Image histogram.

Histology Microscopic study of tissue. A pathologist examining an excised lesion through a microscope makes a histological study of the tissue, resulting in a histopathological report. See also Pathologist, Histopathology

Histopathology Microscopic study of changes in tissue through disease. For a malignant melanoma, the pathologist looks for changes in cell structure and skin structure caused by the proliferation of malignant cells.

Image histogram An image histogram is a bar-graph representing colour distribution in an image. The x-axis has values for each different colour in the image. The y-axis indicates pixel counts, or how many pixels in the image are a particular colour. Also see Segmentation, Thresholding.

Independent variable A variable which explains or effects a dependent variable. For example, asymmetry (independent) may effect the probability of malignancy (dependent). See also Dependent variable.

In-situ 'in place'. A melanoma in-situ is a melanoma that has not begun to grow vertically into the skin.

Lesion(Skin) An abnormal skin feature. Moles, freckles, melanomas and carcinomas are all examples of skin lesions.

Logistic regression A method of regression that is used when the dependent variable is dichotomous (only has two possible values).

Malignant In reference to a skin lesion. A malignant skin lesion is one that has the ability to cause morbidity or mortality in the patient.

Mask A binary image that 'masks' out irrelevant data. For example, a mask image of a skin lesion image would have '1' where the lesion is, and '0' where the skin is. In this way, it is simple to restrict processing to the lesion area.

Melanin Produced by melanocytes. Melanin gives the skin its colour, and is generally thought to be a protection against ultra-violet radiation. The complete function of melanin is not known however.

Melanocyte Cells contained on the basal layer of the skin. These cells produce melanin. With melanoma, it is these cells that become malignant. See also, Melanin, Melanoma.

Melanoma (malignant) The most dangerous cancer of the skin. Sometimes, malignant is placed in front of melanoma to avoid possible confusion about the malignancy of the disease that may be imparted by the 'oma' suffix.

Melanocytic Associated with melanocytes. For example, a melanocytic lesion is a lesion that derives its pigment from melanocytes.

Metastasis The process whereby cancer cells break off the tumour, and are spread through the body, via the lymphatic system. A tumour that metastasises is certainly malignant.

Nodular melanoma A type of melanoma that does not have an appreciable radial growth phase. See Chapter 1, section 1.3 for details.

Ordinal scale A level of measurement that allows cases to be ordered by degree with respect to a variable.

Pathologist A medical practitioner who specialises in the examination of removed tissue for the purpose of assessing disease type, spread etc.

Radial growth (phase) Radial growth refers to the spreading out of a lesion over the surface of the skin (horizontally) without the lesion becoming significantly deeper. The radial growth phase refers to the period where the lesion undergoes radial growth, before the more life threatening vertical growth phase occurs. See also Vertical growth

Resolution The number of individual picture elements that make up the image. The more picture elements, the more detailed the image and the larger the image size. Resolution is generally expressed in dots per inch (DPI). See also Dots Per Inch.

Screening The process of testing for the presence of disease. In the melanoma case, current screening methods use expert clinicians to observe lesions. All screening programs produce false-negative results.

Screening system A system that can be used in the process of screening lesions. In this research, two different approaches to implementing a screening system are looked at, firstly the 'diagnosis' system which has been the subject of much research, and secondly, the 'dermatologists assessment' system, which attempts to replicate the perception of dermatologists.

Segmentation The process of separating an image into two or more distinct areas. In this project, segmentation separates an image into lesion and non-lesion areas. See also Image Histogram, Thresholding.

Sensitivity refers to the proportion of all cases of histologically confirmed melanomas that were clinically identified as melanoma. In the context of automated systems, the definition of sensitivity changes. For the diagnosis problem, sensitivity is the proportion of melanoma that the system classified as melanoma. For the 'dermatologists assessment' problem, sensitivity refers to the proportion of 'excised' lesions that the system classified as 'excised'. See also Specificity.

Specificity refers to the proportion of all cases histologically proved to be benign that were clinically diagnosed as benign. In the context of automated systems, the definition of specificity changes. For the diagnosis problem, specificity is the proportion of benign lesions that the system classified as benign. For the 'dermatologists assessment' problem, specificity refers to the proportion of 'not-excised' lesions that the system classified as 'not-excised'. See also Sensitivity.

Skin The skin is the largest organ in the body. It has numerous important functions, including regulation of body temperature, protecting against injury and infection, storing water, fat, and vitamins, and of course a sensory device. The skin is comprised of a number of distinct layers, including the stratum

corneum, the epidermis, the basal layer, and the dermis. See also Dermis, Epidermis

Stratum corneum The top layer of dead, scaly skin cells.

Superficial Spreading Melanoma The most common form of malignant melanoma. See Chapter 1, section 1.3 for details.

Surface microscopy See Epiluminescent Microscopy.

Thresholding A method of segmentation. Thresholding examines the histogram of an image and attempts to find one or more points which identify area boundaries in the image. In this project, only one thresholding point is found. This point is then used to segment the image into skin (below the point) and lesion (above the point).

True negatives Benign lesions correctly identified as benign

True positives Melanoma that are identified as melanoma. See also False positives, True negatives, False negatives.

Tumour Literally, swelling. May be ordered (benign) or disordered (malignant), primary or secondary.

Vertical growth (phase) Vertical growth refers to the growth of the lesion both into and out of the skin. As the lesion progresses, it starts to grow both downward into the dermis and vertically out of the skin, so the lesion becomes notably raised. Sometimes referred to as the nodular growth phase, hence nodular melanoma, which only have a vertical growth phase (and not a radial growth phase).

Figure 1: Nodular melanoma with metastasis. J. L. Alibert, Nosologie Naturell Paris 1817 (reproduced from Altmeyer et al. 1997).

Chapter 1
Introduction

Cancer is a word which may bring to mind the worst connotation of disease, that of being incurable. It is a problem affecting the health of humans around the world, and affects a wide range of the population. With the exception of a few obvious risk factors for certain types of cancer (such as smoking and lung cancer), identifying specific at-risk groups is generally difficult. Considerable research has gone into a cure for cancer, but despite innovations such as chemo- and radio- therapies, this goal remains.

Much of the difficulty in finding a cure stems from the nature of cancerous cells. Cancer cells are not foreign to the body, but rather body cells that have 'gone wrong'. Most normal cells do not reproduce constantly (there are exceptions), and only begin to reproduce when prompted. An example of this is wound healing, where chemicals are released from the damaged cells, prompting neighbouring cells to multiply. The reproducing behaviour of cells is typically very regular and orderly.

In some cases however, this order breaks down. The cell may begin reproducing and not stop. As a cell reproduces by splitting in half, all of the sibling cells have the same characteristics as the parent, and therefore, the sibling cells also begin reproducing. This reproduction eventually causes a tumour (from Latin, literally, a swelling). Some tumours are benign, while others are malignant. In general, factors such as structure of the tumour, the rate of tumour growth, invasive growth, and growth by metastasis, define whether or not a tumour is malignant.

Invasive growth refers to growth of the tumour outside of its normal boundaries. For example, skin cells are only found in the skin. The skin is their boundary. Malignant skin cells however, can begin to grow outside the skin boundary, and 'invade' other areas. Similarly, growth by metastasis is only found in malignant tumours. This term refers to the dissemination of the tumour by cells breaking off the primary tumour and being distributed via bodily fluids to other areas of the body. Once this stage has been reached, it is difficult to treat the cancer locally (for example, by excision), and more global methods of treatment must be used.

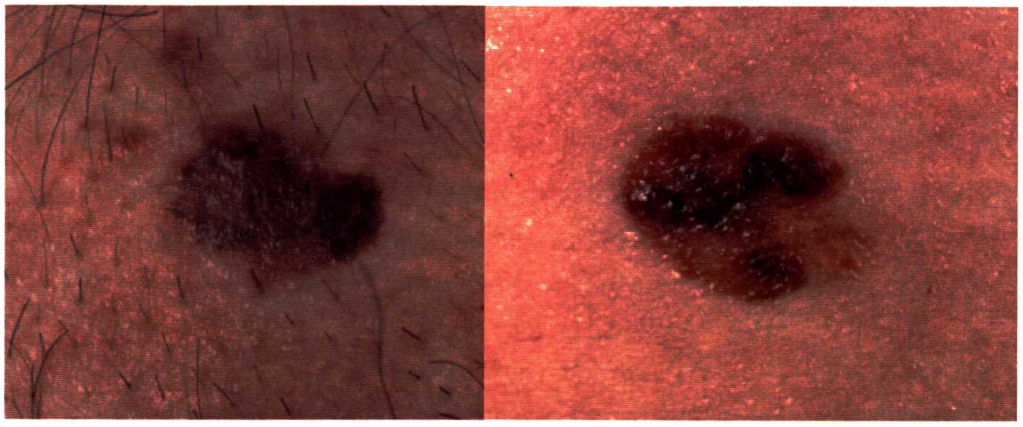

Figure 1.1: Melanomas seen with the naked eye.

This thesis describes research relevant to a particular type of cancer, called melanoma. Melanoma is a skin cancer, that is, it is generally found on the skin. This cancer is a particular problem in New Zealand, with more and more people dying from the disease (Elwood & Glasgow 1993, Skegg 1994). Melanoma is simple to treat if it is detected early. But it appears that for approximately 200 New Zealanders each year, and thousands of other people worldwide, their cancers are not being detected early enough.

1.1 What is Melanoma, and Why is it a Problem?

Melanoma is a cancer of the skin. It occurs when melanocytic cells, the cells that produce the chemical that gives skin its colour, become cancerous. Melanoma is potentially fatal if it is left untreated. However, if melanoma is treated early, there is only a slight possibility of the cancer recurring. The other characteristic that distinguishes melanoma from other cancers is its appearance. Generally, melanoma is observable on the skin surface, in the form of a haphazardly growing mole (Figure 1.1).

> "melanoma writes its message in the skin with its own ink and it is there for all of us to see. Some see but do not comprehend" Dr. Neville Davis (Quoted in Friedman et al. 1985).

Melanoma is increasing rapidly in incidence throughout countries with predominantly white-skinned populations. It kills over two hundred people in New Zealand each year, and more than one thousand Australians (Elwood & Glasgow 1993). The factors contributing to melanoma have not been definitively identified. However, sun exposure and sunburn has been identified by epidemiological evidence as a main factor in the development of melanoma, although the actual relationship is unknown.

In New Zealand, sunburn is a common occurrence. The Cancer Society of New Zealand states:

> "In popular culture, New Zealanders have tended to value the sun-tan as a sign of health, affluence and good times. Both the tough, hardy brown of the outdoors type and the carefully cultivated ...(look) of the 'Baywatch' set are images which effortlessly attract admiration in New Zealand society as they do in other countries. The difference in New Zealand is that geographic and environmental factors make the pursuit of a tan highly dangerous" (Cancer Society of New Zealand Melanoma Awareness Campaign 1988-1995).

Statistics indicate that melanoma incidence in New Zealand is increasing rapidly. In 1992-1994 incidence rose thirty percent, although it is likely that some proportion of this is due to voluntary reporting of cancer before 1994. Melanoma is also reported to be "by far the most common tumour in adults between 20 and 44 years of age in New Zealand, accounting for 30% of all registered cancers in this age group" (Skegg 1994). A similar situation is reported in other countries, most notably, Australia (Thursfield et al. 1995, Giles & Thursfield 1997). It is thus apparent that the problem of melanoma is significant and getting worse.

1.2 How are Melanomas Detected?

So given that a skin lesion may be a melanoma, how can we tell? The typical route to diagnosis is shown in Figure 1.2. The lesion-owner may have noticed something of concern in a lesion, or have it noticed by another person. The next stop would be a general practitioner, who would assess the lesion for malignancy. The general practitioner would then refer suspicious lesions to a dermatologist (or plastic surgeon), who would again assess the lesion for malignancy. Lesions thought to be suspicious would then be excised, and the tissue would be examined by a pathologist, and a diagnosis given.

At this stage, we need to define diagnosis. Diagnosis is the definite assessment of the malignancy of the lesion, and only takes place once the pathologist has assessed the lesion. Dermatologists and general practitioners on the other hand, make an assessment as to what degree the lesion resembles a melanoma. Although dermatologists may suspect that a lesion is a melanoma, they have no way of knowing for certain until the results of the pathologist's analysis are obtained. Even histopathological analysis can be considered another 'assessment of suspiciousness' test, as there is no method of definitely proving malignancy of a lesion through histopathological ex-

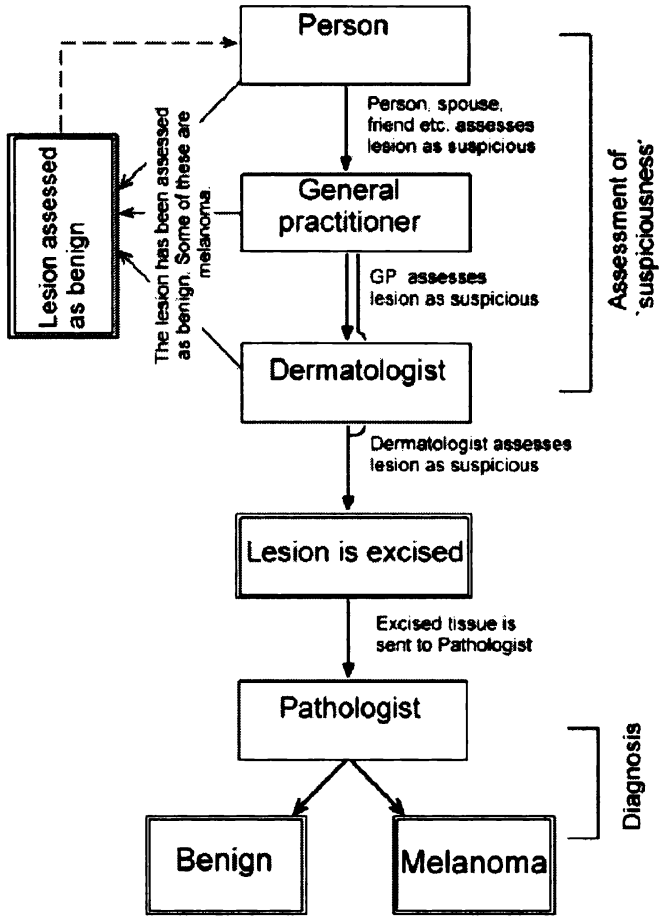

Figure 1.2: The diagnosis cycle of a lesion

amination. Indeed, Swerlick & Chen (1996) hypothesise that the rapidly increasing incidence of melanoma may be a result of biologically benign lesions being classed as melanoma by pathologists, rather than any increase in excised melanoma.

Ignoring this debate, which is beyond the scope of this discussion, Figure 1.2 shows the levels of assessment commonly applied to a particular lesion. At every step, some lesions are assessed as benign, and removed from the cycle (possibly to re-enter at a later date). The remainder are assessed as suspicious and continue to the next level, culminating in histopathological analysis, which provides a diagnosis for the lesion. The point to note is that all previous steps in the cycle are 'assessment of suspiciousness' steps, and diagnosis is left to the pathologist.

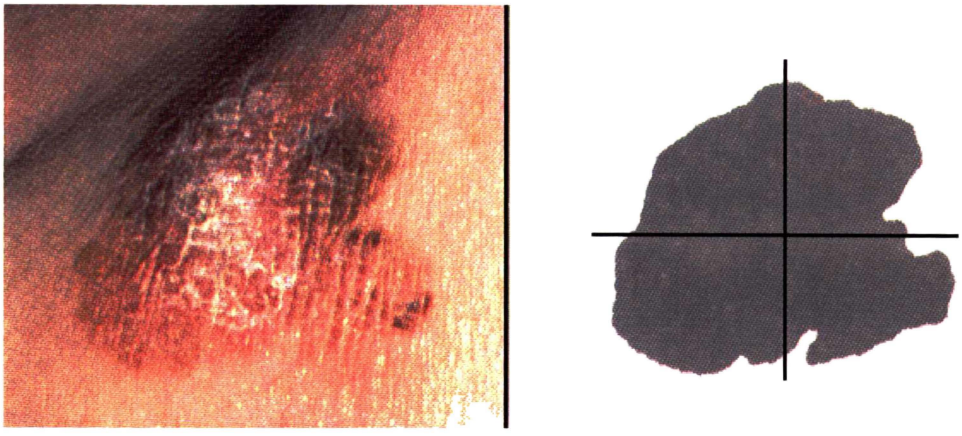

Figure 1.3: Clinical-view Asymmetry

1.2.1 How are Skin Lesions Assessed?

The assessment of lesions by dermatologists is based primarily on visual assessment of lesion attributes. The disease usually gives some visual indication of its presence and progress, and specialists attempt to recognise these visible indicators and make an assessment of the malignancy of the lesion. So what are the indicators?

Several sets of visual indicators have been proposed for use in recognising melanoma. These indicators can be divided into two sets, based on the techniques required to view them. The first set of indicators are intended for lesions viewed with the naked eye. This viewing method is referred to as the Clinical-view. The second set of indicators is for lesions viewed under epiluminescent microscopy (ELM), where the oil-covered lesion is viewed through a hand lens. Indicators for each of these views are described below.

Clinical-view

For Clinical-view assessment, perhaps the most popular set of indicators is the ABCD checklist (proposed by Friedman et al. 1985). This checklist was developed to highlight features of malignancy that are apparent at the Clinical-view, and as such are ideal for use by the public. The four characteristics making up the ABCD checklist are Asymmetry, Border Irregularity, Colour Variegation, and Diameter. Images below (Figures 1.3-1.5) are from the bookmark produced by the Cancer Society of New Zealand.

The first of the checks, Asymmetry, is a subjective judgement of the (lack of) symmetry of a lesion, based on the visible area. It is usually found by splitting the lesion in two using an imaginary mirror line, and comparing the two halves in terms of area (Figure 1.3). If a lesion is asymmetric, it is more likely to be malignant.

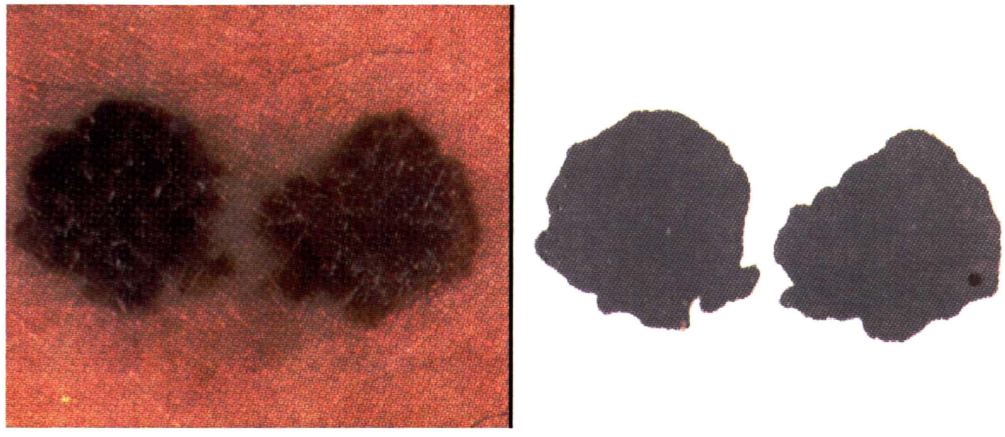

Figure 1.4: Clinical-view Border Irregularity

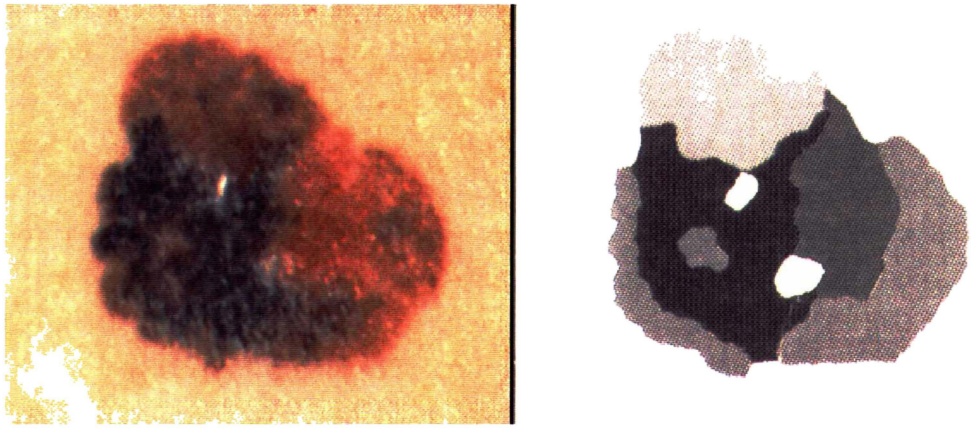

Figure 1.5: Clinical-view Colour Variegation

The second of the checks, Border irregularity, is a measure of the 'roughness' or 'jaggedness' of the border of the lesion image. Benign lesions tend to have comparatively smooth boundaries, while melanomas tend to have more jagged or notched boundaries (Figure 1.4).

Colour variegation refers to the number of different colours seen in the lesion, and may include tan, browns, blacks, reds, white/grey and blue. Three or more colours in a lesion may indicate malignancy. Benign lesions tend to have only one or two colours (Figure 1.5).

Diameter refers to the largest diameter of the lesion. The usual guideline is around 5-6mm. Benign lesions tend to be smaller than 6mm in diameter, although this is far from defining, as many benign lesions will be bigger than this. Melanoma, due to the initial radial growth phase exhibited by several types, are most frequently recognised when they are larger than 6mm in diameter. In many cases the radial growth phase occurs before the life-threatening vertical growth phase, and thus the lesion can be easily removed. This circle ◯ is approximately 6mm in diameter, so in general, melanomas appear quite large.

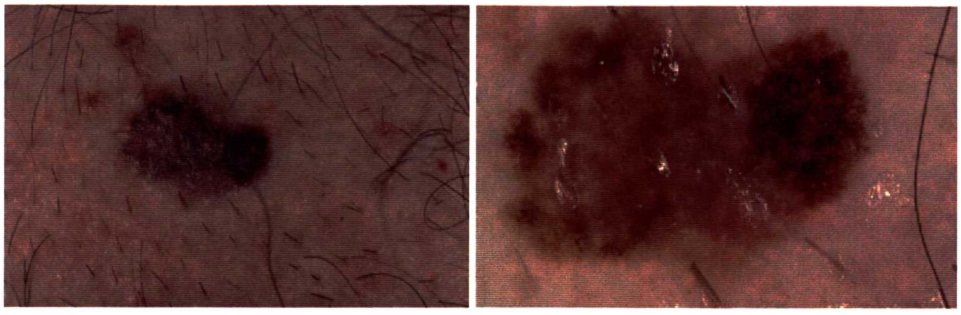

Figure 1.6: A melanoma seen with the naked eye (left) and under epiluminescent microscopy(right).

The ABCD Clinical-view checklist is intended for public use, partly due to the unaided nature of observation. For specialists, other methods utilising visual aids for viewing lesions exist. One of the more popular methods is referred to as epiluminescent microscopy (ELM). ELM refers to the use of an incident light magnification system to examine skin lesions. With ELM, the lesion is first covered in oil, and a glass plate is placed against the lesion. The lesion is then viewed through a low power microscope. This technique removes the normal light reflection of the top layers of skin and allows detailed examination of the morphological structures in the lesion (Figure 1.6). Consideration of these structures (coupled with suitable training) enhances the clinical identification of most lesions.

ELM View

As with the Clinical-view of skin lesions, criteria exist for the detection of malignancy in lesions viewed with ELM. Stolz et al. (1994) and Menzies et al. (1996) amongst others, have proposed lists of indicators for lesions viewed with ELM. We focus on the ABCD method of Stolz et al. (1994), as this method is simpler to illustrate, and both methods share a number of similar features.

Asymmetry for the epiluminescent ABCD criteria is similar to the Clinical-view ABCD criteria. The lesion is assessed on its axis for differences in shape, colour and structure distribution. For example, the lesions in Figure 1.7 are rated a) Symmetric, b) Asymmetric on axis 2 due to colour distribution, and c) Asymmetric on both axis.

The second check in the ELM ABCD checklist, Border Contrast, refers to whether or not the border of the lesion gradually fades into the surrounding skin (a criteria for benign-ness), or whether the border is sharply demarcated (a possible indication of malignancy). The lesions in Figure 1.8 are rated a) 0-no demarcation, b) 4-four segments sharply demarcated, and c) 8-all segments sharply demarcated.

Similarly to the Clinical-view checklist, colouring is important to identifying melanoma

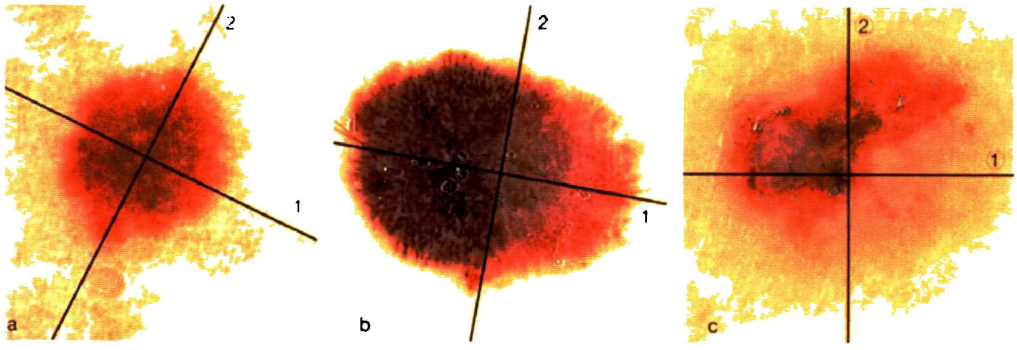

Figure 1.7: ELM Asymmetry

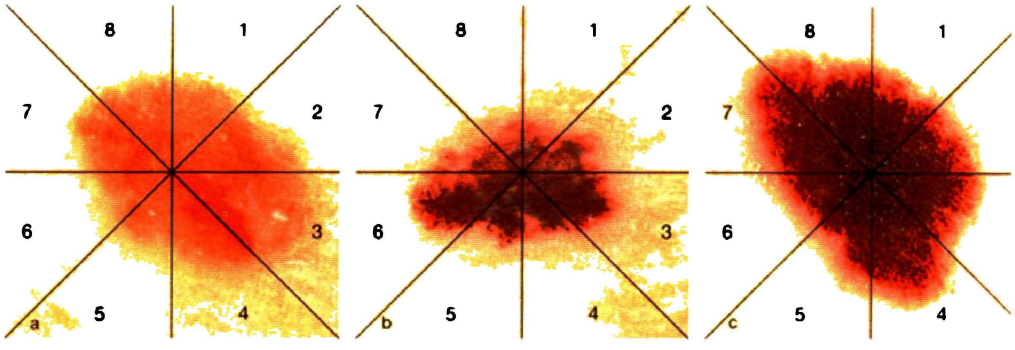

Figure 1.8: ELM Border Contrast

when viewed under ELM. Stolz et al. (1994) define six colours (light brown, dark brown, black, red, white, slate-blue). The more distinct colours in the lesion, the higher the likelihood of malignancy. The lesions of Figure 1.9 are rated a) 2 colours, light and dark brown, b) 4 colours, light and dark brown, slate blue, and black c) all six colours are present here

The final check, Differential Structures, examines the structures of the lesion. Differential structures include dots, globules, structureless areas, network, and branched streaks. These components are apparent in most lesions. However, the more of these structures a lesion has, the higher the likelihood of malignancy. The lesions in Figure 1.10 are rated a) 1 - Only structureless area, b) 3 - structureless areas ($*$), network ($\mapsto$), and branched streaks ($\rightarrow$) c) all five components appear, structureless areas, network ($\mapsto$), branched streaks ($>$), dots ($\rightarrow$) and pigment globules ($\Rightarrow$).

Once these four criteria have been evaluated, the scores for each are combined into a 'total dermatoscopy score'(TDS) using Equation 1.1. If the resulting TDS is above 5.45, the lesion is highly suspicious for melanoma. Stolz et al. have also found melanoma with TDS scores as low as 4.75.

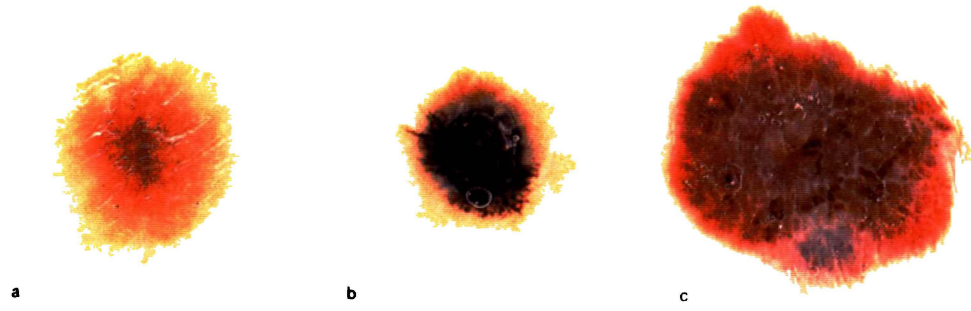

Figure 1.9: ELM Colour Variegation

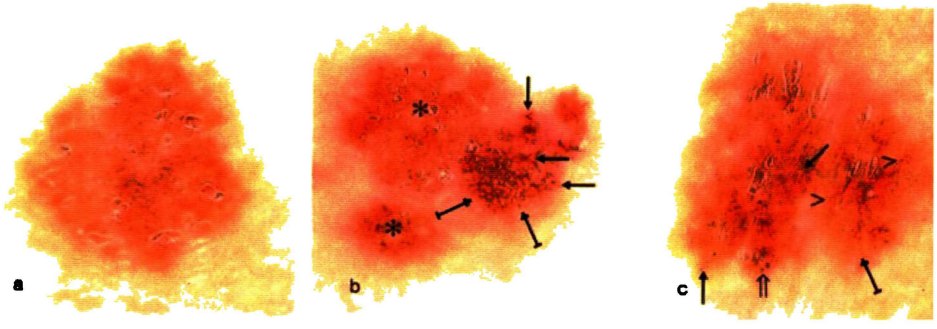

Figure 1.10: ELM Differential Structures

$$A \times 1.3 + B \times 0.1 + C \times 0.5 + D \times 0.5 = TDS \tag{1.1}$$

It should be noted that the ABCD criteria for ELM only applies to lesions that are derived from melanocytes (the cells that create skin pigment). See Section 1.3 for further explanation.

Menzies et al. (1996) presents another, related set of guidelines for differentiating malignant lesions from benign. These guidelines consist of negative features, which melanoma almost certainly will not have, and positive features that may indicate melanoma. This set of criteria is more particular than that of Stolz et al., and includes features such as blue-white veil, radial streaming and pseudopods. Blue-white veil is a feature highly suggestive of melanoma, and is represented by a blue-white coloring over some part of the lesion. Colour and asymmetry also play an important role in this method.

Although recognising the visual aspects of the disease is the main method of identifying potentially malignant lesions used by clinicians, it is difficult to reproduce the precise techniques in a stepwise fashion. Most identification tends to be based on the clinicians' past experience, and the entire process is "more of an art" (Personal Communications: Oakley 1997).

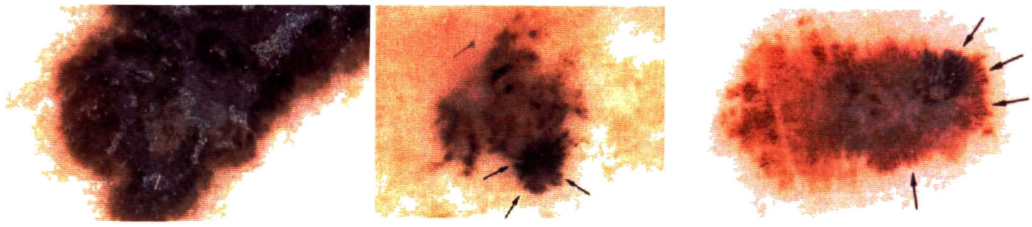

Figure 1.11: Examples from the criteria of Menzies et al. (1996). The first picture shows an example of blue-white veil. The second illustrates pseudopods (indicated by arrows), while the third shows radial streaming (arrows).

1.3 Skin Lesions

This section will describe the Clinical-view and ELM characteristics of the most common lesion types, including melanoma. The data from this section is based on several books (MacKie 1989, Stolz et al. 1994, Habif 1996, Menzies et al. 1996). The images used to illustrate the lesions are from the image set gathered for this project unless otherwise stated.

Lesions can be broken into two groups, depending on whether or not melanocytes are involved in the pigmentation of the lesion. Melanocytes are the cells that produce melanin, and thus give skin its colour. A dark skinned person will have more melanin in their skin than a light skinned person. Moles are clusters of melanocytes (not melanin). Melanoma occurs when these cells become malignant.

For example, the first three lesion types presented below are examples of non-melanocytic lesions. The pigmentation displayed is not due to melanocytes and as such, there is no possibility of these lesions becoming melanoma. However, some of these lesions can cause considerable confusion with melanoma.

The second group of lesions owe their pigmentation to melanocytes. These are the lesions that have some possibility of becoming melanoma. These lesions include junctional and compound naevi (common moles) as well as lentigo and of course the melanomas.

Seborrhoeic Keratoses Seborrhoeic keratoses are very common lesions, especially in older people, but it is unknown why they develop. These lesions display enormous variation in appearance, although one of the most common forms is shown in Figure 1.12. They are not melanocytic lesions, that is, they do not involve melanocytes, although they are often similar in colour to some melanocytic lesions. It is exceedingly rare for seborrhoeic keratoses to become malignant.

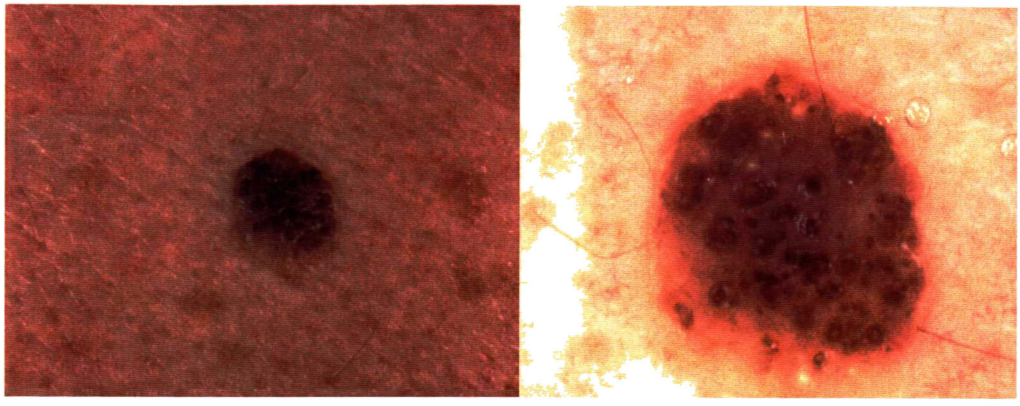

Figure 1.12: Seborrhoeic Keratosis.

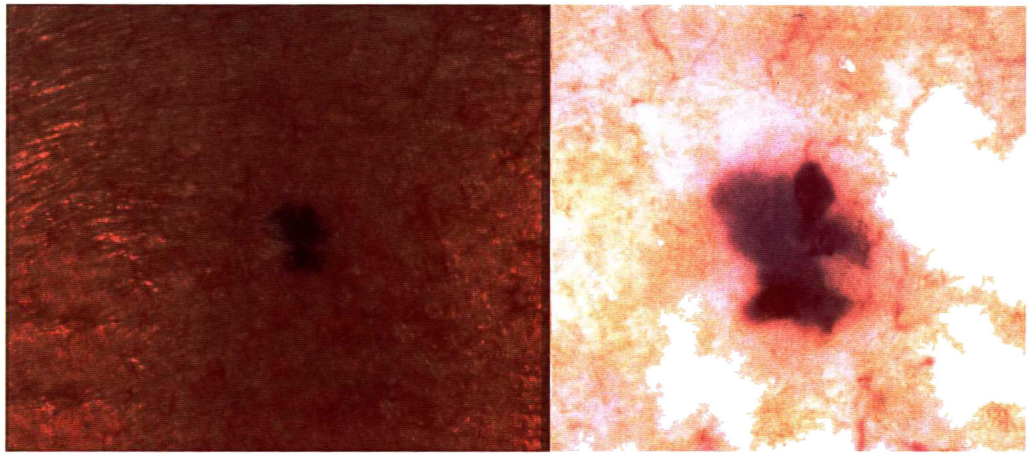

Figure 1.13: Haemangioma.

Cherry Angioma (Figure 1.13). Cherry Angiomas (sometimes referred to as haemangiomas) are small lesions which occur in most people after age 30 (Habif 1996). They consist of dilated skin capillaries and have a distinctive red colour, which may tend to blue with deeper lesions (Fitzpatrick et al. 1993). They have no melanocytic component, and hence no likelihood of malignancy. Cherry Angioma are a subtype of haemangioma.

Basal Cell Carcinoma (Figure 1.14). Basal cell carcinomas (BCCs) are the most common skin cancer. BCCs rarely metastasise and mortality rates for this type of cancer are low. Most BCCs are found on the face. Typically, a BCC is initially seen as shiny or translucent raised nodules, which grow slowly, although more advanced tumours may vary significantly in appearance. The development of this carcinoma is related to cumulative ultra-violet radiation exposure. BCCs often show pigmentation although they are not melanocytic in origin.

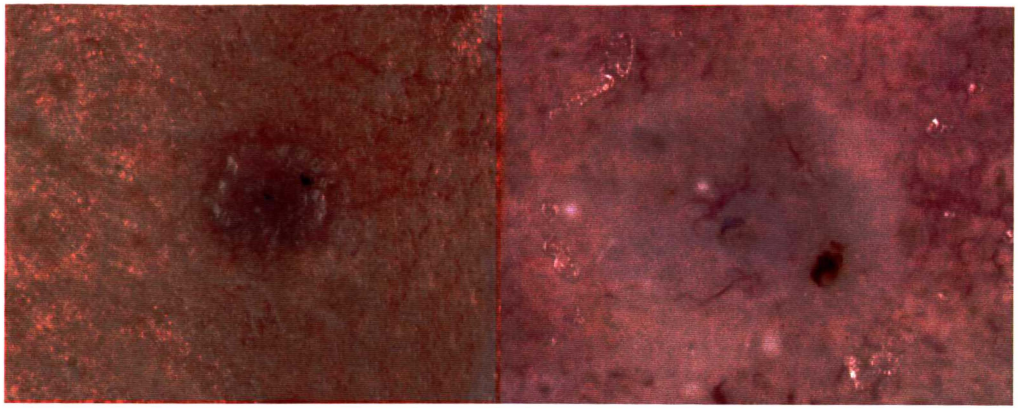

Figure 1.14: Basal Cell Carcinoma.

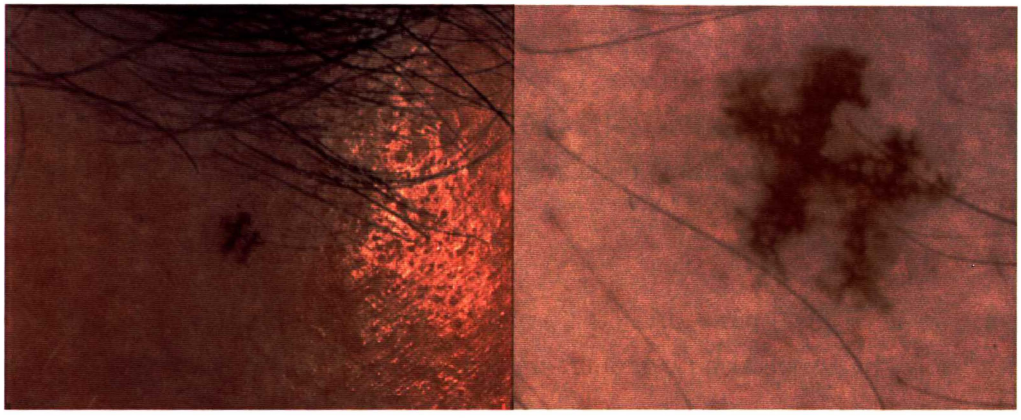

Figure 1.15: Lentigo.

Ephelides Ephelides or freckles are common in most people with slightly sun-damaged skin. They appear as small, lightly pigmented lesions that may occur in large numbers. They are completely benign. The light pigmentation is due to the melanocyte cells producing an increased amount of melanin in response to exposure to UV light, and without this exposure, the lesions fade.

Lentigo (Figure 1.16). Lentigo are similar to ephelides in morphology, but are generally darker and persist in the absence of UV exposure. Lentigo are caused by an increased number of melanocytes at the junction of the dermis and epidermis. These lesions are generally quite small (<2mm) and tend to appear on sun-exposed areas of the skin. In some cases, they can cause confusion with lentigo-maligna melanoma.

Congenital Melanocytic Naevi (Figure 1.16). These lesions are present at birth. They range in size from small (under 1.5 cm) to the giant 'garment' variety, which may cover large portions of the body. They are seen as a possible precursor to melanoma, although they share few of the clinical features of melanoma. The

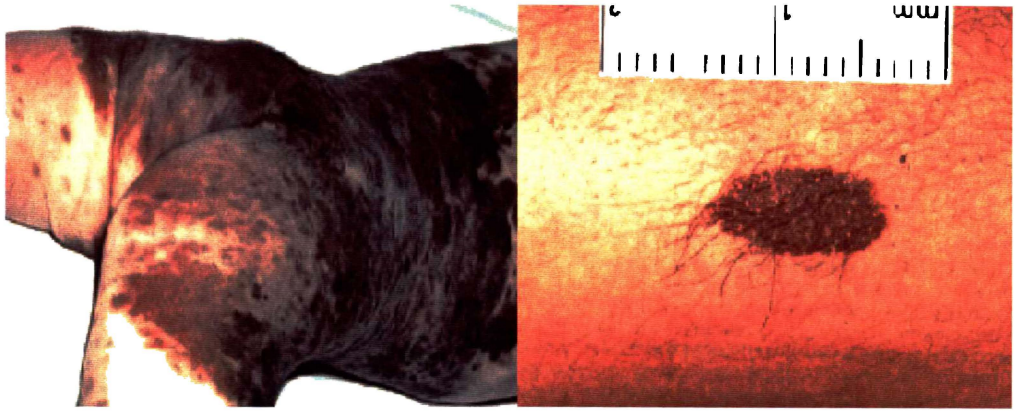

Figure 1.16: Congenital Melanocytic Naevi. The left hand image shows a garment or giant congenital naevus. Both are shown at the Clinical-view . Images ©Habif (1996)

larger lesions represent a significant risk factor for development of melanoma. It has been reported that 8.1% of melanoma developed from congenital naevi (page 556 Habif 1996).

Acquired Melanocytic Naevi (Figure 1.17). Acquired (that is, naevi acquired after birth) melanocytic naevi are benign tumours consisting of melanocyte cells. These lesions can develop throughout at any stage in life, and are commonly referred to as moles. There are three major sub-types that exhibit slightly different characteristics. The first sub-type, the junctional naevus, occurs when melanocyte cells cluster and proliferate along the junction of the epidermis and dermis, causing a flat or slightly elevated, light-brown to black regularly pigmented lesion. These lesions can form into a compound naevus, when the proliferating naevus cells extend into the dermis. These lesions therefore consist of a compound of junctional and dermal melanocytes. They appear as brown or flesh coloured, and elevated. They may be smooth or warty, and are uniformly symmetrical. Compound naevi mature into dermal naevi, where all of the naevus cells are contained in the dermis and the expansion of the lesion stops. These lesions range from black through to flesh coloured, and are commonly dome shaped. Figure 1.20 shows the difference in structure of these three lesions.

The three lesion types may appear quite similar, although junctional naevi tend to be darker than compound naevi, which in turn tend to be darker than dermal naevi. In general, all three are considered benign melanocytic naevi.

Atypical or Dysplastic Naevi (Figures 1.18 and 1.19). Atypical naevi are acquired melanocytic naevi that exhibit some form of irregularity at a clinical or cellular level. For example, the lesion may be larger than normal, and have an irregular

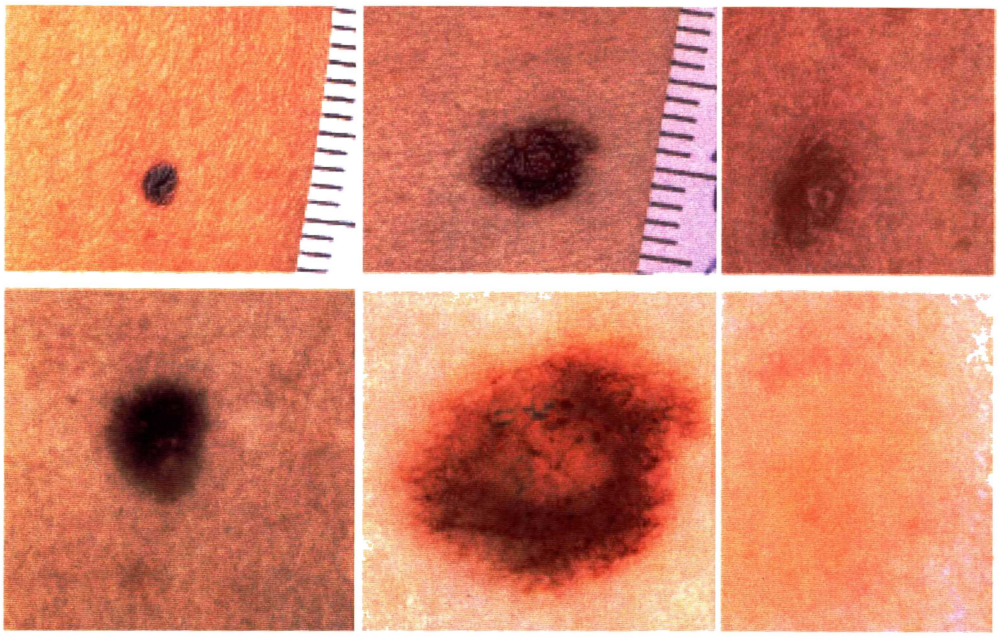

Figure 1.17: Examples of acquired melanocytic naevi. The left hand image shows a junctional naevus, the middle image is a compound naevus and the right hand image is a dermal naevus. Images from Menzies et al. (1996). It should be noted that the appearance of these naevi can differ markedly from those shown here.

border. Or the cells composing the lesion may exhibit some form of abnormality. In many cases, these lesions appear very similar clinically to melanoma, and it has been suggested that these lesions may be precursor lesions to melanoma.

It must be noted that many different meanings have been given the term dysplastic naevi. In current practice, atypical naevi is the preferred term because the definition is more precise. In this research however, a number of images were identified as dysplastic naevi by original contributors, and therefore the terms 'dysplastic' and 'atypical' are used interchangeably.

Melanoma Melanoma is the most aggressive of skin cancers. If a melanoma is left for long enough, (sometimes only a few months, other times many years), it is highly likely to metastasise. There are several clinical sub-types of melanoma: superficial spreading melanoma, nodular melanoma, lentigo maligna and lentigo maligna melanoma, and acral lentiginous melanoma. Figure 1.20 shows an abstract view of melanocytic skin lesions. Notice that the melanocytes of the melanoma are spreading both horizontally (radial growth) and vertically (vertical growth). The melanocytes of the benign lesions are contained.

Superficial spreading melanomas (SSM) are the most commonly reported sub-type of melanoma (Figure 1.21). They are potentially invasive tumours that have a radial growth phase as the initial stage of development, before the tumour begins to

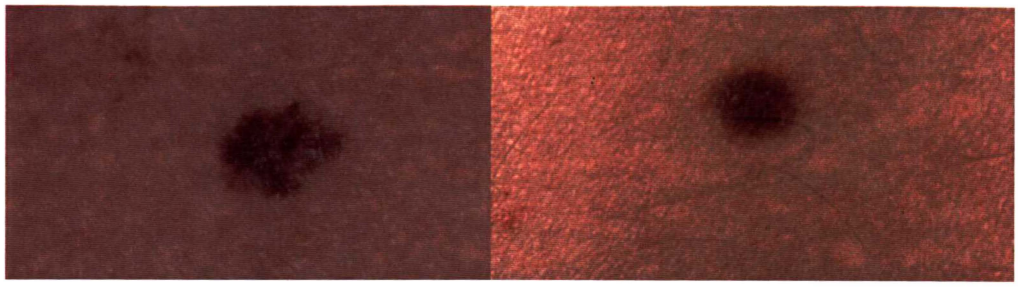

Figure 1.18: Clinical-view Atypical Naevi (left) compared to normal naevus (right). Note the difference in shape and pigment between the two lesions. Images from Dr. Scott Menzies and Health Waikato Ltd. respectively.

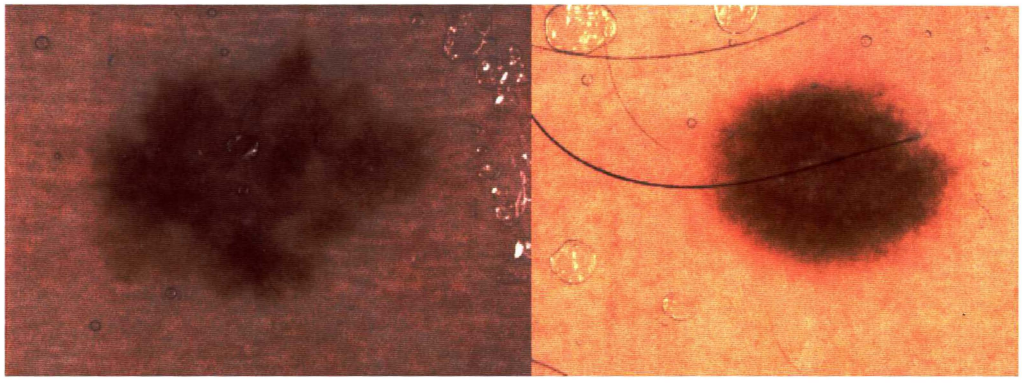

Figure 1.19: ELM Atypical Naevi (left) compared to normal naevus (right). These ELM images are of the Clinical-view lesions presented in Figure 1.18

invade the dermis. This radial growth phase may last anywhere between months to a decade.

Nodular melanomas (Figure 1.22) are also invasive, but unlike SSM, do not have an appreciable radial growth phase. Because of the lack of the radial growth stage, nodular melanomas tend to be the most aggressive of all melanoma. SSM may develop a nodular component, which may make differentiating between these two types difficult. Some authors suggest such differentiation is unnecessary (Menzies et al. 1996).

Another type of melanoma is lentigo maligna and the more advanced lentigo maligna melanoma. Lentigo maligna is an in-situ (confined to the epidermis) melanoma which usually occurs on the head and neck. They are also most common in older people. If lentigo maligna progresses to an invasive stage (i.e. penetrating the dermis), it becomes known as lentigo maligna melanoma. Both of these appear as highly irregular lesions (more so than superficial spreading melanoma), and often have very indistinct boundaries. They often enlarge very slowly.

The final sub-type of melanoma is acral lentiginous melanoma. These melanomas are found on the palms, soles, or under the nail bed. Because of their location, these

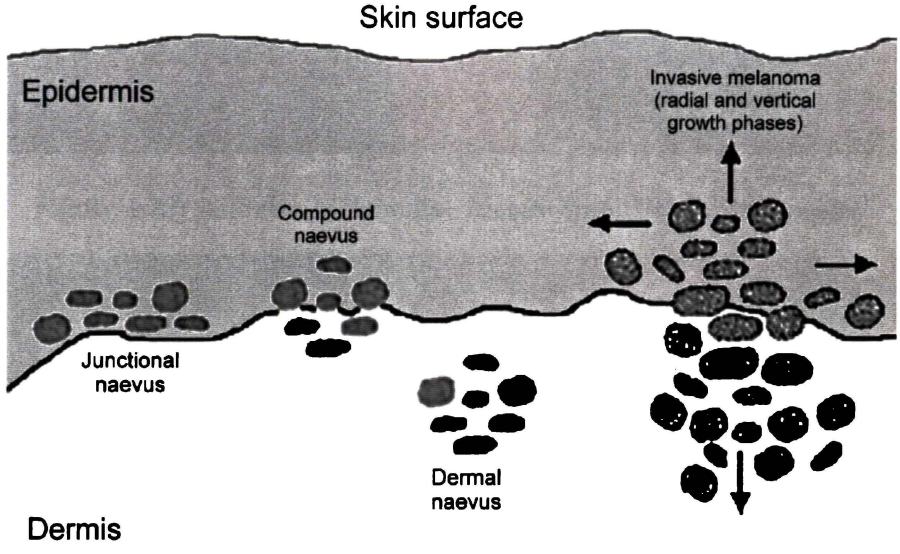

Figure 1.20: The location of melanocytic skin lesions in the skin. Junctional naevi are located on the junction between the dermis and the epidermis, with all cells contained in the epidermis. Compound naevi have cells in both the dermis and epidermis, whilst dermal naevi are entirely contained in the dermis. The cells of these three lesions are all grouped together. The melanoma on the other hand, grows both horizontally along the epidermis, and vertically into (out of) the skin.

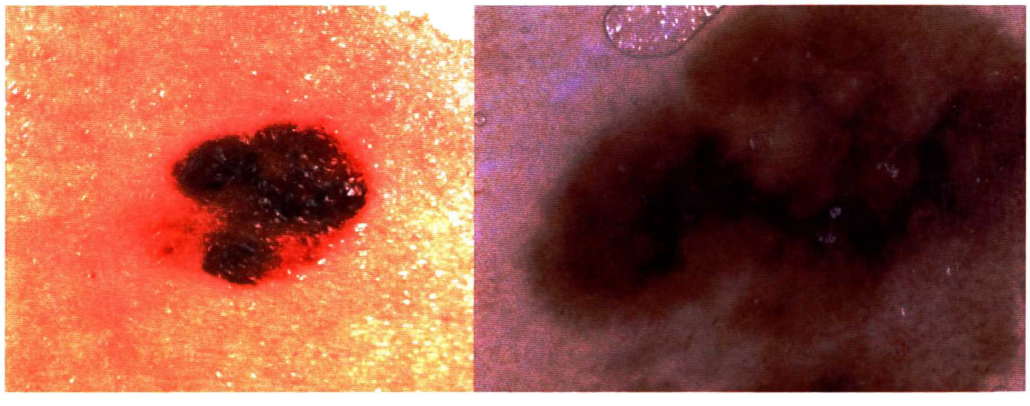

Figure 1.21: Superficial Spreading Melanoma

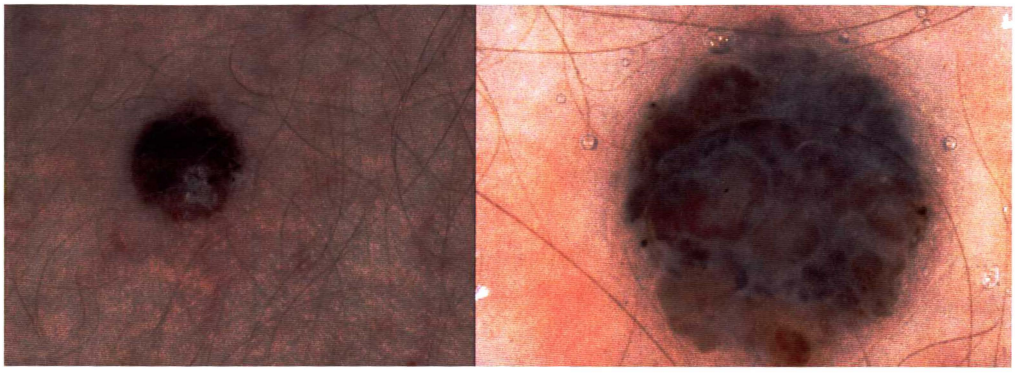

Figure 1.22: Nodular Melanoma. Images from Dr. Scott Menzies.

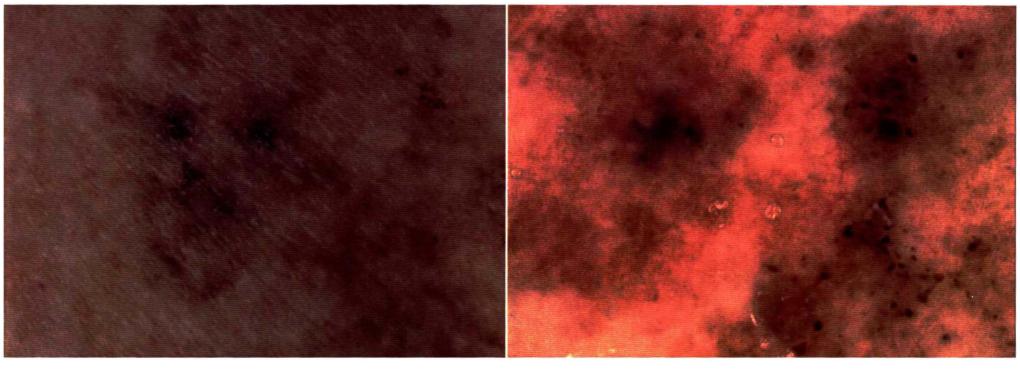

Figure 1.23: Lentigo maligna

lesions may appear quite different to other melanoma.

The above introduction is not exhaustive, but is intended as an introduction to the major types of lesions. It is apparent that the appearance of skin lesions varies enormously, not only between lesion types, but also within a type.

1.4 Outline of the Thesis

This thesis is organised into nine chapters, as indicated in Table 1.1. This first chapter presented an introduction to the context of the research, namely melanoma, and in particular methods of detecting this disease. It also shows examples of some of the most common skin lesions.

The next chapter reviews the literature concerning the context. The problem of melanoma is established from the literature, and methods that may assist in combating this disease are examined. From this review, an area of research is identified, and we develop a thesis for investigation.

Chapter 3 covers literature associated with automated systems for identifying melanoma, which is our area of interest. The fourth chapter details the method used in the

project to assess the validity of the thesis, while Chapter 5 describes the investigations carried out to assess the thesis. The sixth chapter lists the results obtained from the investigations, which leads into Chapter 7, where an analysis of these results is performed.

Chapter 8 presents possible implications from this research on the wider research field, and the research is summarised in the final chapter, together with directions for future work.

1.5 Chapter Summary

This chapter has introduced the problem of melanoma. Melanoma is the most deadly form of skin cancer. It is apparent that in countries with predominantly white skinned populations, melanoma is a serious problem. Melanoma is of particular concern in New Zealand and Australia.

We have looked at the current methods used to detect melanoma. In particular, we have presented techniques used at the Clinical-view, and also using ELM. ELM is currently the "state of the art" for melanoma detection. We then looked briefly at different types of skin lesions. Both a Clinical-view and an ELM view of the lesions are presented, so that the differences between these views may be appreciated. Some of these lesions are straightforward to distinguish from melanoma, while others are much more difficult.

The next chapter reviews the literature associated with this field. From this literature an area of interest is found, and a thesis argument put forward.

Table 1.1: Contents of the thesis at a glance

	Chapter Title	Main Contents
1	Introduction	Introduces the major context of the research - melanoma and skin lesions. Describes how lesions are identified, and introduces the Clinical-view and ELM concepts. Also shows examples of main types of skin lesions.
2	Literature Review	Examines research concerning the problem of melanoma. Establishes the problem, discusses methods of reducing the problem. From this discussion, a gap in knowledge is identified, and the context of the thesis, automated melanoma detection systems for skin lesions, is found. From this context, the thesis argument is presented.
3	Automated Systems Review	Looks at research concerning the three components of automated diagnosis systems, segmentation, feature analysis and classification.
4	Method	Having reviewed techniques used in past research in the previous chapter, this chapter is concerned with presenting the methods used in this research. Emphasis is on feature analysis algorithms and classification.
5	Investigations	Describes the investigations performed in this research. The investigations are: the diagnosis problem, the 'dermatologist assessment' problem, and the human comparison investigation.
6	Results	Presents the results of the investigations.
7	Analysis	Interprets the results of the investigations and discusses the meaning of the results for the thesis. Support for the thesis argument is shown here, and conclusions regarding the thesis argument are drawn.
8	Implications	Presents implications of this research for the wider research field.
9	Main findings, limitations & further work	Summarises the research, and presents the major contributions associated with this work. The further work section discusses immediate extensions to the work that could be performed.

Chapter 2
Literature Review

As outlined in the previous chapter, melanoma incidence is increasing around the world. The first section of this chapter presents more evidence concerning the melanoma problem, and looks at methods proposed to reduce the problem, in particular, early detection. The concept of early detection is examined from the point of view of the lay-person, detailing initiatives, and examining how aware lay-persons are of melanoma. The methods and accuracy of clinicians are then reviewed. Finally in this section, the concept of population screening for melanoma is examined. Such screening is used for breast and cervical cancers, and we look at whether a similar system could be successful for melanoma. We discover that population screening for melanoma is unlikely to be utilised, and surmise that perhaps automating the screening process may make screening more cost-effective. The remainder of this chapter looks at current automated research, and locates an area for investigation. The thesis argument is then proposed, and the thesis content is discussed.

2.1 Melanoma is a Problem

Melanoma is a problem that has received an increasing amount of attention in recent years. Many researchers are investigating different aspects of the disease, from management to trends and solutions. One of the trends that appears the most prominent in the literature is that the incidence of melanoma in countries with mainly white populations is increasing rapidly. In New Zealand, Skegg (1994) reports that one in 31 people born in 1994 will develop melanoma given current rates, and that by the year 2005, one in 14 people are likely to contract the disease. In unpublished research, Rademaker & Zainal (1997) estimate the risk to white New Zealanders' in the Waikato region of New Zealand in 1997 was nearly one in 12. This result is backed up by the latest Waikato statistics (Health-Waikato 1995), which report that 326 new cases of melanoma were recorded in 1995, from a population of European descent of 244,181. These results correspond to one in 10 people in the

Waikato developing melanoma, and confirm the results of Rademaker & Zainal.

In 1993, the New Zealand Cancer Society and the Department of Health Working Group reported on the melanoma situation in New Zealand. This report, Elwood & Glasgow (1993) details the problem of melanoma in New Zealand, and describes proposed plans to control this cancer. The report states that New Zealand has one of the highest incidence rates of melanoma in the world, and that melanoma "is the most common tumour in 20 to 39-year old adults, and the incidence is rapidly increasing". In Australia, the trend is repeated. The Anti-Cancer Council of Victoria reports that the incidence of melanoma is growing "at a faster rate than any other cancer in Australia" (Thursfield et al. 1995). Lifetime risk was assessed at one in 45 and one in 50 for men and women respectively. This report also shows comparative international incidence rates, with Australia and New Zealand regions taking the top four places in incidence. These results may be superceded by more recent data, for example Rademaker & Zainal (1997), but the implications for both New Zealand and Australia are clear. Saxe et al. (1998) report on the South African situation. Although they do not report change in incidence, they conclude that "results... indicate a high incidence rate of melanoma in white South Africans, comparable with that of Australia".

Interestingly, Giles et al. (1996) reports that mortality from the disease in Australia has plateaued. They report that some age-groups had increased mortality rates, whilst other groups fell. Those groups showing increases were generally older people born before 1930. Death rates were stable in those born between 1930 and 1950, while the death rate fell for younger groups. They conclude that the death rate from melanoma is stable, and based on the age group data, can be expected to fall in coming years. Most other research from around the world reports a linear increase, and it may be that more recent data from those countries will show the death rate reaching a similar plateau.

Friedman et al. (1985) and Friedman et al. (1991) reported on the situation in the United States. The first of these papers found that one United States citizen in 150 was likely to develop melanoma in 1985. The second paper reported that this rate had increased to one in 105 by 1991, a more than 40% increase in less than a decade. These papers also introduce the ABCD criteria for evaluating lesions, and are recommended reading as an introduction to the problem of melanoma.

In more recent American literature, Rigel et al. (1996) report a worsening situation. In particular, they state that one in 87 Americans will develop melanoma (compare to Friedman et al. 1985, Friedman et al. 1991). They claim that this result is accurate, and not simply due to changes in surveillance or detection methods. They also report that annual mortality from melanoma continues to show a linear increase,

contrary to the plateau reported by Giles et al. (1996).

Obviously melanoma is a worsening worldwide problem. From Giles et al. (1996), some progress in reducing mortality appears to have been made, at least in Australia. Incidence is still rising however. Melanoma is easy to cure if it is detected early, and so emphasis on treating this disease is placed on early detection. The following section looks at early detection from the point of view of both lay-persons and clinicians.

2.1.1 Early Detection is Vital

Early detection of melanoma is reported in the literature as being very important to improving the prognosis of this disease. Friedman et al. (1991) state "In sum, the current death rate from malignant melanoma can be reduced to nearly zero through early detection coupled with prompt surgical removal".

Both lay-persons and clinicians have a role to play in early detection. The next section describes some of the reported initiatives intended to increase lay-person awareness, and also techniques intended to give lay-persons the ability to identify melanoma. We then look at early detection from the point of view of clinicians. This inspection details the techniques clinicians use to identify melanoma, and also reports on the accuracy of clinicians at identifying these lesions.

Early Detection 1: Lay-persons

Activities concerned with reducing the mortality rate from melanoma come in three forms, primary, secondary, and tertiary. Primary activities are intended to reduce the number of people getting melanoma, secondary activities are concerned with promoting early detection of melanoma, and tertiary activities are those techniques that can help treat advanced disease. Promotion of covering up and sun-screen use are examples of primary activities, while skin self-examination, promoting awareness of the appearance of melanoma, and population screening are all examples of secondary prevention measures. Figure 2.1 shows the relationship between these activities (reproduced from Champion et al. 1998).

The ABCD checklist of Friedman et al. (1985), together with the seven point checklist (Table 2.1) of suspicious features of MacKie (1985), have been mainstays of secondary education activities. For example, the Cancer Society of New Zealand produces a bookmark containing the ABCD checklist for distribution to the public. Other educational programs have also been used in New Zealand. These include Spot Check days (where members of the public can have moles checked for free

by experienced clinicians), mass-media awareness programs and school education programs. Such programs generally have both primary and secondary prevention effects. Elwood & Glasgow (1993) reviews these programs from the New Zealand perspective. Similar initiatives have been used in other countries, but the efficacy of such programs is not well known. In reviews of education campaigns, such as MacKie & Doherty (1988), Thursfield et al. (1995) and Giles & Thursfield (1997), an increased number of moles being presented is commonly reported, which would be expected if the program is having a secondary effect. However, primary effects may take much longer to establish. From the rapidly increasing incidence of melanoma around the world, it appears that primary prevention activities are either not having the desired effect, or that delay exists between the program and the apparent primary effect.

In the New Zealand case at least, it appears some progress has been made concerning secondary prevention, in particular, lay-persons awareness of the disease. The Public Health Commission's report to the New Zealand Minister of Health states, "New Zealanders' awareness of melanoma has increased markedly since the first campaigns in 1978. By 1989, more than half the population could describe melanoma accurately,

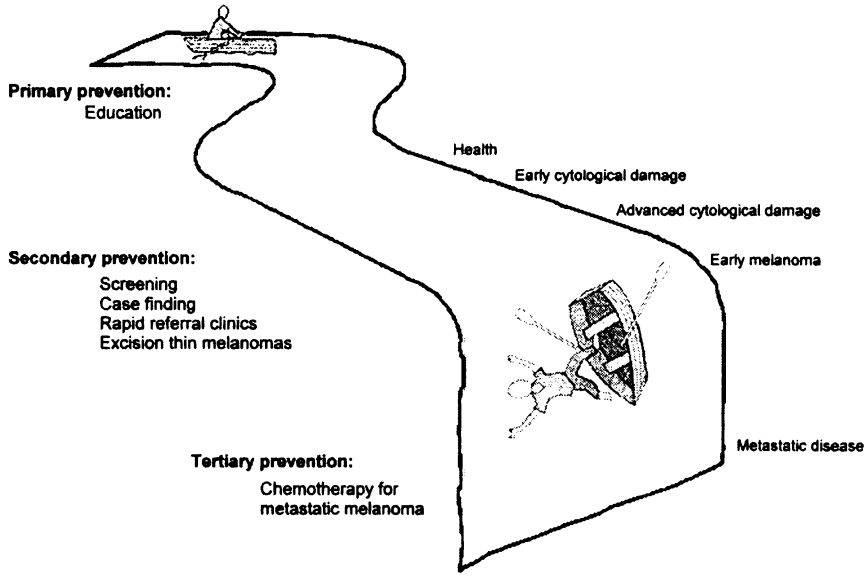

Figure 2.1: The relationship between intervention and stages in disease progression, reproduced from Champion et al. (1998). Champion et al. state "Disease prevention in a serious condition such as melanoma is much more sensible than treating sick individuals with expensive drugs at the end of a long chain of irreversible pathological events".

Table 2.1: Seven point checklist of suspiciousness of MacKie (1985)

1	Minor itch or other change in sensation.
2	A lesion greater than 1 cm in largest diameter.
3	A history of growth or other change in a pigmented lesion in an adult.
4	An irregular outline.
5	Irregular and varied colors.
6	Inflammation in or at the edge of the lesion.
7	Bleeding or crusting.

and by 1992, 83 percent" (Skegg 1994).

These results suggest high levels of awareness in New Zealand. It is not clear however, whether such high levels of awareness will continue. It is also not clear whether this awareness of melanoma results in people being concerned about skin lesions. That is, people may be aware of the problem of melanoma, but this awareness does not necessarily mean they use preventive measures, or seek medical advice early when they have identified a potentially malignant lesion. Further research may be needed to clarify this point. If people do not seek medical advice even with a high level of awareness, it is possible that educational activities are missing the mark, and different techniques may be required.

Australians, perhaps not surprisingly, also seem to have high awareness. An editorial by Robin Marks introduces the steps used in Australia (Marks 1994) to increase awareness. Borland et al. (1992) reports that over 90% of Australians' surveyed had heard the term melanoma, and 95% believe it to be a serious disease, but many were confused about the visual characteristics of the disease. These results suggest that perhaps New Zealanders and Australians are quite similar in terms of knowledge of melanoma.

Similarly, Martin's (1995) research on risk factors, knowledge and preventive behaviour in Australia also reports a high level of knowledge and awareness. Martin found that although awareness was high, a significant proportion of respondents were unaware of risk factors, particularly those associated with 'Celtic' heritage (blue eyes, fair or red hair), and those at high risk were unlikely to know they were. Jackson et al. (1999) present similar findings for the United Kingdom.

In contrast to these results, it appears awareness of melanoma is low in the United States. Miller et al. (1996) presents recent research concerning Americans knowledge of melanoma. According to this research, of the 1001 people surveyed, around 50% of men and 35% of women did not recognise the term melanoma. They also report that awareness is related to levels of education and income. In general, the higher the education and income levels, the higher the knowledge. This finding is reproduced to a degree in Scottish research by MacKie & Hole (1996), who conclude that although

the incidence of melanoma is higher among people on higher incomes, the mortality and morbidity rates are less than those on lower incomes. This finding suggests that those on higher incomes are more likely to be aware of the problem, and thus MacKie & Hole state that "early diagnosis campaigns should be targeted particularly to less affluent men...".

At this stage, it is apparent that awareness of melanoma by lay-persons is variable by country. New Zealanders and Australians appear quite aware of the problem, and this result is encouraging for reducing mortality from melanoma. Once a person becomes aware of a possibly malignant lesion, the next step is most likely to be a visit to a general practitioner or dermatologist. The general practitioner will usually refer 'suspicious' lesions to a dermatologist. The next section reviews the methods for identifying melanoma available to clinicians. Research is also presented concerning the ability of both specialist and non-specialist medical personnel to identify melanoma.

Early Detection 2: Clinicians

Clinicians also have a role to play in early detection, and therefore, the ability of clinicians to detect melanoma is important. In New Zealand, the evidence in the literature suggests that the ability of medical personnel to recognise a possibly problematic lesion is quite high. McGee et al. (1994) performed a survey of 900 general practitioners around New Zealand, regarding their ability concerning melanoma detection. 35 dermatologists were given the same survey as a comparison sample. The survey asked respondents to suggest a diagnosis for 12 lesions, including three melanomas. The results indicated that general practitioners in New Zealand had a reasonable level of skill in identifying melanomas, and were also quite successful in identifying the need for biopsy. 67% of the lesions were correctly identified by the general practitioners, and 83% of the lesions were correctly referred for biopsy. Dermatologists were found to be significantly more accurate (average 87% correct). McGee et al. concluded, "the generally good results of this survey suggest a high degree of expertise among New Zealand general practitioners".

In Australia, MacKenzie-Wood et al. (1998) report that dermatologists achieved an overall accuracy rate of 65.6%, lending further emphasis to the above results. This value was found through analysis of 61 suspected melanoma excisions over the period of one year. Morton & MacKie (1998) also report on this topic. They found diagnostic accuracy rates were highly dependent on the time the clinician had been practicing. Sensitivity (the proportion of actual melanoma identified as melanoma) rates varied between 79% for registrars to 91% for consultant dermatologists. Similarly, diagnostic accuracy (the proportion of cases in which the clinician was correct

in their clinical diagnosis of melanoma) varied between 56% and 80% for registrars and consultants respectively.

In related research, Ramsay & Weary (1996) present a summary of results for the accuracy of dermatologists compared to non-dermatologists when presented with a range of skin lesions. They quote data from the United States, England, Australia, as well as the data of McGee et al. (1994) for New Zealand. Although their aim was to highlight the lack of ability of non-specialists concerning general skin disease, they conclude that the ability of dermatologists to identify skin lesions is very high. General practitioners were significantly lower. It should be noted that this result concerns the ability of medical practitioners to identify melanoma, not their ability to recognise which lesions need to be looked at by a dermatologist.

Other studies report similarly on the accuracy of clinical detection of skin lesions. Grin et al. (1990) reported on the accuracy of clinicians regarding skin lesions in the United States over the period 1955 through to 1982. Using the computerised database of the Oncology Section of the Skin and Cancer Unit of New York, they investigated the sensitivity, specificity and predictive value positive of diagnosis of melanoma. Sensitivity refers to the proportion of all cases of histologically confirmed melanomas that were clinically identified as melanoma. Specificity refers to the proportion of all cases histologically proved not to be melanoma that were clinically identified as not melanoma, and predictive value positive is the proportion of all cases clinically identified as melanoma which were histologically confirmed to be melanoma. High values for sensitivity and predictive value positive indicate that a large proportion of melanomas were identified prior to confirmation by biopsy. Grin et al. reported that sensitivity has increased over the period, from 63% to 84.5%. They conclude that in the United States a substantial percentage of melanoma were identified prior to biopsy, but still a significant number of melanoma eluded recognition. Del-Mar et al. (1994) also report on an investigation into clinical accuracy in the United States. They measured the percentage of malignant, pre-malignant and potentially malignant lesions out of a set of lesions excised. Only eight percent of the nearly two thousand lesions recorded were in one of these categories. They report that this percentage increased with age. Using this percentage as a crude tool to measure diagnostic accuracy (or perhaps more accurately, excision accuracy), they conclude that the data suggests poor specificity of clinical excisions, contrary to the results previously reported by Grin et al. (1990). It should be noted that the data may be skewed by non-medical excisions, for example those removed for cosmetic reasons.

The above reports were all based on viewing the lesions at the Clinical-view. Another method of viewing lesions used by clinicians is Epiluminescent Microscopy, or ELM. ELM is gaining popularity as a primary tool for melanoma detection. A large amount

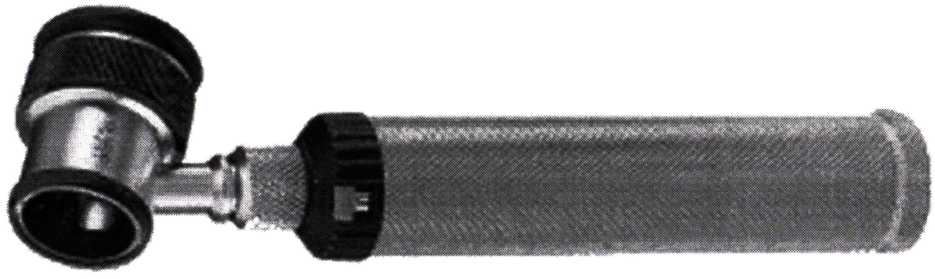

Figure 2.2: A hand-held dermatoscope

of literature describes the method, for example Pehamberger et al. (1987), Steiner et al. (1987), Kenet et al. (1993), Nilles et al. (1994), Stolz et al. (1994), and Menzies et al. (1996). The method is reproduced here from Kenet et al. (1993).

> "The technique consists of placing a thin layer of mineral oil on the skin and inspecting pigmented structures below the skin surface, typically with a 6X to 40X magnification using a hand-held lens... The oil eliminates surface reflection due to the refractive index mismatch between air and skin. This renders the stratum corneum transparent, enabling the in-vivo visualization of pigmented anatomic structures of the (skin)".

In practice, ELM simply means covering the lesion in oil, placing a glass plate directly on the lesion and then viewing the lesion through a magnifying lens (Figure 2.2). Section 1.3 in the previous chapter showed examples of the differences in appearance between Clinical-view images and ELM images. It has been reported that ELM allows trained specialists to achieve a higher diagnostic accuracy rate than simply using the Clinical-view of lesions, for example Pehamberger et al. (1987), Steiner et al. (1987), Pehamberger et al. (1993), and Binder et al. (1995) amongst many others. However, there is data to suggest that dermatologists who are not formally trained in the area may in fact decrease their clinical accuracy (Binder et al. 1995). For an introduction and review of ELM literature, the reader is referred to Argenyi (1997).

To assist the development of clinical expertise with ELM, several researchers have described ELM-criteria thought to indicate malignancy. Kenet et al. (1993), Stolz et al. (1994), and Menzies et al. (1996) all present work in this area. Binder et al. (1999) have recently re-evaluated the ABCD criteria of Stolz et al. (1994) using seventeen dermatologists ranging in experience from first year registrars to dedicated skin lesion clinicians with many years of ELM experience. They found that the method generally enhances clinical accuracy. Sensitivity and specificity results varied

considerably however, given the different groups of dermatologists.

As the ELM technique becomes more widespread, and more clinicians become familiar with it, it is likely that the current known indicators of malignancy will become more refined, and the ability of dermatologists to recognise melanoma may therefore improve. Other techniques, such as ultrasound (Dummer et al. 1995), have also been proposed to be used in conjunction with ELM, and may lead to a further increases in the ability to detect melanoma. However, ELM is not a panacea for early detection of melanoma. Steiner et al. (1987) wrote "Although epiluminescence microscopy does not-given the biologic variability of pigmented skin lesions-cannot provide an absolutely reliable diagnosis for all lesions, it does considerably improve the overall diagnostic accuracy at the clinical level".

From these reports, it appears that dermatologists are likely to be quite good at recognising skin lesions, although discrepancies exist. The results for general practitioners are less clear, but one study (McGee et al. 1994) reports that general practitioners do have a good ability to recognise the need to send skin lesions for further inspection.

Perhaps the overwhelming (and understandable) attitude of dermatologists can be summed up by MacKenzie-Wood et al. (1998). They state "...it must be said that the key decision surrounds whether or not to biopsy a suspicious lesion. Clinical accuracy is not vital if the suspicious lesion has been removed". That is, the accuracy of dermatologists in recognising melanoma is less important than the ability to recommend a lesion be excised. It is apparent that although dermatologists tend to be quite adept at identifying melanoma, their primary function is to recommend lesions for excision.

2.1.2 How About Screening?

Thus far, we have established that melanoma is an increasing problem. Research was presented concerning the methods used to increase awareness in lay-persons, focussing on secondary methods, that is, those methods encouraging early detection and treatment. The ABCD criteria and danger signs reported in Friedman et al. (1985) are a major focus for early detection campaigns. We have also looked at methods used by clinicians to recognise melanomas, and how accurate these clinicians are. A relatively new technique, epiluminescent microscopy, has been shown to improve the accuracy of clinicians, and is becoming an increasingly important tool, although it is apparent that care must be taken to ensure proper training.

In order to utilise the expertise shown by clinicians, the patient must first report with a suspicious lesion. However, MacKie & Doherty (1988) have reported that

the most significant delay for a patient with melanoma occurred before that person saw a medical practitioner. This delay indicates the need for intervention before medical practitioners are consulted. Population screening, similar to that already undertaken for cervical and breast cancers, is one of these proposed interventions.

Several agencies have looked at population screening of skin lesions as an initiative to reduce the death rate from melanoma, by increasing the rate of early detection. Population screening involves regular visits to a specialist for examination. It is an appropriate measure in cases where early detection is vital to improving prognosis, or when pre-cancerous conditions can be readily identified. At first glance, it appears both of these conditions are true for melanoma. However, in the New Zealand situation, Elwood & Glasgow (1993) have said that:

> "Recommending a routine general skin examination involves a major commitment of health care resources. In New Zealand, we estimate that if general practitioners offered an annual spot check to all patients over age 30, the check would consume 5 per cent of their total clinical time" (Elwood and Glasgow 1993, page 20)

Five percent of the clinical time of all general practitioners represents a large diversion of resources. If such a program were to go ahead, the benefits would have to be correspondingly large. In order to assess the benefits of a screening program, Grob (1997) proposed three assumptions that need to be fulfilled for gain from melanoma screening. The first assumption is that screening will detect melanoma faster than no screening. For a tumour such as melanoma that can grow rapidly, he postulates that "(screening) twice a year would be a minimum for an efficient screening (of melanoma)". The second assumption is that the participants in the screening program are high-risk, that is have a high incidence of melanoma. The final assumption is that the screening test (in this case clinical examination) is accurate and reliable. He reports that clinical examination is not accurate enough, and therefore this assumption cannot be justified. The conclusion from this report is that melanoma screening of the general population is "certainly not cost effective".

In a similar vein, the report of the United States Preventive Services Task Force (DiGuiseppi et al. 1997) presents a discussion concerning routine screening for skin cancer. The report concludes "there is insufficient evidence to recommend for or against either routine screening for skin cancer by primary care providers or counselling patients to perform periodic skin self-examinations". This conclusion summarises the lack of data concerning aspects of skin cancer treatment, especially results of techniques improving mortality and morbidity.

Jackson et al. (1998) puts the case for more targeted screening. Patients are asked

to complete a risk factor flow chart proposed by Rona M. MacKie. This flow chart incorporates four risk factors for melanoma, freckling, greater than 20 moles, atypical naevi, and history of severe sunburn. Patients who score four (out of four) are said to be high-risk. Such targeting is intended to reduce the high costs associated with population screening, but debate exists over whether such methods are valuable, and some papers suggest the possibility of negative effects such as decreased vigilance on the part of those not targeted (for example, Sinclair 1998).

The papers reviewed above indicate the costs involved in screening skin cancer through traditional methods. Given the current situation, it appears that population screening is unlikely to be implemented, due to high costs and uncertain benefits. However, if a low cost screening program could be found, benefits from screening become relatively more attractive. Given that most of the expense associated with screening is associated with the (valuable) time of medical personnel, an automated screening system able to be operated by non-specialists may be more cost-effective. This use of an automated system may increase the number of melanomas excised, and thus decrease the mortality rate.

2.1.3 Section Summary

Several points should be gleaned from the above review. Melanoma is a serious problem around the world, and especially in New Zealand. Early detection is the easiest and best way of reducing morbidity and mortality from the disease, and the chance of a cure increases proportional to the earliness of detection.

It is apparent that dermatologists are quite accurate when recognising melanoma, and new techniques such as ELM have the ability to further improve accuracy. The previous papers suggest that the majority of melanoma are recognised and treated once referred to a specialist, and also that general practitioners are well aware of the need to biopsy suspicious lesions. Therefore, we may conclude that if a lesion is seen by a general practitioner, in most cases, a melanoma will be referred to a dermatologist. Once referred to a dermatologist, most melanomas are likely to be excised.

From Elwood & Glasgow (1993), DiGuiseppi et al. (1997), Grob (1997), and Sinclair (1998), it could be concluded that it is unlikely that population screening for skin cancer utilising general practitioners and dermatologists will become a reality until further cost/benefit analysis is provided, or at least a more accurate method of melanoma detection is available. A cheaper method of screening may be more likely to be implemented. An automated system may be useful in this situation.

2.2 Automated Solutions - call in the computer!

Detection of melanoma by computer seems to be a fairly obvious research field, and indeed, numerous research papers have been published chronicling development of such systems. In this section, we look at some of this work. The research is presented in roughly chronological order, to give some idea of the progression of ideas in this field. Table 2.2 presents a summary of this research. This review is summarised in Day & Barbour (2000).

The earliest research found concerning automated systems for melanoma detection was that presented by Dhawan (1988). Dhawan presented an instrument (termed 'Nevoscope') developed from ideas described in his Ph.D. thesis. The concept behind the Nevoscope is to transilluminate the skin lesion a number of times, and use the resultant images to develop a simple three dimensional model of the lesion. From this 3D model, a set of features is measured, for example thickness, 3D size, skin and lesion color, lesion margin, together with boundary, shape and surface characteristics. These measurements, together with patient history data, are used as input to a rule-based analysis and interpretation system, that attempts to classify the image. This paper reported a work in progress, and as such does not report results.

Green et al. (1991) developed an automated system using a video camera connected to a computer. A sample of 89 images was captured using the video camera, together with five clinical details for each lesion. These five details were: diameter (mm), colour of lesion (uniform light brown, uniform dark brown, uniform black, variegated), regularity of outline (very, moderately, very irregular), blurriness of edge (clearly defined, blurred) and whether the lesion was palpable (yes, no). The images were then algorithmically analysed. Image analysis consisted of colour means and variances, infrared colour data, as well as area and perimeter data. Once the image analysis was complete, a software package was used to attempt to separate three classes of lesion, namely melanoma, melanocytic naevi and 'other' pigmented lesions. They report that 76% of the lesions were correctly classified. One melanoma out of five was incorrectly classified using all the criteria, four out of 53 lesions were falsely classified as melanoma, and six out of 12 'other' lesions were incorrectly classified. This accuracy dropped to 73% percent when infrared measurements are left out. Interestingly, this paper showed that there was little correlation between the clinical details and the corresponding image analysis results, excepting lesion size.

Cascinelli et al. (1992) present a similar automated system, termed SkinView. The system analysed lesion images and measured colour, shape, and texture features (somewhat strangely assigning binary numbers to this data, rather than real numbers that would allow a continuum of values). Data from clinical assessment by medical

Table 2.2: Summary of automated melanoma detection system research. CV = Clinical-view . Notes: * Bostok et al. only used silhouettes of Clinical-view lesions. † Specificity was not calculated in research. This figure was derived from other figures and may be incorrect. ‡ Gutkowicz-Krusin et al. fixed sensitivity in their system to be 100%. ¶ Schindewolf et al. (1993b) compare Clinical-view and ELM images. * Schindewolf et al. (1994) looks at direct digitised images versus digitised slides.

Author	Date	Sensitivity	Specificity	Total images	Melanoma images	Image type
Dhawan	1998	Not avail.		Not avail.		Other
Green et al.	1991	80	72	70	5	CV
Cascinelli et al.	1992	83	60	169	45	CV
Schindewolf et al.	1993a	94	88	353	215	CV
Schindewolf et al.	1993b	Accuracy=81(CV)¶		320	194	CV
		Accuracy=78(ELM)¶		320	194	ELM
Bostok et al.	1993	92	68	124	68	CV*
Ercal, Chawla et al.	1994	≈82	≈83	240	120	CV
Green et al.	1994	83	82	164	18	CV
Ercal, Lee et al.	1994	96	≈ 62†	399	135	CV
Schindewolf et al. *	1994	Accuracy≈80		404	240	CV slides
		Accuracy≈80		309	80	CV direct
Andreassi et al.	1995	Unknown		430	50	ELM
Hintz-Madsen et al.	1996	59	?	180	60	CV
Menzies et al.	1997	93	67	170	75	ELM
Gutkowicz-Krusin et al.	1997	100‡	61	104	30	ELM
Horsch et al.	1997	Unknown		118	60	ELM
Seidenari et al.	1998	93	95	917	65	ELM
Bischof et al.	1998	80-100	80-84	221	45	ELM
Binder et al.	1998	90	74	120	39	ELM
Landau et al.	1999	Accuracy = 92		71	7	CV
Seidenari et al.	1999	100	92	424	37	ELM

practitioners was also included. The major conclusion contained in this report was that algorithms could capture the clinical judgement of the expert physician to some degree. However, no evaluation into how well this judgement could be captured was attempted. They also identify the possible use of an automated system to "identify lesions that need further investigation". In a follow-up to this paper, Sober & Burstein (1994) describe their experience with the SkinView system. They report that the system had difficulty discriminating between benign and malignant lesions in their trial. For example, use of shape features classified 82% of melanoma as 'suggestive of melanoma', whereas the same features classified 85% of non-dysplastic benign naevi as 'suggestive of melanoma'. One of the reasons proposed for the poor results was that pre-training (in Milan) used more advanced lesions, skewing the model. This failure indicates the difficulty involved in developing a system that will work on the population of lesions.

Schindewolf et al. produced three papers concerning automated classification of melanoma. The first, Schindewolf et al. (1993*a*) described an automated system that analysed a set of 353 lesions. 138 were benign, with the remaining 215 melanoma. The best classification results reported indicate sensitivity of 94%, and specificity of 88%. 16 benign and 12 malignant lesions were mis-classified. This paper is detailed in its description of techniques, and also presents an evaluation of the image features used. The second paper, Schindewolf et al. (1993*b*), evaluated two different types of images in the context of an automated system. Using the system described in Schindewolf et al. (1993*a*), Clinical-view and ELM classification systems were compared. They report that the Clinical-view classifier produced slightly better cross-validated results than the ELM classifier. Interestingly, they also created a classifier that used both Clinical-view and ELM features. This classifier performed slightly better than both the Clinical-view and ELM classifiers. The final paper, Schindewolf et al. (1994), reports on a comparison between digitised slides and directly digitised images of skin lesions in a skin lesion classification system. The methodology was similar to the previous two papers. They report cross-validated accuracy of 80% for both the slide-based system and the directly digitised image system. However, the lesions analysed by the two systems are not identical, and it may be that the results are a reflection of the difference in image sets rather than the systems performing identically.

A follow-up to Green et al. (1991) is presented in Green et al. (1994). This system again utilised a colour video camera and frame grabber-mounted on a computer. The video camera and frame grabber captured lesion images, which were then analysed in a similar way to their previous paper. 164 images were collected and analysed by the system, and a success rate of 89% was reported, compared with a rating of 83% percent based on the clinical grading of the lesion characteristics. The system

classified 16 of the 18 melanomas correctly.

A system using neural networks is described in Ercal, Chawla, Stoecker, Lee & Moss (1994), and also presented subsequently in Lee (1994). In this study, a neural network was used to classify skin lesion images based on a number of characteristics including border irregularity, colour variances, relative chromaticity, and asymmetry. These characteristics were based primarily on the ABCD characteristics of Friedman et al. (1985). They reported a best success rate of 86%, although this result was obtained without inclusion of dysplastic naevi in the test set. The success rate inclusive of dysplastic naevi approached 83%. From their work, they note a number of findings. Firstly, that colour data is crucial to the process of identifying melanoma. Secondly, that boundary irregularity and asymmetry are important for distinguishing melanoma from benign lesions, and finally, they noted the difficulty of distinguishing dysplastic naevi from melanoma.

Ercal, Lee, Stoecker & Moss (1994) presented a continuation of the previous paper that used more advanced neural network techniques (again subsequently reported in Lee 1994). The simple neural network was re-tested with a larger and more varied data set, which resulted in 65% success rate for melanoma. Using the new neural network, they report an improved success rate for melanoma of 88%. The differences between the results of the simple system reported in these two papers again highlights the difference that a change in image sets can make.

The above systems all use Clinical-view images. In the area of epiluminescent microscopy, the hand-held Dermatoscope (Figure 2.2) was only just beginning to be widely used, and most skin lesions were assessed using the Clinical-view of the lesion. After this point (around 1995), however, much more effort was put into ELM-based techniques (See Table 2.2).

This progression of ideas from Clinical-view to ELM is seen quite clearly in the review papers of this area. Stoecker & Moss (1992) presents one of the early reviews of dermatological applications of computers. The overview at this early stage indicates the focus on Clinical-view images, although Wilhelm Stolz et al. from the University of Munich were reported to be experimenting with digital analysis of ELM images. A further paper from the same authors, Stoecker et al. (1995), titled "Non-dermatoscopic digital imaging of pigmented lesions" reports progress in Clinical-view imaging. The title suggests that dermatoscopic digital imaging was becoming more important. The review carried out in Sober & Burstein (1994) also concentrates on Clinical-view based systems. However, mention of the work of Schindewolf et al. (1993*b*), which compared classification rates of ELM and CV images using image analysis, together with the computer enhancement of ELM images shown in Kenet et al. (1993) indicate the first steps in this area.

From 1995 however, ELM research reports become much more prolific. Andreassi et al. (1995), Menzies et al. (1997), Bischof et al. (1998), Binder et al. (1998), Seidenari et al. (1998), and Seidenari et al. (1999) all report on 'smart' systems based on the ELM view of images. MoleMax II by DermaInstruments, Austria is a commercial system based on ELM images with a function for identifying melanoma. Research into Clinical-view 'smart' systems becomes less apparent. The editorial by Kopf et al. (1997) reviews the dermatoscopy and digital imaging literature and is a good place to start for a more recent look at developments in this field. This review shows more emphasis on ELM-based systems. However, the timing of the paper excluded coverage of yet more recent advances concerning the use of ELM-images in automated systems. The following literature review comes from the post-1994 period.

The first research in this period, Andreassi et al. (1995), details the development of ELM-based software termed DBdermoMIPS. In similar fashion to the papers above, the lesion image is segmented and analysed with image analysis algorithms. Both ELM and magnified Clinical-view images were captured. The image analysis algorithms used were based on the ABCD criteria, almost certainly that proposed in Friedman et al. (1985), rather than the ELM-based criteria of Stolz et al. (1994). However, there is little evidence that these criteria are valuable for images captured under epiluminescence. Andreassi et al. state that because of the lack of light reflection observed with ELM, "this method (ELM) was therefore more reliable with our system". It may therefore be assumed that the results quoted apply to the ELM images. The correctness of this assumption is far from clear in the paper, and throws uncertainty over the results. Another problem with this paper concerns the application of algorithms based on the ABCD criteria of Friedman et al. It appears strange to apply Clinical-view based algorithms to epiluminescence. In spite of these possible problems, Andreassi et al. report significant differences between melanoma and benign lesions for most of the algorithms investigated. The differences were as expected, for example melanomas had higher asymmetry than benign lesions. Unfortunately, sensitivity and specificity results for this system are not reported.

Gutkowicz-Krusin et al. (1997) reports on an investigation into automated melanoma detection based on ELM images, similar to previous reports. Little new work is presented although they describe perhaps the first set of feature algorithms definitely based on ELM-criteria, specifically, that reported in Stolz et al. (1994). The image data set contained 30 images of melanoma and 74 atypical naevi. They report that the use of a linear classifier trained on a subset of 76 images returned 100% sensitivity when tested on the remaining 28 images (five melanoma and 23 atypical naevi). Specificity was 61%. It should be noted that these results are from attempts to distinguish between melanoma and atypical naevi, which are normally very difficult

to separate, even at the ELM view.

Horsch et al.'s (1997) research describes another ELM based system. In this system, algorithmic equivalents of the 'B' and 'C' components of the ABCD checklist of Stolz et al. (1994) are used to classify images. Few details on actual implementation of the algorithms are reported, and results are sketchy. They report that the system "bought a quality of classification near to that of our human experts". Little of substance is reported in this paper, although the work itself may be worthy.

In Menzies et al. (1997) and Bischof et al. (1998), the automated system developed in conjunction with CSIRO, the Sydney Melanoma Unit and Polartechnics Ltd., is described. The system is based on ELM images, and proof of concept began in mid-1994. The first phase system (Menzies et al. 1997) utilised digitised images of ELM slides and logistic regression techniques. They obtained a sensitivity of 93% percent and a specificity of 67% percent for this phase, using an image set consisting of 75 melanoma and 95 atypical benign lesions. Details of the individual feature algorithms are not presented. In the next phase of this project, a video-capture system was added to allow real-time processing. Interim results from this prototype are reported to be between 89 and 100% for sensitivity, and 80 to 84% for specificity. The image set for this phase included 45 melanoma and 176 atypical benign lesions. In-situ melanomas, pigmented basal-cell carcinomas and acral-lentiginous lesions were not included. Again, little detail of the algorithms is reported.

Binder et al. (1998) report on an ELM-based 'smart' system that uses an artificial neural network as a classifier. The major contribution of this research is an investigation into how well lesions could be separated into three classes, melanoma, dysplastic naevi and common naevi. Most other papers use two classes (melanoma and benign), and Binder et al. also present data for the two class problem. The neural network was trained using 83 randomly selected lesions, and tested on the remaining 29 lesions. For the two class problem (melanoma versus benign), sensitivity/specificity was reported as 90%/74%, which are similar results to previous research in this field. For the three class problem (melanoma versus dysplastic naevi versus common naevi) however, melanoma results (sensitivity/specificity) dropped to 38%/63%, with dysplastic naevi and common naevi results being similarly poor (62%/38%, 33%/67% for dysplastic and common naevi respectively). Binder et al. suggest that such poor results may be indicative of the difficulty of distinguishing between dysplastic naevi and melanoma. However, they obtained adequate results for the two class problem. The two class problem results suggest that melanoma were being distinguished from dysplastic and common naevi and therefore the difficulty appears to be in distinguishing between dysplastic naevi and benign lesions.

Seidenari et al. (1998) presents similar research with few new contributions. 917

lesions were included in the study, although only 65 of these were melanoma, and results are only reported for a 90 image (31 melanoma) subset. The lesions were segmented and analysed using dedicated image analysis software. Their methodology is possibly flawed however, due to the image analysis features chosen. They use the DBDermoMIPS(r) software, similarly to Andreassi et al. (1995). The problems noted previously with the work by Andreassi et al. therefore also apply. It should be noted that Seidenari et al. state "We used a program based on the translation of the ABCD concept from Nachbar et al. (1994)". Andreassi et al. (1995) however, do not mention either of these references, although seemingly use the same software. Unless the software has been upgraded, there exists an interpretation discrepancy between these two papers. Although there is some overlap in the ABCD criteria proposed by Friedman et al. (1985) and the ELM-based criteria of Stolz et al. (1994), it cannot be expected that software (DBDermoMips) developed for Clinical-view criteria accurately replicates another set of criteria intended for ELM use.

However, although the features chosen are not obviously based on recognised ELM features, it is not to say that the features are without merit. As noted above, there is some overlap in the two ABCD criteria, for example, the asymmetry and colour features. Seidenari et al. report sensitivity of 93% compared with 81% for an experienced observer and 74% by the inexperienced observer over the same image set (90 images, 31 melanoma). It appears to be that the features chosen may be relevant for analysis of ELM images, based on the results shown.

Seidenari et al. (1999) reports on further developments to the system described by the previous paper. Using discriminant analysis, they train a classifier on a set of 59 naevi and 19 melanoma. This classifier is then tested using a set of 365 naevi and 18 melanomas. They report sensitivity on the test set of 100% and specificity of 92%. The major concerns with this paper are similar to those described above. With the low number of melanoma, it appears unlikely that these cases represent the range of features displayed by this disease.

The final paper discussed in this section, that of Landau et al. (1999) is something of a anomaly, a report on a Clinical-view based system in 1999. The project is similar in concept to those presented previously, in that pigmented skin lesions are identified by computer. 71 Clinical-view images are analysed, of which seven are melanoma. They use the standard deviation of hue as the only feature by which lesions are classified, and report 92% accuracy in identification of lesions. However, there are several problems apparent with this paper. Firstly, the low number of images, and especially the low number of melanoma is a concern. In research such as this however, lesion image sets are often difficult to obtain. Secondly, there is little justification for their choice of feature, standard deviation of hue, other than the importance of colour variegation in clinical assessment of skin lesions. It must be

noted that hue in a technical sense does not mean colour. For a discussion of colour, see Chapter 4, Section 4.2.1. The third problem exists because they use standard deviation of hue as the only feature. Care should therefore be taken to ensure that the images are obtained under similar conditions. Unfortunately, only rudimentary measures are taken, and in particular the distance from the camera to the lesion is not fixed. This lack of standardisation may have an effect on the standard deviation of hue, and may skew results to an unknown degree. There are obvious problems with the research reported in this paper, and these problems should be kept in mind when interpreting the reported results.

2.2.1 Screening? Diagnosis?

Implicitly, all of the automated systems presented above were concerned with diagnosis of melanoma, that is, reproducing the perception of pathologists. This is apparent as only histologically diagnosed lesions were used to train the systems, and results were reported on the basis of whether the systems reproduced those histopathology results. Such systems can be referred to as 'automated diagnosis systems'. However, Section 2.1.2 made the case for an automated screening system. Here, we look at the two concepts, screening and diagnosis, and discuss the differences.

Current methods of population screening would use a dermatologist or experienced general practitioner to identify lesions that may be malignant. Pathologists are not used, as the lesion would need to be excised before pathological examination could be conducted. A screening system would be required to perform similarly to the dermatologists, namely to identify lesions that may be malignant. A screening system therefore, is any system that can be used to identify lesions that may be malignant. From this definition, it should be apparent that the automated diagnosis systems reviewed previously are also screening systems.

However, diagnosis systems do not attempt to replicate what occurs in the clinical setting, namely the assessment of lesions by dermatologists. Diagnosis systems attempt to reproduce the results of pathologists, that is, perform the function of diagnosis. This emphasis on reproducing histopathological results (diagnosis systems) can be considered the current state of automated melanoma detection research (and hence, automated screening research).

The problem that a diagnosis system is therefore attempting to solve is to distinguish between melanoma, and benign lesions that 'look like' melanoma. If a benign lesion did not look like melanoma, it appears unlikely that the lesion would have been excised (excluding other reasons for excision). Therefore, no obviously benign lesions will be used to train a diagnosis system. When the system is presented with an

obviously benign lesion, there is no guarantee that the lesion will be classified as benign. Often however, good results on the arguably more difficult problem of separating suspicious benign lesions and melanoma is used to infer good results on the lesion population.

For example, consider the words of Bischof et al. (1998) when reporting the results of an investigation into a computer-based diagnosis system. They state that "the results are for a set of lesions that have already been pre-filtered by an expert. The non-melanomas... were considered sufficiently atypical (as judged by that expert) to be excised... So the percentage of non-melanomas correctly identified by our procedure will be much higher than the specificity figures reported here". In other words, if 'typical' benign lesions were available, the system would perform better. This conclusion, that if obviously benign lesions were added to the set they would be easily classified *may or may not* be true. In fact, it is entirely possible that the addition of 'easy to diagnose' lesions may make the classification task more difficult, and hence adversely affect the results of the system.

In a diagnosis system, the task of identifying obviously benign lesions is being left to the clinician, a role which the literature suggests clinicians are extremely capable at performing. The task of identifying malignant lesions may then be performed with the assistance of the diagnosis system, which has been trained on lesions of this nature. In general, the role intended for diagnosis systems is to assist the clinician with a difficult diagnosis.

For a screening system however, no clinician is available for the task of identifying obviously benign lesions. The screening system is intended to replace expensive human expertise, as one of the arguments against screening is that it requires experienced (expensive) medical personnel. The task of identifying obviously benign lesions is therefore being left to the screening system. If a diagnosis system was used in this position, we would have to also train the diagnosis system to recognise lesions that are obviously benign, as well as lesions that are benign and look like melanoma. Obviously, the requirement for histological examination for each lesion makes such a task difficult. Ignoring this difficulty, it appears likely that melanoma have more visual characteristics in common with benign lesions that appear malignant, than benign lesions that appear malignant have with obviously benign lesions. Therefore, the task to separate melanoma from all 'benign' lesions may be an incredibly difficult one.

It appears likely that there is a much larger distinction between obviously benign lesions and other lesions (melanoma and benign lesions that look like melanoma). This distinction is what is being modeled by a 'dermatologists assessment' system. What we are suggesting in this research is that developing a model that can separate

obviously benign lesions from lesions that appear as melanoma (both benign lesions and actual melanoma) could be useful in the context of a screening system, especially given the different roles of screening and diagnosis systems.

To summarise, we have shown the need for an automated screening system. The current use of computers to diagnose skin lesions was reviewed, and we have proposed an complementary methodology for identifying potentially malignant lesions. It is proposed that attempting to reproduce the assessment of dermatologists may be adequate as the basis of a screening system.

2.2.2 Summary of the Thesis Argument

As an aid to early detection of melanoma, an automated screening tool may be useful. Literature was presented to show the development of ideas in automated diagnosis systems, and we saw evidence of a shift towards ELM-based research.

It is not clear why this shift occurred. Perhaps it occurred simply because more clinicians were becoming experienced with ELM techniques. It may also be that researchers were becoming frustrated with the limitations of analysing Clinical-view images. These reasons are pure conjecture however. The shift itself may be justified by research indicating that clinicians experienced in using ELM were more capable of recognising melanoma. However, it does not necessarily follow that computer systems will replicate this success.

Dr. Hugues Talbot of the CSIRO in Australia who is part of a team developing an ELM-based diagnosis system (described in Menzies et al. 1997, Bischof et al. 1998) wrote:

> "The mainstream opinion is that ELM is vastly superior to Clinical-view in terms of what an automated vision system can do with the images..." (Personal Communications: Talbot 1999)

However, there is little evidence in the literature to support this claim. Schindewolf et al. (1993*b*) presents early work comparing the results of a system using the two types of images. Results were inconclusive, with the Clinical-view based system giving slightly better results than the ELM-based system. So the thesis question for this research is: are Clinical-view or ELM images more useful for an automated screening system for skin lesions? In other words, since the more recent research focuses on ELM images, ELM images must be better than Clinical-view images in an automated screening system for skin lesions, given the current state of the field.

The difficulty in establishing a definitive answer to such a thesis question may be apparent. It would be very difficult to categorically establish the relative usefulness

of Clinical-view and ELM images over the entire population of lesions. Obviously such a task is beyond the scope of this research, and probably beyond the scope of possibility.

Although an answer to the above thesis question may be impossible given resource limitations, and in particular the limitations of Ph.D. research, a less comprehensive question may be addressed, and useful answers obtained. In particular, we limit the scope of the research in two major ways. Firstly, we limit the lesion population to those lesions obtainable in suitable image sets. These image sets are described fully in the following chapter. The makeup of the image sets limits the scope of the results to those particular image sets, and perhaps those image sets that exhibit a similar distribution of lesion features.

Secondly, we limit the techniques used to create the classification systems in this research. As there are a large number of techniques that can be used to examine the data present in images, with no strict guidelines as to suitability, we must choose algorithms that appear to be relevant to the context of skin lesions, and limit these to a managable number within the scope of the research. To achieve this, image analysis techniques are obtained from the current published state of the field or developed based on recognised human guidelines for identification of malignant lesions.

The refined question can now be rephrased into a thesis argument:

Given the current published state of the field and a limited set of real world images, ELM-images are more useful in an automated screening system for skin lesions than Clinical-view images.

As has been pointed out, current work is focussed on diagnosis systems. However, the question of Clinical-view or ELM images is also important if we are considering a 'dermatologist assessment' system, as described above. Therefore, we look to investigate this thesis in both the context of diagnosis, similarly to the work of Schindewolf et al. (1993*b*), and also in the context of 'dermatologist assessment', which is the replication of dermatologist's assessment of 'suspiciousness'.

2.2.3 Thesis Discussion

It may be asked at this stage why it matters which image type is used. The answer to this question is simply that the goal of these systems is that they perform as accurately as possible. If one type of image is preferred over another (such as ELM over Clinical-view), the assumption is made that the preferred image type (ELM) allows more melanomas to be detected. There is no evidence in the literature to support such an assumption, and as such, the numbers of melanomas detected by these systems may be lower than necessary. Another point to note is if one image

type can be categorically shown to produce better results, efforts in automated melanoma detection can focus on that image type. Such a reduction in effort may speed advancement in this field. If however, no such result can be shown, research into both areas should continue.

As we noted previously, Schindewolf et al. (1993*b*) looked into a similar question. However, this research extends on that reported by Schindewolf et al. Most importantly, the question of Clinical-view versus ELM is also looked at in the context of reproducing dermatologists' assessment. This investigation marks a major addition to the focus of current mainstream thinking, and is a significant contribution of this research. Also, there are several associated issues with the research by Schindewolf et al. Firstly, the date of the paper is 1993. Most of the work in computer-based diagnosis of Clinical-view and ELM images is not covered by this paper. Schindewolf et al. report no other research into ELM-based image analysis taking place at this time, and Clinical-view based research is similarly not well represented in the paper.

Another issue with this paper is the lack of reporting of the image-analysis algorithms used. No details are presented and it is unknown whether the results are simply an artifact of the features chosen. Finally, Schindewolf et al. report that Clinical-view images perform marginally better (80.7% correct classification) than dermatoscope images (78.1%), using cross-validated results. Given the predominance of post-1994 ELM-based literature for diagnosis systems, this conclusion appears unlikely, and indeed contradicts the thesis argument.

For these reasons, it is therefore considered that this research will make a valuable contribution to this field, a statement backed up by Dr. Scott Menzies of the Sydney Melanoma Unit.

> "While I have no doubt that ELM will be superior to CV images in diagnostic algorithms it has never been proven, although Schindewolf examined it to some extent. So I think your study is worthwhile" (Personal Communications: Menzies 1998)

2.3 What was actually done in this research?

This research was concerned with comparing the use of Clinical-view and ELM images in an automated screening system. In order to establish (or otherwise) the thesis argument, we need to provide two different classifications systems, one for classifying lesions based on Clinical-view images, and the other for classifying lesions using ELM images. These systems are based on a set of image analysis algorithms which provide the data for the classifier. Using these two systems, we can investigate

the two different techniques of implementing a melanoma screening system that we have identified.

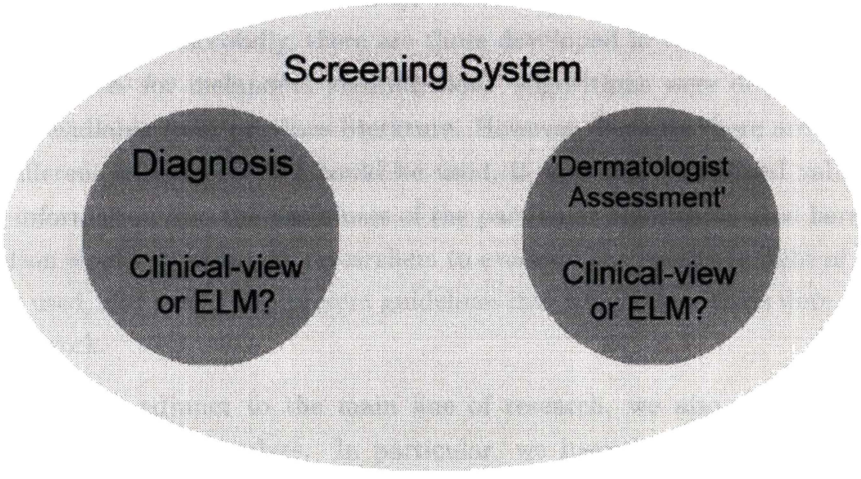

Figure 2.3: Pictorial summary of the thesis. The question of Clinical-view versus ELM is looked at in two contexts, automated diagnosis and automated 'dermatologist assessment'.

The first technique was the standard diagnosis method reported on by previous literature. That is, how well histologically confirmed melanoma can be distinguished from histologically confirmed benign lesions using an automated system. The problem of automatically classifying lesions into benign and melanoma groups is known as the diagnosis problem, and is the subject of the first investigation.

The second method was concerned with reproducing the perception of dermatologists. As has been pointed out previously in this chapter, the major function of dermatologists is to recommend lesions for excision, not to diagnose lesions. What we wanted to see was firstly how well we could reproduce the perception of dermatologists regarding the decision to excise a skin lesion, and secondly, whether this reproduction could be better achieved with the Clinical-view or ELM system. This problem, of reproducing the assessment of excision reached by dermatologists, is referred to here as the 'dermatologist assessment' problem, and is the subject of the second investigation in this research.

If it can be shown that differences in the performance of the Clinical-view and ELM systems exist, then we can answer the thesis question with regard to each of these

two investigations. However, as will be discussed in Chapter 5, there are a number of limitations that must be placed on the results of this research. Because of these limitations, the nature of the results will be indicative, rather than definitive.

An additional investigation is also performed in this research. The image analysis techniques consist of two types. Firstly, we have those that have been presented in previous research. Secondly, there are those developed in this research based on human guidelines for melanoma identification. Algorithms were developed where none were available from previous literature. However, because there are any number of different techniques that could be used, it was also considered valuable to provide information into the usefulness of the particular algorithms used here. This information would allow future researchers to evaluate the results in light of the algorithms used, and would also present guidelines into which algorithms were suitable for future work.

Therefore, as an adjunct to the main line of research, we also present research into the algorithms themselves. In particular, we investigate the relationship of the algorithms to the human perception on which they are based. The algorithms used in this research are mostly based on human criteria, in particular the ABCD criterias of Friedman et al. (1985) (Clinical-view) and Stolz et al. (1994) (ELM view). Therefore, it was considered useful to evaluate how well the algorithms reproduced human perception of these criteria. This was the third investigation carried out in this research, The results of this investigation will indicate whether the data contained in both Clinical-view and ELM images is being captured by the image analysis algorithms, and will also indicate which set of image analysis algorithms is 'better', when compared with the perception of human experts.

To summarise the major steps in this research:

1. Two image sets were obtained for use in developing the systems.
2. Two sets (Clinical-view and ELM) of image analysis algorithms were developed. Most of these algorithms are based on human guidelines, such as those presented in Chapter 1. The Clinical-view algorithms are mostly obtained from previous literature. The ELM algorithms are original variations on previous algorithms. All together, 20 Clinical-view features and 30 ELM features were algorithmically measured.
3. The algorithms were applied to the images and results were classified. We could therefore establish whether in fact the ELM system performed better than the Clinical-view system.
4. Two sets of investigations were performed. The first investigation set tests the performance of the Clinical-view and ELM systems when applied to the

diagnosis problem, that is the problem of identifying melanoma in a set of images. The second set of investigations assesses how closely each of the systems can reproduce the assessment of dermatologists. In particular, the systems are attempting to reproduce the decision to excise a lesion. This problem is referred to as the 'dermatologist assessment' problem.

5. The results of the systems are compared within each investigation. A determination of whether significant differences in the results of the Clinical-view and ELM systems for each investigation is then made. The thesis argument is then resolved given this data, firstly for the diagnosis problem, and then for the 'dermatologists assessment' problem.

6. The image analysis algorithms based on human perception are tested to evaluate whether or not the algorithms reproduced the perception of dermatologists.

2.4 Chapter Summary

In this chapter, some of the relevant literature concerning melanoma has been reviewed. It is clear from the literature that melanoma is a problem worldwide, and the incidence of melanoma is increasing. Early detection of melanoma is crucial to decreasing mortality rates. Population screening has been proposed, but little support is gathered in the literature due to the nature of melanoma, costs involved, and the lack of a reliable method of detection.

However, if an automated method of screening could be developed, costs would decrease as less clinical time was used. Therefore, benefits from screening would become more attractive. Hence, it may be useful to "call in the computer!". From the review of computer applications in this field, we saw that ELM-based systems are more "in vogue" currently. However, little reason for this has been proposed, other than the fact that humans have the ability to perform more accurately using ELM. It is certainly not clear whether this improvement is similar for automated systems. We therefore discovered the thesis of this research, which is an investigation into whether ELM images are better than Clinical-view images for a screening system. This question is looked at in the traditional context of diagnosis systems, and also in the original context of a system based on reproducing the assessment of dermatologists.

Now that the context and focus of the research has been established, we need to look at how automated systems are built. The next chapter continues the review of the literature, concentrating on research concerning the components of an automated system for classification of skin lesions. In particular, the components of the systems reported in Table 2.2 are examined, and specific techniques used in those papers are reported.

Chapter 3
Automated Techniques Review

All of the automated systems presented in the previous chapter share three components. These components are: Image segmentation, the removal of irrelevant image data; Feature analysis, the identification of the features on which to base the classification; and Classification, the method of classifying the features obtained previously (See Figure 3.1).

These three components are the essential parts of an automated system for melanoma screening. To build such a system, we require some idea of how to go about building these three components. This chapter reviews the automated diagnosis system literature presented in the previous section, focussing on the techniques used for each of these three components. As an introduction, the review papers by Hall et al. (1995) and Stoecker et al. (1995) are an excellent starting point.

3.1 Image Segmentation

The first step a human clinician takes in the detection of melanoma is to actually find the lesion! Computer systems need to locate the lesion as well. For a human,

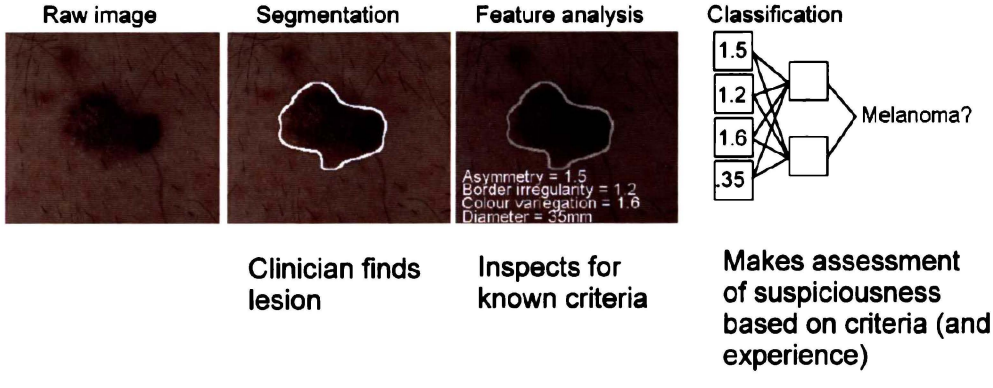

Figure 3.1: Comparison of the steps involved in automated detection of melanoma with the equivalent steps carried out by a clinician.

the task of finding the lesion is usually simple, and it takes place without thought. For a computer on the other hand, finding the lesion in a skin lesion image is a very difficult problem.

To understand why, remember that digitised images are simply large arrays of numbers. So the task is to actually identify one group of numbers (those that represent lesion) from another group of numbers (those representing skin). The general problem of grouping similar areas in images is called image segmentation.

As would be expected for such a fundamental problem, there are a large number of proposed techniques. Numerous texts contain introductions to these techniques, for example Rosenfeld & Kak (1976), Castleman (1979), and Gonzalez & Wintz (1987). Reviews of this literature which try to assess the 'best' technique are also common, for example Pal & Kak (1993) and Glasbey (1993). It is apparent however, that no 'globally effective' method of segmentation exists.

In most cases, techniques are quite context-specific and researchers try to use knowledge about the context to develop a suitable segmentation scheme. An example in this context is reported in Golston et al. (1990), who proposed a 'Radial Search' algorithm that makes use of luminance as a means of detecting the boundary of the lesion. The algorithm locates a point inside the lesion boundary, and then searches along lines at constant angle intervals for sustained increases in luminance. The use of context-knowledge allowed Golston et al. (1990) to determine that on average, lesions are darker than the surrounding skin. Most of the techniques in skin lesion image segmentation make use of this assumption. Golston et al. report 85% success rate for their technique (17 out of 20 images). However, the algorithm has an obvious restriction. One of the criteria for a sustained increase in luminance is that the radial lines can only cross the boundary point once. This restriction means that excessively irregular lesions may have segments of their boundary missed.

Another technique for lesion image segmentation was proposed by Ercal et al. (1993). This technique used a preprocessing step to roughly identify the two areas of the image (lesion and skin). Once both these areas were located, they were analysed to provide the final segmentation rule. This rule is used to identify the actual areas of skin and lesion. The paper reports an 82% success rate for this approach on 61 images.

Ngan & Coombs (1994) present similar work in the area of segmenting gray-scale skin lesion images. Again, they assume that skin lesions are generally darker than the surrounding skin. If a gray-scale image is modelled as a topographical surface, darker regions can be seen as 'pits' or 'basins' in the surface. When this surface is illuminated by 'algorithmic light', shadow regions containing the region-of-interest can be observed. Preliminary example images were presented, but no definitive

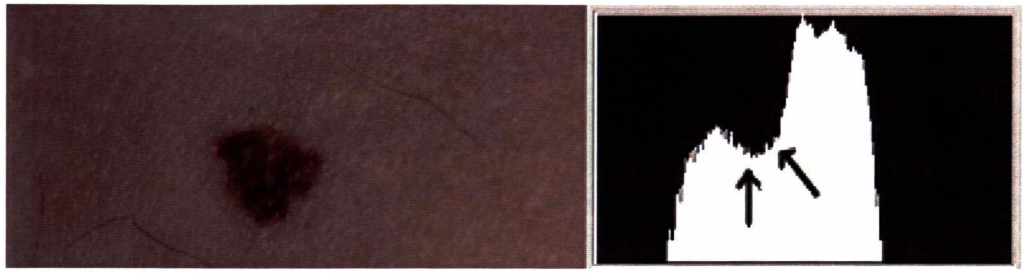

Figure 3.2: Thresholding example. The histogram represents the luminance component of the lesion image. The question in thresholding is where to place the arrow, so everything to the left of the arrow is lesion, and everything to the right is skin. How do we decide where the arrow should go?

results were available.

One of the most common techniques for segmentation is based on thresholding. With thresholding, it is assumed that the areas to be separated are different colours (or intensities for grey-scale images). Thresholding techniques are concerned with locating the optimal separation colour (threshold) between the areas. For example, it is normally assumed that lesions are darker than skin, so a thresholding algorithm would look for the 'best' colour signifying the difference between lesion and skin (Figure 3.2).

Day & Barbour (1996) presented an initial investigation into thresholding techniques for segmentation of skin lesion images. Four techniques based on thresholding algorithms were developed. The best results were obtained by a technique based on Minimum Error Thresholding (Kittler & Illingworth 1986). This technique achieved an average deviation of 9.6% from the hand-drawn boundary on a set of 16 images. A similar comparison was reported in the review paper by Stoecker et al. (1995). Six different techniques were examined on a set of 66 lesion images. Results varied, but the best techniques, the Radial Search algorithm (described above) and the PCT/Median Split algorithm (Umbaugh 1990) showed only a small proportion of lesions with less than 20% deviation from a pre-specified 'ideal'. These poor results highlight the difficulty of segmenting lesion images.

It is apparent that all of the diagnosis systems presented previously require some method of segmentation. Below, we examine each of these systems to find out how they segmented the lesion images (Table 3.1).

Dhawan (1988) used a custom method of segmentation based upon adaptive thresholding. Two windows were found, one within the lesion and one outside, similar to Ercal et al. (1993). Based on analysis of these regions, a region-growing method (see Russ 1999) was utilised to develop a binary mask of the image. The performance of this algorithm was not described, and it is not clear how many images

Table 3.1: Summary of segmentation in automated melanoma diagnosis system research. "Partial" means some details of the procedure are given, but not enough to reproduce the technique. Notes: * Schindewolf et al. (1994) uses the same procedure as described in Schindewolf et al. (1993a and b). † Two sets of images were used in a comparative study between digitised slides and directly digitised images.

Author	Date	Number of Images	Segmentation Description	Described?
Pre-1995 research				
Dhawan	1988	Not avail.	Automated	Partial
Green et al.	1991	70	Automated	Partial
Cascinelli et al.	1992	169	Automated	No
Schindewolf et al.	1993a	353	Automated	Yes*
Schindewolf et al.	1993b	320	Automated	Yes*
Bostok et al.	1993	124	See Claridge et al. (1992)	Yes
Ercal, Chawla et al.	1994	240	Manual	Yes
Green et al.	1994	164	Automated	Partial
Ercal, Lee et al.	1994	399	Manual	Yes
Schindewolf et al.	1994	404/309†	Automated	Yes*
Post-1994 research				
Andreassi et al.	1995	430	Automated	No
Hintz-Madsen et al.	1996	180	Not Described	No
Menzies et al.	1997	170	Automated	No
Gutkowicz-Krusin et al.	1997	104	Automated	Yes
Horsch et al.	1997	110	Automated	No
Seidenari et al.	1998	917	Automated	No
Bischof et al.	1998	221	Automated	Partial
Binder et al.	1998	120	Automated	Partial
Landau et al.	1999	71	Manual	Yes
Seidenari et al.	1999	461	Automated	No

were processed.

Green et al. (1991) and Green et al. (1994) used thresholding techniques, where the histograms of the red, green and blue planes of the lesion image were analysed for the ideal thresholding point. Their technique is only partially described, and reproducing the method would be difficult. Again, no results concerning the success of the technique are provided, although 70 lesions were analysed with their system, which implies that segmentation was successful on these images.

Similarly, Cascinelli et al. (1992) give few details of their technique. They report that the image is subject to edge-enhancement algorithms which "increases contrast and detects the edge of the lesion". Again, no results concerning segmentation are reported, but 169 lesion images were analysed.

Schindewolf et al. (1993*a*) and Schindewolf et al. (1994) use the CIE-DIN standard colour system for segmentation. Two values were obtained for each pixel, brightness (Y) and chromaticity (z). This data, graphed for the entire image created a two-dimensional scattergram, where the cluster with small Y and z values represented the lesion. For this cluster, standard deviation (SD) was calculated and a 'threshold line' was obtained within one SD of the cluster centre. This line divided the graph into two, one area representing lesion, and the other skin. They report that using this technique, 56% of the borders were initially correct. After one interactive step where the threshold line was moved, a correct border was found for 86% of lesions.

Lee (1994) and hence, Ercal, Chawla, Stoecker, Lee & Moss (1994) and Ercal, Lee, Stoecker & Moss (1994) used dermatologist- drawn boundaries. A low (80%) success rate for automated drawing is the reason given for the use of manual boundary drawing. No description of the drawing technique is given. Landau et al. (1999) also use manual borders, because "(obtaining borders) automatically could fail to define unclear borders with sufficient precision".

As stated previously, there are few examples of ELM-based segmentation algorithms. In most cases, the images and algorithms used are commercially sensitive. However, some details are available. In Bischof et al. (1998) for example, two images are captured. The first contains the lesion and some skin area, while the second shows only skin. The second image is then analysed for colour statistics, which are then used to segment the first image. Bischof et al. analyse an image set of 221 images, suggesting reasonable success for this technique.

Gutkowicz-Krusin et al. (1997) used a simple technique based on thresholding. The blue plane of the images was analysed, and the minimum value between the two peaks in the histogram (one peak for skin, the other for lesion) was used as a threshold. There were no results reported that describe the generality of this technique,

although Gutkowicz-Krusin et al. used 104 lesion images in their system. Binder et al. (1998) also use thresholding techniques, but again the technique is only partially described. They report that only 5% of the 120 lesions segmented were not segmented correctly.

Horsch et al. (1997) state that the problem of lesion image segmentation has "been solved very successfully by means of histogram analysis in a transformed colour space". They report that of 110 ELM images, 70% were correctly segmented, with a further 17% requiring manual correction. The remaining 13% were not segmented correctly. The colour space and segmentation technique used were not described. Again, poor reporting prevents evaluation of the usefulness of the technique.

Andreassi et al. (1995), Seidenari et al. (1998) and Seidenari et al. (1999) also give few details of their segmentation procedure. They state that an edge following algorithm is used. Unfortunately, the use of an edge following algorithm requires some judgement to be made as to what constitutes an edge. It is this judgement which is the difficult part of the problem, and it is unfortunate that this important step is not reported. They do state that complex cases require correction after segmentation, although again, it is not clear from the report how this correction occurs.

3.1.1 Section Summary

It is apparent that most of the lesion image segmentation techniques presented here have some difficulty in segmenting Clinical-view lesion images. This difficulty is illustrated by the comparison of techniques reported in Stoecker et al. (1995), which shows quite poor results, even for the best techniques. Thresholding techniques are popular for use in skin lesion image segmentation, with a number of the reviewed papers using some form of thresholding. Several of these papers however (for example, Green et al. 1991, Green et al. 1994) fail to include any results of the segmentation phase, other than implying that segmentation was successful. The use of thresholding in this context shows the application of a-priori knowledge, namely that skin lesion areas tend to be darker than the surrounding skin.

However, in regard to thresholding techniques, Golston et al. (1990) state:

> "Although adaptive thresholding can be highly successful in certain domains, features such as variegated colouring in malignant melanoma ... make this an unreliable method for differentiating skin tumours from normal skin"

It should be noted that the Radial Search algorithm proposed by Golston et al.

(1990) is amongst the most successful segmentation techniques for skin lesion segmentation. Other papers however appear to use thresholding techniques successfully and there is little actual evidence to support Golston et al.'s (1990) statement.

Most of the segmentation techniques examined in this section are intended for use on Clinical-view images. It is apparent that the problem of segmenting Clinical-view lesion images is difficult. For ELM images however, further issues arise. In the ELM ABCD criteria proposed by Stolz et al. (1994), the 'B' stands for 'sharp border contrast'. 'Border contrast' is measured by assessing how sharply the lesion fades into the skin (Figure 1.8 on page 9). A sharp demarcation in any of the octiles is a possible indicator of malignancy. The corollary of course, is that a blurred border, or an area that is difficult to separate from the surrounding skin is a sign of benignancy. This corollary has important implications for segmentation algorithms, as the majority of lesions analysed are expected to be benign, and hence have blurred borders.

Once the lesion has been isolated, by whatever method, the area of the lesion can be analysed to find features that may be of use for classifying the lesion. The next section introduces literature related to feature analysis, and reviews some of the features used previously in skin lesion classification.

3.2 Feature Analysis

In order to classify an image, we need to be able to identify relevant features that suggest a class, for example melanoma or benign. When humans wish to classify a lesion, one of the techniques used is to investigate for criteria known to be relevant. For example, a clinician may investigate the lesion for the ABCD criteria of Friedman et al. In this way, the clinician is identifying features (the ABCD criteria) and assigning 'values' to each feature. These values could be 'very asymmetric', 'little border irregularity' and 'high colour variegation' for example.

An automated system needs to perform a similar function. The system must analyse the image of the lesion for specific features, and obtain values for each of these features, prior to attempting classification. The clinician already knew what features of the lesion were most relevant due to past research and publication. Automated systems do not yet have this advantage.

Therefore, we must look at the available techniques and attempt to obtain some algorithmic measures that are relevant to the identification process.

Castleman (1979) states:

> "If we desire a system to distinguish objects of different types, we must first decide which parameters, descriptive of the objects, will be measured... Proper selection of the features is important, since only they will be used to distinguish the objects".

Most introductory image processing texts, for example, Castleman (1979), Rosenfeld & Kak (1982), Gonzalez & Wintz (1987), Russ (1990) and Russ (1999), present a range of feature analysis algorithms. Russ (1999) is especially thorough, presenting a large number of disparate methods. So which of these algorithms should be implemented?

Bischof et al. (1998) present a discussion on the approaches to selecting relevant features in the skin lesion context. They identify two approaches to selecting features, the "expert system" approach and the "computational" approach. The "expert system" approach focuses on algorithmic reproduction of techniques used by clinicians for melanoma detection. The "computational" approach relies on algorithms that are "easily extracted and measured by computer". Bischof et al. suggests that the drawbacks of the expert system approach is "that extracting expert knowledge is a notoriously difficult problem" and that "an expert system, by it's very nature, cannot hope to do better than the best skin specialist". The drawbacks they associate with the computational approach is that "it is difficult to design a set of features that will effectively diagnose melanoma". They recommend compromising and using techniques derived from both approaches. Although the use of an algorithm simply because it is "easily extracted" suggests that classification of objects is domain independent when it is obviously not, it does not hurt to implement a large number of algorithms. However, some robust method of measuring how useful an algorithm is would then be required, in order to weed out those algorithms that are irrelevant.

In skin lesion research, most feature analysis algorithms are developed using the "expert system" approach, in theory at least. For example, in the Clinical-view context, algorithms are for the most part based on the ABCD criteria of Friedman et al., despite this set of criteria not being proven clinically. Similarly in ELM view, algorithms (when reported) seem to be derived from the large set of ELM specific criteria available, for example those described in Stolz et al. (1994) and Menzies et al. (1996). An advantage of these criteria is that they tend to have been proven clinically to aid the adoption of the ELM technique. Algorithm designers therefore have a known target to aim for.

The literature concerning feature extraction from skin lesion images is reviewed below, again presented in roughly chronological order. We focus on the diagnosis systems presented previously.

As we have seen, the ABCD criteria of Friedman et al. (1985) is a popular basis for

Clinical-view algorithms. For example, Stoecker et al. (1992) and Gutkowicz-Krusin et al. (1997) present a method to measure asymmetry in binary images. Russ (1999) also describes this method in a general context (Orientation - page 518). Stoecker et al. report that this technique agreed with the dermatologist in 93% percent of cases. The area and diameter of the lesion are also easy to measure, and details can be seen in Schindewolf et al. (1994) and Green et al. (1994) for example. Some papers do not implement these two measurements, as they are highly dependent on the resolution of the image (for example, Ercal, Chawla, Stoecker, Lee & Moss (1994) and Ercal, Lee, Stoecker & Moss (1994) use dermatologist rating for area).

Border irregularity has also been well covered by the literature. Claridge et al. (1992) presents three algorithms designed to measure border irregularity. One of the most popular methods of measuring lesion border irregularity is the ratio of the area of the lesion to the perimeter of the lesion (referred to as Irregularity index in this research). Green et al. (1991), Cascinelli et al. (1992), Green et al. (1994), Ercal, Chawla, Stoecker, Lee & Moss (1994) and Ercal, Lee, Stoecker & Moss (1994) have all used this measure to describe border irregularity. Golston et al. (1992) investigated the use of this algorithm as a method for measuring border irregularity and found that 87% of lesion images analysed were classed as irregular both by the algorithm and by a dermatologist. Horsch et al. (1997) also investigate this method. Their research however, evaluated this method on both Clinical-view and ELM images, and concluded that border irregularity of melanoma is greater than that of atypical moles. They also conclude that this difference is clearer using ELM images, rather than Clinical-view images.

In general, skin lesion diagnosis systems use a wide variety of features to measure relevant aspects of the lesions. Below, we examine the features used in Clinical-view diagnosis system research, and find that most of the features are loosely based on the ABCD criteria of Friedman et al. (1985). The ELM literature is then examined.

Green et al. (1991) presented a lesion classification system that used a number of different features based mainly on the ABCD criteria of Friedman et al. (1985) They state that the mean and variance of red, fragmentation index (identical to irregularity index), area and perimeter features were the most important features in classifying lesions. Not surprisingly, the last three features were quite significantly correlated to each other, suggesting an overlap in information.

Green et al. (1994) used similar features in their follow-up work, although their feature set also included: means and standard deviations of the colour gradient at the boundary points; the change in area when thresholding point is reduced by ten percent; and the change in the fragmentation values at these two thresholds.

The usefulness of each feature was obtained, and again, the most useful features were the size features (area and perimeter) and irregularity index. They also show colour and gradient standard deviations to be useful to the classification problem. Green et al. (1994) also report algorithmic colour features were correlated with clinical perception of colour, but not greatly, while perimeter and fragmentation were significantly correlated with clinical perception of border irregularity. This work is relatively unusual in that an attempt was made to investigate how well the algorithms reproduced human perception.

The ABCD criteria again formed the basis for the features used in Cascinelli et al. (1992). These features included shape features such as the Circularity Index (identical to Irregularity Index) and the minimum covering rectangle. Texture features, almost unique to this paper are also included, and are found using Fourier Analysis. They do not describe how useful each of the features was to the classification process.

The ABCD criteria of Friedman et al. also forms the basis for the features used in Schindewolf et al. (1993*a*), and are similar to those presented previously. However, they also describe the implementation of texture features. These features were based on the well-known Sobel Operator and are intended to measure the 'roughness' of the lesion surface. Usefully, Schindewolf et al. (1993*a*) report on the relative importance of thirteen of the most useful features. Asymmetry, shape and border features were the most important, with colour and texture only contributing one feature each.

In what is perhaps the most comprehensive Clinical-view research in this field, the thesis by Lee (1994) describes over twenty different features extracted from skin lesion images. This thesis forms the basis of Ercal, Chawla, Stoecker, Lee & Moss (1994) and Ercal, Lee, Stoecker & Moss (1994). The feature set is mainly based on the ABCD criteria, but other "computational" features are also included. The first two features, Irregularity index and Asymmetry, are described above. For colour, Lee used the variances of red, green and blue planes (equivalent to the standard deviation measures used in Green et al. 1994) to measure variegation, but also includes a large number of other colour features, including sixteen relative colour features. Relative colour features are intended to allow for differences in colour between lesion images. Such differences may come about because of different image acquisition techniques. These features can be broken into three sections, chromaticities, ratios, and colour differences. Chromaticities of red, green and blue describe the difference between the normalised average tumour colour and the normalised average skin colour. Red, green and blue ratios were found by dividing the average red, green or blue tumour colour by the average red, green or blue skin colour. Colour differences measure the difference in colour between the lesion and skin. The final features used by Lee are measures of area and elevation.

Lee (1994) then assessed the usefulness of these features using Pearson correlation coefficients. This statistic is perhaps not adequate, when you consider that Pearson correlations are intended to measure the linear relationship between two continuous (or sometimes ordinal) variables. In the case where one of the variables is dichotomous (for example, melanoma and benign), the Pearson correlation coefficient can obtain values greater than |1|. To avoid such spurious results, and potentially meaningless correlation values, it may be more suitable to use difference of mean tests (such as t- or Mann-Whitney U tests) to measure the strength of the relationship. However, bearing these deficiencies in mind, Lee reports that Irregularity index, Asymmetry, Variance red and Variance green are highly correlated to melanoma. This thesis is an excellent introduction to a wide variety of image analysis features used in the context of skin lesions.

Tomatis et al. (1998) present recent research into feature analysis of 40 Clinical-view skin lesion images. They do not attempt a classification, but report a number of features showing significant differences between benign and malignant lesions. In particular, average hue and average saturation values were higher for melanoma than for benign lesions. Interestingly, they measure 'roundness' ($\frac{4\pi}{P^2}$ which is equivalent to the Irregularity index measure reported previously) for lesion silhouettes from each of the red, green, and blue planes, and report that none of these values are significantly different at the 99% level. This result suggests that Irregularity index is not useful, contrary to the results presented by Lee (1994), Lee (1994) and Green et al. (1994).

Finally for Clinical-view research, the work by Landau et al. (1999) uses hue standard deviation for classification. They report significant difference between benign naevi and malignant (melanoma and basal cell carcinomas) for this feature (p=0.05). In particular, they conclude that hue standard deviation is significantly higher for malignant lesions. However, doubt must be raised about their conclusions due to the lack of standardisation of the images, and the low number of images in their dataset (52 benign naevi, 7 melanomas, and 2 basal cell carcinomas).

It is apparent that Clinical-view details of algorithms are reasonably abundant. The same can not be said of the ELM literature. Although Andreassi et al. (1995), Gutkowicz-Krusin et al. (1997), Seidenari et al. (1998), Binder et al. (1998) and Seidenari et al. (1999) all provide some descriptions of their algorithms, these descriptions tend to be lacking in detail. This literature is reviewed below. As noted previously, Andreassi et al. (1995), Seidenari et al. (1998) and Seidenari et al. (1999) (which are reports on the same system) appear to be based on Clinical-view criteria rather than those intended for ELM use.

The features used by Gutkowicz-Krusin et al. (1997) do not have this problem. Their

feature set appears based on the ABCD criteria for ELM reported in Stolz et al. (1994), although this is not explicitly stated. In summary, algorithms for measuring asymmetry (as described above), border contrast, border irregularity, texture and colour are presented. Of interest are the unique algorithms presented, namely border contrast, colour and texture algorithms. Unfortunately, inadequate reporting of the algorithms makes evaluating their suitability impossible, although Gutkowicz-Krusin et al. report some good results.

Similarly, Horsch et al. (1997) use algorithmic equivalents of the ABCD criteria of Stolz et al. (1994). To date, only the 'B' and 'C' components have been implemented, but no details of the implementation of these algorithms are reported. Binder et al. (1998) also report few details of their algorithms. Most of the algorithms appear to be of the "easily extracted and measured" type, rather than based on any set of ELM criteria. They do present a rough attempt to assess the contribution of each feature, and report that the number of colours, and border measurements contribute the most to classification. In a similar situation, the final ELM papers from Table 2.2, Menzies et al. (1997) and Bischof et al. (1998), also do not report their algorithms. Both of these papers concern the development of the Skin PolarProbe(tm), and reporting is therefore constrained by commercial interests.

3.2.1 Section Summary

The choice of image features is very important to the classification process, as these features represent all the data by which the image is to be classified. It is apparent from the above review that there are many different features that can be extracted from skin lesion images, and it may be difficult to establish which are most useful. In choosing features, there should be some justification for selection. In most of the papers presented above, the algorithms are justified by being based on human criteria, for example the ABCD criteria of Friedman et al. (1985). However, little work has gone into establishing whether or not these algorithms reproduce human perception of those criteria, with exceptions being the work by Golston et al. (1992) and Stoecker et al. (1992). Another basis for selection may be the results of previous research. To date however, there have been few attempts to assess features for suitability in this context.

Once relevant features have been decided upon, and the images analysed to obtain this feature set, a classification method (classifier) is required. The classifier assigns each lesion to either a malignant or benign (or 'excised'/'not excised') group on the basis of the feature set.

3.3 Classification

There are a large number of techniques associated with classifying patterns into sets, or matching patterns to corresponding outputs. Classification of patterns and matching patterns to outputs will be collectively referred to as the classification problem. This problem is a standard problem of artificial intelligence, as indicated by the large volume of related literature. Ripley (1996) presents a wide variety of methods, including machine learning techniques, statistical classifiers and neural networks. Bischof et al. (1998) present a concise introduction to the classification problem from the automated diagnosis point of view. We look at the classification problem in detail in the next chapter. This section again focuses on techniques used previously in automated skin lesion diagnosis systems. The research shown in Table 2.2 is grouped into three major sections, Artificial neural networks, Statistical methods, and Rule induction classifiers depending on the primary classification technique used. These are discussed in turn below.

3.3.1 Artificial Neural Networks

Artificial neural networks (ANNs) represent a popular paradigm in machine learning. Since their rediscovery in the 1980's artificial neural networks have been applied to a wide variety of disparate problems, often with great success. ANNs are networks of units that provide a non-linear mapping between input and output. The network requires training, where a large number of input patterns are shown to the network. In the case of supervised training, weights within the network are modified iteratively until the network can produce the correct output from a given input. 'Supervised' refers to the 'correction' of the network when the wrong output pattern is produced. In unsupervised training, the network is left to group the input patterns without correction. The strength of neural networks lies in their ability to match patterns with noisy input, something that makes them favoured when dealing with real-world data.

In the context of automated skin lesion diagnosis, several papers have reported using ANNs. The papers derived from Lee (1994), namely Ercal, Chawla, Stoecker, Lee & Moss (1994) and Ercal, Lee, Stoecker & Moss (1994) have been looked at previously in this chapter. They present a diagnosis system using a neural network as the classification technique. In Ercal, Chawla, Stoecker, Lee & Moss (1994), a single neural network was used, while in Ercal, Lee, Stoecker & Moss (1994), a hierarchy of networks, each trained to diagnose a sub-group of lesions, as well as a rule based 'pre-screen', were used. The pre-screen attempted to screen out those lesions that were obviously benign based on rules supplied by a dermatologist. As

noted previously, results from both of these papers were good, with the hierarchical system improving significantly over the single neural network in terms of ability to classify melanoma. Probably the most significant contribution of these papers was the use of multiple neural networks being combined to produce a single output. Such a strategy may have wide benefits for classification of skin lesions, due to the natural division of lesions based on subtype. This strategy is discussed more thoroughly in Chapter 8.

Further research concerning skin lesions and neural networks was presented by Hintz-Madsen et al. (1995). The research described by this paper concerns the design and optimising of neural networks in order to classify skin lesions. They discussed a method of optimisation for neural networks, termed optimal brain damage, and the resulting architecture is then applied to the problem of classifying skin lesions. They report a success rate of 66% in the classification task. In a further paper by Hintz-Madsen et al. (1996), the techniques described above were used to classify a number of skin lesion images. Twenty-one features were used as input for the network, although these were not described. They report results of 54.6% classification of the test set for a fully connected network, and 58.9% for the network optimised through the optimal brain damage technique. Although results were comparatively poor, these two papers represent perhaps the only research in the context of skin lesions looking primarily at the contribution of the classifier in skin lesion identification.

The final piece of research that utilises artificial neural networks is by Binder et al. (1998). They train the neural network using back-propagation with a set of 83 images. The network was then tested on a set of 29 images. Two different problems were looked at, the first distinguishing between benign naevi and melanoma, and the other allowing lesions to be classed as common naevi, dysplastic naevi and melanoma. The network performed satisfactorily on the first test, but failed on the second. No reasons for the choice of artificial neural network as a classifier were given, and from these results, it is not clear whether other techniques may be more valuable, such as the statistical and rule induction classifiers described below.

3.3.2 Statistical Methods

Most methods of classification have some basis in statistics. However, the statistical methods referred to here are those methods that utilise well-known statistical techniques. Obvious candidates include regression, cluster analysis and discriminant analysis. For an introduction to these methods, see Tabachnick & Fidell (1996). In skin lesions research, Green et al. (1991), Green et al. (1994), Seidenari et al. (1998), and Seidenari et al. (1999) use discriminant analysis to good effect. No justification is presented for this choice of classifier in any of this work (although on inspection it

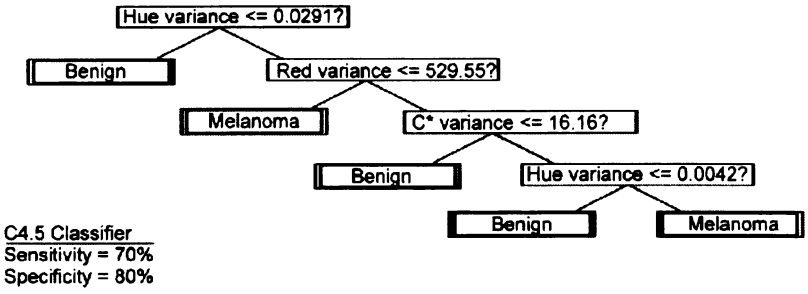

Figure 3.3: Example decision tree for melanoma classification. This tree is the result of applying C4.5, a popular rule induction algorithm, to the Sydney Image Set ELM data. See Section 4.1.1 for a description of this image set.

appears a reasonable choice, see page 32 Tabachnick & Fidell 1996). Similarly, Menzies et al. (1997) and Landau et al. (1999) use logistic regression, with no attempt to justify the choice of method (again though, it appears a reasonable choice).

Gutkowicz-Krusin et al. (1997) used a multivariate linear regression variant as a classification method. The independent variables were the image features, with the dependent variable L being a linear combination of these features. Gutkowicz-Krusin et al. then obtained a threshold on L, such that sensitivity was 100%, that is, all melanoma had L-values above this threshold. Specificity, or the number of misclassified benign lesions were recorded. The choice of linear regression as a classification method is perhaps surprising, given that linear regression requires both independent and dependent variables to be continuous. However, Gutkowicz-Krusin et al. go on to test how generalisable the results obtained are, using resubstitution techniques and blind testing on an independent image set.

3.3.3 Rule Induction Classifiers

Rule induction techniques, sometimes known as decision trees, are popular alternatives to the approaches described above. These techniques algorithmically decide on a set of rules through which every case of the training set can be classified. The rules are generally hierarchical in nature, starting at a 'root' rule, and depending on the answer, descending the tree to the 'leaf' nodes, which assign class. A significant advantage of these classifiers is that the rules produced are generally interpretable by humans, given a set of features that have human equivalents. An example tree is shown in Figure 3.3.

Tree-based methods have been used in a number of papers. Two examples of these are Schindewolf et al. (1993*a*) and Bischof et al. (1998). Bischof et al. give a good introduction into types of classifiers, and use RPART (Recursive Partitioning Tool).

Schindewolf et al. (1993*a*) use the same software, but do not justify their choice.

3.3.4 Section Summary

It is apparent from the literature presented in this section that little consensus concerning a 'best' classification function has been reached. Further, in most cases, little attempt to justify the classifier chosen has been made, and in several cases it appears as though practical issues of classification have been ignored. For example Green et al. (1991) classifies 70 lesions using discriminant analysis. They use 11 image analysis features for the classification. Tabachnick & Fidell (1996) say that such a high feature to cases ratio suggests that the model may be overfitted.

Perhaps the most important result of classification is to have some idea of how generalisable the results are. If 90% accuracy is obtained on an image set, can we expect this result given a different image set? Techniques such as cross-validation are popular for ensuring that results will tend to generalise to a larger population. Cross-validation is discussed in more detail in Section 4.3.3 on page 95 in the following chapter. The 'leave-one-out' method is a special case of cross-validation. However, such techniques are not commonly used in previous research. For example, Green et al. (1991),Green et al. (1994), Seidenari et al. (1998) and Landau et al. (1999) do not use any method of assessing the generalisability of their results, raising considerable doubt about the usefulness of their results. Cascinelli et al. (1992), Ercal, Chawla, Stoecker, Lee & Moss (1994), Ercal, Lee, Stoecker & Moss (1994), Binder et al. (1998) and Seidenari et al. (1999) use blind testing on a single independent set of images. Such testing is an improvement, but results may still be an artifact of the training and test sets.

3.4 Chapter Summary

This chapter has reported on the techniques that have been used previously in automated diagnosis systems for melanoma. These techniques were broken into three sections, corresponding to the major components of such systems. The first section, segmentation, reviewed the methods of identifying the lesion from the skin in an image, while the second section examined the techniques used to quantify aspects of a skin-lesion image (feature analysis). The final section looked at classification techniques used in previous automated melanoma detection research.

The next chapter describes the methods used to implement the systems in this research. Firstly, the acquisition and pre-processing of the image sets is described, and we briefly examine the problem of segmentation. The details of the feature analysis

algorithms are then presented, and finally the choice of classifier is considered.

Chapter 4
Method

As was shown in Chapter 3, a large number of techniques are used in previous automated melanoma diagnosis research. In this chapter, the techniques used in this research are described. Because the purpose of the research is to investigate whether ELM images are more use than Clinical-view images in an automated system, it is necessary to develop two separate sets of techniques for dealing with the two different types of images. In essence, two different diagnosis systems were developed, although in practice, several of the methods are duplicated in the two systems.

This chapter is organised into four major sections. The first section describes the image sets used in this research, and explains how they were obtained. The problem of separating images into skin and lesion areas, or segmentation, is also discussed. The second section describes the image-analysis algorithms used in this research. Firstly, the Clinical-view algorithms are presented. These algorithms are mostly taken from previous literature. Subsequently, the algorithms for ELM-image analysis are presented. These algorithms are mostly new variants on the Clinical-view algorithms, as few details of algorithms used in previous research for ELM-image analysis exist.

Once the lesion images have been analysed, some method of classifying the images is required. The classifier takes the features found by image analysis, and groups the lesion images. The classifier is the subject of the third part of this chapter.

The statistics that are used in this work make up the fourth and final section of this chapter. Firstly, statistics concerning the classifier are presented, and then other statistics used in this work are detailed.

4.1 First Steps

This section describes the steps that were undertaken before the image analysis algorithms could be applied. The focus of this section is on the image sets used in this study. The problem of segmentation is also examined.

4.1.1 Image Sets

Computer image analysis is generally very demanding on the standards of image capture. Previous research in this field has been based on image sets obtained through clinical practice, and there tends to be one for each research project. The standard of images varies considerably, as does the distribution of the lesions in each image set. This disparity in image sets between research projects makes interpretation and comparison of results difficult.

With the advent of dermatoscopy, and technologies such as the DermaPhot(tm) camera which takes standardised ELM photographs, capturing standardised ELM images became a much simpler process. In this section, the image sets used in this research are described.

Sydney Image Set

There are two image sets used in this study. The first set was obtained from Dr. Scott Menzies of the Sydney Melanoma Unit. It consists of one-hundred lesion images. Fifty of these were melanoma, while the remaining fifty images were of atypical naevi of various diagnoses. Both Clinical-view and ELM-view images were obtained. The Clinical-view images were mostly at 1:1 magnification (and the magnification of the others was recorded). When magnification was not 1:1, rescaling in software was conducted. This rescaling resized the image by the desired amount, using a pixel-based method that duplicates pixels already in the image (PaintShopPro 5.01). This technique does not introduce new colours into the image. It should be noted that rescaling of the image does not produce a large quantity of extra information, as most of the measurements gathered are independent of size (with diameter measurements being the only exception).

The quality of the Clinical-view images varied widely (for example, Figure 4.1). In this image set, 14 Clinical-view images were rejected because of poor quality or indistinct lesion boundaries. No basal cell carcinomas were included (three images). Five superficial spreading melanomas, three dysplastic naevi, two seborrhoeic keratoses and one lentigo maligna were also excluded. The ELM images however, were taken with a Heine Dermaphot camera, and therefore were standardised. Ten ELM lesions were removed from this image set because of the difficulty involved in segmentation, and finally, 83 lesions were left with both Clinical-view and ELM images. These 83 lesions make up the Sydney Image Set used in this project. Table 4.1 shows the breakdown of the Sydney Image Set.

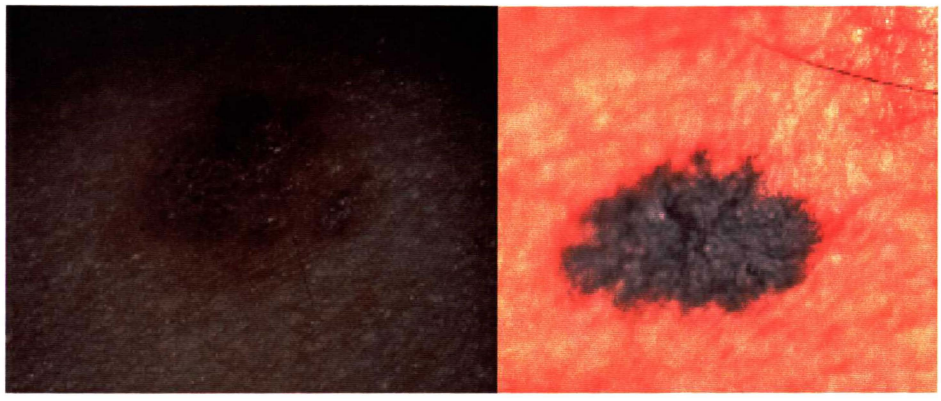

Figure 4.1: Two images from the Sydney Image Set. The figure on the left is under-exposed with shadows, while the figure on the right is over-exposed and blurred.

Table 4.1: Breakdown of the Sydney Image Set

	Melanoma	Melanocytic Naevi	Seborrhoeic Keratoses	Other	Rejected
Clinical-view	50	37	6	7	14
ELM	46	33	5	6	10
Both	42	41 benign lesions			

University/Health-Waikato Image Set (UHWIS)

The second set of images was obtained through a specific image capture project conducted at the Dermatology Department of Health Waikato. For approximately fourteen months, I attended a skin-lesion clinic every week, and took photographs of lesions. Ethical approval for this project was obtained from the Waikato Ethics Committee on July 20^{th} 1998. Approval required informed consent from all participants. Patient confidentiality for participants was also assured, with identifying data only available within the Dermatology Department of Health-Waikato.

For this image set, both Clinical-view and ELM-view images were obtained for most lesions. The Clinical-view images were obtained with the same equipment (Nikon F70 camera, AF Micro Nikkor 60mm lens, Nikon Macro Speedlight Ring flash), at the 1:1 magnification. Ambient lighting conditions were similar for all of the photographs. Several photos (generally two) of each lesion were taken, and the 'best' photograph selected. 'Best' was determined by visual inspection of the slides with a hand-lens. The ELM-images were again obtained with a Dermaphot camera, resulting in standardised images.

This image set contains a wide variety of lesions, from the clearly benign, to unmistakable melanomas. Histology data was not available for the majority of lesions.

In total, 113 Clinical-view images and 94 ELM images were obtained (Table 4.2). It must be noted that in a number of cases, both Clinical-view and ELM images were not available, due to difficulties of a technical or logistical nature. Of these lesions, a number were removed from consideration, primarily due to difficulty in segmentation, and in the case of ELM, lesions exceeding the boundary of the slide. Finally, 73 lesions had both Clinical-view and ELM images. Seven of these lesions were melanoma, with the rest being various other naevi, including three basal cell carcinomas. Sixteen of these lesions were excised by the dermatologists. It should be noted here that different dermatologists were involved in deciding whether to excise the lesions.

Table 4.2: Breakdown of the University/Health-Waikato Image Set. The numbers in brackets are the lesion counts after lesions that exceeded the slide boundary are removed.

	Melanoma	Other Naevi	Rejected
Clinical-view	11	102	9
ELM	11(7)	74(71)	8
Both	7	66	

This image set is contained on the CD found inside the back cover of this book. It is being included in this thesis in an effort to provide a world-wide basis for comparing different image analysis techniques in this context. The major drawback with this image set (apart from the low number of melanoma) is that histological diagnosis was not available for every lesion. In fact, only 16 of the 73 lesions that were used from this image set were excised, and hence had histological data. The reason for this lack of data is simply because the number of benign lesions included in the image set is large. Therefore, excision was not an option for most of these cases.

At this stage, it is worth noting the possible reasons that a lesion may be excised. One of the consulting dermatologists stated:

> "Skin lesions are removed for medical and non-medical (cosmetic) reasons. The publically funded Health Service primarily removes lesions for medical reasons. Such reasons include: malignancy, suspicion of malignancy (i.e. appearance or history consistent with melanoma or other skin cancer), inflammation or irritation. We also may remove lesions which are significantly unsightly or to provide the patient with peace of mind. We are unlikely to remove an obviously benign lesion which is not causing the patient any symptoms or concern." (Personal Communications: Oakley 2000).

The apparent possibility of malignancy is a primary cause of excisions. Other reasons include the cosmetic appearance of lesions and whether the lesion is causing irritation for the patient. It is important for this image set to be reasonably confident of the reasons for excision. If lesions are excised for different reasons, there exists no one 'baseline' set of criteria that can be used to decide whether a lesion should be excised. For example, if one lesion is identified as potentially malignant by its colour distribution and excised, while another is removed for cosmetic reasons, for example due to its location, any automated system would be required to reproduce both of these responses. This requirement would complicate classification enormously, and it would not be clear what the results of such a system would mean. If we obtained 80% sensitivity, does that mean that lesions excised for cosmetic reasons were being identified correctly, or that lesions excised for malignancy reasons were being identified correctly? Obviously, without data concerning the reason for excision, we cannot tell. Additionally, it may be that if the reason for excision is not based on malignancy considerations, the data used to make the decision may simply not be available to the classifier. For example, lesions may be excised for cosmetic reasons based on their location, rather than their appearance. This data is unlikely to be available to a classifier, and will again cause uncertainty regarding the results.

In the case of the UHWIS, we are confident that for the majority of the sixteen excised cases, malignancy concerns were the reason for excision. The rationale for this confidence is as follows: Firstly, seven of the sixteen lesions were melanoma and these lesions were excised for malignancy concerns. Of the remaining lesions, two were excised for other than malignancy concerns. One of these however, was considered 'suspicious' in appearance by at least one dermatologist. The remaining lesions were all excised for malignancy concerns, although none were thought to be melanoma at the time of excision. For this image set therefore, we can be reasonably sure that malignancy concerns was the reason for excision in all but two cases, and that in one of these cases, the lesion was considered 'suspicious' in appearance.

4.1.2 Guidelines for Images

With the UHWIS, I had control over the image capturing process. Therefore, it was possible to specify a set of guidelines for the images captured. These criteria were designed to minimise the range of features displayed by the image set, thereby simplifying the classification task to some degree.

- **Lesion position**. Lesion images should be taken from a flat surface, for example, the back, rather than curved areas, such as the face, hands, and feet. This guideline is likely to rule out classification of lentigo maligna and acral-lentiginous melanoma.

- **Lesion pigmentation** Lesions should be pigmented. This constraint rules out most basal cell carcinoma and also amelanotic melanomas, which do not exhibit pigmentation.
- **Nodular Lesions**. Excessively nodular lesions are not catered for, particularly at the Clinical-view. This constraint may rule out nodular melanoma, some seborrhoeic keratoses, as well as other nodular lesions.
- **Artifacts**. Hair, rulers and other introduced artifacts should be minimised. Bubbles (for ELM images) should be minimised where possible.

To summarise, the images used in this research, both clinical and ELM, are in general examples of pigmented lesions located on flat areas of skin.

Scanning

The Sydney Image Set was already scanned when it was made available. The original slides were not available, and no direct comparison could be made. With the UHWIS, scanning of the slides using a MicroTek 35T slide scanner resulted in some scanned images that were considerably different to the original slides. In particular, each image was much darker than the original. Because relative colour features are used (described later in this chapter), differences in lighting should not be significant. Relative colour features "should equalize any variations caused by lighting, photography/printing, or the digitization process" (Umbaugh 1990). However, differences in colour balances between images were also of concern.

To test for colour balance differences, a number of reference white photographs (both clinical and ELM) were obtained under the same conditions as the original slides. The lesion slides were adjusted using the 'corrective adjustments' procedure of Herbin et al. (1990). This procedure involved finding the mean red, green and blue values for the reference white images ($\bar{R_w}, \bar{G_w}, \bar{B_w}$). Each pixel in each of the images was then adjusted by:

$$X = \max\{\bar{R_w}, \bar{G_w}, \bar{B_w}\} \tag{4.1}$$

$$R_{new} = R_{old} * \frac{X}{\bar{R_w}} \tag{4.2}$$

$$G_{new} = G_{old} * \frac{X}{\bar{G_w}} \tag{4.3}$$

$$B_{new} = B_{old} * \frac{X}{\bar{B_w}} \tag{4.4}$$

This correction (Equations 4.1 - 4.4) is referred to by Herbin et al. (1990) as "the Global Correction" (Page 263 Herbin et al. 1990). This correction is intended to normalise the colour balance over the set of images. For example, in the UHWIS ELM-images, the white reference photo had a much larger blue component than either red or green (the average value of the white reference photograph was (169,185,237) for red, green and blue respectively). Given that white is meant to be achromatic, it would be expected that the red, green and blue components would all be (roughly) equal. The equations presented above 'shift' the red and green components to equal the blue component for the white reference. This shift is then applied to each of the lesion images. In the Clinical-view case however, little difference was noted between the components of the white reference, and thus the above equations had negligible effect.

Image Size

One of the major difficulties with ELM images is the tendency for the lesion to be greater in size than the boundaries of the slide. This is especially true in cases of melanoma and congenital naevi, which tend to be larger than other lesions.

To deal with this problem, the image analysis algorithms developed for the ELM images were examined. They can be broken into two categories, those that depend on shape, and those that are shape independent. The major shape dependent algorithms are the asymmetry algorithms. The Border contrast algorithm is also shape dependent as the true border of the lesion is required. The remainder, that is colour and differential structures, are shape independent. Shape dependent algorithms require the entire lesion to produce acceptable results. For example, it would be misleading to obtain asymmetry features for lesions exceeding the slide boundary, as a large portion of data may be missing. However, shape independent features can produce reasonable results from these images.

Menzies et al. (1997) and Bischof et al. (1998) restrict their image set to those lesions that fit completely in the slide. The image sets in those two papers were much larger than is available here. In this research, the SIS contained a large number of lesions that were larger than the slide boundary. Therefore, ELM shape dependent features were not used for the SIS image set. The University/Health-Waikato Image Set was also restricted to lesions that were mostly contained within the boundary of the slide, as only a small proportion of lesions were larger than the slide boundary. All features were used for this image set.

4.1.3 Segmentation

As was seen in the last chapter, no accurate method of skin lesion image segmentation exists to date. Segmentation methods are often not reported, or if they are, they are reported so poorly that reproducing the algorithm is impossible. This treatment undermines the importance of image segmentation.

Following a small investigation reported in Day & Barbour (1996), we noticed several important effects that the choice of segmentation technique may have. Perhaps most importantly, different techniques may alter the results of image analysis algorithms, by producing significantly different lesion areas. Another effect may be that of pre-screening. In most related research, the images in the image sets are chosen by the ability of the segmentation method to segment the image. If the segmentation method fails, the image is not included in the set. This process occurs in Green et al. (1991), Ng & Lee (1996), and Binder et al. (1998), and may occur more frequently than is reported. The lesions are therefore 'pre-screened' by the segmentation method. Pre-screening skews the image set towards the easily segmented lesions to some unknown degree.

Because of the "image set specific" capabilities of most segmentation algorithms in this context, as well as the problems outlined above, it was decided that the lesion boundaries would be hand-drawn. Hand-drawn boundaries have been used previously in Ercal, Chawla, Stoecker, Lee & Moss (1994), Ercal, Lee, Stoecker & Moss (1994) and Landau et al. (1999). Hand drawn boundaries have a number of advantages, including minimising the pre-screening and algorithm result modification. Introducing errors from automated segmentation into the classification process is also avoided. The boundaries were outlined for each lesion using the freehand selection tool of PaintShop Pro 5.01. These boundaries were not subject to scrutiny. In a number of cases, especially in ELM images, the boundary was very indistinct, making the process of segmentation difficult.

Perhaps it would be expected that the 'best' boundaries would be identified by dermatologists. Such methods are used in Ercal et al. (1993). However, in consultation, it was found that dermatologists have great difficulty definitively identifying boundaries, as it was "not what we do" (Personal Communications: Oakley 1996). Therefore, it appears that little extra validity can be gained from the use of dermatologists in the time-consuming process of hand-segmenting a large number of images.

4.2 Image Analysis

Once the lesion component is segmented from the skin component of the image, analysis of the lesion can take place. This section describes the algorithms used to analyse the lesion images. Because there are two different types of lesion images, Clinical-view and ELM, two sets of algorithms had to be developed. The algorithms developed for analysis of the Clinical-view images are described in the first section, followed by a description of the algorithms for use with ELM images.

The choice of analysis algorithms is of fundamental importance to the classification task. The results of the algorithms make up the only data the classification system has to differentiate between the various classes (for example, melanoma/benign or 'excised'/'not excised'). If the algorithms do not define data that is of importance to the classification task, the results of classification will be poor.

Therefore, there should be some method in selecting 'relevant' algorithms. But what constitutes a 'relevant' algorithm is not straightforward. As was shown in the previous chapter, much of the work in image analysis algorithms in skin lesion research has attempted to recreate the diagnostic rules developed by human clinicians. A good example of a set of diagnostic rules is the ABCD criteria proposed by Friedman et al. (1985). This set of criteria for melanoma detection formed the basis of most of the algorithms designed for use in Clinical-view based lesion diagnosis systems, for example Cascinelli et al. (1992), Lee (1994), and Green et al. (1994). The attractiveness of the criteria is quite obvious, as the individual components conceptually lend themselves to computer implementation.

This approach, to select human criteria and reproduce them algorithmically, appears appropriate, and indeed, is probably the best approach to the problem of selecting 'relevant' algorithms. This approach is adopted in this research.

In the Clinical-view case, the algorithms implemented have been selected from the literature, and most correspond to the ABCD criteria of Friedman et al. (1985). It should be noted however, that the ABCD criteria is simply a set of guidelines that may indicate malignancy. To my knowledge, there has been no research that describes the clinical usefulness (in terms of sensitivity and specificity) of these rules. Therefore, although the image analysis algorithms may reproduce these criteria perfectly (they obviously cannot), it is not clear what accuracy will result.

In the case of ELM algorithms, there is certainly no shortage of human criteria available for reference, as seen in Kenet et al. (1993), Stolz et al. (1994), and Menzies et al. (1996). There is however, a shortage of published research into algorithmic equivalents of these criteria. Therefore, the solution in the ELM case is to create new algorithms, based on a set of criteria, and test these algorithms for suitability.

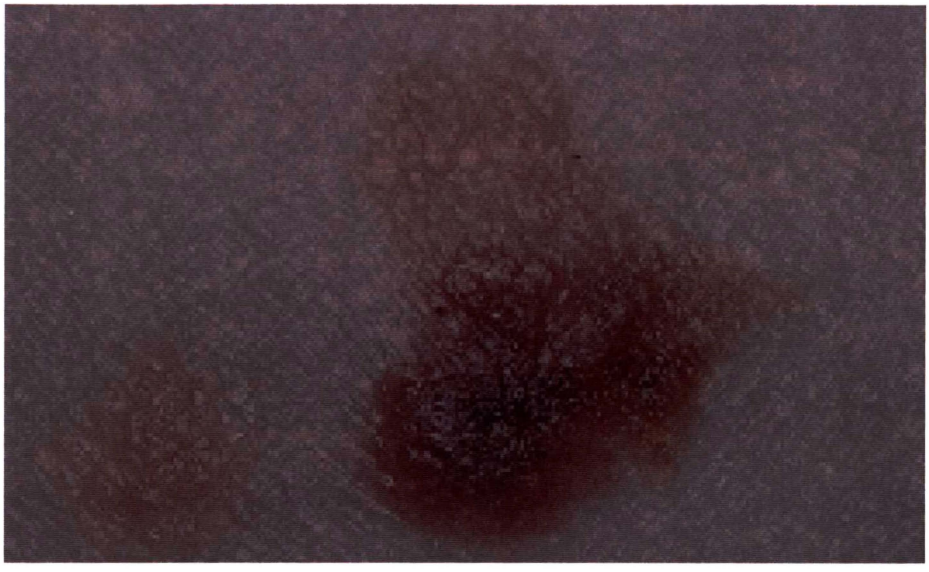

Figure 4.2: Example lesion. Superficial Spreading Melanoma

That was the approach taken for this research. New algorithms based on the ABCD criteria of Stolz et al. (1994) were developed. Both sets of algorithms were then tested concerning how well the algorithms reproduce the clinical rules on which they are based. These tests are described in the next chapter.

4.2.1 Clinical-view Image Analysis Algorithms

In most cases, the Clinical-view algorithms are either taken directly from previous research, or derived from related literature. Previous research has reported success using these algorithms. Therefore in some sense, the algorithms constitute the 'best' set of algorithms available for use at the Clinical-view. Figure 4.2 shows the example lesion used to illustrate the various algorithms.

Asymmetry

Symmetry is a concept describing the number of degrees of reflection that exist within a given shape. If there are no degrees of reflection within the lesion, the lesion is said to be asymmetric. Friedman et al. consider asymmetry of a lesion to be an important indicator of malignancy.

In this research, the technique described in Stoecker et al. (1992) and also in Gutkowicz-Krusin et al. (1997) is utilised. This technique locates the principal axis of inertia of the image and uses this axis as a mirror line. The technique is reproduced below.

An intensity moment $< f(x,y) >$ on a binary image can be defined by the following:

$$< f(x,y) > \equiv \frac{\sum_x \sum_y f(x,y) I_L(x,y)}{\sum_x \sum_y I_L(x,y)} \tag{4.5}$$

where $I_L(x,y) = 1$ inside the lesion, and 0 elsewhere. For example, consider the task of finding the centroid (x_c, y_c) of a binary figure. In this case, $x_c = < x >$, and $y_c = < y >$. These values are calculated by substituting x for $f(x,y)$ and y for $f(x,y)$ in Equation 4.5 for x_c and y_c respectively.

In order to measure lesion asymmetry, the mirror line must first be found. Stoecker et al. (1992) define the mirror line as the principal axis of inertia. This axis passes through the centre of the lesion. The following formula is used to find the orientation of the axis of inertia (4.6).

$$\tan 2\theta = \frac{2 < (x - x_c)(y - y_c) >}{< (x - x_c)^2 > - < (y - y_c)^2 >}, \; x_c =< x >, \; y_c =< y > \tag{4.6}$$

Once the angle of the principal axis of inertia is found, the binary lesion image is rotated to make the principal axis parallel to the x-axis. The two halves of the image are then overlapped using an XOR function (Gutkowicz-Krusin et al. (1997) subtracted each pixel from its reflection and took the absolute value of the remainder. For binary images, this is the equivalent of an XOR function). Any non-zero pixel remaining indicates a pixel that only appeared in one half of the image. These pixels, representing the area difference between the two halves of the lesion, are counted (Equation 4.7). This technique is repeated, using the axis of inertia orientated at right angles to the principal axis of inertia (Equation 4.8).

$$A_x = \frac{\sum_n \sum_y f(x_c + n, y) \quad \text{XOR } f(x_c - n, y)}{\sum_x \sum_y f(x,y)} \tag{4.7}$$

and

$$A_y = \frac{\sum_x \sum_n f(x_c, y_c + n) \quad \text{XOR } f(x, y_c - n)}{\sum_x \sum_y f(x,y)} \tag{4.8}$$

Gutkowicz-Krusin et al. (1997) use a modification of the above technique to deal with colour asymmetry. They use a colour intensity function (as opposed to the

Figure 4.3: Clinical-view asymmetry example. The left-hand image shows asymmetry in the X-axis, the second in the Y-axis.

binary intensity function shown) in Equations 4.5, 4.7 and 4.8 to sum differences in colour. This extension is not useful for Clinical-view images, but may be useful for ELM images. It is covered in more detail in the ELM Asymmetry section below.

Stoecker et al. (1992) use the minimum of these two values to obtain an Asymmetry index. This practice follows the guidelines presented in Friedman et al. (1985), where a lesion is considered symmetric if it can be reflected in one axis. However, Gutkowicz-Krusin et al. (1997) sum the two results to obtain an Asymmetry Factor. The technique of Stoecker et al. (1992) is followed here.

Border Irregularity

The second criterion defined in the ABCD criteria is border irregularity. This characteristic describes the tendency for melanoma to have haphazard boundaries, due to the uncontrolled growth of the tumour. There are a number of different techniques for describing the irregularity of a border identified in the literature. One of the most popular measures is the ratio of the square of the perimeter length of the object divided by the area of the object (or the area multiplied by 4π). This algorithm is referred to as the Irregularity index in this research. Symbolically:

$$\text{Irregularity index} = \frac{P^2}{4\pi A} \tag{4.9}$$

The equation is minimised in the case of the circle, and is also known as Circularity index. This measure was first proposed as a measure of skin lesion border irregularity by Golston et al. (1992), although also used by Cascinelli et al. (1992).

Clinical-view diagnosis literature reports extensive use of this algorithm (for example, Green et al. 1991, Schindewolf et al. 1993*a*, Ercal, Chawla, Stoecker, Lee & Moss 1994, Green et al. 1994). However, there may be problems with this approach, as stated in Young et al. (1974) "Measures such as $\frac{P^2}{A}$ are not robust... ...they yield similar numerical values for contours that are significantly different". Golston et al. (1992) also state "digitization affects the computation of irregularity. As sampling resolution increases, the perimeter of a digitized region often grows exponentially while the area approaches a finite limit, thus causing (the Irregularity Index) to approach infinity". They do not offer a solution to this problem. In our research, Irregularity index values of less than 1 (theoretically impossible) were occasionally noted. These errors were attributable to the digitisation process and the method of measuring the perimeter of the lesion. No adequate solution to this problem was apparent, as no 'best' level of digitisation has been proposed. For these reasons, two other methods of calculating border irregularity are also developed. They are Box counting and Convex hull.

Box Counting

Box counting is a simple and popular method of finding the fractal dimension of an object. The concept of fractals was developed by Mandelbrot (1983) to describe irregularity of natural objects, in particular, self-similar irregularity. The fractal dimension of a shape describes its space-filling capabilities. For example, a straight line would have a fractal dimension of 1, whereas more complicated lines with some degree of self-similarity will have higher fractal dimensions. See Mandelbrot (1983) for a detailed discussion with many pictorial examples.

Box counting is a measure of fractal dimension that has been examined previously in this context by Ng & Lee (1996), who state "Among all the methods, the simple box counting... still performed reasonably well....". They found that out of the five fractal techniques investigated, box counting consistently placed second to a higher order 'multi fractal'. However, Cross et al. (1995) also investigate the use of box counting, and conclude that there was "no significant difference between the fractal dimension of melanocytic naevi and malignant melanomas (p=0.18)". There is obviously uncertainty concerning the use of this measurement.

The Box counting method of estimating the fractal dimension of a shape involves covering the image with a lattice of squares with edge size r. For images, r is measured in pixels. For example, Figure 4.4 shows two lesion images with the lattice overlaid. The number of squares containing part of the lesion, N_r, is then counted. In the first case (the left hand lesion in Figure 4.4), N_r is equal to 107. This procedure is repeated for varying r values. An example is shown in the right

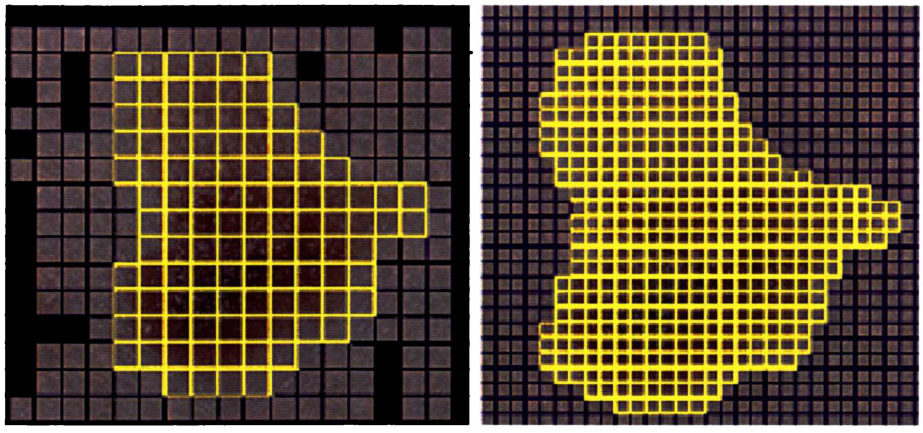

Figure 4.4: Box count example. The image on the left is covered 20×20 pixel squares. The number of squares (N_{20}) containing lesion is 107. On the right, the image is covered in 10×10 pixel squares, with the number of squares containing lesion (N_{10}) being 327. These figures are obtained for a number of square sizes (r). Subtracting the slope of $\ln(N_r)$ vs $\ln(r)$ from 1 gives the fractal dimension of the figure.

hand lesion of Figure 4.4, where $N_r = 327$. The fractal box dimension measure, D_b, is given by a power law, $N_r = K * r^{D_b}$. D_b is assessed by measuring the slope of the line $\ln(N_R)$ vs $\ln(r)$. This slope is obtained using the least squares method. r was varied between 10 pixels and 28 pixels at two pixel intervals.

Convex Hull

The third method of calculating border irregularity is Convex hull. The convex hull of a regular polygon is the convex polygon that encloses all of the points in the polygon (Figure 4.5). The difference between the lesion polygon and its convex hull gives a measure of how notched the border is. In this research, the popular Graham Scan algorithm is used to calculate the convex hull of the lesion. Many introductory algorithm texts (for example, Cormen et al. 1990, Sedgewick 1992) contain details of this algorithm. The paragraph below describes the algorithm.

To begin the Graham scan algorithm, a point is located that is definitely on the convex hull (an extreme left, right, top or bottom point is the usual choice). From this point, p_1, all the other points are ordered with respect to the angle each point makes with p_1. The algorithm then iterates through the set of points in a direction d (either clockwise or anti-clockwise), adding each point to the convex hull. If the new point makes a turn that is opposite to d, the algorithm removes the last point added from the convex hull, and then adds the new point.

Once the convex hull has been located, the areas of the original boundary polygon and the convex hull are calculated, using the technique shown in Arvo (1991). The

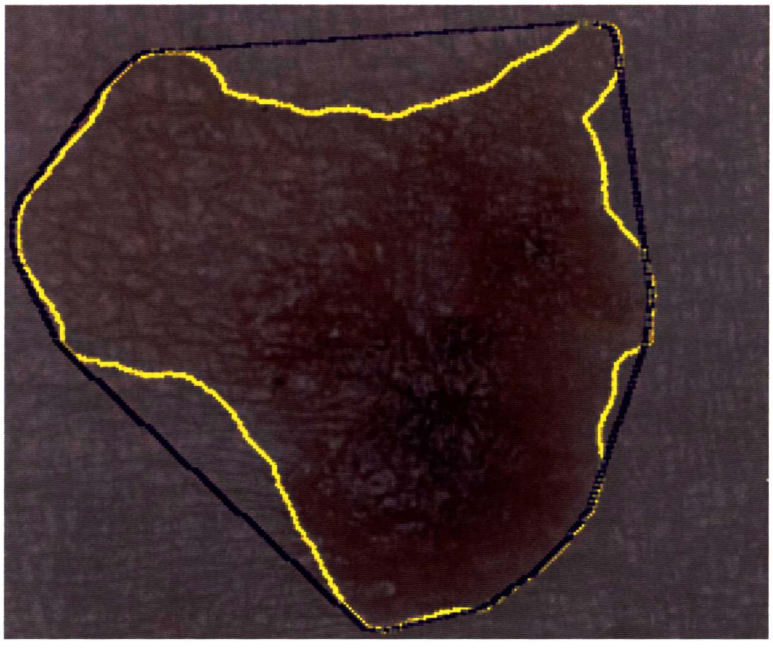

Figure 4.5: Convex Hull example. The blue line shows the convex hull of the lesion boundary (yellow line).

ratio of the convex hull area to the original area is then calculated and used as a measure of border irregularity.

The Convex hull technique is reported in Hall et al. (1995) as a possible test for notching, and hence border irregularity. However, to my knowledge, this technique has not been investigated previously in this context. In this research, the technique was tested as a method of measuring border irregularity, but may hold some merit as a separate criterion in its own right.

A Note on Smoothing and Digitisation

It is important to note that all the above measures of border irregularity are dependent to some degree on the amount of smoothing inherent in the border of the lesion. Digitised images have a degree of jaggedness introduced through the digitising process. Often, this jaggedness can cause errors in the results of irregularity measures. As images are digitised with smaller picture elements (i.e. more dots per inch), the area of the digitised lesion approaches the actual area of the lesion. However, perimeter measurements do not behave in the same way. "the perimeter ... will usually grow exponentially as (the size of picture elements) goes to zero" (Rosenfeld & Kak 1976). Rosenfeld & Kak go on to state "it is important to choose (the resolution) appropriately when we digitize a picture...". It is not clear what the optimal resolution for skin lesion images is, and it may be the case that what is

optimal for one set of measurements may not be optimal for another.

Furthermore, segmentation algorithms can leave jagged borders. These borders are either accepted as-is (for example Gutkowicz-Krusin et al. 1997), or smoothed (for example Schindewolf et al. 1993*a*, Ercal et al. 1993). This smoothing can alter the results of irregularity algorithms significantly. In summary, differences in image capture and segmentation can cause difficulties in replicating results of algorithms, in particular the border irregularity algorithms. This fact should be kept in mind during implementation.

Colour

Colour is a familiar concept to all of us, yet it is quite difficult to actually define what is meant by 'colour'. It is quite clear from the literature that colour has an extremely important role to play in determining whether a particular lesion is malignant (for example Friedman et al. 1985, Menzies et al. 1996). In particular, the greater the number of colours in a lesion, both for clinical and ELM views, the greater the likelihood of malignancy.

Although the concept of colour is very familiar to humans, no exact representation of how humans perceive colour is available. Umbaugh (1990) wrote, "There is no general method (of representing colour) that has been developed that is applicable to all domains".

In this research, two colour spaces are utilised. A colour space is a "geometric and mathematical representation of colour" (Umbaugh 1990). Each of these colour spaces has been used similarly in previous literature in this field. The first colour space is the RGB (red, green and blue) colour space, whereby each colour is represented by a three-dimensional vector (or set of tristimulus values). The components of the vector describe the amount of each of the three primary colours in the colour.

The second colour space is a transformations of the RGB colour space, proposed in Umbaugh (1990). Umbaugh wrote his Ph.D. dissertation on the subject of colour spaces for skin lesion analysis, in particular, the use of "color metrics" for skin lesion identification and segmentation of skin lesion images. He presents a new colour space, the $L\alpha\beta$ colour space, specifically designed for these purposes. The equations are presented below. This colour space has been used previously in this context by Lee (1994).

$$L = \sqrt{R^2 + G^2 + B^2} \tag{4.10}$$

$$\alpha = \cos^{-1}\left[\frac{B}{L}\right] \tag{4.11}$$

$$\beta = \cos^{-1}\left[\frac{R}{L\sin\alpha}\right] \tag{4.12}$$

These two colour spaces, RGB, and $L\alpha\beta$ are used in this research. When an algorithm produces results based on colour, for example colour variance and colour asymmetry measures, separate results for both colour spaces are returned.

It should be noted that the different colour spaces are not primarily included for dermatological reasons. Different colour spaces may or may not assist classification. The $L\alpha\beta$ colour space was proposed by Umbaugh (1990) to be used in this context, and may therefore hold some merit. In general however, these are values that fit into the category of features described by Bischof et al. (1998) as "easily extracted and measured by computer".

In several cases, a measure of the difference between two colours is required. Here, colour difference is described by the euclidean distance between the two colour vectors in the colour space. For example, Equation 4.13 shows the euclidean equation for distance between two points in a three-dimensional space. In this case, the three dimensional space represents the RGB colour space. Because the $L\alpha\beta$ colour space is a linear transform of the RGB colour space, $L\alpha\beta$ difference values are not calculated, as reverse transformation is required in order to calculate distance measures. The results would therefore be identical to the results from RGB.

$$D(R_0G_0B_0, R_1G_1B_1) = \sqrt{(R_0 - R_1)^2 + (G_0 - G_1)^2 + (B_0 - B_1)^2} \tag{4.13}$$

Colour Variance: The colour features calculated in this research are in three groups, variance, chromaticities and gradients. Variance is a common statistical measure of the spread of a population. It is used extensively in previous literature as a measure for colour variegation, for example Green et al. (1994), Lee (1994), Tomatis et al. (1998) and Landau et al. (1999) amongst others. In general, this calculation is simply applied to the lesion image on a pixel level, first calculating the mean value for each of the tristimulus values in the colour space. Then, the difference between the mean and every pixel is squared and summed. The result is the variance for that tristimulus value. Equation 4.14 presents the variance calculation for the red tristimulus value. x and y are the image width and height respectively. R_i is the red component of the i^{th} pixel, while $\bar{R}$ is the mean red value of the lesion. Variance

of the other tristimulus values are calculated similarly.

$$\text{Variance}_{red} = \sqrt{\sum_{i=0}^{xy} \frac{(\bar{R} - R_i)^2}{n}} \tag{4.14}$$

It should be noted that variance calculations such as that presented in Equation 4.14 are pixel level operations. It appears unlikely that human perception of colour variegation works at this level of detail. It may be more likely that higher scale variance is more important. To investigate this proposal, the image is sub-divided into squares of arbitrary size. These squares are treated as large 'pixels', in that each square has a single colour, and this colour is used in the variance calculation. Deciding on the colour of each square requires some thought however. The naive approach would be to use the mean colour of the square, treating each colour plane separately. However, mean calculations may introduce colours not originally in the image. Therefore, the median colour for each square was obtained. The median is the 'middle' colour of all the pixels in the square, and as such is guaranteed to exist in the image. To calculate the median colour, the colours in the square need to be ranked by some method. In this case, each pixel in the square is ranked by its absolute euclidean distance from (0,0,0). The distance calculation is detailed above in Equation 4.13. The conjecture of higher level colour perception is tested in Part 4 of the Investigations, described in the following chapter.

Chromaticities: The second group of colour features are relative chromaticity features. Lee (1994) discussed relative chromaticity in this context. Chromaticity coordinates are calculated from tristimulus values and represent a method of representing colour in a two- dimensional colour space. For example, given a pixel consisting of (R,G,B) components, the chromaticity of red is given by Equation 4.15. The other chromaticities are calculated similarly, and $C_R + C_G + C_B = 1$. Relative chromaticity of a lesion is calculated in Lee (1994) as the chromaticity of the average skin tristimulus value subtracted from the chromaticity of the average lesion tristimulus value, where R_l, G_l and B_l represent the mean red, green and blue values of the lesion, while R_s, G_s and B_s are the mean skin values of red, green and blue (Equation 4.16). Ercal, Chawla, Stoecker, Lee & Moss (1994) state that "these features are important in discriminating melanoma from seborrhoeic keratoses and intradermal naevi".

$$C_R = \frac{R}{R+G+B} \tag{4.15}$$

$$RC_R = \frac{R_l}{R_l + G_l + B_l} - \frac{R_s}{R_s + G_s + B_s} \tag{4.16}$$

Ercal, Chawla, Stoecker, Lee & Moss (1994) state that relative chromaticity features "provide less discriminating power between melanoma and dysplastic naevi", but report better results when dysplastic naevi cases were removed. However, as was discussed in the previous chapter, the method of measurement of discriminating power, the Pearson correlation coefficient, is likely to be misleading. Therefore, the true value of these features is not known.

Skin-lesion gradients: Umbaugh (1990) used skin-lesion gradients to assist in differentiating between different lesion types. Similar measures are also used in Lee (1994). Skin-lesion gradients are simply the average tumour value of a tristimulus value (for example, average red value of the tumour), which is subtracted from the average skin value of the same tristimulus value. For example, we might find the red skin lesion gradient, which is the average red skin value minus the average red tumour value. Skin-lesion gradients are calculated for each of the tristimulus values for both of the colour spaces described above.

The colour features described above are examples of relative colour features. Relative colour features are features where the difference in colour is captured, rather than absolute measures of colour (such as mean colour values, for example, mean red tumour value). Umbaugh (1990) states three reasons for using relative colour features: 1. To equalize any variations caused by lighting, photography/printing, or the digitisation process; 2. It should equalise variations in normal skin colour between individuals; 3. The human colour system works on a relative colour system. These three reasons and the difficulty in obtaining standardised images, especially Clinical-view images, suggests that the use of relative colour features is a reasonable approach in this research. Even in the case of the ELM images, it is not certain that images were captured and digitised in a standardised manner. This uncertainty rules out the use of any absolute colour features.

Diameter

The last of the ABCD criteria of Friedman et al. (1985) is diameter. This value was found by finding the longest straight line that passed through two boundary points and the centroid. Friedman et al. state that any lesion with a diameter of more than six millimeters is suggestive of malignancy. Here, no differentiation between diameters over this threshold and those under is performed.

Summary

This section has presented the methods to measure the Clinical-view image features used in this research. Table 4.3 summarises these features. The choice of these features was somewhat arbitrary because of the enormous number of possible features available in a colour image. All of the features used in this research are based on work reported in related literature, or guidelines developed in medical literature, in particular the ABCD criteria proposed by Friedman et al. (1985). As has been noted, the ABCD criteria has not been proven clinically, and is more a popular rule of thumb for lay-person use than a serious diagnostic tool. However, without more appropriate Clinical-view guidelines, there exists little alternative. This lack of choice is evident throughout the literature, where the ABCD criteria form the basis of most algorithms used in this context.

Descriptive data for each of these feature algorithms after application to the two image sets are shown in Appendices A and B. This data consists of mean and standard deviations for both classes for each of the classification problems, together with the results of the Mann-Whitney U Test which is used to show whether significant differences exist between the values for each class. For example, are melanomas more asymmetric than benign lesions? Two sets of data are presented, one for the Sydney Image Set (and the diagnosis problem), and one for the University/Health-Waikato Image Set ('dermatologist assessment' problem). The data for the Sydney Image Set is contained in Table A.1, while the University/Health-Waikato Image Set data is shown in Table B.1.

Table 4.3: List of Clinical-view Features

	Shape algorithms
1	Asymmetry Index
2	Irregularity Index
3	Box Count
4	Convex Hull
5	Diameter
	Colour variance algorithms
6-8	RGB Tumour Variances
9-11	$L\alpha\beta$ Tumour Variances
12-14	Relative Chromaticities of Red, Green and Blue
	Colour gradient algorithms
15-17	Skin-lesion gradients for Red, Green and Blue
18-20	Skin-lesion gradients for $L\alpha\beta$

4.2.2 ELM Image Analysis Algorithms

Similarly to the Clinical-view algorithms, the ELM algorithms are derived from criteria available for human use. Several such sets of criteria for ELM have been published, for example Kenet et al. (1993), Stolz et al. (1994), and Menzies et al. (1996). This research focuses on the ABCD criteria of Stolz et al. (1994). This set of criteria is similar in concept to the Clinical-view ABCD criteria proposed by Friedman et al., but is designed to take features specific to ELM into account.

Usefully, this set of criteria has been investigated in a clinical context. Stolz et al. (1994) report that the criteria give sensitivity of 98%, and specificity of 90% when used by experienced clinicians. Nachbar et al. (1994) and Binder et al. (1999) report on the accuracy of the ELM ABCD criteria in daily routine. Nachbar et al. reports sensitivity of 93% and specificity of 90% for the ELM ABCD criteria. Interestingly, Binder et al. re-evaluate the ABCD criteria, to observe the diagnostic abilities with this method in dermatologists with a range of experience. They find considerable variation in results depending on the experience of the clinician. Using the cutoff points proposed by Stolz et al. (see Chapter 1), average sensitivity/specificity results were 73% /90% for the 5.45 cutoff. For the 4.75 cutoff, sensitivity/specificity results were 81% /77%. These results indicate the results that can be obtained using this set of criteria, and allow developers to evaluate how well their algorithms can reproduce the criteria set.

To date however, little published work concerning algorithmic equivalents of the ELM ABCD criteria are available. Therefore new algorithms were developed for this research. For the most part, the algorithms presented here are simple extensions of previous work in this field, either ELM-based diagnosis systems such as that reported in Gutkowicz-Krusin et al. (1997), or aspects of Clinical-view based research, such as that reported in Lee (1994).

Asymmetry

Asymmetry is an important indicator of malignancy for ELM-view lesions, as with the Clinical-view. However, the definition of asymmetry is expanded from the Clinical-view shape asymmetry to include asymmetry of colour, and asymmetry of differential structures. This expanded definition makes it easier for a lesion to be labelled asymmetric. For example, a lesion can be symmetric in shape and colour, and still be labelled asymmetric if differential structures are not distributed symmetrically.

There are therefore three components of ELM asymmetry, namely shape, colour and differential structures. For shape, the Clinical-view shape asymmetry algorithm

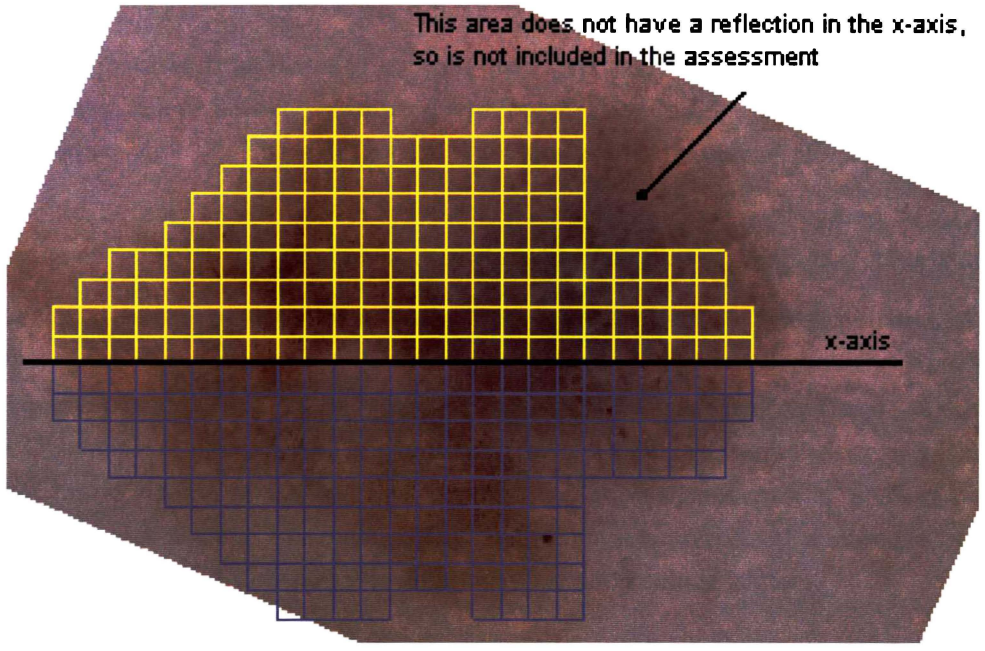

Figure 4.6: Colour and Structure Asymmetry example. The lesion is divided into squares, and reflected through the principal axis of inertia. Differences in colour (and variance) between each square (yellow) and its reflection (blue) are squared, summed, and averaged. Only squares that have a reflection containing lesion are processed, to avoid capturing data already captured by Shape asymmetry.

is used. However, in keeping with the more rigorous definition of asymmetry, the maximum of the asymmetry values is used. The algorithms for the remaining two definitions of asymmetry are also both based on this algorithm.

Colour asymmetry: Gutkowicz-Krusin et al. (1997) implemented a colour asymmetry algorithm for use in ELM by simply replacing the binary intensity function used by the Clinical-view shape asymmetry algorithm (where $I(x, y) = 1$ if pixel(x,y) is lesion and 0 if not), with an intensity function based on actual pixel values. For example, when obtaining Asymmetry Factor of Red, the intensity function was: $I(x, y) = r(x, y)$ where $r(x, y)$ is the red value of the x^{th},y^{th} pixel. Green and blue intensity functions were obtained in a similar fashion.

In this research, a slight variation on this technique was used because the above technique operated strictly at a pixel level, raising similar concerns to those discussed in Section 4.2.1. Using the Clinical-view Shape asymmetry algorithm described in the previous section, the principal axis of inertia was aligned with the x-axis. Then, the lesion area was divided into non-overlapping squares. The size of the squares is found by the investigation into colour box size described above. The complete description of this investigation is found in Section 5.3.2, and the results can be found in Section 6.3.2.

From this investigation, it was found that pixel level calculations were adequate. The difference in colour between each pixel and its reflection in the x-axis was calculated, using Equation 4.13. These differences were squared and summed (Figure 1.7).

It should be noted that only pixels that had a reflection were compared to avoid skewing of the results by data already captured by the Shape asymmetry algorithm.

The algorithm was repeated using the reflection in the y-axis. These two values (A_X and A_Y) were then divided by the number of comparisons (that is, half the number of squares) to give two Colour asymmetry indices for each colour space. Again, the Colour asymmetry index is the maximum of these two values. The resulting algorithm is similar in method to that used by Gutkowicz-Krusin et al., but our algorithm does not separate colours into individual components (for example, red, green and blue), but instead finds differences between colour vectors.

Structure asymmetry: Asymmetry of differential structures is difficult to accurately implement, due to the difficulty of actually detecting whether the structures exist at all, let alone exist in one place and not another. In this research, a simple measure of differential structure asymmetry was developed, based on the technique previously used for calculation of the Colour asymmetry index. This technique is very similar to Colour asymmetry, but calculates the difference in colour variance instead of calculating the difference in colour between two halves of the lesion. The algorithm used squares with a side length of 20 pixels, which was the approximate diameter of an average 'dot', one of the smaller structural features. Again, only squares that had a valid reflection were examined (Figure 4.6). The rationale for this technique is identical to that for the Differential Structures algorithms described below.

Border Contrast

The second of the ELM ABCD criteria is an abrupt cutoff between the lesion and the surrounding skin. Melanoma tend to show a sharp cutoff, while benign naevi are more likely to fade gradually into the skin (See Section 1.2.1 on page 8). Gutkowicz-Krusin et al. (1997) evaluate the colour gradient between lesion and skin in an attempt to measure this feature. The higher this value, the more likely the lesion exhibits a sharp border contrast. It is not clear from their paper how this algorithm was implemented.

In this research a similar method is developed. This method is reported in Day (In press). The centroid (cx,cy) of the lesion is located, and a line radiating from (cx,cy) to a boundary point (bx,by) is obtained. Lightness (L) values (given by Equation

4.10) along this line are read (Figure 4.7), starting from inside the boundary, and finishing outside the boundary. The number of pixels read was arbitrarily set at sixty (thirty inside the boundary and thirty outside), and represents a 'reasonable' number of skin and lesion pixels.

The line of best fit of the lightness values is then obtained, estimated by the least-squares method. The slope of this line is intended to reflect the increase in lightness from the lesion (dark - low lightness) to the skin (light - high lightness). The greater the slope, the higher the border contrast (Figure 4.8).

Once border contrast is measured in various positions around the border, a method of aggregation is required to combine these values into a single measure. Stolz et al. (1994) divide the lesion into eight segments of forty-five degrees each, and assess the cutoff of the border for each of these segments, scoring the lesion between zero and eight for each segment exhibiting sharp cutoff. To perform a similar procedure, some idea of what is meant by 'sharp cutoff' must be defined. Here, the slope of the best-fit line which indicates sharp cutoff is defined as 1.8. This value was derived experimentally by comparison with specialist observation, and is described more fully in Section 5.3.2 of the Chapter 5. Results leading to the use of 1.8 as a cutoff value can be found in Section 6.3.2, Chapter 6.

Using the 1.8 threshold, the border contrast values are aggregated similarly to Stolz

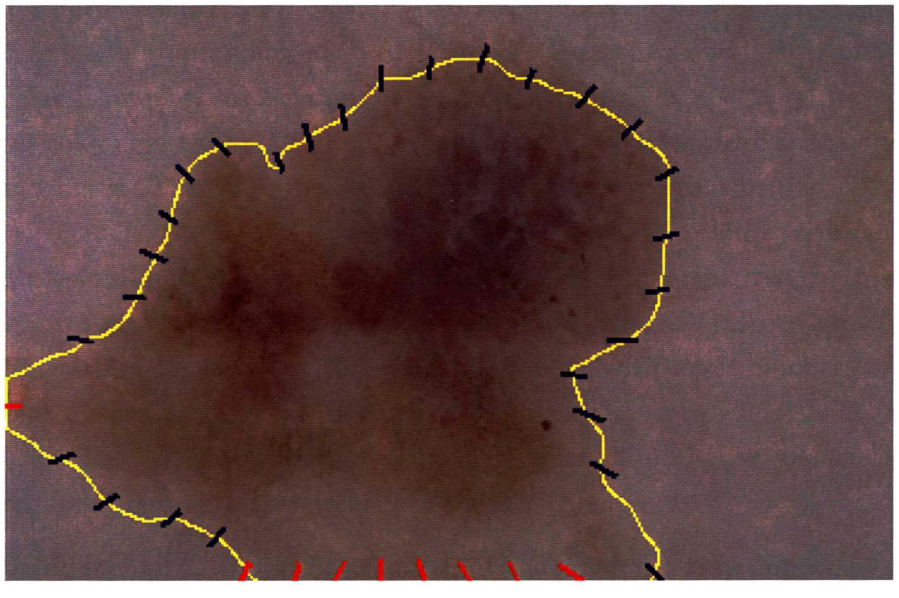

Figure 4.7: Border contrast example. The luminance values of the lesion under each line are recorded, and treated as a cross-section. The slope of this cross section is estimated by least-squares, and the result recorded. The percentage of slopes above the threshold mark makes up the Border Contrast score. Red lines cannot be assessed, as the lesion exceeds the boundary of the slide at these points.

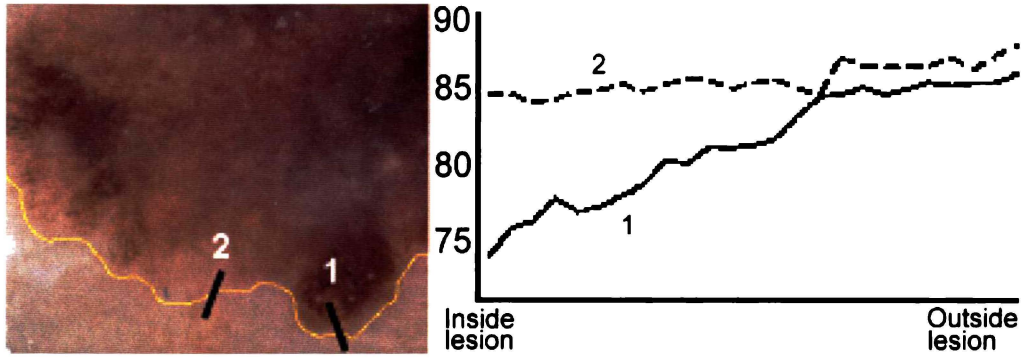

Figure 4.8: Border contrast example. Two areas were analysed. The first shows an edge with high contrast, while the second shows a low contrast edge. The graph on the right shows the lightness values for each of the thirty points. The slope of the '1' line is 0.48 while the slope of the '2' line is 0.13, showing that steeper slope indicates sharper contrast.

et al. (1994). The lesion area is divided into eight sectors, and five contrast values are calculated for each sector. If the maximum of these values is over 1.8, the sector is marked as high contrast. The number of high contrast sectors are counted, and a score between zero and eight is obtained.

Colour

Colour is perhaps the most important indication of malignancy under ELM. Colours appear much more pronounced in ELM images than the same images viewed clinically. For examples, see Section 1.3 on page 11. Stolz et al. (1994) recognise six different colours in their ABCD criteria, light brown, dark brown, red, white, slate-blue and black. These colours are not individually important in the ELM ABCD criteria. That is, Stolz et al. do not associate individual colours with melanoma. As with lesions viewed clinically, it is the number of different colours that is useful. Because of this fact, the colour variegation algorithms developed previously for the Clinical-view images are utilised again. Lesions with lower numbers of colours are likely to have smaller colour variance across the lesion, whilst multi-coloured lesions are likely to exhibit high variance. The other Clinical-view colour algorithms detailed previously were also applied to the ELM images.

Although the number of colours exhibited by the lesion is a very important diagnostic feature, there are also other colour features that are important. In particular, Menzies et al. (1996) reports a particular colour feature known as the 'Blue-white veil'. This feature is very specific to melanoma, although it is exhibited by other lesions (especially basal cell carcinomas). It is defined by Menzies et al. (1996) as an "irregular, indistinct, confluent blue pigment with an overlying white ground glass film, not associated with red-blue lacunes". However, it was clear from discussions

with dermatologists that little agreement existed on what constituted the blue-white veil, given the image set utilised. Therefore, it was decided that describing this feature algorithmically was out of the scope of this project.

Differential Structures

Differential structures are very difficult to implement algorithmically. In many cases, clinicians experience difficulty detecting these structures. Also, the definitions of the structures are not precise, although work such as Menzies et al. (1995) begins to address this issue.

Stolz et al. (1994) attach considerable importance to these structures, similarly to Menzies et al. (1996). In the work by Stolz et al., each differential structure that is detected increases the likelihood of a lesion being classed as malignant by approximately 10% (calculated by dividing the value of one differential structure, 0.5, by the threshold score used by Stolz et al. (1994) to indicate malignancy, 5.45). Therefore, it can be concluded that the number of structures, rather than the type of structures, is important, although this conclusion is not completely supported by Menzies et al. (1996) who identify structures significant to melanoma. However, Menzies et al. still consider the number of structures detected important.

In order to detect a number of different structures, without detecting the individual structures explicitly, we must first consider what distinguishes one differential structure from another. With more structures it would be expected that larger variations in local variance would be exhibited. For example, if one area contained predominantly pigment network, colour variance would be high. If another contained structureless areas, low variance would be expected. Therefore, variance of the colour variance would be large. Dots and globules would also exhibit different local variance patterns, and would also contribute to high variance in local colour variance measures.

The algorithm used to assess these changes is simply a "Variance of variance" calculation. The lesion area is divided into 20×20 pixel squares. For each of these squares, the variance of RGB and $L\alpha\beta$ tristimulus values are found. The algorithm then calculates the variance of each of these variance values over the entire lesion. For example, variance of red variance is found by Equation 4.17, where VR_i is the red variance of the i^{th} square, and $\bar{V_{red}}$ is the average variance of all of the squares.

$$\text{Variance variance}_{red} = \frac{\sum_{i=0}^{n}(VR_i - \bar{V_{red}})^2}{n} \tag{4.17}$$

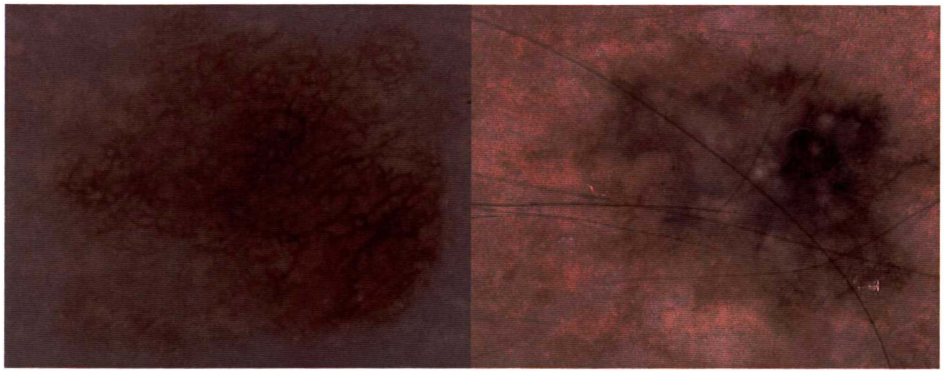

Figure 4.9: Differential Structures example. The left most lesion has a fairly obvious pigment network over most of the lesion. The right hand lesion has a number of different structures including pigment network and structureless areas. The variance of variance values for these two lesions are: Red (812,3602) Green (1870,6348), Blue (1118,3973), L (1781,10150), α (0.04,0.05), β (0.03,0.03). It is obvious that the right-most lesion has higher variance of variance values than the left.

A similar procedure is performed for the remaining colour-space variables. Nine "Variance of variance" values are calculated. It would be optimistic to expect that these values capture all of the differential structures mentioned by Stolz et al. (1994) but it was not considered productive to devote considerable time to developing individual algorithms for each of the criteria, based on the advice of Dr. Scott Menzies of the Sydney Melanoma Unit.

Summary

This section has described the algorithms used to obtain features from ELM-images. The algorithms are modeled on the ABCD criteria of Stolz et al. (1994). We investigate how well this criteria is reproduced by the algorithms in the next chapter. The algorithmic techniques used are quite simple, as there exists little evidence to recommend more complicated techniques. Table 4.4 summarises the features obtained from the ELM-images. The first fifteen features (asymmetry and border contrast) are shape dependent to some degree, and therefore cannot be used to analyse lesions that exceed the boundary of the slide (see Section 4.1.2). Descriptive data for each of these features are shown in Tables A.2 and B.2 for the Sydney Image Set and the University/Health-Waikato Image Set respectively.

Now that both sets of image analysis algorithms have been presented, a re-orientation may be useful. Remember that the purpose of the research is to establish whether Clinical-view or ELM images are more use in an automated system for identifying skin lesions. To do this, two systems for identifying lesions are required, one that uses Clinical-view images, and one that uses ELM images. At this stage in the

Table 4.4: List of ELM Features.

Asymmetry algorithms	
1	Shape Asymmetry
2-4	Variance asymmetry (RGB)
5-7	Variance asymmetry ($L\alpha\beta$)
8	RGB Colour asymmetry
Border contrast	
9	Border contrast
Colour gradient algorithms	
10-12	RGB gradients
13-15	$L\alpha\beta$ gradients
Colour variance algorithms	
16-18	RGB variances
19-21	$L\alpha\beta$ variances
Relative chromaticity algorithms	
22-24	Relative Chromaticities of red, green and blue
Differential structure algorithms	
25-27	Variance of variance (RGB)
28-30	Variance of variance ($L\alpha\beta$)

process, the methods of measuring features of the lesion have been presented. Two different sets of image analysis algorithms were shown above and these algorithms form the basis of the two systems (Clinical-view and ELM). In this next section, the method of classifying the results of the two sets of algorithms is presented.

4.3 Classification

Once analysis of the lesion images is complete, and a set of feature values for each lesion is obtained, some technique is required to classify the lesions on the basis of these values. Generally stated, classification refers to a technique that groups related instances on the basis of some selected characteristics of the instances. We now define the classification problem in the current context.

Every lesion image, L_i, once subjected to image analysis, has an associated feature set, F_i. This feature set contains the results produced by applying the algorithms described previously to the image. For example, a feature vector for a Clinical-view image from the UHWIS would contain thirty features, or $F_i = \{f_0, f_1, f_2, f_3, ..., f_{29}\}$, where each f_i is the result of an image analysis algorithm.

Every lesion image (L_i) also has an associated class (C_i). For example, this class may be melanoma or benign (or 'excised' or 'not excised'). The classification problem is to take F_i for any lesion, apply a function to it, and get C_i. That is, $\hat{C}_i = \text{Classifier}(F_i)$,

and $\hat{C}_i - C_i = \varepsilon$. In other words, the predicted class ($\hat{C}_i$) of L_i is a function of the feature set (F_i). The difference between the predicted class ($\hat{C}_I$) and the actual class (C_i) is the error (ε) of the classifier for the feature set F_i.

The 'best' classifier will be the one that minimizes ε over the population of lesion images. Of course, the population of lesions is unknown, and therefore samples of the lesion population are used, namely the Sydney Image Set and the University/Health-Waikato Image Set.

Ripley (1996) reports that "the task (of a classifier) is to classify an object, which means reaching one of $K + 2$ possible decisions..." where K represents the number of possible classes. The two additional classes represent 'being in doubt', and an outlier, or case definitely not belonging to any of the K possible classes.

Considering Ripley's statement in this context, a pigmented skin-lesion has to be either benign or melanoma (with non-melanocytic malignant lesions, such as the carcinomas being removed from the image set). Therefore, the 'outlier' class can be removed from consideration. A similar situation exists for the 'dermatologist assessment' problem, where a lesion must be either 'excised' or 'not excised'. For the 'doubtful' case, Ripley states that the doubtful classification may result in a possible postponement of "the decision until further measurements are made". There is no ability in this or any other skin lesion diagnostic system to cater for such further measurements. This lack of provision is not to rule out the usefulness of such a technique. Indeed, there may exist many features with very poor sensitivity and specificity for the population of skin lesions, but may be useful for classifying some of the doubtful cases. This idea is considered in Chapter 9.

4.3.1 Classification Considerations

Now that the classification problem has been defined for this context, some choice of classification function must be made. Given the range of classification techniques, a number of points need to be kept in mind. Some of these are well described in Bischof et al. (1998), including the concept of hard or soft classification, amount and distribution of data, and the importance of having interpretable models. Further considerations that should be mentioned are pragmatic considerations, such as the classifiers available and appropriate knowledge.

Hard classification is when the classifier function returns a definite set for each feature set. The opposite is soft classification, where a probability of set membership is returned. In this context, there are possibly serious consequences for error, especially when melanoma are classed as benign. Therefore, we would like the ability to interpret or modify results to minimise these false positive results. Soft classifica-

tion, where cost functions can be applied to the probabilities to make false positive less likely, appears the better choice.

The amount of relevant data is one of the primary concerns in most classification research, as there is never enough! This research is no exception, and the difficulty with which good quality image sets are obtained has been covered previously. It should be understood that the number of features is not the problem, but the *number of instances* available. In other words, how many lesions there are in the set. More instances and less features allow a more generalisable model to be developed. Some classifiers perform better than others on small data sets.

Interpretable models are generally preferred in classification, as the value of each of the features can be identified. Some models, such as artificial neural networks, perform as *black boxes* where it is difficult, if not impossible, to measure the contribution of each feature. This lack of interpretability can be seen in Lee (1994) and Binder et al. (1998), where the features are evaluated independently of the classification function.

Pragmatic considerations are almost completely ignored in previous research. It is assumed that researchers have perfect access to all models, as well as perfect knowledge of how best to apply these models. This is clearly not the situation, and it would be misleading to deny that such considerations were not important. In this case, the difficulty of obtaining and using some of the more advanced models precluded their use.

4.3.2 Classifier Choice

As seen in the above discussion and also in Chapter 3, there is no clear choice of classification model in this context. Some of the more advanced projects, for example, Bischof et al. (1998), give some justification of their chosen classifier. Others appear to simply use the method at hand with little justification for their choice. Early in this research, we had investigated an artificial neural network for classification, similar to that presented in Ercal, Chawla, Stoecker, Lee & Moss (1994). However, the lack of interpretability of the model led us to other methods.

The method chosen for this research, logistic regression, has a number of suitable features. In particular, the model allows soft classification, and has been used previously in related research (Menzies et al. 1997). Also, the pragmatic considerations outlined above were also important. It was also felt that a simple, well-understood classification model was important for this research.

Logistic Regression

There are only two possible outcomes of the classifier for the classification problems in this research. The lesion image is classed as either melanoma (or 'excised') and benign (or 'not excised'). In such a case, the dependent variable is *dichotomous*, or can only take two values. Compare this to linear regression, where the dependent variable is free to take a number of values. When the dependent variable is dichotomous, logistic regression has "become, in many fields, the standard measure of analysis in this situation" (Hosmer & Lemeshow 1989). Much of the following discussion is derived from Tabachnick & Fidell (1996) (an excellent introductory text) and Hosmer & Lemeshow (a detailed look at the application of logistic regression), and it is recommended that interested readers obtain these books.

If we consider a scatter plot of a binary dependent variable (in this case the dependent variable is "melanoma or benign"), it is apparent that the standard linear regression measures will not work. Figure 6.1 on page 117 shows an example scatter plot using diameter as the independent variable. If we ask, "What is the expected value of diagnosis, given any diameter x?", the expected value could range anywhere between $-\infty$ to ∞. A better measure would be a function that is bounded to the two possible values of 'diagnosis' (or 'dermatologist assessment') namely, 0 and 1. This is where the logistic model becomes valuable.

The multivariate logistic regression model is presented in Equation 4.18 (page 6 Hosmer & Lemeshow 1989). It is apparent that this function can only vary between 0 and 1, which is ideal for a binary dependent variable. Fitting the logistic regression model means solving this equation for the 'best' values of β_i. This situation is analogous to the multivariate linear regression model $y = \beta_0 + \beta_1 x_0 + \beta_2 x_1 + ...$, where fitting involves some 'best' values of β_i.

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_0+\beta_2 x_1+...}}{1 + e^{\beta_0+\beta_1 x_0+\beta_2 x_1+...}} \tag{4.18}$$

Discussion of techniques for finding the 'best' values for β_i is not reproduced here, but can be found in Hosmer & Lemeshow (1989). In any case, statistical packages such as SPSS and Minitab are readily available to perform the fit. Once the model is fitted, the Equation 4.18 can be used to find the conditional mean of the dependent variable (the class of the feature set), given the value of the independent variables (the values of the feature set), and the ability of the model to classify can be tested.

4.3.3 Classification Method

The details of how logistic regression is applied are now described. This discussion is broken into three sections, covering limitations of the logistic model, model building and methods of assessing model fit.

Logistic Regression Limitations

One of the best features of logistic regression is the relative lack of restrictions. There exist few limits on what can be included in the model. For example, any combination of discrete and continuous variables can be used. Tabachnick & Fidell (1996) identify some practical limitations of logistic regression analysis, including collinearity and outliers. In reality however, these limitations are not confined to the logistic regression technique, but are common to a number of regression methods. For more information, the reader is referred to Tabachnick & Fidell (1996), page 578.

Perhaps one of the most important limitations of logistic regression (in fact all classification methods) is the problem of low case-to-feature ratio. In other words, a large number of features are available for use in classification, but only a comparatively small number of cases are available. Tabachnick & Fidell (1996) state that "the cases-to- (independent variables) ratio has to be substantial or the solution will be perfect - and meaningless". If this situation occurs, the model is said to be *overfitted.* Overfitting is evident where the model performs well on the training data, but poorly on the test data, illustrating that the model cannot generalise to a larger population well. Such a situation is avoided primarily through the use of large data sets (and especially a 'high' case to feature ratio).

However, in the case where there is little control over the size of datasets available, techniques such as cross-validation can also reduce the problem of overfitting. This process involves dividing the cases into x equal sets, and using $x - 1$ of the sets to build the model. The remaining set is used to test the model. This process is repeated x times, each time using a different set to test the model. In this research, x was ten (that is, ten-fold cross-validation). Results from the x iterations are averaged.

Overall, the goal of model building is to obtain a model that is likely to generalise well to a larger population. Therefore, we are looking to find the most parsimonious model that adequately explains the data. Hosmer & Lemeshow (1989) state:

> "The rationale for minimizing the number of variables in the model is that the resultant model is more likely to be numerically stable, and is more easily generalized'

Table 4.5: Number of model features allowed for each image set. From Tabachnick and Fidell (1996)

Image Set	Number of Images (N)	Number of Features (m)
Sydney Image Set	83	4
UHWIS	73	< 3

Tabachnick & Fidell (1996) suggest that for testing the fitted model, $N \geq 50 + 8m$, where N is the number of cases, and m is the number of independent variables (or features). In our situation, we have little control over N, and therefore the number of features (m) that can be included to make a reasonable model can be inferred. These inferences are described in Table 4.5.

To summarise, perhaps the biggest limitation of logistic regression is common to all classifiers, namely a low case-to-feature ratio. If this ratio is low for a model, it will be difficult to infer that the model can generalise to a larger sample of images. If we cannot infer this, or at least suggest reasons why the model may have this ability, the model will be meaningless. In this research, the adherence to the model sizes described in Table 4.5 and the use of the cross-validation technique form the basis of efforts to ensure reasonable generalisation ability. However, this generalisability cannot be confirmed without further research on a larger set of images.

Model Building

Model building is an essential part of the classification process, and has tended to be poorly documented in previous research. Model building is concerned with building a plausible model that has the ability to generalise well to the population. The above discussion on overfitting is also relevant to this section.

Model building is a complex task with a multivariate data set. Previous research has relied on the classifier to find useful groups of features (for example Green et al. 1991, Schindewolf et al. 1993*a*, Green et al. 1994), or assessed each feature by statistical means, independent of the model (Gutkowicz-Krusin et al. 1997, Seidenari et al. 1998). In this research, model building is begun by using stepwise logistic regression to select features. Stepwise logistic regression is an extension of stepwise model building method common to linear regression. The idea is simply to start with either the constant model to which attributes are added (forward regression), or the model built with all attributes from which attributes are removed (backward regression). Hosmer & Lemeshow (1989) present a detailed explanation of this technique (pages 106-110 Hosmer & Lemeshow 1989). This technique is explained below in the forward case.

Using the forward stepwise technique, we begin with the logistic model built using

the constant (Equation 4.19). We would expect any model using the selected features to perform better than this model, so the constant model can be thought of as a baseline for performance.

$$\pi(x) = \frac{e^{\beta_0}}{1 + e^{\beta_0}} \tag{4.19}$$

A feature is now added to the model. To add the 'best' feature, Hosmer & Lemeshow (1989) find the log-likelihood of the univariate logistic model for each possible feature. Subtracting the log-likelihood of the constant model from the log-likelihood of the model built with each feature will result in the likelihood ratio, which has a chi-squared distribution. Hosmer & Lemeshow add the feature with the smallest chi-squared probability (p-value, that is, the highest significance level. Note that the smallest p-value is required to be under some pre-specified addition threshold, P_{in}).

We now have a model with one attribute and the constant term. The previous steps are repeated to add a further feature. At this stage, the model has two features and a constant term (Equation 4.20). But has adding feature x_2 made feature x_1 redundant? If the first feature is indeed redundant, it needs to be removed from the model.

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2}} \tag{4.20}$$

Removal of features is achieved through similar means to the feature addition process. The first feature is removed and the model is refitted, using only the second attribute and the constant. If the p-value for log-likelihood ratio is not significant (over some pre-specified removal threshold P_{out}), the first feature is removed.

The attribute addition and removal steps continue for each of the possible attributes. At some stage, no attributes will have a p-value under the addition threshold, and therefore no further attributes can be added, and the process stops. It should be noted that in this research, the stepwise procedure was performed using SPSS version 9.0.1, which allows a variety of tests for inclusion and exclusion. The likelihood-ratio test was used as described above.

The above treatment ignores the possibility that several features individually may not contribute much to the model, but together may be very useful. Hosmer & Lemeshow state "One problem with any univariate approach is that it ignores the possibility that a collection of variables, each of which is weakly associated with

the outcome, can become an important predictor of outcome when taken together" (page 86). There is no 'right' method for dealing with such combinations of features. In this research, this issue was not addressed. Due to the low number of cases, and the correspondingly low number of features in the models, it was felt that these features were unlikely to be discovered.

It should be apparent from the above description that the stepwise technique is a hill-climbing variant. Therefore, the problems with hill-climbing also apply here. In particular, the model developed may only represent a local maxima in the model space, rather than the desired global maxima. One technique used for testing for the presence of local-maxima is to perform the same regression, but using backwards stepwise technique to select the features in the model. However, due to the low number of cases and high number of features, backwards stepwise regression almost always resulted in a perfect model with large numbers of features. Further consideration of this aspect was considered out of the scope of this work.

After stepwise regression and modification of the model, the model contains a number of features selected from the feature set. The correlation of each of the features with each other is assessed, using the well-known Pearson correlation coefficient. If any pair of features correlate significantly, the feature with the lowest contribution to the model is removed, in order to minimise collinearity in the model. The Wald statistic and LR Test are used to assess the contribution of the feature to the model. Following this step, further features may be manually included or removed in an effort to find the most parsimonious and useful model. The guidelines suggested in Table 4.5 serve as a guide. Features are evaluated by assessing whether significant differences are noted between the two classes of lesion. For example, if a feature showed significantly different results between melanoma and benign lesions, that feature may be included in the model. Further assessment of such features is performed using univariate analysis in the manner suggested by Hosmer & Lemeshow page 84. In particular, the results of the likelihood ratio test and the Wald statistic are considered.

Assessing Model Performance

Once the model is considered reasonable, it needs to be tested. That is, we would like to know how well the model can be used to predict the dependent variable (diagnosis or 'excision') from a set of image features.

Firstly, we would like some idea of how well the model 'fits' the data. That is, did the model approximate the data well, or were there large amounts of data that were not modeled accurately? There are many methods of assessing the fit of the model. In the first case, the summary statistics available through SPSS are evaluated. In

particular, the basic fit of the model is assessed using the Model-χ^2 value. Further assessment of fit is evaluated using the Hosmer-Lemeshow test (page 140 Hosmer & Lemeshow 1989). Both of these techniques are described below in Section 4.4.

The second method of assessing the performance of the model is through the results of testing. The cross-validation procedure is used to test the model. Ten-fold cross-validation is used to assess the model performance. However, the cross-validation procedure described above may be sufficient for large datasets. In this case however, with the datasets being relatively small, some concerns about the results of cross-validation were raised. In particular, the random selection of the test and training sets may inflate or lower the results. Therefore, the ten-fold cross-validation procedure was repeated ten times, and averaged results from the ten cross-validation runs are presented.

4.4 Statistics

This section presents an explanation of some of the common statistics used in this research. The first section briefly describes those statistics associated with the logistic regression procedure. The second describes further statistics used in the research, details of which can be found in most introductory statistics textbooks.

4.4.1 Logistic Regression Statistics

The following statistics concern the logistic regression model. For a more detailed description of the logistic regression statistics, the reader is referred to Hosmer & Lemeshow (1989).

Log-Likelihood

The log-likelihood can be thought of as the likelihood of observing the actual data from the model. It is a measure of the total error in the model, and is calculated by summing for each case the log of the difference between actual and predicted values (Equation 4.21). Y_i is the actual value (0 or 1 for case i), $\hat{Y}_i$ is the value predicted by the model for case i.

$$LL = \sum_{i=1}^{n} [Y_i ln(\hat{Y}_i) + (1 - Y_i) ln(1 - \hat{Y}_i)] \tag{4.21}$$

Likelihood Ratio Test (model chi-square)

The likelihood ratio value is simply two times the difference in log-likelihoods of two models. In general, the statistic is used to compare a larger model (containing more regressors) with a smaller model containing a subset of the regressors of the larger model. The equation is shown below (Equation 4.22). In this research, the Model-χ^2 value is used as an indicator of goodness-of-fit, by comparing the model to the constant only model. It is also used to measure the contribution of a feature to the model, and is termed the LR Test.

$$\chi^2 = 2[(LL_{biggermodel}) - (LL_{smallermodel})] \tag{4.22}$$

Hosmer-Lemeshow Statistic

The Hosmer-Lemeshow statistic (Hosmer & Lemeshow 1989) is used to assess model fit. In simple terms, the statistic is used by creating ordered groups and comparing the observed number in each group with the number predicted by the model. Commonly, ten groups corresponding to risk deciles are used (giving rise to the term, deciles-of-risk statistic). For example, group 1 contains the sum of all of the cases that have an estimated probability of less than 0.1, that is, in the lowest decile. Group 2 contains the sum of all cases in the second decile and so on. Each group is then subdivided into two, based on the dependent variable (for example, melanoma or benign). If the model fits well, all of the '0' cases (benign) should reside in the lower deciles, and all of the '1' cases should reside in the higher deciles. This statistic reports a non-significant chi square value if the model fits, that is, there is no significant difference between this model and the 'perfect' model.

Wald Statistic

The Wald Statistic is the coefficient of the independent variable divided by the standard error of the independent variable. It is a commonly used measure for assessing the contribution of an individual independent variable to the model. Kleinbaum et al. (1998) state:

> "...its specific purpose is to assess whether the effect of (an independent variable in a model) describes a linear relationship between (the independent variable) and the log-odds of the dependent variable... (Is) a linear effect more plausible than no effect? Such a test is typically referred to as a linear trend test"

So the Wald Statistic is a measure of the strength of linear relationship between the independent and dependent variables. It should be noted however, both Hosmer & Lemeshow (1989) and Tabachnick & Fidell (1996) report difficulties found with this statistic.

In SPSS 9.0.1, the Wald statistic is calculated as: $\left(\frac{B_x}{S.E._x}\right)^2$, which follows a χ^2 distribution.

Standard Error

The standard error (of a coefficient) as used in the Wald statistic is a measure of the error in the estimated coefficient. Coefficients of features in the logistic regression equation (Equation 4.18) are estimated through maximum-likelihood methods, and as such, the estimate has an associated error. The standard error value is commonly printed by logistic regression software, and further discussion is out of the scope of this work.

4.4.2 Other Statistics

Three statistics are described in this section. The first statistic, the Mann-Whitney U Test, is a non-parametric measure of difference in means between two populations. Secondly, the well known Pearson correlation coefficient is used to measure the linear relationship between two independent variables, while Spearman's rank order correlation coefficient is an equivalent non-parametric measure of association. These statistics are calculated using the correlation function in SPSS 9.0.1. The use of each of the statistics is explained below.

Mann-Whitney U Test

The Mann-Whitney U test is a non-parametric test of differences in two population means. In other words, it is a non-parametric version of the t-test. In this research, this statistic is used to measure whether significant differences exist in the results of an algorithm between two different groups. For example, we are interested in whether Asymmetry Index tends to be lower for benign lesions than it does for melanoma. The Mann-Whitney U Test is a suitable test when the assumptions for other 'difference of mean' tests, such as the t-test, cannot be satisfied.

Product Moment Correlation Coefficient

The Product moment correlation coefficient (Pearson's correlation coefficient) measures the linear relationship between two continuous variables. It is used in this research to measure how correlated image analysis features are. If two features are highly correlated, little extra information is supplied by one of the features and it should therefore be removed from the model to minimise overfitting.

Spearman's Rank Order Correlation Coefficient

Spearman's rank order correlation coefficient (Spearman's rho) is a non-parametric statistic for measuring association. It uses the ranks of the data, rather than the data values themselves. Spearman's rho simply performs the product-moment correlation described above on the ranks of the data.

Spearman's rho is used in this research to assess association between dermatologists perception (an ordinal variable) and several of the image analysis algorithms (continuous variables). It should be noted that there are various other methods of performing such an assessment, for example, Kendall's tau and Somer's *d*. However, there is little to suggest one above any other.

Receiver Operator Characteristic Curves

Receiver operator characteristic curves are a well-known method of visualising the results of classification tests. In general, there is always a tradeoff between sensitivity and specificity for any given classifier. In the logistic regression case, a threshold point is used to decide how to classify a given instance. Normally, this threshold is 0.5. All instances above 0.5 are classed as '1', while all below are classed as '0'. If however, this threshold is moved, say to 0.8, we would expect to see sensitivity fall as it became more difficult to class instances as '1'. Correspondingly, specificity would rise as more instances were placed in the '0' category. Conversely, if the threshold was lowered, for example to 0.3, we would expect the opposite to occur.

The ROC curve presents a straightforward method of viewing this tradeoff. The ROC curve is a graph of sensitivity on the y-axis, and the false negative rate (which equals 1-specificity) on the x-axis. The threshold is then iterated through the range of possible values (generally from 0 to 1) and the sensitivity/1-specificity pairs are plotted. For an example, see Figure 6.2.

Further statistics can be calculated from the ROC curve. Particularly common is the area under the curve (AUROC) which can be used as a measure of classifier usefulness. 'Perfect' classifiers obtain an area of 1.0, while classifiers that are no

better than chance obtain an area of 0.5. For an introduction to ROC curve analysis, and in particular the area under the curve, see Hanley & McNeil (1982).

4.5 Chapter Summary

In this chapter, the methods used in this research are presented. Firstly, the image-sets gathered for this research were examined, together with the methods used to obtain, pre-process and segment these images. It may be apparent that the good-quality image sets required for this type of research are not easy to find.

The second section described the image analysis features obtained from the images. The first part of this section looked at the Clinical-view features. Most of these algorithms are taken directly from the literature, together with several new algorithms based on human criteria. The ELM features were mostly new algorithms. It was necessary to create new algorithms because few papers describing ELM automated systems exist that report the algorithms used. Where algorithms were reported, there was considerable doubt as to how they were implemented.

Finally, the classification function was examined. In Chapter 3, the previous classification functions used in this context were reviewed. Little work in finding a 'best' method has been carried out. In this chapter, some of the considerations in choosing a classification function are discussed. Then a discussion of the chosen classification function, logistic regression, is presented.

The system components have now been described. In the next chapter, the investigations conducted in this research are detailed. There are three investigations, each dealing with a separate aspect of the research.

Chapter 5
Investigations

The first three parts of this chapter describe the three investigations conducted in this research. The first investigation examines the diagnosis problem. Histologically proven lesions (the Sydney Image Set) are used to compare the Clinical-view diagnosis system with the ELM system. This investigation provides the first set of data concerning the relative performance of the Clinical-view and ELM systems.

The problem of reproducing dermatologists' assessment of 'suspiciousness' provides the next investigation. In particular, the purpose of this investigation is to evaluate whether Clinical-view or ELM images are more useful to the problem of replicating the decision of dermatologists to excise a lesion.

The results of these two investigations will allow the thesis argument to be evaluated. The third and final investigation is a retrospective look at the image analysis algorithms used in this research. As has been discussed, most of the algorithms are based to some degree on criteria intended for human use. It would be reasonable therefore to expect the algorithms to reproduce human perception of these criteria. This investigation looks at how well the algorithms reproduced the perception of dermatologists, and allows the ability of the Clinical-view algorithms to reproduce human perception to be compared to the ability of the ELM algorithms.

The penultimate section in this chapter identifies variables that may affect the comparison of results of the Clinical-view and ELM systems. In some cases, these variables could be controlled for, and thus their impact on the results could be minimised. In other cases however, there were variations between the Clinical-view and ELM systems that could not be controlled for, most notably the difference in image analysis algorithms, and these variables would have some effect on the results. The variables identified and the possible impact on the results is discussed in this section.

The chapter concludes with a brief look at the software used in the investigations.

5.1 Diagnosis

The first investigation looks at the problem of detecting melanoma in an image set. This problem, the diagnosis problem, is identical to that investigated in previous automated research. The results from this investigation will allow the Clinical-view and ELM diagnosis systems to be compared. Only the Sydney Image Set was used in this investigation, due to the low number of melanomas in the UHWIS. The Sydney Image Set is a collection of 83 atypical naevi and melanoma.

5.1.1 Investigation Method

The general method of this investigation was to fit a logistic regression model to the results of the feature analysis algorithms described in the previous chapter. Two separate logistic regression models were fitted, the first based on the Clinical-view image analysis results, while the second model utilised the ELM image analysis results. The models were then tested using cross-validation.

Model Building

Section 4.3.3 covered model building in detail. Firstly, relevant features are selected using the forwards stepwise logistic regression procedure with a P_{in} value of 0.10 and a P_{out} value of 0.15. That is, only features that have a Likelihood-ratio test significance of < 0.1 may be included in the model. Features in the model are also removed if the change in log likelihood after removal is not significant at least at the 0.15 level.

From this procedure, the number of features in the model is assessed, and if it exceeds the number recommended by Tabachnick & Fidell (1996) (See Table 4.5), the least significant features are removed. The resultant model is then assessed for the possibility of there being a local maxima, and adjustments to the features may be made. Features showing high collinearity as assessed by the Pearson correlation coefficient are also removed. The final model is then tested using cross-validation.

5.1.2 Results

There are a number of results reported from this investigation. The first set of data is a summary of the descriptive data from the algorithms. This data shows which algorithms produce significant differences between melanoma and benign lesions. The results of the logistic regression model are then shown. Firstly, the components of the logistic model are presented, including the coefficients and Wald statistics for

each of the features in the model. Although Wald statistics are intended to assess the contribution of individual features to the logistic regression model, doubts are raised about the usefulness of this statistic. See Hosmer & Lemeshow (1989), page 17 and Tabachnick & Fidell (1996), page 599 for more details. To alleviate this problem, we also provide the Likelihood-ratio test statistic for each of the features in the model.

The second set of results is the summary goodness-of-fit statistics. In particular, the Model-χ^2 and Hosmer-Lemeshow deciles of risk statistics are presented. These statistics indicate how well the data has been modelled. Poor results for these values may indicate that the model has difficulty describing the data, and would raise concerns about the model results.

The ten-fold cross-validation results of the model make up the third set of results. Ten-fold cross-validation was repeated ten times for each model with different (random) cut points, and the ten results for each cross-validation run were averaged. This procedure was performed since a single cross-validation run may not give representative results due to the small size of the data sets.

For each model, the means of the cross-validation results are used to create a ROC curve. ROC curves illustrate the tradeoff between sensitivity and specificity as the threshold representing the cutoff point is varied. The y-axis represents sensitivity, while the x-axis represents 1-specificity (or false-negative rate). As the threshold separating melanoma from benign is increased (towards '1'), sensitivity is likely to decrease, and specificity will increase, as it becomes more difficult to class a lesion as melanoma. On the graph, a threshold increase will be seen as a shift to the left, while a decrease will be a shift to the right. For an example, see Figure 6.2 on page 119.

5.2 Dermatologist Assessment

The previous investigation looked at the problem of automated diagnosis. That is, how well melanoma could be distinguished from non-melanoma. In some senses, the systems were attempting to reproduce the results of pathologists from either Clinical-view or ELM inputs. This investigation is concerned with reproducing the assessment process of dermatologists. Here, we are not concerned with the diagnosis of the lesion, but only with the assessment of the dermatologists as to whether the lesion was excised or not. This classification problem may be a useful alternative to the current emphasis on automated diagnosis.

5.2.1 Investigation Method

This investigation only used the University/Health-Waikato Image Set. The Sydney Image Set is not used because the lesions in this image set are all considered atypical (or suspicious) enough to excise. The format of this investigation is similar to the previous investigation, with the exception of class definition.

Class Definition

This investigation required new classes to be defined. In the previous investigation, the classes, melanoma or benign, were supplied by the pathologist. Here, the perception of dermatologists was used to assign classes to each of the lesions. This assignment was performed on the basis of whether or not the lesion was excised, and therefore, the two classes for this investigation were 'not excised' and 'excised'. 16 of the 73 lesions in the UHWIS were excised by the various dermatologists at Health-Waikato Dermatology Department. The remaining 57 lesions were not considered suspicious enough to excise.

5.2.2 Results

Results are reported similarly to the previous section. It should be noted that sensitivity and specificity figures are related to the lesions classed as 'excised'. For example, sensitivity here refers to the percentage of excised lesions that were classed as 'excised' by the classifier. From the results of this investigation, the suitability of this system as a method for implementing a screening system will be judged.

5.3 Human Comparison

The feature algorithms used in the previous two investigations are mostly based on guidelines developed for the detection of melanoma by humans. For example, most of the Clinical-view algorithms are based on the ABCD criteria of Friedman et al. (1985). Similarly, the ELM algorithms are based on the ABCD criteria of Stolz et al. (1994).

However, little investigation has gone into assessing whether or not these algorithms correspond to human perception in any way. If these algorithms do not reproduce human perception to some degree, perhaps other algorithms would be more useful in this context. This investigation evaluated how well the algorithms reproduced the human criteria on which they were based, and allows some conclusions to be made as to the suitability of the algorithms.

5.3.1 Investigation Method

Three dermatologists were asked to rate 40 lesions for firstly, the ABC of Friedman et al. (1985) for Clinical-view images, and secondly, the ABCD of Stolz et al. (1994) for ELM images. Diameter was not assessed as clinical perception is based on measurement. The images were obtained from the University/Health-Waikato Image Set. The lesions in the Clinical-view set were different from those in the ELM set.

This investigation was carried out over a period of four weeks. The ELM rankings were obtained in the first two weeks, whilst the second two weeks were used to gather the Clinical-view data. Two dermatologists participated for the duration of data gathering. One changed between the Clinical-view and ELM phases.

In each week, approximately 20 lesions were seen at a time. Each lesion was displayed using the original slide on a Kodak Carousel S-AV 1030 slide projector in a darkened room. The dermatologists were asked to rate the lesions on a Likert scale for the Clinical-view images (asymmetry was rated similarly to that presented in Stolz et al. (1994), that is 0, 1 or 2. Border irregularity and colour variegation were rated on a 0-9 scale). For the ELM images, they were asked to reproduce the ABCD ratings of Stolz et al. (1994). It should be noted that the dermatologists were not especially familiar with the ABCD rule of Stolz et al., although all were expert clinicians with significant experience in ELM assessment and had been exposed to this criteria set previously.

5.3.2 Statistics

Two methods of assessing the algorithms were used. Firstly, in the case of asymmetry, the method of Stoecker et al. (1992), who first proposed this method of measuring asymmetry, is used. In that paper, a dermatologist was asked to rate each lesion as asymmetric or symmetric. A threshold is then found, over which a lesion is considered algorithmically asymmetric. The threshold proposed by Stoecker et al. (1992) was 6%, and therefore, if any lesion had a minimum asymmetry value of over 6%, the lesion was considered asymmetric.

In this research, three major differences between the methods used here and that used by Stoecker et al. (1992) are evident. Firstly, three dermatologists rated each lesion, and secondly, the dermatologists stated the number of axes on which asymmetry is present (0, 1 or 2). A rating of '2' was required to consider a lesion asymmetric in the Clinical-view case. Each of the dermatologists was assessed separately. Finally, instead of selecting a 'best' threshold as was done by Stoecker et al. (1992), a ROC curve for each dermatologist was produced, and the area under this curve is reported. The ROC curve is obtained by varying a threshold that marks

the cutoff point for asymmetric/symmetric. For example, Stoecker et al. (1992) decided that a threshold of 6% was optimal. However, varying the threshold produces a different pair of sensitivity/specificity results, which are equally correct. The ROC curve present all of the different sensitivity/specificity combinations obtained through varying the threshold. The area under the ROC curve is a measure of the tests usefulness over the range of thresholds, while the point closest to the top-left corner is proposed as a 'best' combination, in the manner of Stoecker et al. (1992).

ELM-asymmetry is also assessed by this method, although in this case, a rating of '1' was sufficient to signify asymmetry (following the method described in Stolz et al. 1994). This investigation measures how well perception of dermatologists is reproduced by each of the individual algorithms, similarly to the Clinical-view method. It must be noted that in the method of Stolz et al. (1994), asymmetry is a collective assessment of colour, structure and shape asymmetry, not individual measures as is assessed here. Perhaps a better method would be to find a linear combination of the algorithms, covering colour, structure and shape asymmetry. However, because only 40 lesions were rated by the dermatologists, such a combination cannot be ascertained with any degree of certainty. That investigation is regarded as further work.

For the other comparisons between algorithms and human perception, a different method was used. The rating of the dermatologists was an ordinal variable, while each of the algorithms can be considered a continuous variable. It must be noted that there is no single 'best' method of measuring the association between these two types of variables. Here, Spearman's rho was used to measure the association of the algorithms with the relevant human rating. This statistic is a measure of the linear relationship between the ranking of the variables, rather than the actual variable values. For each algorithm and human rating pair, Spearman's rho was calculated using SPSS 9.0.1.

Colour and Border Contrast

The outcomes of the human comparison investigation were two-fold. Firstly, we wanted to see whether dermatologist perception was in fact accurately reproduced by algorithms. Secondly however, this investigation obtained required values for several of the algorithms.

For example, recall that colour variance features were calculated using squares, rather than at the pixel level used previously in the literature. This was based on the premise that dermatologist perception did not operate at the pixel-level. We therefore needed to evaluate what size box allowed the closest reproduction of human colour variegation perception. To do this, colour variance values were cal-

Table 5.1: Matchups between human perception and algorithms.

Friedman et al. (1985) ABCD criteria

Asymmetry	Shape Asymmetry Index
Border Irregularity	Irregularity Index Convex Hull Box Count
Colour Variegation	Variance Red, Variance Green and Variance Blue Variance $L\alpha\beta$

Stolz et al. ABCD criteria

Asymmetry	Shape asymmetry index Variance asymmetry of red, green and blue Variance asymmetry of $L\alpha\beta$ Colour asymmetry of RGB
Border contrast	Border contrast rating
Colour Variegation	See colour variegation above
Differential Structures	Variance of variance (RGB) Variance of variance $L\alpha\beta$

culated using six different box sizes, ranging from 1(pixel level) to 30 pixels. All of these results are compared to dermatologists' perception of colour variegation. Again, this procedure is performed for both Clinical-view and ELM images.

A similar procedure needed to be performed for the Border Contrast algorithm for ELM images. It was stated in the previous chapter that the slope gradient threshold used to class a slope as "sharply defined" was found to be 0.35. This value is the value that maximized the Spearman's rho correlation between border contrast and the dermatologists' perception of border contrast.

5.3.3 Results

Results from this investigation vary depending on the algorithms. The asymmetry results show the amount of agreement between dermatologists and the algorithms using the 'best' threshold. For the Clinical-view asymmetry algorithm, we also tested the 6% threshold proposed by Stoecker et al. (1992).

For the other algorithms, Spearman's rho values are presented for each pair of (dermatologist, algorithm), described in Table 5.1. We also note the statistical significance of the results. In the case of colour variance measures, the maximum rho value obtained over all of the box sizes is presented, along with the box size that attained the maximum. For border contrast, a range of rho values corresponding to differing threshold values are presented.

5.4 Experimental Variables

There are a number of differences between the Clinical-view and ELM systems that may be of importance when considering the results of these investigations. For example, differences in image quality between Clinical-view and ELM may be an important factor in the results of the systems. If the ELM system performs much better, is it because the ELM technique provides more useful information to the classifier, or is it because the Clinical-view images were of poor quality and thus of less use to the classifier?

In this section, we identify a number of such variables, and assess the potential impact of those variables on the results. These variables should be kept in mind when viewing the results and analysis presented in the subsequent chapters.

5.4.1 Algorithms

The algorithms present perhaps the most significant difference between the two types of images. In general, the algorithms are different for each image type, and obtain different information regarding the lesions. It makes little sense to analyse both sets of lesion images with the same set of algorithms, as the important information in the images is likely to be significantly different. For example, differential structures are of significance to ELM images, but not to Clinical-view images, as they are not visible. Evaluating differential structures therefore makes little sense for Clinical-view images, and not evaluating it for ELM images may ignore important information.

It is apparent therefore that the algorithms are required to be different. However, the quality of the algorithm implementation is also important. Consider if the Clinical-view algorithms are of near optimal quality, and measure what they are intended to measure. The ELM algorithms however, are not, and perform quite poorly. If the Clinical-view systems outperform the ELM systems, does this mean that we can conclude that Clinical-view is better than ELM?

Obviously not. The reason for the difference may be simply that the ELM algorithms were not adequate to obtain the information required. The difference in algorithms therefore, although necessary as the information in the different image types is not the same, represents a major variable to this research.

We have pointed out previously that the Clinical-view algorithms are likely to be close to optimal, as they have been proposed in earlier literature and in general have been utilised by a number of different researchers. The ELM algorithms however, are unlikely to represent the 'best' possible algorithms for ELM analysis, as they

have been proposed for the first time in this research. They have not undergone scrutiny from other researchers as have their Clinical-view counterparts. However, due to the lack of publication of ELM algorithms, there are no better sources. We cannot therefore control for differences in algorithm quality, as there is no way of ensuring that the two sets of algorithms are of similar usefulness.

We must state therefore, that the results of the analysis systems can be considered valid, given that *the best available Clinical-view and ELM algorithms were utilised.* Obviously, the best available Clinical-view algorithms are likely to be closer to optimal than those provided for ELM analysis.

5.4.2 Images

The difference in images between the two systems is the difference we are interested in testing. However, the images may be the source of other variables, which may affect the results. The first possible difference concerns the makeup of the two (Clinical-view and ELM) image sets. We have included lesions that have both image types available to avoid biasing the results, and hence this difference (and its impact on the results) has been controlled for.

Image Quality

Another we can be concerned about the difference in image quality between the two types of images. In general, it is much easier to obtain standardised ELM images than it is to obtain standardised Clinical-view images. This was especially apparent with the Sydney Image Set, where several Clinical-view images were of noticeably poor quality.

In this research, we controlled for this variable to some extent by removing those images that were of observedly poor quality. In general, the Clinical-view images were the type most likely to be of poor quality. It could be considered that these images are simply a reflection of the attributes of the image type, that is, it is generally easier to obtain high quality ELM images than Clinical-view images, but this view was not taken in this research.

Also, the Clinical-view algorithms have been used previously on image sets that were obtained in a clinical setting, and had good results obtained. It was considered unlikely that the quality of the remaining images would seriously effect the results of the Clinical-view systems.

Image Size

Another difference between the Clinical-view and ELM images was the tendency of ELM images to be larger than the slide boundary. In some cases, there was no identifiable skin component in the slide. This size problem was especially evident for the Sydney Image Set, and therefore a number of ELM image analysis features could not be used when analysing this image set. Because of this restriction on the algorithms available, the Clinical-view classifier for the Sydney Image set could be considered to have a wider range of available information on which to create a model.

The obvious resolution of this difference is to remove those lesion cases which were too large. This was the solution used for the UHWIS. However, for the SIS, a large number of lesions exhibited this feature. Due to the difficulty of obtaining image sets of this nature, we could not use this solution for this image set, and therefore this variable could not be controlled for. The potential impact of this variable may be to decrease the the ability of the ELM diagnosis system to accurately classify lesions.

5.4.3 Segmentation

The method of segmentation was identical for both sets of lesions. What was not identical was the apparent contrast of the image types. In some cases the lesions showed little contrast, and those lesion were removed from consideration. The major reason for removing these lesions is that segmentation was not necessarily identical for both image types, and therefore one image type may be adversely effected. For example, if a Clinical-view image was easily segmentable, but its ELM equivalent was not, extraneous data could become a source of confusion for the classifier. Thus the results of the ELM classifier would be adversely effected. We were not interested in comparing the abilities of the two images to be segmented correctly, but given that segmentation had occurred correctly, how well each of the image types performed.

5.4.4 Dermatologists Experience

In the case of the human comparison investigation, the experience of the dermatologists was an important factor. In particular, the dermatologists who took part in this investigation were more experienced with Clinical-view assessment than ELM assessment. Also, they were not particularly familiar with the ABCD criteria of Stolz et al. (1994), although all had been exposed to it previously. Therefore, when asked to make judgements about lesions, it is likely that the Clinical-view judgements may

be more accurate than those based on ELM.

5.5 Software

A variety of software was used in the investigations. Principally, SPSS 9.0.1 (SPSS Inc. 1989-99) was used to produce the Pearson correlation values, univariate analyses, and the stepwise logistic regression model. SPSS was also used to produce the Spearman's rho values in the Human comparison investigation. Microsoft Excel 97 (Microsoft Corporation 1985-1996) was used to tabulate the asymmetry data, together with ROC curves from data produced by SPSS and WEKA (see below). AccuRoc 2.1 (Accumetric Corporation) was used to calculate significant differences in the ROC curves.

Since SPSS did not have the ability to produce ten-fold cross- validated results, a separate method of testing the logistic model was required. The Logistic function of WEKA (Waikato Environment for Knowledge Analysis, version 3.1.8) performed this function. WEKA is a machine learning workbench developed at the University of Waikato. See http://www.cs.waikato.ac.nz/~ml/ for further information. SPSS was used to create the initial model, as the WEKA Logistic function was less informative than SPSS, and WEKA did not support stepwise variable selection. WEKA was then used to cross-validate the model and produce data for the ROC curves.

5.6 Chapter Summary

Three investigations were carried out in this research. The first two are designed to obtain support for the thesis argument, firstly in the context of automated diagnosis, and secondly in the context of 'dermatologist assessment'. The third investigation tested whether or not the algorithms reportedly based on human criteria actually reproduced perception of those criteria. The results of these investigations are reported next.

Chapter 6
Results

The first results presented in this chapter were obtained from the diagnosis investigation. Firstly, a summary of the Clinical-view descriptive data is presented. In particular, we tested to see which algorithms show a significant difference between the two classes of lesions (melanoma and benign). For example, is the Irregularity index significantly lower for benign lesions than melanoma? Following this data, the results from the Clinical-view diagnosis model obtained from the Sydney Image Set are presented. This structure repeats for the ELM data. These results allow resolution of the thesis argument in the context of automated diagnosis.

The next set of results concerns the 'dermatologist assessment' problem. The presentation of these results is identical to that of the diagnosis investigation, with the Clinical-view algorithm descriptive data and model presented first, followed by the ELM data and model. This set of results allows the evaluation of the thesis argument in the 'dermatologist assessment' context.

The final section presents the results of the Human Comparison investigation. In this section, the correlation of the algorithms to human perception is reported. This investigation has important consequences for the algorithms used in this research. If algorithms do not correlate well with the human perception they are intended to emulate, research should be expended on new algorithms which may be more useful in this context.

Two conventions are used in the results description. For correlations, **bold** marks correlations significant at the 0.01 level, while *italic* marks correlations significant at the 0.05 level. Secondly, the number of decimal places were reported similarly to SPSS 9.0.1.

6.1 Diagnosis (Sydney Image Set)

The Sydney Image Set was used in the diagnosis investigation, which is a similar investigation to previous research in this field. Four sets of results are presented, the first two from the Clinical-view investigation (Sections 6.1.1 and 6.1.2), and the second two from the ELM investigation (Sections 6.1.3 and 6.1.4).

6.1.1 Clinical-view Algorithms

This section presents a summary of the results of the image analysis algorithms after application to the Sydney Image Set. Table A.1 (Appendix A) contains the full results. These results show the range of values obtained for each of the algorithms, and also whether each of the algorithms produced significantly different values between melanoma and benign lesions. The algorithms that do show a significant difference may be the most use in the subsequent classification model.

Table A.1 shows the mean, standard deviation and difference in significance for each of the Clinical-view algorithms for the Sydney Image Set. Difference in significance is calculated by the Mann-Whitney U test, similarly to Seidenari et al. (1998) and Tomatis et al. (1998). From this table, it is apparent that only three features are significantly different (at the 99% level) between benign lesions and melanoma. A further four features are significantly different at the 95% level. These features are shown in Table 6.1. Although these algorithms show significant differences between benign and malignant lesions, large overlaps still exist. For example, the comparison between the Diameter of benign lesions and melanoma is shown in Figure 6.1. Diameter shows the most significant difference between the two groups (melanoma and benign), but it is apparent that a number of benign lesions have similar diameter values to some melanomas.

Table 6.1: Most 'different' Clinical-view features with the Sydney Image Set.

	Benign Lesions Mean± Std. Deviation	Melanoma Mean ± Std. Deviation	p
Red gradient	*33.194 ± 13.216*	*41.971 ± 20.721*	*0.039*
α gradient	*-1.168 ± 0.062*	*-1.132 ± 0.076*	*0.021*
Red variance	*314.806 ± 157.548*	*454.981 ± 338.362*	*0.049*
α variance	*0.007 ± 0.005*	*0.010 ± 0.006*	*0.031*
Chromaticity blue	**-0.031 ± 0.027**	**-0.010 ± 0.030**	**0.002**
Diameter	**190.456 ± 76.844**	**264.271 ± 93.241**	**0.000**
Box count	**1.727 ± 0.074**	**1.783 ± 0.061**	**0.001**

Of the Clinical-view ABCD criteria, border irregularity, colour variegation and diameter are all represented in Table 6.1. Asymmetry was not significantly different between the two groups, although it should be remembered that the Sydney Image

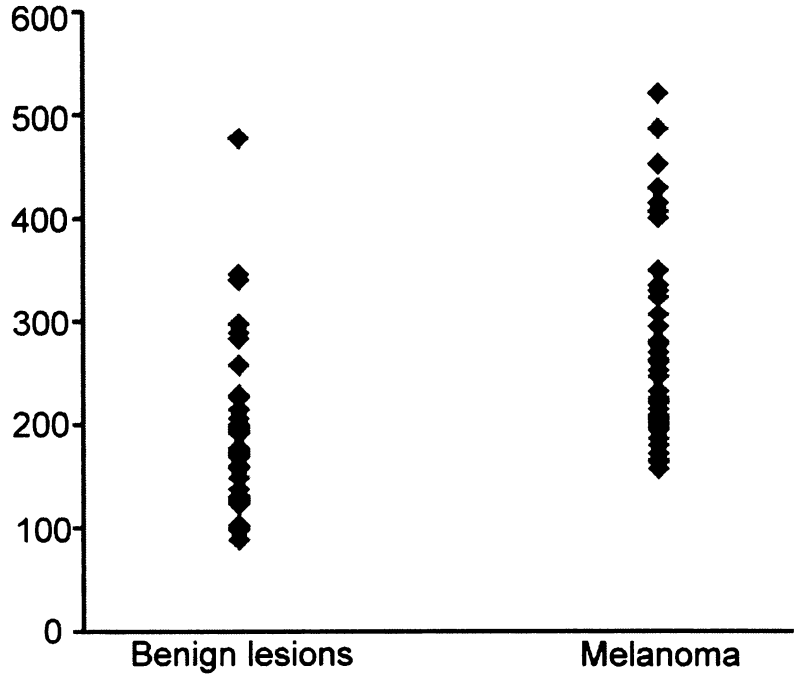

Figure 6.1: Scatter graph of Clinical-view Diameter showing large overlap.

Set consists of atypical naevi and melanoma, and atypical naevi show some features of melanoma.

6.1.2 Clinical-view Diagnosis Model

The results presented in this section show how well the system worked as a Clinical-view diagnosis system. These results form the first part of data that will allow the thesis argument to be resolved in the context of automated diagnosis. A logistic regression model was fitted to the results of the Clinical-view feature algorithms after application to the Clinical-view images from the Sydney Image Set. This section describes the logistic model.

There were 83 images in the SIS, of which 42 were melanoma, and 41 were atypical benign naevi. Stepwise logistic regression using a P_{in} value of 0.10 resulted in a model with three features. The features were Diameter, Chromaticity Blue and Chromaticity Green.

Table 6.2 summarises the model. B is the coefficient for each feature, while S.E. reports the standard error for each feature. The Wald statistic is a χ^2 statistic that measures the importance of each feature in the model. LR Test shows the likelihood-ratio test result for each feature. This value is a function of the difference between the model with the feature and the model without. Again, the statistic follows a χ^2

Table 6.2: Clinical-view diagnosis model summary.

Feature	B	S.E.	Wald	Sig.	LR Test	Sig.
Diameter	0.0111	0.0037	9.0808	0.0026	12.311	0.0005
Chromaticity Green	-28.1004	13.8159	4.1368	0.0420	4.621	0.0316
Chromaticity Blue	23.6414	9.8303	5.7838	0.0162	6.504	0.0108
Constant	-2.3365	1.1628	4.0374	0.0445		

distribution.

From the results of the Wald statistic and LR-test, the relative importance of each feature can be assessed. It is apparent that with a Wald statistic of 9.0808 and a LR-test result of 12.311, Diameter is the most important feature in the model. This result suggests quite strongly that melanoma tend to be larger in this image set than benign lesions, and this fits to some degree what is known about melanoma. Chromaticities of Green and Blue are less valuable to the model, but still significant features.

Table 6.3: Clinical-view diagnosis Pearson correlation coefficients.

	Diameter	Chromaticity Green	Chromaticity Blue
Diameter	1.000		
Chromaticity Green	0.078	1.000	
Chromaticity Blue	**0.284**	0.047	1.000

Table 6.3 shows the Pearson correlation coefficients for the three features in the model. These values show the linear correlation between each of the features. Only the Diameter/Chromaticity blue correlation is significant at the 0.01 level. This correlation is not strong, and therefore collinearity between the features can be ruled out.

Table 6.4 shows the summary goodness-of-fit statistics for the three feature model. Not surprisingly, the degrees of freedom (df) for the Model-χ^2 test is three (as there are three features in the model). Model-χ^2 returns a χ^2 value (25.712) significant at the 0.00 level, indicating that the model is significantly better fitted than the constant only model. The Homer-Lemeshow test has eight degrees of freedom (page 141 Hosmer & Lemeshow 1989), and returns a non-significant value if the model is

Table 6.4: Goodness-of-fit statistics for the Clinical-view diagnosis model.

Goodness-of-fit Statistic	Chi-squared	Degrees Freedom	sig.
Model-χ^2 (Log-Likelihood Test)	25.712	3	0.0000
Hosmer-Lemeshow Goodness-of-Fit	7.2290	8	0.5121

quite close to the perfect model. This model attained a non-significant p-value of 0.5121, indicating that the model fits the data.

The conclusion of adequate fit is further backed up by the results of cross-validation shown in Figure 6.2. The area under the ROC curve was 0.7596 with a standard error of 0.01664. An area of 0.7596 is informally considered to belong to a test with 'good' discrimination.

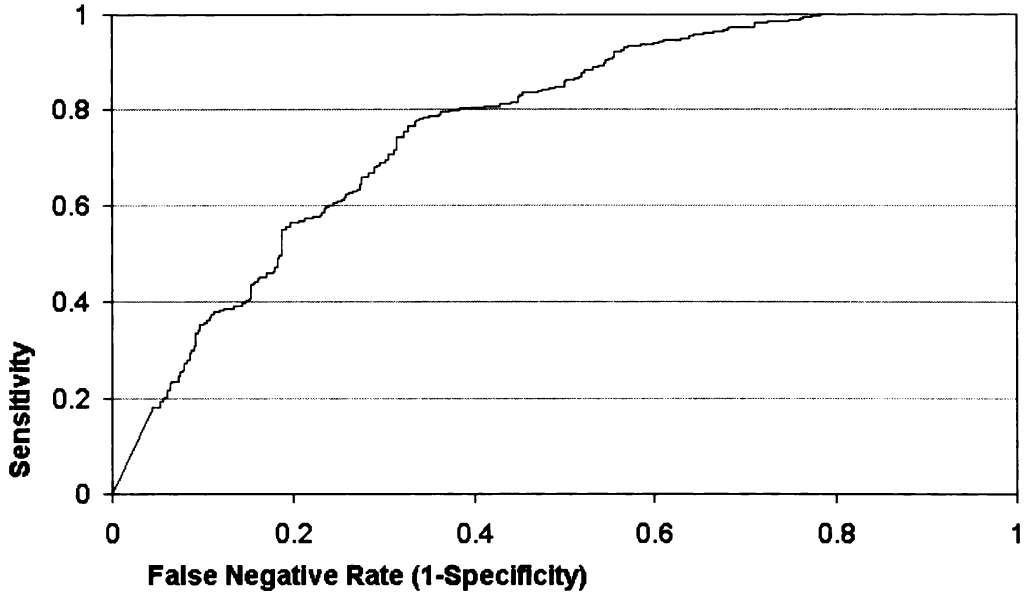

Figure 6.2: Cross-validated ROC curve for the Clinical-view Diagnosis model. The Area under the ROC curve (AUROC) = 0.7596, s.e.=0.01664.

These results appear lower than that reported previously in the literature. In particular, specificity is much lower than results of other research. For example, Schindewolf et al. (1993*a*) report sensitivity of 94% and specificity of 88%, results well in excess of that reported here. Similarly, the research of Green et al. (1994) and Ercal, Chawla, Stoecker, Lee & Moss (1994) report better results. This discrepancy is interesting, given that the image analysis algorithms for the Clinical-view images were taken from previous literature. The most apparent reason for the discrepancy in results between this and previous work is the image sets involved, although factors such as the lack of generalisation methods (for example, cross-validation) may have boosted reported results in previous literature. However, consider the results of Schindewolf et al. (1993*b*), who report a cross-validated accuracy result of 81%. This work is based on the same techniques as their previous research (Schindewolf et al. 1993*a*), who report sensitivity/specificity results of 94%/88%. The difference in results between these two papers indicates the difference a change in image sets can make.

The next section shows the results of the ELM diagnosis model. Again, the results of the algorithms are summarised briefly, and the ELM diagnosis model is then presented. It should be noted here that since thirty images of the eighty-three in this set exceeded the slide boundary by a significant margin, asymmetry and border contrast features were not calculated. Therefore, the assessment was based solely on colour and differential variance features.

6.1.3 ELM Algorithms

A similar set of results to those presented above is presented for the ELM-view of the Sydney Image Set. Table 6.5 shows the abbreviated results (See Table A.2 for the entire table).

Table 6.5: Feature range in the ELM features in the Sydney Image Set.

	Benign Lesions Mean ± Std. Deviation	Melanoma Mean ± Std. Deviation	p
α variance	**0.003 ± 0.002**	**0.006 ± 0.004**	**0.000**
β variance	**0.003 ± 0.003**	**0.005 ± 0.005**	**0.002**
Chromaticity green	**-0.015 ± 0.013**	**-0.024 ± 0.016**	**0.003**

All of the Variance of Variance algorithms except Var. Variance Blue showed significant difference ($p = 0.01$).

All of the Variance of variance algorithms (except blue) showed significant difference between melanoma and benign lesions, indicating that these algorithms may be useful in distinguishing between these lesion types. Similarly well represented were colour variance algorithms, although interestingly, variance of red, green and blue were not significantly different.

6.1.4 ELM Diagnosis Model

This section presents the ELM diagnosis model. The model was built using stepwise logistic regression with $P_{in} = 0.10$, which resulted in a model with six features. However, for a data set of this size, four features is the maximum, from the prescription of Tabachnick & Fidell (1996). Therefore, the two least significant features (as assessed by the Wald statistic and the LR test) was removed. The resultant model was further reduced to a three feature model as removal of the additional feature resulted in a model that was not significantly different ($p = 0.18$). It is also useful for the sake of comparison to have a model that has the same number of features as the Clinical-view model presented in the previous section. The three feature model is described in Table 6.6.

If the Wald statistic and LR test results from Table 6.6 are examined, the most significant feature in the model is α variance, with the other two features being

Table 6.6: Model summary for ELM diagnosis model.

Feature	B	S.E.	Wald	Sig.	LR Test	Sig.
Blue Variance	-0.0014	0.0011	4.7639	0.0291	5.466	0.0194
α Variance	359.9197	127.8006	7.9313	0.0049	10.962	0.0009
Variance of variance (Green)	0.0001	6.550E-05	4.0990	0.0429	6.857	0.0088
Constant	-1.1536	0.6601	3.0542	.0805		

less important. α variance and Variance of variance (green) are significantly higher for melanoma, which is what would be expected. Blue variance however, shows no significant difference between the two groups of lesions (see Appendix A, Table A.2). Table 6.7 shows the Pearson correlation coefficients for the features in the model. With the low strength correlations between all of the features, it was considered unlikely that significant problems with the model would be introduced.

Table 6.7: Pearson coefficients for ELM diagnosis model.

	Blue variance	α variance	Variance of variance (Green)
Blue variance	1.000		
α variance	0.144	1.000	
Variance of variance (Green)	**0.312**	**0.313**	1.000

Table 6.8: Goodness-of-fit statistics for the ELM diagnosis model.

Goodness-of-fit Statistic	Chi-squared	Degrees Freedom	Sig.
Model χ^2 (Log-Likelihood Test)	24.547	3	0.0000
Hosmer-Lemeshow Goodness-of-Fit	5.7782	8	0.6721

Table 6.8 shows the summary goodness-of-fit statistics for the refined model. Again, the log-likelihood test shows a significant improvement on the constant-only model ($p = 0.0000$). The Hosmer-Lemeshow test shows a non-significant p-value of 0.6721, similar to that reported for the Clinical-view model.

Figure 6.3 shows the cross-validated ROC curve from this model. The ROC has an area of 0.7279. Again, these results appear to be significantly lower than previous literature. For example, the paper by Menzies et al. similarly used logistic regression as a classifier, and obtain 92% sensitivity and 65% specificity. Binder et al. (1998) reported results of 90% sensitivity and 74% specificity. The systems developed by Bischof et al. (1998), Seidenari et al. (1998) and Seidenari et al. (1999) report even higher results, although the papers by Seidenari et al. do not report cross-validated results.

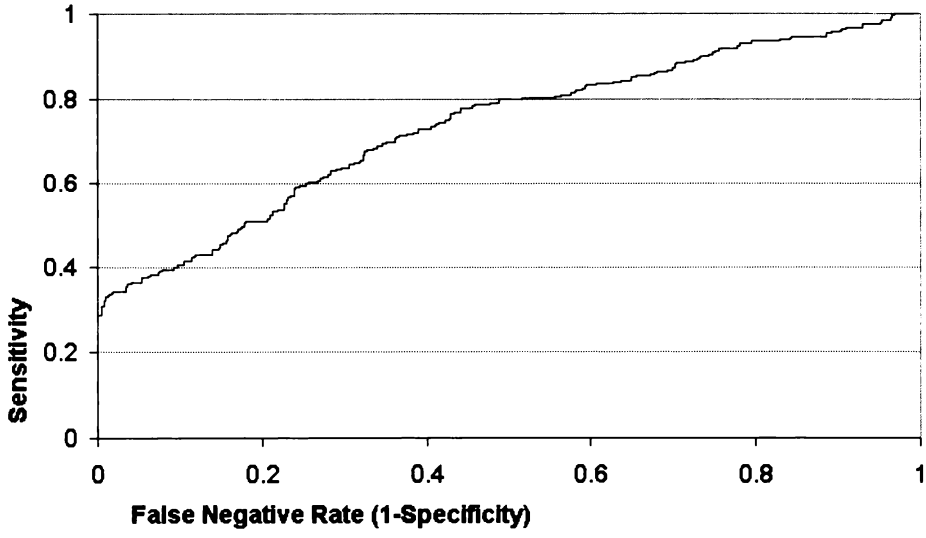

Figure 6.3: Cross-validated ROC curve for the ELM diagnosis model. The curve has an area of 0.7279 with a standard error of 0.01746.

6.1.5 Diagnosis Summary

The results of the diagnosis investigation were presented in two parts, the first describing the Clinical-view investigation, and the second looking at the ELM view. In each section, descriptive summaries of the algorithms showing significant differences between melanoma and benign lesions were presented. The results of the models were then reported.

It was shown that the results for both systems are generally lower when compared to previous literature. However, the difference in image sets and the use of techniques such as cross-validation in this research makes direct comparisons difficult. In the next chapter, these results are used to evaluate the thesis argument in the diagnosis context. The next section presents similar results for the 'dermatologist assessment' investigation.

6.2 Dermatologist Assessment (UHWIS)

The University/Health-Waikato Image Set was used in the 'dermatologist assessment' investigation. This investigation was intended to evaluate how well the decision of dermatologists to excise a lesion can be reproduced by algorithmic techniques, and whether Clinical-view or ELM images are more use for this task.

The structure of this section is identical to the previous. The Clinical-view results are presented first, beginning with a summary of the results of the image analysis algorithms. Again, the purpose of this data is to identify algorithms showing significant differences between the 'excised' and 'not-excised' groups. The Clinical-view dermatologist assessment model is then described, and this structure repeats for the ELM model.

Table 6.9 shows the Clinical-view algorithms that showed differences between excised and non-excised lesions significant at the $p = 0.01$ level. It is of interest to note that Box count and Diameter were the only shape algorithms to exhibit significant differences between the two groups. Also, none of the RGB variance algorithms produced significant differences, while α and β variance produced significant differences. Box count showed the most significant difference between the two groups.

Table 6.9: Feature range in the Clinical-view features in the University/Health-Waikato Image Set.

	Benign Lesions Mean ± Std. Deviation	Melanoma Mean ± Std. Deviation	p
Box Count	**-1.813±0.067**	**-1.868±0.047**	**0.000**
α Variance	**0.004±0.002**	**0.009±0.007**	**0.004**
β Variance	**0.005±0.004**	**0.011±0.007**	**0.001**
Chromaticity Green	**-0.032±0.019**	**-0.046±0.012**	**0.001**
Diameter	**303.967±120.302**	**440.522±132.951**	**0.001**

6.2.1 Clinical-view Dermatologist Assessment Model

The model was initially built using forwards stepwise regression with a P_{in} value of 0.05 and a P_{out} value of 0.10. This process resulted in a model with three features, which is approximately equal to the number recommended by Tabachnick & Fidell (which allows 2.8 features). This three feature model is described in Table 6.10.

From the Wald statistic and LR-Test, Diameter is the most significant feature in the model, suggesting that excised lesions tended to be larger than non-excised lesions. The other features are less important, and suggest that excised lesions have a higher α variance (as would be expected) and lower Chromaticity green values than lesions that were not excised (Table B.1). Table 6.11 shows the Pearson correlation results

Table 6.10: Clinical-view dermatologist assessment model summary.

Feature	B	S.E.	Wald	Sig.	LR Test	Sig.
α Variance	305.4107	127.1468	5.7698	0.0165	11.048	0.0009
Chromaticity Green	-51.8995	24.0794	4.6455	0.0311	5.025	0.025
Diameter	-0.0105	0.0034	9.3055	0.0022	13.375	0.0003
Constant	-9.1088	2.2044	17.0739	0.0000		

for these features. No significant correlation is found, and again, collinearity can be ruled out.

Table 6.11: Pearson coefficients for Clinical-view dermatologist assessment model.

	α variance	Chromaticity green	Diameter
α Variance	1.000		
Chromaticity Green	-0.191	1.000	
Diameter	0.075	-0.061	1.000

Table 6.12: Goodness-of-fit statistics for the dermatologist assessment Clinical-view model.

Goodness-of-fit Statistic	Chi-squared	Degrees Freedom	sig.
Model-χ^2 (Log-Likelihood Test)	31.335	2	0.0000
Hosmer-Lemeshow Goodness-of-Fit	6.9737	8	0.5395

Table 6.12 shows the summary goodness-of-fit statistics. The Model-χ^2 value shows significant improvement over the constant-only model. The Hosmer-Lemeshow statistic shows no significant difference between this model and the 'perfect' model (p=0.5395). However, the low number of lesions in the 'excised' group makes the value of the Hosmer-Lemeshow statistic questionable.

Figure 6.4 shows the cross-validated ROC curve from this model. The area under the ROC curve was calculated at 0.8510. Although these results appear better than those presented in the previous section, no comparison can be made with either those results or previous results presented in the literature, due to the difference in image sets.

6.2.2 ELM Algorithms

Similarly to the Clinical-view algorithms for this image set, most of the ELM algorithms showed significant differences in mean between the 'excised' and 'not-excised' lesions. This data is shown in Table B.2 in Appendix B. To summarise, all of the

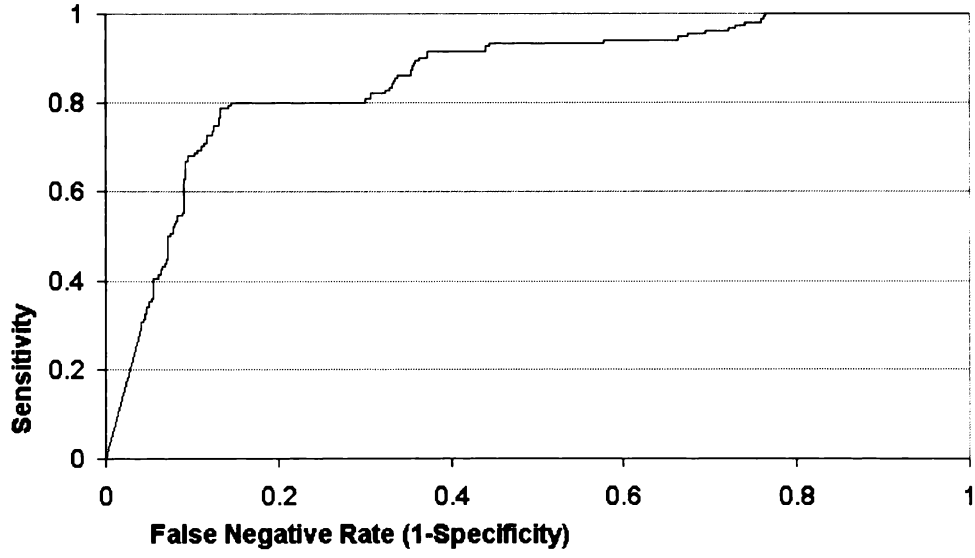

Figure 6.4: Cross-validated ROC curve for the dermatologist assessment Clinical-view model. AUROC = 0.8510, standard error=0.01452.

asymmetry algorithms show significant differences. Of the gradient measures, only β gradient shows a significant difference at the 0.01 level.

Colour variance algorithms in general show significant differences (p=0.05), with the exception of Blue variance. Red and green variances were not significant at the 0.01 level. The chromaticity algorithm results are mixed, with only Chromaticity red showing significant difference at the 0.01 level, while all of the Variance of variance algorithms are significantly different (p=0.01).

6.2.3 ELM Dermatologist Assessment Model

This section describes the ELM dermatologist assessment model. Stepwise regression with a P_{in} value of 0.05 and a P_{out} value of 0.10 resulted in a model with three features. This model is presented below in Table 6.13.

Table 6.13: ELM dermatologist assessment model summary.

Feature	B	S.E.	Wald	Sig.	LR Test	Sig.
RGB Asymmetry	0.0714	0.0287	6.1928	0.0128	6.684	0.0097
Green gradient	-0.0522	0.0196	7.0885	0.0078	9.568	0.002
Chromaticity Green	-56.3967	15.1679	13.8246	0.0002	19.493	0.0000
Constant	-2.9975	1.2195	6.0415	0.0140		

Chromaticity green is the most important feature in the model, and suggests that excised lesions tend to exhibit lower chromaticity values than lesions not excised. Both Green gradient and RGB asymmetry are also important, with the asymmetry

feature showing excised lesions are more asymmetric than non-excised. Table 6.14 shows the Pearson correlation coefficient for the features in this model. There is a significant correlation between RGB asymmetry and Green gradient, and also Chromaticity green and Green gradient, but this correlation is medium strength, and was not considered strong enough to remove either feature.

Table 6.14: Pearson coefficients for ELM dermatologist assessment model.

	RGB Asymmetry	Green gradient	Chromaticity Green
RGB asymmetry	1.000		
Green gradient	**0.430**	1.000	
Chromaticity Green	-0.164	**0.611**	1.000

Table 6.15: Goodness-of-fit statistics for the dermatologist assessment Clinical-view model.

Goodness-of-fit Statistic	Chi-squared	Degrees Freedom	sig.
Model-χ^2 (Log-Likelihood Test)	22.424	3	0.0001
Hosmer-Lemeshow Goodness-of-Fit	4.4594	8	0.8135

Table 6.15 shows the summary goodness-of-fit statistics for the model. The Hosmer-Lemeshow statistic again suggests a well fitted model, but a caution about the low number of 'excised' cases must again be given. Overall however, indications are that the model fits adequately. Figure 6.5 shows the cross-validated ROC curve for this model. The area under the curve was calculated to be 0.7909.

6.2.4 Dermatologist Assessment Summary

The 'dermatologist assessment' investigation represents a new method of screening skin lesions. In this section, results supporting this method have been presented. It has been shown that on a set of 73 skin lesion images, good results can be obtained by both systems, although it appears that the Clinical-view system may outperform the ELM system. These results are examined in detail in the following chapter. It must be noted that no comparison between these results and the results of the diagnosis investigation can be performed, due to the difference in image sets.

6.3 Human Comparison

In most of the previous literature in this field, feature sets have been developed by attempting to reproduce human specialist knowledge. Surprisingly however, little effort has been expended to test whether or not these algorithmic features capture

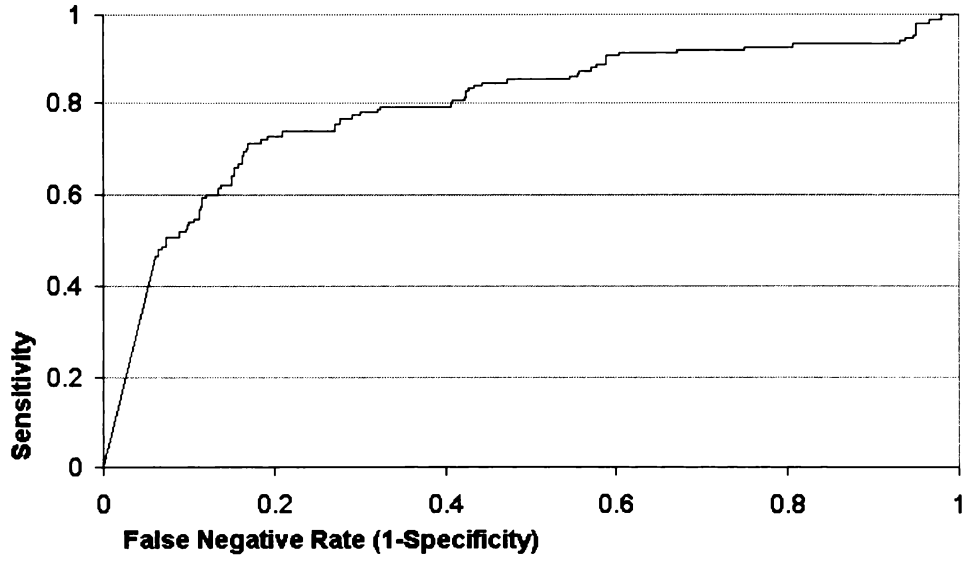

Figure 6.5: Cross-validated ROC curve for the dermatologist assessment ELM model. AUROC=0.7909, standard error = 0.01781.

that knowledge. This investigation is a retrospective look at the algorithms used in this research, to assess whether those that are based on human perception actually reproduce that perception to any degree. If they do not, further work on algorithmic equivalents of human criteria may be required to produce optimal results.

6.3.1 Clinical-view Algorithms

The Clinical-view algorithms used in this research were based mostly on the ABCD criteria of Friedman et al. (1985). Although this set of criteria has not been proved clinically, but is rather a rule of thumb, emulating these criteria may still be useful in identifying possible melanoma. This approach has been used in most previous work in Clinical-view diagnosis systems.

Each of the criteria of Friedman et al. is examined in turn (diameter is excluded, as dermatologist perception is guided by measurement). Asymmetry is assessed using a varient on the method of Stoecker et al. (1992), who first proposed the algorithm. Border irregularity and colour variegation are both assessed using the Spearman rank order correlation statistic (Spearman's rho). We also include figures comparing the three dermatologists, in an effort to highlight discrepancies in perception between dermatologists. However, no conclusions as to reasons behind any discrepancy are presented, as such work is out of the scope of this research.

Asymmetry

The results of the asymmetry correlation investigation are presented below. Table 6.16 shows the results for each of the dermatologists when compared to the shape asymmetry algorithm. Column 2 shows the percentage of lesions rated asymmetric by each dermatologist. For example, dermatologist 1 rated 32.5% of the forty lesions as asymmetric. The next column (column 3) shows the area under the ROC curve for that particular dermatologist.

Table 6.16: Asymmetry Comparison (algorithms and dermatologists) using the AUROC measure. The top-left 'best' point is marked.

	Percent asymmetric	Area under the ROC curve
Dermatologist 1	32.5%	*0.715*
Dermatologist 2	35%	**0.786**
Dermatologist 3	65%	0.667

The ability of this algorithm to match dermatologist perception had already been investigated by Stoecker et al. (1992) when they first proposed the method. This investigation therefore attempted to replicate their work. The first point of note is the discrepancy between the three dermatologists. Dermatologist 3 was more conservative than the other two dermatologists, rating significantly more lesions as asymmetric. On this set of lesions at least, asymmetry perception between dermatologists appeared to differ.

The algorithm also seemed least able to reproduce the perception of dermatologist 3. These two results suggests that dermatologist 3 had a different perception of asymmetry to the other two dermatologists. The perception of dermatologist 2 was reproduced best and the ROC curve for that dermatologist is shown in Figure 6.6.

Overall, agreement for all of the dermatologists was considerably lower than the figure of 93% reported by Stoecker et al.. The 'best' point on the ROC curve of dermatologist 2 as shown in Figure 6.6, had a sensitivity of 86% and a specificity of 70%, an overall accuracy of 75%. Therefore, we can conclude that although the method proposed by Stoecker et al. has been reported to agree with dermatologist perception, this result was not replicated to the same extent here. The difference in images and the use of different dermatologists may have contributed to the disparity in results. The results however, cast doubt on the conclusions of Stoecker et al. (1992).

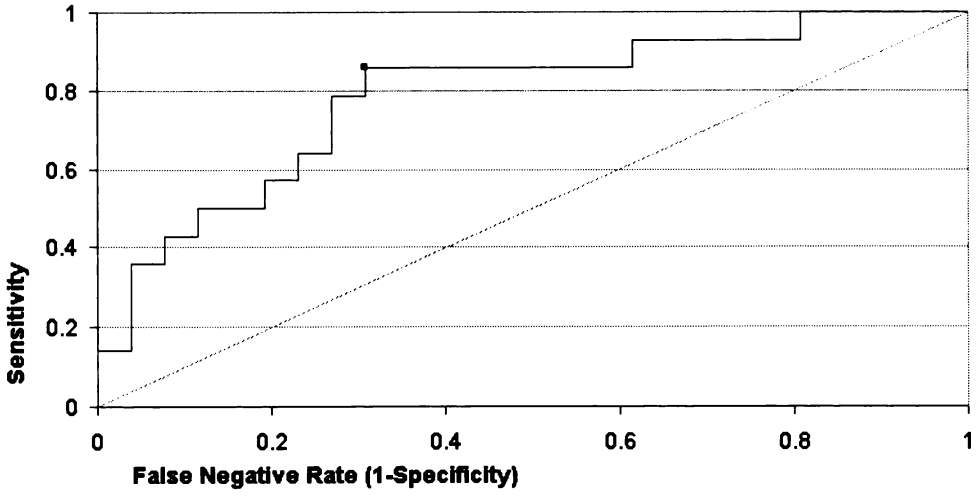

Figure 6.6: Asymmetry ROC Curve for Dermatologist 2. The 'best' point on the curve is marked

Border Irregularity

The three algorithms compared to dermatologist perception of border irregularity are Irregularity index, Box count and Convex hull. Table 6.17 shows the correlation between each of the dermatologists, and also each of the dermatologists and the algorithms. The negative correlation for Box count and Convex hull is to be expected, due to the implementation of the algorithms. Both of these algorithms produce higher values for more regular borders (less negative values in the case of Box count), whereas Irregularity index produces lower values for more regular borders.

Table 6.17: Correlation between dermatologist border irregularity measured by Spearman's rho rank order correlation.

	Dermatologist 1	Dermatologist 2	Dermatologist 3
Dermatologist 1	1.000		
Dermatologist 2	**0.717**	1.000	
Dermatologist 3	**0.430**	**0.515**	1.000
Irregularity index	**0.653**	**0.799**	0.279
Box count	**-0.414**	*-0.378*	-0.060
Convex hull	**-0.662**	**-0.817**	**-0.494**

The border irregularity algorithms also had varied results. Amongst dermatologists, correlation varied between 0.717 and 0.430. Dermatologists 1 and 2 were in reasonable agreement (0.717, $p = 0.01$), but dermatologist 3 was not (0.430 and 0.515 for dermatologist 1 and 2 respectively). These results indicate that perhaps the concept of border irregularity is not as defined as would be expected amongst dermatologists.

It is also apparent that of the three, dermatologist 3 produced the most outlying results, and comparisons with that dermatologist should be treated with caution.

When correlating algorithms with the dermatologists, several results were apparent. Firstly, the results of dermatologist 3 did not correlate strongly with any of the algorithms. Convex hull was the most correlated of the three algorithms investigated (-0.662, -0.817 and -0.414), which is perhaps a surprising finding given that this algorithm has not been utilised in previous research. Irregularity index also correlated well, although not with dermatologist 3 (0.653 and 0.799 for dermatologists 1 and 2). This result confirms the work of Golston et al. (1992) and validates its use in previous diagnosis systems. Finally, Box count did not correlate strongly with any of the dermatologists, although some correlation was shown. This result suggests that Box count is not suitable for reproducing expert perception of border irregularity.

Colour

As discussed in the previous chapter, the colour investigation served two purposes. Firstly, to see whether dermatologist perception was reproduced by the colour variance algorithms, but also to see which box-size resulted in the best correlation. In Chapter 4, it was surmised that dermatologist perception was unlikely to operate on a pixel level, and that algorithms that used a higher view may show more correspondence. The colour algorithms were therefore designed to operate on boxes, or squares containing the median colour of all of the pixels contained within. The size of this square was varied between 6 sizes, 1 (pixel level), 5, 10, 15, 20 and 30. The results presented in Table 6.18 show the maximum rho value for each of the colour variegation algorithms for these six box sizes. The box size that was used to obtain this result is shown in brackets. Table 6.18 also shows how well the colour variegation perception of dermatologists correlated. All of the correlations between dermatologists are significant at the 0.01 level.

Table 6.18: Correlation between dermatologist colour variegation measured by Spearman's rho rank order correlation.

	Dermatologist 1	Dermatologist 2	Dermatologist 3
Dermatologist 1	1.000		
Dermatologist 2	**0.859**	1.000	
Dermatologist 3	**0.604**	**0.727**	1.000
Red variance	**0.473**(1)	**0.499**(1)	**0.469**(5)
Green variance	0.257(1)	0.298(5)	0.306(5)
Blue variance	*0.352*(1)	0.305(5)	*0.343*(10)
L variance	**0.428**(1)	*0.448*(1)	**0.447**(20)
α variance	*0.334*(1)	0.218(1)	0.108(15)
β variance	0.239(1)	0.176(1)	0.093(1)

The most apparent result in Table 6.18 was that none of the algorithms correlated strongly with any of the dermatologists. This result suggests that the concept of Clinical-view colour variegation as interpreted by dermatologists is not being reproduced by colour variance algorithms, and perhaps the variance technique is insufficient for this application. Conversely, considerable agreement amongst dermatologists as to what constitutes colour variegation is found. The discrepancy shown between dermatologists concerning previous measures is not apparent here.

Regarding box size, little difference was found between the different box sizes, and the pixel level (box size 1) was chosen as most appropriate. For all of the algorithms, the results at the pixel level were within 1 standard deviation of the maximum rho value, and for the majority of algorithms, box size 1 had the 'best' correlation. Therefore, colour variance algorithms were applied at the pixel level, and the above results suggest that using pixel level algorithms to measure colour variance is reasonable.

6.3.2 ELM Algorithms

The ELM algorithms were similarly matched with human perception. Asymmetry is assessed in a similar manner to the Clinical-view asymmetry presented above, while again correlations of the other criteria are based on Spearman's rho.

Asymmetry

Table 6.19 shows the summary results for the investigation into how well the asymmetry algorithms reproduce dermatologist assessment. The table is similar in content to that presented previously for the Clinical-view asymmetry algorithm, but restructured to take account of the numerous comparisons.

Table 6.19: Summary of ELM asymmetry algorithms reproduction of human perception. The values show the area under the individual ROC curves for each dermatologist/algorithm combination. Bold/italics indicates whether the area is significantly different from 0.5

	Dermatologist 1	Dermatologist 2	Dermatologist 3
Red Variance	*0.721*	0.597	0.574
Green Variance	*0.726*	0.621	0.626
Blue Variance	0.667	0.560	0.580
L Variance	*0.712*	0.600	0.588
α Variance	0.667	0.568	0.522
β Variance	*0.715*	0.651	0.615
RGB Variance	**0.912**	**0.803**	**0.780**
Shape Asymmetry	**0.849**	*0.747*	**0.791**

The results in Table 6.19 indicate that only two algorithms of the eight investigated reproduced the perception of all dermatologists to any degree, namely RGB asymmetry and Shape asymmetry. The RGB asymmetry algorithm performs particularly well when compared to dermatologist 1, where it showed almost perfect agreement. Shape asymmetry, identical to the Clinical-view algorithm, also shows good reproduction of dermatologist perception, indicating perhaps that colour and structure asymmetry only rarely occur without shape asymmetry. None of the variance asymmetry algorithms, designed to measure structure asymmetry, reproduced dermatologist perception to any degree. It must be noted that the concept of ELM asymmetry is a combination of colour, structure and shape asymmetry, and as such, it would be optimistic to expect the individual algorithms to reproduce this perception.

Border Contrast

Similarly to the colour algorithms, border contrast also required dermatologist perception to find a final value. This value was the threshold over which a border was considered sharply contrasted, and was the value for which the Spearman's rho correlation was maximised. This threshold was found by an iterative procedure. Table 6.20 shows the variation in rho for the different threshold levels and is broken into three parts. The first part of the table shows the agreement between dermatologists concerning border contrast. The second part indicates the percentage of lesions that each dermatologist rated as '0' or having no sectors with sharp contrast. The third part of the table shows the variation in rho for the different threshold levels. The highest rho value occurs around the 0.35 threshold for each of the three dermatologists.

Overall, the algorithm exhibits medium to low strength correlation with the dermatologists using the 1.8 threshold. However, consideration of the nature of the dermatologists' responses is warranted, in particular, the number of '0' responses. Each dermatologist rated at least half of the images as having no border contrast in any sector. This feature of the image set used raises doubts about whether enough examples of border contrast were available. The accuracy of this measurement is therefore in some doubt.

The concern raised about the number of images with zero ratings is not easily overcome without preselecting images, which again may give skewed results. Perhaps the only reasonable solution is to use a larger number of lesions with similar distribution, and perform separate analysis on lesions rated zero and the lesions rated ≥ 1. The cost in terms of clinical time to produce such results is likely to be significant however. This algorithm must be considered early work.

Table 6.20: Spearman's rho value for different border contrast thresholds. All correlations are significant at the 0.05 level.

	Dermatologist 1	Dermatologist 2	Dermatologist 3
Dermatologist 1	1.000		
Dermatologist 2	0.641	1.000	
Dermatologist 3	0.645	0.618	1.000
Percentage of '0' results	75%	57.5%	52.5%
Threshold 1.0	0.377	0.211	0.246
Threshold 1.2	0.583	0.336	0.361
Threshold 1.4	0.571	0.388	0.372
Threshold 1.6	0.539	0.365	0.375
Threshold 1.8	0.553	0.416	0.395
Threshold 2.0	0.573	0.349	0.377
Threshold 2.2	0.540	0.406	0.440
Threshold 2.4	0.435	0.353	0.415
Threshold 2.6	0.364	0.172	0.189
Threshold 2.8	0.333	0.088	0.191
Threshold 3.0	0.217	0.131	0.133

Colour

As the ELM colour algorithms are the same as the Clinical-view algorithms, the box-size investigation described above needed to be repeated. Table 6.21 shows the results in the same format as the Clinical-view colour investigation.

Again, it is apparent that a high degree of agreement exists between dermatologists on colour perception. In the ELM ABCD criteria of Stolz et al., 'C' refers to the number of specific colours identified, and the specific nature of this criterion may have contributed to the high agreement between dermatologists. The algorithms themselves were also more correlated to dermatologist perception than the Clinical-view colour variance algorithms, suggesting that variance algorithms may be more applicable at the ELM view than the Clinical-view, possibly because of the more standardised images and relative lack of artifacts such as flash reflection.

Most of the algorithms had a medium strength relationship to perception of colour by the dermatologists. Red variance and the $L\alpha\beta$ variances are the algorithms most strongly correlated with all three dermatologists, and in the case of dermatologist 3, Red variance approaches the correlation achieved between the dermatologists. In general however, the correlations can at best be considered medium strength.

Finally, it can be concluded that similarly to the Clinical-view results presented above, the pixel level is an appropriate level for colour variance algorithms. Overall, little difference existed between the different box sizes.

Table 6.21: Correlation between dermatologist colour variegation measured by Spearman's rho rank order correlation.

	Dermatologist 1	Dermatologist 2	Dermatologist 3
Dermatologist 1	1.000		
Dermatologist 2	**0.779**	1.000	
Dermatologist 3	**0.799**	**0.754**	1.000
Red variance	**0.637**(1)	**0.479**(1)	**0.701**(1)
Green variance	0.304(5)	0.308(1)	*0.340*(1)
Blue variance	0.035(15)	0.123(15)	0.060(15)
L variance	**0.425**(1)	*0.355*(1)	**0.509**(1)
α variance	**0.530**(1)	**0.423**(1)	**0.560**(1)
β variance	**0.557**(1)	**0.437**(1)	**0.527**(1)

Differential Structures

Perception of differential structures is perhaps the most difficult of the ELM ABCD criteria to reproduce algorithmically. Here, the variance of variance algorithms are compared with the perception of dermatologists. It should be noted that the variance of variance algorithms do not reproduce individual structures, but were instead designed to measure the change in colour variance that may result from numerous differential structures in the lesion. Correlation of these algorithms with dermatologists' perception is shown in Table 6.22.

Table 6.22: Correlation between dermatologist differential structure perception measured by Spearman's rho rank order correlation.

	Dermatologist 1	Dermatologist 2	Dermatologist 3
Dermatologist 1	1		
Dermatologist 2	*0.375*	1	
Dermatologist 3	**0.839**	**0.466**	1
Var. var. red	**0.481**	**0.493**	*0.383*
Var. var. green	*0.347*	*0.319*	0.221
Var. var. blue	*0.327*	0.285	0.190
Var. var. L	*0.354*	*0.391*	0.222
Var. var. α	**0.453**	*0.376*	*0.348*
Var. var. β	*0.333*	*0.402*	0.192

The Variance of variance algorithms proposed to measure differential structures were considered unlikely to have strong correlation with the perception of the dermatologists. The main reason for this expectation is that dermatologists were asked to count the number of specific structures visible in the lesion, while the algorithms were measuring change in variance over the lesion. It was hoped that numerous differential structures would effect the change in variance values, but high correlations were not expected.

From Table 6.22, it appears that the expectations for these algorithms have been produced. None of the algorithms show a particularly strong correlation to dermatologists perception of differential structures, although several, such as Variance of variance red and Variance of variance α, show medium strength correlations. Therefore, it can be concluded that these algorithms, although they show some tendency to reproduce perception of differential structures, cannot be said to detect these structures adequately.

It must be noted however, that the dermatologists themselves showed some disparity. In particular, dermatologist 2 showed marked differences in perception than the other dermatologists, while dermatologists 1 and 3 were in strong agreement.

6.3.3 Summary

The above results detail the correlation of algorithms to the perception of dermatologists, as measured on a set of forty images. Given the nature of this problem, these correlations must be viewed as indicative rather than general, as the results may not apply to another image set. However, several interesting results are noted, and these are discussed more fully in the following chapter.

6.4 Chapter Summary

This chapter has presented the results of the investigations carried out in this research. The results of the diagnosis and 'dermatologist assessment' systems were presented, firstly the Clinical-view results, followed by the ELM results. The final part of this chapter looked at the results concerning the 'human comparison' investigation, and it was shown that some of the algorithms reproduced expert perception quite well, while other algorithms did not reproduce that perception at all. All of these results are examined in the next chapter, with a view to supporting (or otherwise) the thesis argument.

Chapter 7
Analysis

Now that the results of the investigations have been presented, these results are examined with a view to supporting (or otherwise) the thesis statement. This chapter is divided into five sections. First, a comment on the methodology is made, noting limitations on generalisability that should be kept in mind during the remainder of the chapter. The second section (7.2) looks at the diagnosis problem, and attempts to find support for the thesis from the results presented in the previous chapter. Similarly, Section 7.3 examines the thesis statement from the point of view of the 'dermatologist assessment' problem. The following section (Section 7.4) integrates these two analyses, and draws an overall conclusion about the thesis statement. In the final section (7.5), the feature analysis algorithms implemented for this work are examined, and indications for future researchers about the suitability of each of the algorithms are presented.

7.1 A Note on Generalisation

In a field such as this, it is difficult to produce results that can be considered generally applicable. Such generalisability must come from repeated testing of the techniques on various (large) datasets captured under controlled conditions. Before the analysis of the thesis statement is begun, it is worth keeping in mind some of the limitations that are a feature of this work. These limitations were described in Section 5.4.

The major effect of these limitations is to restrict the generalisability of the results. That is, the results are applicable to the image sets on which they were obtained, and not necessarily to any other image set (including the population of skin lesions). Techniques such as cross-validation, and the strict adherence to limits on the size of each logistic regression model suggests that the results are generalisable to a larger population of lesions. This suggestion cannot be confirmed however until further research is performed. Overall, care must be taken when reading the remainder

of this chapter not to interpret the results out of the context in which they were obtained.

> The results from each of the investigations cannot be generalised outside of the image and feature sets analysed. The methodology and adherence to guidelines restricting the size of the models tends to suggest that the results may generalise well to a larger population of lesions. However, this conclusion cannot be supported without further research.

7.2 Diagnosis

Bearing these limitations in mind, we begin with an analysis of the results of the diagnosis investigation. The purpose of the diagnosis investigation was to evaluate the problem addressed in Schindewolf et al. (1993*b*), namely, are Clinical-view or ELM images more use in an automated diagnosis system?

It appears from Figures 6.2 and 6.3 that both the Clinical-view and ELM systems performed quite well on the Sydney Image Set. Following the model building phase, the logistic regression models for both systems used three features to classify lesions. The Clinical-view features were Diameter, Chromaticity Green, and Chromaticity Blue. For the ELM model, the features used were α variance, Chromaticity Blue and Variance of Variance Green. All features showed significant differences between the two classes of lesion with $p = 0.05$.

It should be noted here that the logistic regression models derived for each system are not necessarily optimal, but represent 'good' models. From the results shown in the previous chapter, we may make some conclusion about whether ELM or Clinical-view is better for diagnosis using this image set.

Figure 7.1 shows the comparison in cross-validated ROC curve results for the two systems. The grey ROC curve in this graph shows the results of the ELM system. From inspection of Figure 7.1, two features are apparent regarding the combined ROC curves. Firstly, the ELM system outperforms the Clinical-view system at very high specificity levels, and correspondingly low sensitivity levels. The two curves cross at the point (sensitivity=51%, specificity=82%), and from this point the Clinical-view system produces the best results. The second feature is that in general, the ROC curves tend to show little separation. The most separation after sensitivity rises above 85% and specificity is correspondingly below 50%. Before this point, differences between the two curves are small, suggesting that the two systems are performing similarly. However, after this point, a clear advantage is obtained by the Clinical-view system, contrary to the thesis statement.

In terms of significant differences, we can assess the difference in area underneath

each ROC curve and calculate whether or not the difference between the areas is significant. However, assessing the difference between the two areas is not guaranteed to be completely accurate, as different ROC curves may have similar areas. It also does not take into account the fact that different sections of the ROC curve may be of more interest than others. In this case, it would be quite reasonable to specify that higher sensitivity figures were of more interest than specificity figures. Therefore, we cannot place undue emphasis on these results and they must be treated as indicative only.

Table 7.1: Areas under the ROC curves (AUROC) for Clinical-view and ELM diagnosis systems.

System	AUROC	Std. Err
Clinical-view	0.7596	0.01664
ELM	0.7279	0.01746
Difference	.0317	.0241
P-value (difference in area)	0.0944	
Estimated Normal 95% confidence interval (difference in areas)	-0.0156	0.0790

From the analysis of areas shown in Table 7.2, we can see that although the Clinical-view area is larger than the ELM area, the difference between the two areas is not significant ($p = 0.05$). However, it should be noted that this figure is probably understated due to the crossover of the two curves. Overall, it appears that the Clinical-view system is performing marginally better than the ELM system.

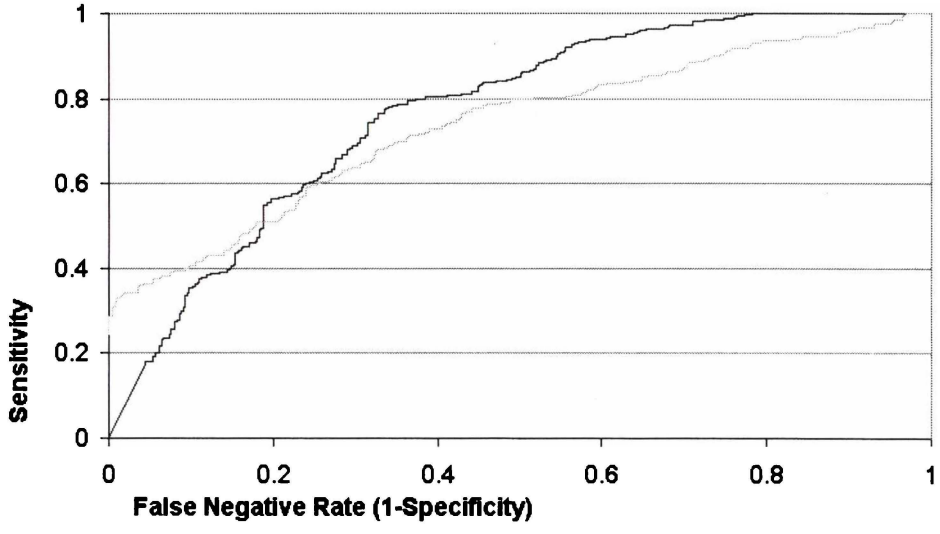

Figure 7.1: Cross-validated results for the Clinical-view and ELM diagnosis systems. Clinical-view results are shown in black, ELM results are shown in grey.

It is of interest to note that the clinical accuracy of dermatologists exceeds the results

reported for the Clinical-view system by a large margin. For example, Grin et al. (1990) and Morton & MacKie (1998) report sensitivity of around 85% with specificity of approximately 99%, which is far in excess of that for the Clinical-view system. However, in calculating specificity, all non-melanomas were considered, including non-melanocytic naevi. A large number of the lesions were carcinomas (either basal cell or squamous cell) which generally appear quite different to melanoma. Inclusion of these lesions in specificity calculations is likely to have artificially increased the specificity to an unknown degree.

Similarly for the ELM case, Nachbar et al. (1994) and Binder et al. (1999) report on the clinical accuracy of the ELM ABCD criteria of Stolz et al. (1994). Nachbar et al. reports that sensitivity for melanoma was 92.8% and specificity was 91.2% using this set of criteria on a set of 172 lesions. Similarly, Binder et al. (1999) report on the abilities of 17 dermatologists using the ELM ABCD criteria. Five were first year residents, eight were certified dermatologists with between 4 and 15 years experience, and four of the raters were expert dermatologists working primarily on pigmented skin lesions. Ability to recognise melanoma was very dependent on experience, but the average sensitivity/specificity for the 17 dermatologists was 80.1%/78.6%. These results are again superior to those produced by the ELM system. Both systems therefore are achieving well below the results obtained by human experts in previous research.

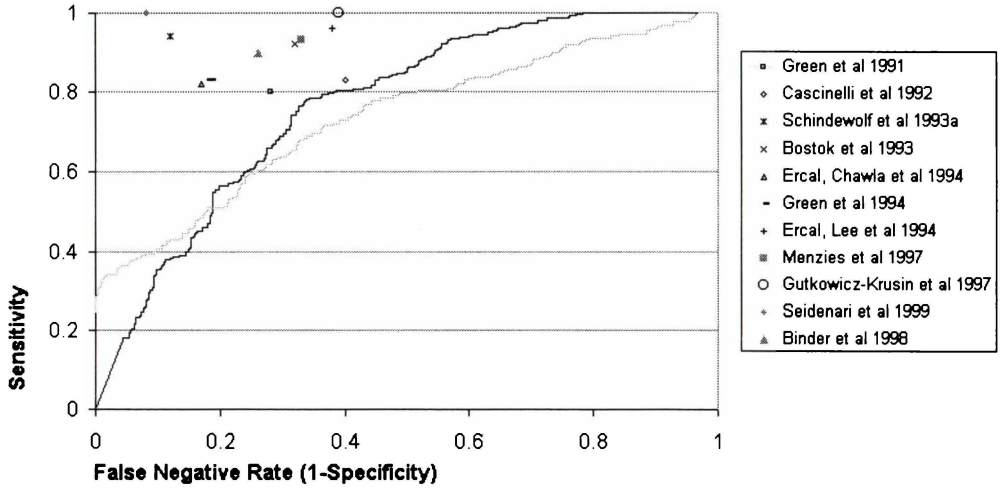

Figure 7.2: Comparison of the diagnosis systems of this research and previous research systems

As we indicated in the previous chapter, the results of the Clinical-view and ELM systems tend to be considerably lower than those of previous automated systems. Figure 7.2 shows a visual comparison. There may be a number of reasons why the results we have shown are lower than those presented in previous research. In fact,

any of the differences in the methodologies may be enough to explain the difference in results. For example, the difference in image sets between projects may explain the difference in results. Similarly, image analysis algorithms differed widely between projects. Other factors, such as image quality and the use of techniques such as cross-validation are also likely to have differed widely. We cannot state for certain that the results obtained here are worse in general than the results of any of the other research projects in this area, but given the disparity in results, and the comparative smallness of the image set used here, such a conclusion is likely. Examining why the results of the systems presented here were worse than previous efforts may be useful to make the results of this investigation more useful. To summarise:

> Overall, the results of the systems are similar. However, the Clinical-view system outperforms the ELM system, contradicting the thesis statement. The difference in results was found not to be significant when assessed using the difference in ROC curve area technique, although the significance is likely to be understated. The results achieved by both systems are lower than that achieved by experienced clinicians, and tend to be under that reported for previous research.

7.2.1 Why?

We have shown in the above section that the results of the two diagnosis systems indicate that the Clinical-view system is performing better than the ELM system, although the results were not statistically significant. This result contradicts the thesis statement to some degree. We have shown no indication that the ELM techniques used in this research have more benefit than the Clinical-view system. This result is similar to that of Schindewolf et al. (1993*b*), and therefore must raise some doubt over the emphasis on ELM images in current automated diagnosis literature.

However, we cannot assume that these results are generalisable to the lesion population as a whole, any more than the results of Schindewolf et al. (1993*b*) were generalisable. There may be a number of reasons that the systems performed as they did, and these reasons are examined in this section. To recap on the differences between the two systems that were identified in Chapter 5:

1. Clinical-view and ELM algorithms.
2. Image quality.
3. Clinical-view and ELM images.

We now look at each of these differences in turn, and examine their potential impact on the results presented above.

Clinical-view and ELM algorithms

One of the most obvious differences between the two systems was the choice of image analysis algorithms. In the Clinical-view case, we were presented with relatively clear guidelines on which algorithms to implement, based on previous work in the field. In the case of ELM however, algorithms had to be developed from scratch, due to the scarcity of ELM algorithms reported in the literature.

This discrepancy, not so much the intent of the algorithms, but the quality of implementation, is an uncontrolled variable in this investigation. The results may be an artifact of the difference in algorithms quality, rather than strictly a result of the difference in images used by each system.

Image quality

Quality of images was an issue that may go some way to explaining the result. Although the Clinical-view images were generally of 'poorer' quality due to the lack of standardisation, the Clinical-view algorithms have been used previously on non-standardised images. Removal of the most obvious poor quality images controlled for this feature, and it is considered unlikely to hinder the Clinical-view results by a large degree.

The major quality issue with the ELM images was that they were often larger than the slide area, and were therefore cropped. This cropping restricted the types of image analysis algorithms that could be used for analysis. In particular, asymmetry and border contrast algorithms could not be used. The lack of these algorithms represents a large loss of data to the classifier. The likely effect of this restriction is to reduce the accuracy of the ELM system to some degree. This feature could not be controlled for in practical terms, and may cause major effects on the results of the ELM system.

Clinical-view and ELM images

This difference is the difference we are interested in exploring, namely which image type is more use in an automated system. We have shown above that Clinical-view images appear to be slightly more use than ELM images in this context. However, this finding contradicts what we know from clinical experience, namely that the ELM technique allows more accurate separation of the two groups of lesions.

The finding here implies that ELM algorithmic analysis is currently not capable of utilising the information known to exist in ELM images. We may conclude therefore that advances in algorithmic analysis techniques may increase the results of ELM

based systems. The same cannot be said for the Clinical-view algorithms, which have been looked at over a number of years.

7.2.2 Diagnosis Summary

In summary, five possible explanations for the results we have shown have been identified:

1. The results reflect the difference in algorithm quality between Clinical-view and ELM image analysis algorithms.
2. The results reflect the limitations on feature choice imposed by the size of ELM images.
3. The results reflect limitations derived by the quantity of experimental data.
4. The results reflect the relative superiority of Clinical-view images in this context.
5. The results reflect some combination of these four points.

The first two options reflect limitations derived from the uncontrolled variables in this research. The results presented above can only be considered valid when allowances have been made for these limitations. Option 3 will always be a restriction on this type of research, as no data sets that are representative of the population of lesions are available.

In our opinion, option 4 is unlikely when considering the literature regarding the superiority of ELM in the clinical setting, although it is possible that Clinical-view images hold as much useful data as the ELM images. This situation would only occur when the lesions were advanced to a degree where the diagnosis becomes apparent at both views. We do not consider this a likely prospect given the composition of the Sydney Image Set.

We may therefore state that these results are valid, given within the limitations imposed by the Sydney Image Set, and the implemented algorithms.

In order to make the results more general, the remaining two variables shown above (algorithm quality, and image size) would need to be controlled. Such control was out of the scope of this research, and are important considerations for future research.

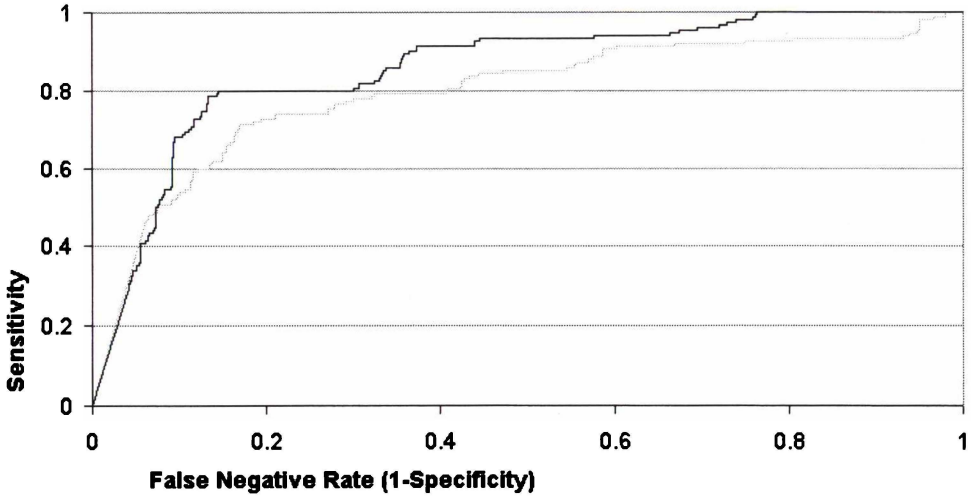

Figure 7.3: Comparison of the cross-validated ROC curve results for the Clinical-view and ELM 'dermatologist assessment' systems. Clinical-view results are shown in black and ELM results are shown grey.

7.3 Dermatologist Assessment

The 'dermatologist assessment' investigation evaluated how well the decision of dermatologists to excise a lesion could be reproduced. In lay-terms, we are trying to get the computer to perform the assessment function of a dermatologist, rather than the diagnosis function of the pathologist as was the case in the diagnosis investigation.

This analysis is presented similarly to the previous section. Three features were used in both of the logistic regression models, and all showed significant (p=0.05) differences between the two classes. In the Clinical-view model, the features were α variance, Chromaticity green and Diameter. For the ELM model, the features were RGB asymmetry, Green gradient and Chromaticity green. RGB asymmetry was the only feature based on human guidelines, and it showed significant correlation with the perception of dermatologists regarding asymmetry. The results of the systems are compared in Figure 7.3. The line in black shows the Clinical-view ROC curve, while the line in grey is the ROC curve of the ELM system. Sensitivity refers to the percentage of lesions that were classed as 'excised' by the classifier that were excised by the dermatologists . Specificity is the percentage of lesions classed as 'not excised' by the classifier that were not excised by dermatologists.

It is apparent from Figure 7.3 that in general, the Clinical-view system outperforms the ELM system. Overall, although the differences between the two curves are small, the results of the difference in area analysis suggests that the difference is significant at the 99% level. Therefore, we can conclude that it is very likely that the Clinical-view system is performing more accurately than the ELM system.

Table 7.2: Areas under the ROC curves (AUROC) for Clinical-view and ELM diagnosis systems.

System	AUROC	Std. Err
Clinical-view	.8510	0.01452
ELM	.7909	0.01781
Difference	.0601	.0230
one tailed p value	0.0045	
Estimated Normal 95% confidence interval (difference in areas)	0.0151	0.1051

The results of this investigation also indicate whether or not the 'dermatologist assessment' system presents a plausible method of screening lesions. To summarise, the Clinical-view system in general performed better than the ELM system. In the 'best' case, the Clinical-view system achieved average sensitivity and specificity results of 91% and 63% respectively. Therefore, 91% of the decisions to excise could be replicated by the Clinical-view system, while 63% of the decisions not to excise were also reproduced. From these exploratory results, it can be concluded that reproducing dermatologist's assessment of excision is certainly possible.

The Clinical-view dermatologist assessment system in general outperformed the ELM system. Although differences were slight, they are statistically significant ($p = 0.01$). However, the composition of the image set is a major limitation of this result, and the result must be interpreted cautiously.
It appears from the results that reproducing the assessment of excision is a plausible goal, although the problem has not been solved in this research.

7.3.1 Why?

Given the exploratory nature of this investigation, it is difficult to postulate ideas why the two systems performed as they did, in particular, why the Clinical-view system performed slightly better than the ELM system. The excised lesions were all observed under ELM, and we could expect that the ELM view would take precedence over the less sensitive and specific Clinical-view. The results of the two systems however, suggest that this is not the case.

Similarly to the diagnosis result, this result contradicts the thesis statement. However, the early nature of this investigation makes definite conclusions impossible. At best, this result can be viewed as indicative. In this section, we recap the differences between the Clinical-view and the ELM dermatologist assessment systems. The identified differences are similar to those presented previously for the diagnosis investigation, and are shown below:

1. Clinical-view and ELM algorithms.
2. Clinical-view and ELM images.

Clinical-view and ELM algorithms.

The difference in algorithm quality is again likely to be an issue. Because the image analysis algorithms play such an important role in these types of systems, any discrepancies between the quality of the algorithms is likely to have a large effect on the results. The only solution to this problem is firstly to acknowledge the potential impact of the algorithm difference, and secondly to ensure that algorithms are published to allow peer review and improvement.

Clinical-view and ELM images.

Given the thesis statement, we assumed that the ELM images would allow more accurate assessment of which lesions to excise. In the diagnosis investigation, the assumption that ELM would be more use than the Clinical-view was justified by results showing just that in the clinical setting. However, there is reason to believe that this assumption is not justified for the 'dermatologists assessment' investigation. When queried about the use of Clinical-view in the decision to excise a lesion, one of the dermatologists involved stated:

> "Clinical view is what we're all trained in and have confidence in. Many dermatologists have not had adequate training in dermoscopy. So most lesions that could be melanomas are excised on the basis of clinical view. But even in inexperienced hands, ELM can save a few unnecessary excisions especially of seborrhoeic keratoses as they often have easy-to-recognise benign features. But very experienced dermoscopists will have greater confidence in their ELM skills and will therefore base more decisions on their findings in this medium.
>
> Moles are frequently removed for cosmetic reasons - ELM is irrelevant to the decision-making as the excision is at the request of the patient. Typical reasons: the mole looks unsightly, the mole gets caught when combing the hair/getting dressed, psychological fears...." (Personal communications: Oakley 1999)

This suggests that lesions are more likely to be excised based on the Clinical-view appearance, either because of dermatologists having more experience with Clinical-view assessment, or perhaps due to lesions being removed for cosmetic reasons.

These two reasons may partially explain the results obtained by the 'dermatologist assessment' systems.

The second of these reasons may be a source of concern. If the excised lesions in the dataset were not excised for malignancy considerations, there exists no standard basis for excising the lesions. At the very least, this lack would cause the classification system difficulty, and make even the tentative conclusions given here subject to considerable doubt. As we have pointed out previously, we consider it likely that the majority of excised lesions in this image set were excised due to malignancy concerns. Firstly, seven of the sixteen lesions were melanoma and it appears likely that these lesions were excised for malignancy concerns. Of the remaining nine lesions, eight were considered 'suspicious' by the dermatologists in clinical surveys of the slide images. Only one was not considered suspicious in these surveys. Care must be taken in future studies of this nature to ensure that the excised lesions were excised primarily for malignancy concerns.

7.3.2 Dermatologists Assessment Summary

Again, we can identify several possibilities to explain the results:

1. The results reflect the difference in algorithm quality between Clinical-view and ELM image analysis algorithms.
2. The results reflect limitations derived by the quantity of experimental data.
3. The results reflect the relative superiority of Clinical-view images in this context.
4. The results reflect some combination of these three points.

As we have pointed out, possibility 1 is a consideration for both investigations, and is one of the major uncontrolled variables in this research. Possibility 2 is likely to be significant when the composition of the UHWIS is considered. Only sixteen of the lesions were excised, and therefore the image set cannot be considered representative of the lesion population to any great degree.

It should also be noted that the class of the lesion was based on whether or not the lesion was excised. The decision to excise was made by individual dermatologists during the course of skin lesion clinics. We do not have a single dermatologist on which we can train, and therefore the classifier may be confused by different perceptions of what should be excised, and what should be left. It is undesirable to classify on the basis of more than one human opinion, but that is a restriction of the University/Health-Waikato image set. In future research, care should be taken

to ensure that the opinion of one expert only should be used, in order to avoid the difficulties associated we training on the opinion (possibly contradictory) of multiple human specialists.

Finally however, we have shown that is it quite plausible, and indeed quite likely that the Clinical-view images were in fact more use for this classification task, and hence the third explaination may be quite valid in this context. Future work would be required to test this proposition.

7.4 Thesis Revisited

In this section, the thesis statement is revisited in light of the previous results and analysis. The thesis statement is reproduced below.

> Given the current published state of the field and a limited set of real world images, ELM-images are more useful in an automated screening system for skin lesions than Clinical-view images.

Two questions have been asked by this thesis statement. Firstly, are ELM images more use than Clinical-view images in an automated diagnosis system? Secondly, are ELM images more use than Clinical-view images in a system intended to reproduce the decision to excise a lesion?

From the above analysis, these two questions can be answered. In the case of the first, there is no evidence that ELM images are more use than Clinical-view images in a diagnosis system. In the case of the second, the answer is similarly no, there is no evidence to suggest that ELM images are more use than Clinical-view images in the 'dermatologist assessment' system. There is some slight evidence to the contrary, but given the exploratory nature of the research, the question cannot be answered conclusively.

There was also an implied question associated with the thesis, namely, can algorithmic techniques reproduce the dermatologists decision to excise a lesion? From the results presented previously, it can be stated that there is early evidence that suggests that algorithmic techniques can be used to reproduce the decision to excise.

> For the diagnosis problem, results suggest that ELM images are not more useful than Clinical-view images, although the differences between the two systems were not statistically significant. Therefore, no support is found for the thesis statement when considering the diagnosis problem.
> For the 'dermatologist assessment' problem, results indicate that the Clinical-view system performed better than the ELM system. Therefore, the thesis statement cannot be supported for the 'dermatologist assessment' problem.

7.5 Algorithms

This section looks at the results collected concerning the algorithms implemented for this research. There were two types of algorithm results collected. The first results were descriptive data for each of the algorithms. This data showed the means and standard deviations for melanoma and benign (or 'excised' and 'not-excised') lesions. The significance of the difference in means was also calculated. These results indicate which algorithms produced different values for the two classes in both of the classification problems. For example, were melanomas more asymmetric than benign lesions? Were 'excised' lesions larger than lesions that were 'not-excised'?

The second set of results concerned the performance of some of the algorithms in reproducing the perception of dermatologists. This dataset allows some judgement to be made as to whether the algorithms based on human criteria reproduce dermatologists perception of those criteria. All of this data is summarised in Table 7.3 (Clinical-view) and Tables 7.4 and 7.5 (ELM). The significance of these two sets of results for the algorithms is examined here, firstly in the Clinical-view case and secondly, for ELM.

7.5.1 Clinical-view

The Clinical-view algorithms for the most part have been used in previous automated diagnosis system research. Therefore, it would be reasonable to expect most of the algorithms to show significant differences between melanoma and benign naevi, together with high correlation with human perception for algorithms based on the ABCD of Friedman et al. (1985). However, from Table 7.3, it is apparent that the majority of the algorithms did not distinguish well between the two classes for the diagnosis problem. In particular, some of the more popular algorithms, such as Irregularity index, Asymmetry index and Colour variances of red, green and blue showed no significant difference between the two sets. Such results are interesting, considering the use of these algorithms in previous literature, although it must be remembered that the Sydney Image Set consists of atypical benign naevi and melanoma. Few obviously benign naevi were included in the image set, and this may have had an effect on the usefulness of the algorithms. We look at previously reported results of these algorithms below.

Overall however, algorithms showing significant differences between melanoma and benign lesions tended to match what is known about melanoma. For example, the variance algorithms all showed higher variance for melanoma. Box count was more negative (representing more irregularity) for melanoma, and melanomas tended to have larger diameters than benign lesions.

Table 7.3: Clinical-view algorithms at a glance.

Algorithm	Diagnosis		Dermatologist Assessment		Human Correlations comment
	$p < 0.05$	$p < 0.01$	$p < 0.05$	$p < 0.01$	
Colour Gradient Algorithms					
Red gradient	Y		Y		
Green gradient					
Blue gradient					
L gradient					
α gradient	Y				
β gradient					
Colour Variance Algorithms					
Red variance	Y				Red, green and blue variance showed slight to medium correlation.
Green variance					Red variance was highest of the variance algorithms.
Blue variance					
L variance					L variance was the next most correlated variance algorithm
α variance	Y		Y	Y	after Red variance ($\bar{p} \approx 0.441$. α and β showed little correlation with
β variance			Y	Y	human perception.
Relative Chromaticity Algorithms					
Chromaticity Red					
Chromaticity Green			Y	Y	
Chromaticity Blue	Y	Y			
Shape Algorithms					
Diameter	Y	Y	Y	Y	
Irregularity Index					Correlated well with two dermatologists ($\bar{p} \approx 0.723$)
Box Count	Y	Y	Y	Y	Slight correlation with two dermatologists ($\bar{p} \approx -0.396$)
Convex Hull					Most correlated border irregularity algorithm ($\bar{p} \approx 0.740$)
Asymmetry Index					Reasonable correlation was noted. Perception of dermatologists differed

For the 'dermatologist assessment' problem, more algorithms produced significant differences between the 'excised' and 'not-excised' lesions. In particular, all of the colour variance algorithms produced significant differences between 'excised' and 'not-excised' lesions. Shape algorithms on the other hand, including Asymmetry index, Irregularity index and Convex hull, again did not produce significant differences between 'excised' and 'not-excised' lesions. Similarly to the diagnosis problem, those features that did produce significant differences between the algorithms tended to match expectations of what constitutes 'suspiciousness'. For example, colour variance and diameter were higher for 'excised' lesions, while Box count indicated that 'excised' lesions were more irregular (as assessed by box counting) than 'not-excised' lesions.

Reproduction of dermatologists perception of the ABCD criteria of Friedman et al. (1985) was relatively poor for the Clinical-view algorithms. With the exception of the border irregularity algorithms (in particular, Irregularity index and Convex hull), only slight to medium correlations were noted, suggesting that these algorithms are not reproducing the perception of dermatologists accurately.

Clinical-view Comparison to Literature

We have seen above how several of the algorithms used previously in Clinical-view diagnosis systems produced no significant difference between melanoma and benign lesions in this research. Some of these algorithms have been assessed in a similar manner in previous research, and it is of interest to contrast those results with the results found here. Only Clinical-view results of the Sydney Image Set could be compared in this manner, because the ELM algorithms in this research are different from those presented previously, and only the diagnosis problem has been investigated previously.

When Stoecker et al. (1992) first proposed the asymmetry algorithm used here, they reported 93% agreement between the asymmetry algorithm and the dermatologist for their data set. We did not reproduce this level of agreement in our research, as shown in Section 6.1.1. Looking at the same algorithm, Ercal, Lee, Stoecker & Moss (1994) stated that "88% of the melanomas... have an asymmetry percentage above 8%". A much lower percentage was found for benign lesions. Again, our results do not replicate their findings. No significant difference was found in asymmetry values between benign and malignant lesions in the Sydney Image Set.

We also find some discrepancy with border irregularity algorithms. For the Irregularity index algorithm, Golston et al. (1992) reported that 87% agreement between algorithm and dermatologist was obtained. Although their method of assessment was different to the Spearman's rho method used here, we also found high correlation with

dermatologists' perception of irregularity. However, when the melanoma and benign Irregularity index values were examined, our results did not show significant differences between melanoma and atypical moles for this algorithm. This is in contrast with previous literature. For example, Huang et al. (1996) reported on differences in border irregularity for Clinical-view images of atypical moles and melanomas. The t-test was used to evaluate the significance of the difference in means, and they report a significant difference in Irregularity index between melanoma and atypical moles. Green et al. (1991) also reports significant differences in Irregularity index ($p = 0.05$), although only five melanoma were analysed. Conversely, Tomatis et al. (1998) investigate Irregularity index in each of the red, green and blue planes, and find little evidence of significant difference in Irregularity index.

Similar results are shown for Box counting. Cross et al. (1995) also looked at box counting as a means of measuring border irregularity. They report no significant difference in the values of box counting for melanocytic naevi and melanoma. Fifteen melanoma and twenty-one melanocytic naevi were analysed in their research. In both classification problems in this research, Box counting showed significant differences between the two classes.

Clinical-view colour variegation is a difficult concept to accurately reproduce through algorithmic means. Colour variance algorithms were used in this research, similarly to previous literature. However, the utility of RGB variance algorithms has not been proven. For example, Green et al. (1991) report no significant difference between melanoma and benign lesions for either Red or Green variance. Tomatis et al. (1998) also replicates these results. Our research partially disagrees with these findings, showing significant differences in Red variance between atypical naevi and melanoma, (p=0.05). No significant differences were found for Green or Blue variance.

Finally, Green et al. (1994) compared several Clinical-view algorithms to the perception of clinicians, similarly to the human comparison investigation performed here. In particular, they found that RGB colour variance algorithms and Irregularity index correlated quite well with clinicians' assessment of these features. We concur with the Irregularity index finding, which we also showed to reproduce dermatologists' perception well. However, our results show only medium correlation between dermatologists and colour variance algorithms.

In comparing our algorithm results to those reported previously in the literature, several of the results obtained in this research did not correspond with previous research. The most obvious reason for these discrepancies is simply that the image sets used were different for each research project, and therefore no conclusions as to which of the results are correct can be made.

7.5.2 ELM View

The ELM algorithms also produced some interesting results. These algorithms were for the most part unique to this research, and were based on the ABCD criteria for ELM images proposed by Stolz et al. (1994). Some algorithms, such as the colour algorithms, were taken directly from the Clinical-view system. The data presented in this section is obtained from Section 6.3.2 and Tables A.2 and B.2.

For the diagnosis problem, 12 of the 21 algorithms showed significant differences between benign and melanoma cases (recall that border contrast and asymmetry algorithms were not used on the Sydney Image Set). All of the Variance of variance algorithms (which were proposed to measure differential structures) showed significant difference between the two groups. Although these algorithms did not reproduce dermatologists' assessment of differential structures, it appears as though they may hold some validity for this problem. Colour variance algorithms similarly showed significant differences between the two groups, although the RGB colour variance algorithms again did not separate the two classes well. The outcome suggests once more the usefulnesxs of considering colour spaces other than RGB. Overall, if we consider what is known about melanoma and the results of the algorithms, few surprises are found. Colour variance and Variance of variance algorithms all showed higher results for melanoma than benign lesions.

For the 'dermatologist assessment' problem, all of the algorithms were able to be used. 26 of the 30 algorithms showed differences between 'excised' and 'not-excised' lesions. Particularly significant once again were the Variance of variance algorithms, together with the asymmetry algorithms. Border contrast also showed significant differences between the 'excised' and 'not-excised' sets, suggesting that this algorithm may have usefulness in this domain.

Variance asymmetry was in general higher for 'excised' lesions than 'not-excised', although Variance asymmetry (blue) indicated the reverse. Similarly, colour and shape asymmetry were higher for 'excised' lesions as could be expected. Border contrast indicated that 'excised' lesions had on average 2.2 sectors showing 'sharp' contrast, in comparison with 0.8 sectors for 'not-excised' lesions.

L variance results indicated that 'excised' lesions had *lower* variance than 'not-excised', but the other variance algorithms contrasted with this result. Finally, the variance of variance algorithms all indicated that 'excised' lesions were more likely to show higher variance of variance in all colour spaces than 'not-excised' lesions.

For the human comparison investigation, the ELM algorithms generally did better than the Clinical-view algorithms. The colour variance algorithms for example, better correlated with human perception when applied to ELM images than they did

Table 7.4: ELM algorithms at a glance: Part 1

Algorithm	Diagnosis		Dermatologist Assessment		Human Correlations comment
	$p < 0.05$	$p < 0.01$	$p < 0.05$	$p < 0.01$	
Asymmetry Algorithms					
Shape Asymmetry	NA	NA	Y	Y	Only two of the asymmetry algorithms showed any ability to
RGB Colour Asymmetry	NA	NA	Y	Y	reproduce the perception of dermatologists. RGB
					Colour Asymmetry was perhaps the best of the algorithms.
Var. Asymmetry Red	NA	NA	Y	Y	Shape asymmetry was also quite well correlated with the dermatologists
Var. Asymmetry Green	NA	NA	Y		suggesting that this algorithm also has use for ELM analysis.
Var. Asymmetry Blue	NA	NA	Y	Y	It should be noted that these algorithms were designed to measure
Var. Asymmetry L	NA	NA	Y	Y	one aspect of asymmetry, and that some of these algorithms
Var. Asymmetry α	NA	NA	Y	Y	may be useful in reproducing human perception when
Var. Asymmetry β	NA	NA	Y	Y	considered in combination.
Border Contrast					
Border Contrast	NA	NA	Y	Y	
Colour Gradient Algorithms					
Red gradient					
Green gradient			Y		
Blue gradient			Y		
L gradient	Y		Y		
α gradient			Y		
β gradient			Y	Y	
Colour Variance Algorithms					
Red variance	Y		Y		Red variance showed good correlation with human perception
Green variance			Y		with a mean Spearman's rho value of 0.606. Neither green
Blue variance					or blue variance correlated to any degree.

Table 7.5: ELM algorithms at a glance: Part 2

Algorithm	Diagnosis		Dermatologist Assessment		Human Correlations comment
	$p < 0.05$	$p < 0.01$	$p < 0.05$	$p < 0.01$	
L variance			Y	Y	The $L\alpha\beta$ variance algorithms showed medium strength
α variance	Y	Y	Y	Y	correlation with dermatologists' assessment of ELM
β variance	Y	Y	Y	Y	colour. Average Spearman's rho values were between 0.4 and 0.6
Relative Chromaticity Algorithms					
Chromaticity Red			Y	Y	
Chromaticity Green	Y	Y	Y		
Chromaticity Blue	Y				
Differential Structure Algorithms					
Var. Var. Red	Y	Y	Y	Y	All of the Variance of variance algorithms showed medium to
Var. Var. Green	Y	Y	Y	Y	low correlations (rho < 0.5)
Var. Var. Blue	Y		Y	Y	indicating that these algorithms are not reproducing the
Var. Var. L	Y	Y	Y	Y	concept of differential structures to any great degree.
Var. Var. α	Y	Y	Y	Y	
Var. Var. β	Y	Y	Y	Y	

with Clinical-view images. The asymmetry algorithms showed some ability to reproduce the perception of dermatologists. The variance of variance algorithms showed medium to slight correlation with the difficult problem of reproducing the perception of differential structures. Border contrast also showed some ability to reproduce the perception of sharp contrast at the lesion border. It should be noted that the dermatologists used in the human comparison investigation were not particularly expert with the ABCD criteria of Stolz et al. (1994). This lack of expertise may influence the results of the human comparison investigation. However, all dermatologists were familiar with the ABCD criteria of Stolz et al.

7.5.3 Algorithm Summary

Results for the two sets of algorithms were mixed. More ELM algorithms showed significant differences between the two classes for both classification problems. Several of the Clinical-view results did not replicate the results of similar investigations in previous literature, suggesting that the image sets used could alter the results of these algorithms substantially. The ELM algorithms also showed more ability to reproduce the perception of dermatologists.

Interestingly, the results of the Clinical-view human comparison investigation did not appear to have a significant impact on whether a particular algorithm would exhibit significant differences between the two classes for each classification problem. For example, the Irregularity index and Convex hull algorithms both showed good ability to reproduce dermatologists perception of border irregularity. However, neither of these algorithms showed significant difference between the two classes for either of the classification problems. These results were not repeated to the same degree in the ELM case.

Overall, it appears that the ELM algorithms are 'better' than the Clinical-view algorithms. More ELM algorithms produced significant differences between the two classes for both classification problems, and the ELM algorithms also appeared to reproduce the perception of dermatologists to a higher degree. These results may tend to support the thesis statement, but no strong conclusions as to the validity of the thesis statement can be reached on the basis of these results. It may be the case that with a different method of model building, different classification techniques, and/or a different image set, the 'better' ELM algorithms may have more of an impact on the results of the systems described previously. However, such conclusions cannot be supported without further research.

7.6 Chapter Summary

This chapter has presented an analysis of the results from Chapter 6. Firstly, the results of the diagnosis investigation were examined, followed by a similar examination of the 'dermatologist assessment' results. From this examination, the thesis statement was resolved. Overall, no support could be found for the thesis statement in the context of diagnosis. Similarly in the context of 'dermatologist assessment', no support for the thesis is gathered, and some contrary evidence was found. The results of the algorithms were then covered in detail, looking at which algorithms firstly showed significant differences between the two classes of both classification problems, and secondly, which algorithms correlated well to human perception. Some evidence that the ELM algorithms were 'better' than the Clinical-view algorithms was shown, firstly that more ELM algorithms produced significant differences between the two sets of classes, and secondly that the ELM algorithms were better able to reproduce human perception. However, no definite support for the validity of the thesis could be inferred from these particular results.

In the next chapter, some of the possible implications of the results of this research are described, including implications of the thesis, 'dermatologist assessment', and algorithms used in this work.

Chapter 8
Implications

The last chapter presented an interpretation of the results obtained from the investigations. In this chapter, a wider analysis is presented, and some possible implications for the research field are identified. We look first at the question of Clinical-view images versus ELM images in the context of automated skin lesion screening systems, and provide some directions for research with respect to the choice between the two image types. The 'dermatologist assessment' problem is then examined, and possible directions for this aspect of the field are proposed. We then look at the results of the algorithms used in this research, and suggest which algorithms may not be useful for similar research. Finally, the usefulness of a standardised image set is raised, and reasons for providing this image set are stated. We then propose a specification for a 'perfect' image set.

8.1 Clinical-view versus ELM

We have stated previously the reasons why this question should be researched (Section 2.2.3). In particular, proving one image type to be better than another will focus research efforts on the more suitable image type. Furthermore, the current emphasis on ELM research, if unwarranted, may cause results of automated systems to be lower than necessary. Based on the results we have shown, it appears as if the current trend towards ELM-based diagnosis systems may not be justified. However, several caveats must be noted about this result. Firstly, the scope of the results is limited to the image set and algorithms that were used to obtain them. Although we have made efforts to promote generalisability, different techniques may produce different results. This topic was discussed fully in the previous chapter.

For the 'dermatologist assessment' case, where the system was required to reproduce the decision to excise a lesion, the results obtained in this research again suggested that Clinical-view images were more use than ELM images. One reason may be that the decision to excise a lesion was not necessarily based on the potential malignancy

of a lesion, but may have been based on issues such as cosmetic appearance. If the set of excised lesions was restricted to those excised on the basis of malignancy considerations, where ELM may play a more important role, ELM images may be relatively more useful. However, such suppositions cannot be supported from the results to date, and are an area for future research.

Because we have indicated that Clinical-view images may be useful for both classification problems, perhaps the most useful line of research would utilise information from both types of image. Data from both image types could be input to a classifier which would then use both types of information to classify an image. Work such as the combined model of Schindewolf et al. (1993*b*) only looks at this problem at a superficial level. In their research, they combined the two sets of features (Clinical-view and ELM) and passed all of the features to the classifier. They report better results than either the Clinical-view or the ELM systems that they developed. Another method of combining the two sets of data would be to develop a 'stacked' classifier. Stacking is a well known procedure in the machine learning literature, and was first proposed by Wolpert (1992). As an example, consider the classifier shown in Figure 8.1. The first classifier (base classifier 1) takes features from Clinical-view images as input. Either of the Clinical-view classifiers described previously could be used. The second classifier (base classifier 2) takes features from ELM images as input, similarly to the ELM classifiers described in Chapter 6. These two classifiers are 'stacked' above a third classifier, which produces the output of the system, depending on the classification task. In this way, the third classifier can weight the Clinical-view and ELM results as required, and may be the best method of utilising information from both Clinical-view and ELM algorithms. Preliminary investigations into such a classifier have produced results 'averaging' those of both the Clinical-view and ELM systems. For example, if the Clinical-view system was more sensitive and less specific than the ELM system (as was the case in the diagnosis investigation) the stacked classifier produced results that were less sensitive and more specific than the Clinical-view system, and more sensitive and less specific than the ELM system. Overall accuracy was therefore higher than either the Clinical-view or ELM systems. Further investigation of this concept would be a simple extension to the work presented in this thesis.

8.2 Dermatologist Assessment

Perhaps the most significant contribution of this research to the field is the proposal that reproducing the assessment of dermatologists using computer based methods may be sufficient for an effective automated screening system. In the previous chapters, exploratory research into this concept was presented. Here, the logic behind

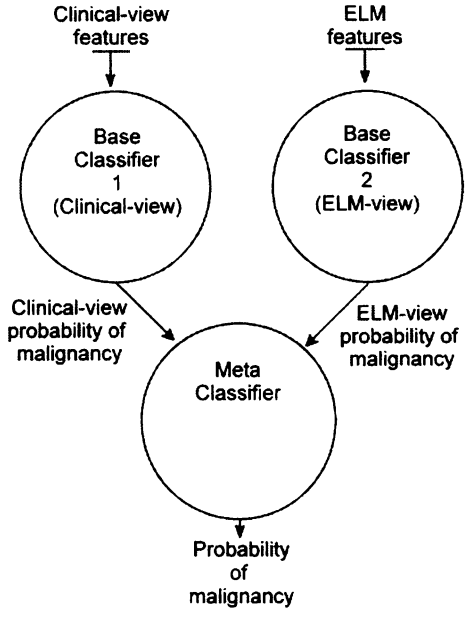

Figure 8.1: Example of a 'stacked' classifier combining Clinical-view and ELM features.

this proposal is reiterated and possible alternatives to the current emphasis on diagnosis systems are indicated.

Research to date has focussed on automated diagnosis systems. We proposed the reproduction of dermatologist's assessment as a method of screening skin lesions because of difficulties applying the current approach to the screening context. Recall that the diagnosis system approach to automatically identifying melanoma is based on attempting to reproduce the results of pathologists, as pathologists are the only people to diagnose lesions. However, this approach requires lesions that have been excised and subject to histopathological examination. Obviously, it is unacceptable to excise obviously benign lesions, and therefore these lesions will never be analysed by a diagnosis system. This feature of diagnosis systems is acceptable when a clinician is on hand to screen out those obviously benign lesions. In the context of screening however, one of the major arguments against screening is the high cost/benefit ratio. A major proportion of these costs can be attributed to the clinical time. A screening system therefore, would need to be used without clinical assistance, and therefore be trained to recognise these benign lesions.

Because a system is required that can be trained to recognise obviously benign lesions, we proposed that reproducing the assessment of dermatologists may be a useful adjunct to diagnosis systems. In this research, the viability of reproducing the decision to excise has been tentatively demonstrated. However, we have concerns

about the methodology used in this investigation. In particular, we are concerned that the 'dermatologist assessment' system was not trained on the perception of one dermatologist, but on the perception of a number of dermatologists. The use of multiple dermatologists may serve to confuse the classification problem, as the opinions of dermatologists regarding lesion excision are likely to differ. For example, less experienced clinicians may be more (or less!) conservative than their more experienced colleagues. Because of this feature of the image set, we cannot make definite conclusions about this problem. It is important that future work on this problem attempt to remove this source of uncertainty by training on the opinion of one clinician only.

8.2.1 Alternatives

We have identified and discussed shortcomings in the current methodology for automated melanoma detection from the point of view of screening, and have examined the question of reproducing the perception of dermatologists regarding the need to excise a lesion. Our results have suggested that such a system may be quite achievable and would be a useful addition to the current emphasis on automated melanoma diagnosis systems. Table 8.1 shows both roles for computers in melanoma detection, and looks at advantages and disadvantages of each.

The current research emphasis is on diagnosis systems. Considerable research has gone into these systems and some good results have been obtained. However, the identified difficulties with this method when applied to screening may restrict the results achievable with such systems. In this research, we have proposed that the reproduction of the perception of dermatologists may be a viable basis for a screening system, and have presented exploratory research into that concept. At a 91% level of sensitivity, 63% specificity was achieved. These results indicate that a system based on reproducing the decision of dermatologists regarding lesion excision may be a useful basis for a screening system. However, the problem of producing such a system is complicated by the fact that different dermatologists may produce markedly different decisions regarding any particular lesion. Of course, similar criticism could be levelled at the histopathology data required for a diagnosis system. For borderline cases, pathologists may produce different decisions regarding lesion malignancy.

Overall, it would certainly be advantageous to produce a system that could replicate the perception of an experienced dermatologist, not only for general practitioners, but also for less experienced dermatologists. Such a system may also be more achievable than the current emphasis on reproducing the results of pathologists, as the system would be analysing the same data as, and reproducing the results of, the

Table 8.1: Possible computer roles in melanoma detection

	Diagnosis system	'Dermatologist assessment' system
Input	Observes in-situ lesions at the Clinical or ELM views	Observes in-situ lesions at the Clinical and/or ELM views
Desired Output	Results of pathologist	Results of dermatologist
Major advantages	Large body of existing research. If desired output can be obtained, the screening problem is also solved.	Uses same input as human experts making the problem more realistic to solve. Sensitivity may be higher than diagnosis systems, as dermatologists err towards caution.
Major problems	Reproducing results of pathologist may not be possible given inputs. Cannot be trained on obviously benign lesions as histology data is required for all lesions.	Reproducing results of experts may result in melanomas being misclassified. New data sets with human ratings of 'suspiciousness' or excision data are required. Decision to excise may not be based only on malignancy issues.
Reproduces which human expert?	Uses input of dermatologist to reproduce results of pathologist	Dermatologist

dermatologist. Conceptually, such a system could also function as a 'pre-screen' for a diagnosis system, similarly to how a dermatologist acts as a screen for pathologists (Figure 8.2).

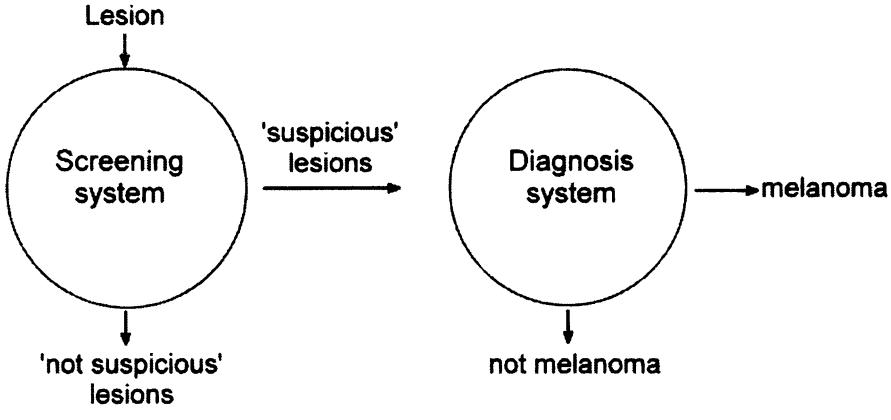

Figure 8.2: The future of automated analysis of skin lesion images?

Although in this research we have focussed on the decision to excise a lesion, it may also be useful to investigate other 'assessments of dermatologists'. For example, it may be of use to produce a system that replicates some minor assessment of 'suspiciousness'. Dermatologists could be used to identify lesions that show some minimum level of atypia. If a computer based system could then reproduce this perception, such a system may be of use as a first-contact screening system. Those lesions identified as 'suspicious' by this system could then be subject to analysis by a more advanced screening system, similar to that implemented for this research.

8.3 Algorithms

The results of the algorithms used in this research have important implications for the field. In particular, both useful and useless algorithms were identified, which may prevent replication by future researchers. Two sets of algorithm results have been presented, the first describing significant differences between the classes of the two different classification problems, and the second looking at human correlation with the perception of dermatologists.

It has been shown that many of the algorithms did not produce significant differences between the classes of either of the classification problems. These results suggest that these algorithms will not be useful in subsequent research into either of these problems, and that efforts should be concentrated on those algorithms that did obtain significant differences. Table 8.2 and Table 8.3 summarise the Clinical-view and ELM algorithms that did not show significant differences (at least p=0.05) between

classes for each of the classification problems.

Table 8.2: Clinical-view algorithms showing little distinction between classes for each of the classification problems.

Diagnosis	Dermatologist Assessment
Green gradient	Green gradient
Blue gradient	Blue gradient
L gradient	L gradient
Green variance	α gradient
Blue variance	β gradient
L Variance	Red variance
β Variance	Green variance
Chromaticity red	Blue variance
Chromaticity green	Chromaticity red
Irregularity Index	Chromaticity blue
Convex Hull	Irregularity Index
Asymmetry Index	Convex Hull
	Asymmetry Index

Interestingly, Table 8.2 list several of the algorithms we previously showed to be well correlated to human perception. Algorithms such as Irregularity index, Convex hull and Asymmetry index were shown as not being useful to differentiate lesions, but being reasonably well correlated with human perception. How can these algorithms, that are reproducing human perception, be considered useless for these two classification problems? The most obvious reason for these results in the case of diagnosis is that the ABCD criteria of Friedman et al. (1985) is not intended to be able to reliably distinguish between atypical naevi and melanoma. Friedman et al. (1991) state "dysplastic (atypical) naevi have one or more of the clinical features of malignant melanoma - i.e., asymmetry, border irregularity, color variegation, and a diameter greater than six mm". Therefore algorithms that reproduce the ABCD criteria may not be particularly useful for the diagnosis problem. This conclusion is interesting as we have stated previously that the ABCD criteria of Friedman et al. (1985) forms the basis of algorithms for most previous Clinical-view research.

Another reason may be that the lesions are not advanced enough to show the types of characteristics covered by the ABCD criteria of Friedman et al. (1985). Also, for the 'dermatologist assessment' problem, the removal of lesions because of non-malignancy issues, such as cosmetic appearance, may have skewed these results. Therefore, although we have indicated that several of the algorithms shown in Table 8.2 do reproduce dermatologists perception well, this ability was not useful when differentiating between lesions in the two image sets here.

In the ELM case, such results were less common, perhaps because asymmetry and

Table 8.3: ELM algorithms showing little distinction between classes for each of the classification problems. Note that asymmetry and border contrast algorithms were not evaluated for the diagnosis problem.

Diagnosis	Dermatologist Assessment
Red gradient	Red gradient
Green gradient	Blue variance
Blue gradient	Chromaticity blue
α gradient	
β gradient	
Green variance	
Blue variance	
L Variance	
Chromaticity red	

border contrast algorithms were not used for the diagnosis problem (Table 8.3). Of the algorithms used, the least useful algorithms were mainly gradient algorithms.

8.4 Image Sets

It may have become apparent that some of the major limitations on the results of this research (and all previous research in this field) are due to the characteristics of the image sets used. For example, the results obtained for the diagnosis investigation can only be attributed to the Sydney Image Set. It is hoped that such results are generalisable to the lesion population as a whole, but of course, there is no way of establishing whether or not this is true. Techniques such as cross-validation allow us to be more confident about the generalisation ability of the system, but ultimately, the range of variation in the image set limits the applicability of the findings.

Similarly with the dermatologist assessment investigation, the results cannot be compared to the diagnosis system results, which would be the ideal outcome of this research. Because the image sets used are made up of different lesions, with differing image characteristics, any differences between the diagnosis and dermatologist assessment systems may simply be an artifact of the image sets, rather than actual differences in classification performance between the two systems. Again, due to limitations imposed by the image set, such a comparison is out of the scope of this research.

On a more fundamental note, we have already stated that comparing research projects may be difficult if different image sets are used. For example, if Bischof et al. (1998) achieve $\approx$90% sensitivity, and specificity of $\approx$80%, does this mean that the system of Seidenari et al. (1999) is a better system, with reported results of

sensitivity = 100% and specificity of 92%?

The answer is not necessarily. It may that Bischof et al. have the better system, but the image set used was more difficult to classify. Consider that the image set of Bischof et al. consisted of 45 melanomas and 176 atypical non-melanomas. Seidenari et al. however, analysed 18 melanomas and 365 benign naevi "including common naevi and clinically dysplastic (atypical) naevi". Perhaps the extra 27 melanomas analysed by Bischof et al. made the classification task more difficult. Perhaps the use of clinically atypical naevi by Bischof et al. complicated the classification task more than the benign naevi "including common naevi and clinically dysplastic (atypical) naevi" used by Seidenari et al. It is apparent that no reliable comparison can be made between these two pieces of research. The situation is similar for all other research on automated melanoma detection.

It is likely that the provision of a readily available image set consisting of a wide variety of lesions and a significant number of melanoma would alleviate a number of the problems described above. In machine learning research, repositories of data sets exist, allowing machine learning researchers to test new algorithms and compare their results to those already published, for the same data sets (for example, see http://www.ics.uci.edu/~mlearn/MLSummary.html). Such a situation would be advantageous to research into automated melanoma detection, by allowing new researchers to immediately evaluate the utility of different techniques proposed by different researchers. Below, requirements for such an image set are proposed.

8.4.1 Proposal for a 'Perfect' Image Set

A 'perfect' image set would minimally require:

1. Standardised images. Standardisation covers a number of different steps, including adherence to prescribed values for magnification, colour calibration, and the digitisation process.

2. Lesions reflecting the range of features found in the population of skin lesions.

3. A significant number of lesions. This requirement is associated with the previous item. Obtaining an image set containing the range of features found in the population requires a large number of lesions.

4. A significant proportion of lesions should be melanoma. Because melanoma are highly variable in appearance, a large number should be collected. Consequently, a large number of benign lesions is also required.

5. Lesions should not be limited to those having been excised as is the case currently. The entire range of lesions, from obviously benign to obvious melanomas, is required to give as much flexibility as possible to researchers.

6. Data concerning histopathology results (where available), whether a dermatologist recommended an excision, and associated patient data should be obtained for each lesion.

7. Lesions should be captured using both Clinical-view and ELM. The ELM images should be sufficiently large that the vast majority of lesions fit into the field-of-view.

8. The lesions and patient data should be collected under an appropriate Code of Ethics.

However, gathering such an image set is a time consuming task. For example, the University/Health-Waikato Image Set was created over the course of 14 months. Approximately 140 individual lesions were obtained, and of those lesions, only 11 were melanoma. Given the variability shown by skin lesions and melanoma in particular, a sample of 11 melanoma is not enough to completely describe the population of melanoma. However, it is not clear what number would suffice. Furthermore, pathological data together with associated patient data and 'dermatologist assessment' data should also be gathered. Unfortunately, lesion image sets are not something that can simply be measured, like the length of an abalone or the marital status of adults (two examples from data sets in the UCI machine learning repository). Images require calibration and standardisation which again are non-trivial tasks, and a significant amount of time and effort must go into obtaining such a collection. In summary, the gathering of image sets is a non-trivial chore, and therefore the expense (in terms of time) associated with such a project has been, and is likely to continue to be, prohibitive. Provision of such an image set is however, very important for progress in this field.

8.5 Chapter Summary

This chapter has reviewed some of the implications for the research field arising from this work. The question of Clinical-view versus ELM has been looked at, and we conclude that no evidence could be obtained from our research that suggests that ELM is more use in the context of automated skin lesion identification. From the results, it appears as if both image types may contain useful information, and we therefore looked at the possibility of combining Clinical-view and ELM data in a single classifier, and proposed a method of developing such a classifier using stacking.

The implications of the 'dermatologist assessment' proposal were then looked at. We have shown that reproducing the assessment of dermatologists may be a viable alternative to the emphasis on diagnosis. Reproducing the assessment of dermatologists opens up a whole new avenue for research, and therefore the implications for the research field are far reaching. Many different aspects of dermatologists perception regarding lesion 'suspiciousness' can be investigated, from the problem investigated here, reproducing the excision decision, to reproducing some minimal level of 'suspiciousness'. Of course, investigating such a problem requires the acquisition of additional data, and in particular, a set of benign lesions that are not clinically suspicious, to ensure completeness of the image set.

The results of the algorithms utilised in this research also have implications for the field. We have attempted to measure not only the possible contribution of each algorithm to a classifier, by means of detecting significant differences between the two classes, but also to measure how well algorithms reproduce human perception. Several popular algorithms were of no use to the classifiers, but did correlate well with human perception. Such a result suggests that the criteria being reproduced by the algorithms may not be valuable for the problems investigated by the classifiers. Other algorithms that were not based on human criteria (such as several of the gradient algorithms) were found to be of no use to the classifiers. Such results suggest that these algorithms are of little use in the context of melanoma detection.

Finally, we looked at the need for a standardised image set that was freely available to all researchers. Comparisons between systems and techniques (not to mention Clinical-view versus ELM comparisons!) by different researchers cannot be made currently with any degree of accuracy, simply because the image sets used are not similar. Any such comparisons would require reproduction and testing of reported methods, which would be extremely time consuming. A standardised image set would alleviate this problem by allowing researchers to test their techniques on the same image set as every other researcher. Comparisons between systems would become straight forward. A short prescription for such an image set was given, and the difficulty in obtaining such an image set was noted.

The next chapter reviews the research, and presents the main findings, the major contributions of this work, and some avenues for future work.

Chapter 9
Findings, Contributions, and Future Work

The death rate from melanoma is increasing in light skin populations around the world. Early detection and treatment of this skin cancer are essential to reducing the death rate. A number of methods have been proposed to increase rates of early detection, including population screening, similar to existing programs for breast and cervical cancers. As was seen in Chapter 2, it appears unlikely that population screening for melanoma will be implemented due to high costs and uncertain benefits. However, if an automated system to screen lesions could be developed, the relative ratio of costs to benefits could be improved.

Current methods of population screening would use dermatologists to identify lesions that may be malignant. An automated screening system would perform the same function. A large amount of research has already gone into automated diagnosis systems for melanoma and in the decade of research that has taken place, a number of good results have been achieved. Diagnosis systems however, have two shortcomings that may inhibit their results. Firstly, such systems are intended to reproduce the diagnosis of pathologists, but using Clinical-view or ELM images. Such results may not be possible to achieve. Secondly, because all lesions used to train and test the system must have histopathology results, obvious benign lesions will not be analysed. This means that a representative sample of the lesion population is not available to train and test the system.

Because of these problems, a new method of screening lesions was proposed in this research, namely the reproduction of 'dermatologist assessment'. In particular, we were looking to reproduce the decision to excise a lesion. The rationale for this method is straightforward. Dermatologists, rather than pathologists, are the recognised experts at identifying malignant skin lesions in-situ, and reproduction of this expertise may produce a system that can operate without trained clinical supervision, as a screening system would be required to do.

In the early period of diagnosis system research (up to 1995), Clinical-view images were the most common image type used. Good results were reported, especially by Lee (1994) and Schindewolf et al. (1993*a*). However, at the end of 1994, research into Clinical-view based systems almost ceased. Systems based on ELM images became the norm, and this is the situation up until the present time. However, little research has investigated the comparative abilities of these two types of systems. In the only previous research to date, Schindewolf et al. (1993*b*) showed better results for the Clinical-view system. This result is at odds with the current emphasis on ELM images, and therefore, research was required to clarify the question of image choice.

This research looked at the problem of comparing the two types of images in an automated screening system. This question is important for this field, as consideration of the two types of image may slow the advance of the field. Similarly, ignoring the Clinical-view in favour of the ELM view may restrict the results of automated systems. To investigate this question, two systems were developed, one using Clinical-view and the other using ELM images. The results of the systems were compared, both in the traditional context of an automated diagnosis system for melanoma, and in the original context of reproducing 'dermatologist assessment' of the need to excise a lesion.

This chapter summarises and discusses the research, and is presented in four major sections. The first section outlines the major findings from the research, and the status of the thesis position is reported. The next section presents the major contributions of this research, while the third section looks at areas of possible future work arising from this research, focussing on immediate extensions to this work. Finally, the chapter closes with consideration of the direction of research in this field.

9.1 Main Findings and Thesis Summary

In this section, the major findings from the research are summarised, and the thesis statement is re-examined. We begin with a summary of the major findings of this research. These findings are reproduced in the order in which they appeared in the text. References to previous sections are also shown.

1. Overall, the Clinical-view diagnosis system slightly outperformed the ELM system. However, the difference was not shown to be statistically significant, although the method of comparing the AUROC's would tend towards this finding (Sections 6.1 and 7.2).

2. In the 'dermatologist assessment' investigation, the Clinical-view system tended to be more accurate than the ELM system. This finding was shown to be statistically significant. However, both systems were reasonably successful in performing this classification task (Sections 6.2 and 7.3).

3. From the results of the 'dermatologist assessment' investigation, it was concluded that reproducing the assessment of dermatologists with regard to lesion excision is a viable basis for a screening system (Sections 6.2 and 7.3).

4. Several algorithms reported previously in the literature did not differentiate well between melanoma and atypical benign lesions, while most of the algorithms show significant differences between 'excised' and 'not excised' lesions (Appendices A and B).

5. Results confirmed that several previous algorithms reproduced human perception of human criteria. Several new algorithms based on either the Clinical-view ABCD criteria or the ELM ABCD criteria also correlated well with dermatologists' perception of those criteria (Sections 6.3 and 7.5).

6. The lack of an image set accessible to all researchers to test such systems limits the applicability of the research. Comparisons between systems, both of techniques and results, are ultimately limited by the difference in image sets (Section 8.4).

From these findings, the thesis statement can be concluded. The thesis statement is reproduced here:

> Given the current published state of the field and a limited set of real world images, ELM-images are more useful in an automated screening system for skin lesions than Clinical-view images.

No evidence was found to support the thesis in the case of the diagnosis problem. For this problem, the Clinical-view system performed more accurately than the ELM system, although the difference was not shown to be statistically significant. This result agrees with the research performed by Schindewolf et al. (1993*b*).

For the 'dermatologist assessment' problem, the results of the Clinical-view system were higher than those of the ELM system, and therefore again tended to contradict the thesis statement. Due to the exploratory nature of this research, it is not clear whether this result can be generalised to a larger image set, or even a different set of dermatologists. Further research is required to clarify the nature of this results.

9.1.1 Scope of the findings

Although we have not found any support for the thesis statement from the previous investigations, it must be noted that these findings are significantly restricted by the limitations of the research. The first limitations are those encompassed by the thesis statement, principally the limitations imposed by the image sets used, and those imposed by the choice of image analysis techniques. It may be that ELM images are more use given another set of lesions, or image analysis algorithms. However, the lack of published algorithms and the difficulty in obtaining suitable image sets make this statement pure supposition.

Arguments concerning the quality and size of the image sets may also be raised. In the case of the UHWIS, some uncertainty exists about the reasons for excision, although we are confident that this was not an issue in this research. Also for this image set, the excised lesions were not excised based on the perception of only one dermatologist, but on a number of different dermatologists. These features of the image sets introduce uncertainty into the generalisability of the results obtained through this work.

The second set of restrictions concern the uncontrolled variables that differ between the two systems. These variables can be roughly classed in two major groups, Algorithms, and Images. The Algorithm variables concern the differences in image analysis algorithms used in the two systems. While the Clinical-view algorithms have been obtained from previous research, the ELM algorithms are experimental and derived for this research. Therefore it is likely that the quality of ELM algorithm implementation does not match that of the Clinical-view algorithms. This difference would occur simply because the ELM algorithms have not been subject to peer scrutiny and refinement as the Clinical-view algorithms have. Therefore, it is difficult to assess whether the lower ELM results are indicating that ELM images are less use in the investigations, or whether the results are an artifact of poor image analysis algorithms.

The Image variables can be considered in two. The first concerns the quality of the images, which was especially an issue with the Sydney Image Set. The Clinical-view images were not obtained under standardised conditions, and although efforts were made to control for this variable by removing the poorest images, some effect may still be included in the results.

The second variable concerns the tendency of lesions viewed under ELM to exceed the boundaries of the slide. This was a problem with the Sydney Image Set in particular, and required the removal of a large number of image analysis algorithms (the 'A' and 'B' of the criteria of Stolz et al.). This restriction on the image analysis

data undoubtedly caused some impact on the results of the ELM diagnosis system in particular.

These restrictions have been covered in detail in Section 5.4. It is important to recall these limitations when considering the results reported in this research, and the major findings described above.

All of these restrictions should be noted when assessing the results of the research. The results are not conclusive and should not be treated as such. What they are is indicative of the *position of the field at a point in time.* Arguments could be raised suggesting that the results here have understated the ability of ELM based diagnosis systems, and this argument is certainly appears valid given the more successful results in the literature. However, the scarce details concerning the methodology of these pieces of research throws uncertainty over the claims, and until such time as the methodology (or at least the systems) become available for independent testing, the argument becomes irrelevant. The results presented here represent those obtained given the current (published) state of the field, on a set of real world images.

9.2 Major Contributions

These findings are part of the contributions made by this research to the research field. Firstly, and perhaps most significantly, we have proposed a new framework for screening skin lesions - reproducing dermatologists' assessment of lesion 'suspiciousness'. This contribution provides a new angle on skin lesion classification, and models more accurately what occurs in clinical practice. It has several advantages over the current emphasis on diagnosis systems, and considerably widens the scope for research in this field.

The results of the investigations present further contributions to the research field. We have evaluated the premise that ELM images are more use than Clinical-view images in an automated diagnosis context. No support was found for the thesis statement in the diagnosis context, although the restrictions of the research make definitive conclusions impossible.

Similarly, the premise that ELM images are more use than Clinical-view images for reproducing the decision of dermatologists regarding lesion excision cannot be supported. The indications were in fact that Clinical-view images are more useful for this classification problem.

A number of new algorithms for the evaluation of ELM images based on the well known ABCD set of ELM criteria (Stolz et al. 1994) have been proposed. These algorithms are unique to this research, and represent amongst the first algorithms

specifically based on a set of ELM criteria to be reported. Such algorithms may be further evaluated and refined by future researchers.

Furthermore, the algorithms used in this research have also been tested. Firstly, each algorithm that was based on a human criterion was investigated for correspondence to human perception. Secondly, each of the algorithms in this research was assessed for significant differences between melanoma and benign groups for the diagnosis problem, and 'excised' and 'not-excised' groups for the 'dermatologist assessment' problem. This contribution gives future researchers insight into which algorithms may be useful to replicate. This step is particularly important for the new ELM algorithms presented in this research.

As can be seen, a number of significant contributions have been made by this research. Some, such as those concerning the algorithms, are likely to be evaluated and superceded in the near future. Others, such as the proposal of reproduction of dermatologist's assessment and the results of the Clinical-view versus ELM debate, are much more significant for the field. In particular, the concept of reproducing dermatologist's assessment opens up significant new areas for research, and may lead to the provision of an effective automated method of screening skin lesions.

9.3 Possible Directions for Further Work

The current state of this research field presents numerous opportunities for further work. This section is confined to three immediate extensions of the research presented here. The first looks at the 'dermatologist assessment' problem and illustrates how other definitions of 'suspicious' may be useful. It is also recommended that research comparing the results of a diagnosis system be compared to results from a 'dermatologist assessment' system. The second extension, classification, looks at simple extensions to the classifiers using well known techniques of improving classification ability. Finally, we look at possible directions for algorithm research.

9.3.1 Dermatologist Assessment

Perhaps the most significant contribution of this work is the proposal of a new framework for melanoma screening. This framework involved the reproduction of dermatologists' assessment into lesion 'suspiciousness'. The definition of suspiciousness investigated in this research was 'a lesion that was excised'. We contend that such a system may be a useful adjunct to the current emphasis on diagnosis systems for the reasons outlined previously.

However, although such a system has been implemented here and good results

achieved, it was not possible to contrast the results of the 'dermatologist assessment' systems with the results of the diagnosis systems. Therefore, research is needed that contrasts the two types of systems, in order to establish whether a system based on reproducing the assessment of dermatologists can be more successful at classifying melanoma than a diagnosis system. It could be expected that the 'dermatologist assessment' system would be more sensitive and less specific than a comparative diagnosis system, as dermatologists would tend to err on the side of caution. Such speculation is unproven however, and needs to be verified.

Another area of extension that would need to be performed before any definitive conclusions about 'dermatologist assessment' systems could be made involves restricting the image set to those lesions assessed by one dermatologist only. As was pointed out in the previous chapter, using lesions that were assessed by different clinicians may confuse the classification task. Unfortunately, due to the difficulty in obtaining suitable image sets, this was a feature of the UHWIS. In future work, care should be taken to ensure that only lesions examined by one dermatologist be used. In that case, the results could be presented in the following manner: $x\%$ of the clinician's decisions were accurately reproduced, and that the clinician had a sensitivity and specificity of $y\%$ and $z\%$ when compared to the results of histopathology. In this way, we could obtain some idea of the usefulness in terms of sensitivity and specificity of a 'dermatologist assessment' system. Such an extension would be produce valuable results concerning the usefulness of a 'dermatologist assessment' system.

Further work is also required to explicitely restrict the images used in 'dermatologists assessment' systems to those that were excised for malignancy concerns only. In this research, although we were certain that that vast majority of lesions were excised for malignancy considerations, two lesions were included that were possibly excised for other reasons. In future research, the set of lesions should be restricted to those that are firstly assessed by a single dermatologist, and secondly, those that the dermatologist considered were potentially malignant.

Another research avenue involves the definition of 'suspiciousness' used in the 'dermatologist assessment' system. In this research, the system was designed to replicate the decision of dermatologists to excise the lesion. However, many other definitions of 'suspiciousness' may be used. For example, dermatologists could be used to consider whether a lesion meets criteria for a minimum grade of 'suspiciousness'. Any lesion that showed some minor irregularity would be classed as suspicious, while the very obviously benign lesions would be classed as 'not-suspicious'. A system that could reproduce this perception may be of use as a 'first contact' screening system and may encourage people to seek medical advice.

Finally, the dermatologists' assessment of 'suspiciousness' needs to be based on pos-

sible malignancy, rather than be influenced by irrelevant details such as cosmetic appearance. Further research should make efforts to acquire lesion images that were excised due to malignant considerations, to avoid skewing the data set.

9.3.2 Classification

It is difficult to recommend any classification method as being the 'best' for a particular domain, and that is certainly the case in the context of skin lesions. However, research into classifiers in the domain of skin lesions is yet to take advantage of some of the more advanced techniques in artificial intelligence and machine learning.

As was mentioned in Chapter 8, stacking of Clinical-view and ELM classifiers may be a useful method of improving results. Preliminary results into such a system have found that the 'stacked' classifier performs better than both the Clinical-view and ELM models which are used as input. Another simple extension to the classifier is the use of ensemble methods of classification. Two of the most well known methods of ensemble classifier generation are boosting (for example, Schapire 1999) and bagging (Breiman 1996). Ensemble classification involves taking a classification algorithm (for example, logistic regression) and training multiple instances of the algorithm on different training sets from the same set of instances.

In a similar vein, we now look at the possibility of 'cascading' classifiers. (Ripley 1996) mentioned that "the task (of a classifier) is to classify an object, which means reaching one of K+2 possible decisions..." where K represents the number of possible classes. One of the two extra classes, as described in Chapter 3, was 'doubtful'. This class represents those cases where the classifier is unable to place the feature vector into any of the other K+1 classes. This outcome may occur for example, when the classifier cannot determine whether the feature vector represented a benign or malignant lesion. Ripley (1996) goes on to say that in the case of a 'doubtful' classification, perhaps other measurements should be made. In the context of skin lesion classification, this technique could be extremely valuable.

A simple example of such a measurement is the mean hue value, which could be used to rule out haemangioma and other vascular conditions, but is likely to contribute little to the majority of lesions. For example, if the hue of the lesion is predominantly red, the lesion is likely to be a haemangioma. Other hue features could also be used to confirm this classification. If the initial classifier identified a haemangioma, the classification process would stop. If an 'other' lesion was indicated, new features would be passed to the 'main' classifier and classification would be performed as previously (Figure 9.1). Such a system would be simple to implement, given a large enough image set.

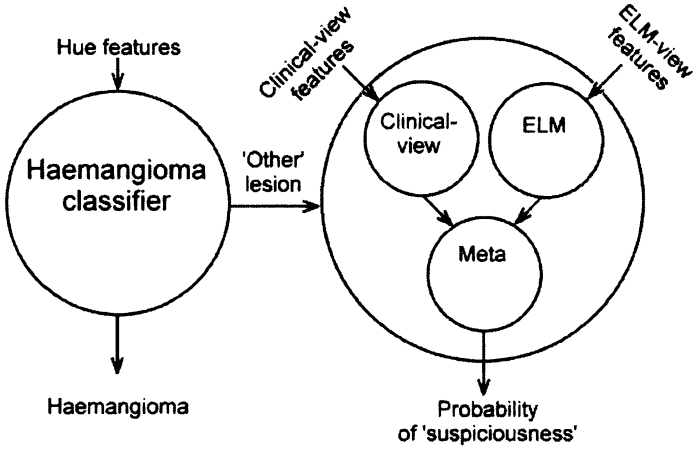

Figure 9.1: Example of a pre-screen classifier. The task of the 'haemangioma' classifier is to screen out most haemangiomas. Other lesions are passed to the main classifier for processing as before. This technique may reduce the load on the main classifier, as it will be expected to classify a smaller range of lesions.

9.3.3 Algorithms

In this research, the ELM algorithms were based on the ABCD criteria of Stolz et al. (1994). We have shown some success in replicating human perception of these criteria, but further work in improving and testing the algorithms remains to be done. Particular emphasis could be given to investigating differential structures, which are often highly indicative of melanoma. Also, other factors should be considered. Menzies et al. (1996) propose a related method of recognising melanoma under ELM, similarly to those developed by Stolz et al. (1994). Some of the features identified, such as blue-white veil, are very significant indicators of malignancy. Including algorithmic measures of these features may be useful to improve the results of the ELM system.

9.4 Concluding Thoughts

The field of automated melanoma detection is still in its infancy. Much work remains to be done in this field, and in this research we have looked at establishing whether or not ELM images were more use than Clinical-view images in an automated screening system for melanoma. We have looked at this problem in the context of diagnosis, and also examined it in the context of an alternative classification problem, namely the reproduction of the assessment of dermatologists. Such a system is likely to

be of more use in the context of screening where expert medical personnel are not available to identify obviously benign lesions.

However, regardless of the classification problem attempted, there exist several restrictions on progress in this field. Perhaps the major restriction is the quality and availability of skin lesion image sets. Simply, the field requires an image set with a large number of lesion images, captured using standardised techniques, and containing a large proportion of melanoma. The image set should feature the range of images, from the very obviously benign to obvious melanoma. Both Clinical-view and ELM views of each lesion should be obtained, and techniques that allow the capturing of the entire lesion under ELM should be adopted. Such techniques are already in use by Bischof et al. (1998) for example. The lesion set should also have histological data where available, and dermatologist assessments made on the lesions that were not excised.

The primary function of this image set would be to allow comparisons to be made between different techniques and systems. Currently, the field is restricted by the current, rather ridiculous, situation where no direct comparisons can be made between different research results. The results of the research become meaningless, as they cannot be placed in context with each other, nor can the results be reliably replicated.

Such an image set would also allow comparisons to be made amongst different classification problems. Currently, with the exception of this research, the only classification problem being investigated is the diagnosis problem. With an adequate image set, new classification problems, such as the 'dermatologist assessment' classification problem proposed for the first time in this research could be examined. Further, the results of such research could be compared with the diagnosis research that has already taken place.

However, gathering such an image set is a non-trivial chore, and appears unlikely to become a priority in the near future. It is hoped that the provision of the University/Health-Waikato Image Set with this thesis may represent the first steps towards an easily obtainable image set. Until such an image set is obtained, and made freely available to researchers around the world, this field will continue to be hampered by efforts "re-inventing the wheel". In a field such as this, where the results may directly affect the well-being of human beings, such delay is objectionable. For too many people around the world, delay in melanoma diagnosis and treatment means death.

Appendix A
SIS Algorithm Results

Table A.1: Feature range in the Clinical-view features in the Sydney Image Set

	Benign Lesions Mean±Std. Deviation	Melanoma Mean±Std. Deviation	p
Red gradient	*33.194 ± 13.216*	*41.971 ± 20.721*	*0.039*
Green gradient	37.764 ± 10.095	43.089 ± 13.085	0.052
Blue gradient	35.048 ± 10.614	34.386 ± 12.748	0.548
L gradient	-121.142 ± 31.625	-119.530 ± 34.708	0.709
α gradient	*-1.168 ± 0.062*	*-1.132 ± 0.076*	*0.021*
β gradient	-0.501 ± 0.078	-0.501 ± 0.081	0.949
Red variance	*314.806 ± 157.548*	*454.981 ± 338.362*	*0.049*
Green variance	307.930 ± 102.460	373.210 ± 162.664	0.101
Blue variance	285.569 ± 112.552	352.927 ± 177.787	0.057
L variance	755.282 ± 283.214	982.590 ± 571.725	0.101
α variance	*0.007 ± 0.005*	*0.010 ± 0.006*	*0.031*
β variance	0.009 ± 0.007	0.011 ± 0.008	0.070
Chromaticity red	0.062 ± 0.036	0.049 ± 0.038	0.059
Chromaticity green	-0.032 ± 0.020	-0.039 ± 0.020	0.059
Chromaticity blue	**-0.031 ± 0.027**	**-0.010 ± 0.030**	**0.002**
Diameter	**190.456 ± 76.844**	**264.271 ± 93.241**	**0.000**
Irreg. Index	1.128 ± 0.192	1.160 ± 0.195	0.367
Convex hull	0.935 ± 0.030	0.929 ± 0.032	0.397
Asymmetry Index	0.076 ± 0.026	0.079 ± 0.031	0.757
Box count	**1.727 ± 0.074**	**1.783 ± 0.061**	**0.001**

Table A.2: Feature range in the ELM features in the Sydney Image Set

	Benign Lesions Mean±Std. Deviation	Melanoma Mean±Std. Deviation	p
Red gradient	36.612 ± 21.289	45.308 ± 23.988	0.119
Green gradient	54.482 ± 17.910	59.456 ± 20.142	0.266
Blue gradient	67.930 ± 21.475	64.330 ± 20.959	0.418
L gradient	*-196.199 ± 37.000*	*-175.848 ± 46.441*	*0.036*
α gradient	-1.184 ± 0.048	-1.160 ± 0.064	0.113
β gradient	-0.562 ± 0.048	-0.548 ± 0.066	0.353
Red variance	600.876 ± 397.850	741.131 ± 435.553	0.158
Green variance	559.460 ± 308.580	590.929 ± 245.461	0.291
Blue variance	510.918 ± 297.503	475.503 ± 183.538	0.985
L variance	1474.648 ± 817.013	1582.374 ± 706.360	0.348
α variance	**0.003 ± 0.002**	**0.006 ± 0.004**	**0.000**
β variance	**0.003 ± 0.003**	**0.005 ± 0.005**	**0.002**
Chromaticity red	0.078 ± 0.022	0.074 ± 0.033	0.308
Chromaticity green	**-0.015 ± 0.013**	**-0.024 ± 0.016**	**0.003**
Chromaticity blue	*-0.064 ± 0.024*	*-0.049 ± 0.029*	*0.018*
Var. Var. Red	**3437.893 ± 3234.040**	**6276.576 ± 5590.988**	**0.002**
Var. Var. Blue	*6079.587 ± 3828.142*	*10499.090 ± 9285.340*	*0.014*
Var. Var. Green	**4120.655 ± 2576.613**	**8255.181 ± 8167.847**	**0.001**
Var. Var. L	**9114.061 ± 6515.964**	**18776.692 ± 19914.889**	**0.002**
Var. Var. α	**0.046 ± 0.045**	**0.136 ± 0.147**	**0.000**
Var. Var. β	**0.044 ± 0.075**	**0.201 ± 0.339**	**0.000**

Appendix B
UHWIS Algorithm Results

Table B.1: Feature range in the Clinical-view features in the University/Health-Waikato Image Set

	Not excised lesions Mean±Std. Deviation	Excised Lesions Mean±Std. Deviation	p
Asymmetry Index	0.067±0.038	0.077±0.027	0.078
Irregularity Index	1.115±0.267	1.119±0.148	0.521
Box Count	**-1.813±0.067**	**-1.868±0.047**	**0.000**
Convex Hull	0.948±0.044	0.943±0.029	0.156
Red Gradient	*42.373 ± 21.448*	*55.509 ± 21.015*	*0.031*
Green Gradient	46.017±17.736	53.858±27.862	0.530
Blue Gradient	41.306±24.582	40.765±30.616	0.643
L Gradient	71.540±28.271	84.891±37.186	0.199
α Gradient	-0.077±0.081	-0.035±0.092	0.101
β Gradien	0.113±0.057	0.134±0.054	0.219
Red Variance	491.770±446.765	646.358±423.774	0.083
Green Variance	414.730±357.267	444.539±354.136	0.692
Blue Variance	321.581±329.541	305.191±270.030	0.924
L Variance	1024.146±847.626	1171.106±767.800	0.375
α Variance	**0.004±0.002**	**0.009±0.007**	**0.004**
β Variance	**0.005±0.004**	**0.011±0.007**	**0.001**
Chromaticity Red	0.066±0.039	0.058±0.048	0.486
Chromaticity Green	**-0.032±0.019**	**-0.046±0.012**	**0.001**
Chromaticity Blue	-0.034±0.038	-0.012±0.045	0.088
Diameter	**303.967±120.302**	**440.522±132.951**	**0.001**

Table B.2: Feature range in the ELM features in the University/Health-Waikato Image Set

	Not excised lesions Mean±Std. Deviation	Excised Lesions Mean±Std. Deviation	p
Shape Asymmetry	**0.103±0.056**	**0.122±0.067**	**0.001**
Var. Asymmetry Red	**21.534±30.091**	**23.279±14.609**	**0.000**
Var. Asymmetry Green	*28.030±29.395*	*27.376±19.702*	*0.013*
Var. Asymmetry Blue	**29.789±33.681**	**27.116±18.797**	**0.004**
Var. Asymmetry L	**52.200±57.825**	**54.574±39.745**	**0.002**
Var. Asymmetry α	**0.000±0.000**	**0.000±0.000**	**0.000**
Var. Asymmetry β	**0.000±0.000**	**0.000±0.000**	**0.001**
RGB Asymmetry	**37.343±14.295**	**44.144±11.744**	**0.000**
Border Contrast	2.345±2.653	2.600±2.444	0.554
Red Gradient	34.322±22.742	48.725±25.736	0.127
Green Gradient	*68.207±28.481*	*70.250±29.086*	*0.050*
Blue Gradient	*75.335±26.772*	*63.566±29.257*	*0.016*
L Gradient	*90.890±35.645*	*95.138±37.983*	*0.021*
α Gradient	*-0.155±0.068*	*-0.141±0.084*	*0.050*
β Gradient	**0.153±0.069**	**0.202±0.101**	**0.008**
Red Variance	*622.472±652.012*	*759.579±402.059*	*0.018*
Green Variance	*1102.830±841.583*	*829.002±438.759*	*0.029*
Blue Variance	846.446±608.452	507.572±373.482	0.106
L Variance	**1907.715±1515.799**	**1608.333±667.840**	**0.009**
α Variance	**0.005±0.003**	**0.006±0.002**	**0.000**
β Variance	**0.007±0.005**	**0.011±0.006**	**0.004**
Chromaticity Red	**0.098±0.036**	**0.121±0.062**	**0.007**
Chromaticity Green	*-0.028±0.025*	*-0.058±0.037*	*0.016*
Chromaticity Blue	-0.069±0.031	-0.062±0.040	0.211
Var. Var. Red	**9921.485±23210.186**	**17818.431±13339.101**	**0.000**
Var. Var. Blue	**14547.103±19933.662**	**24245.537±20660.742**	**0.000**
Var. Var. Green	**13262.061±20514.873**	**24369.150±21358.906**	**0.000**
Var. Var. L	**23460.086±50346.178**	**43714.197±42117.816**	**0.000**
Var. Var. α	**0.103±0.264**	**0.269±0.225**	**0.000**
Var. Var. β	**0.109±0.281**	**0.381±0.385**	**0.000**

Appendix C
Contents of the CDROM

The CDROM contains several datasets, including the 73 lesion University/Health-Waikato Image Set. The Java code for all of the image analysis algorithms is also included, together with instructions for use. Figure C.1 shows the directory structure of the CDROM.

C.1 Datasets

The first directory (`'cvcode'`) stores the image analysis algorithms used on the Clinical-view images. The algorithms are written in Java. The second directory (`'cvimages'`) contains the 73 Clinical-view images from the University/Health-Waikato Image Set. The third directory (`'cvpeople'`) stores the data obtained from the Clinical-view human correlation investigation. This data was used to obtain the results shown in Section 6.1.1. The fourth and fifth directories (`'siscvdat'` and `'sissmdat'`) contain the results of applying the image analysis algorithms to the Sydney Image Set, Clinical-view and ELM images respectively. The sixth directory (`'smcode'`) contains the ELM image analysis algorithms, while the seventh directory (`'smimages'`) stores the 73 ELM images from the UHWIS. The `'smpeople'` directory stores the results of the ELM human correlation investigation, which were used to obtain the results shown in Section 6.3.2. The `'Targa'` directory contains Java code which is used to load, display and manipulate Targa files. This directory also contains the `RGBPoint.java` file, which contains the methods used to manipulate individual pixels (for example, changing from RGB to $L^*a^*b^*$ colour space). The second to last directory (`uhwisdat`) contains both the Clinical-view and ELM algorithm results for the UHWIS. The final directory (`util`) stores utility java code, including code to transform a mask Targa file into a `.msk` file, a quicksort routine, and the algorithm used to perform the 'global adjustment' of Herbin et al. (1990).

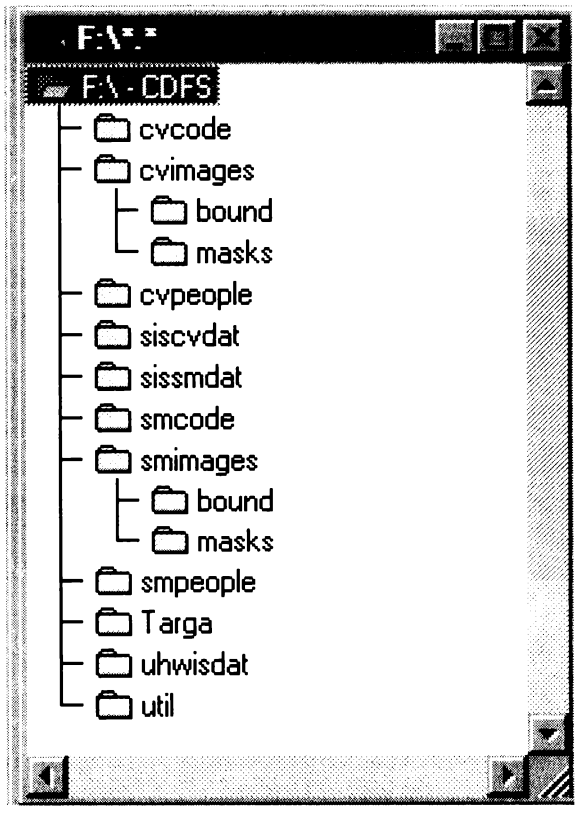

Figure C.1: Directory structure of CDROM

C.2 Code

The code for each of the image analysis algorithms is stored in one of four subdirectories, (`cvcode,smcode,Targa,util`). The first two store the Clinical-view and ELM code respectively. The second two directories contain related code, including algorithms to trace contours and other utilities (`'util'`) as well as algorithms for using Targa files (`'Targa'`).

Firstly, we assume the CDROM directory has been copied to the root directory of drive C (other drives may be substituted). To compile the code, the following method should be used:

```
C:\CDROM\<code directory>\javac -classpath ..\ <java file>.java
```

For example, to compile Boundary.java, the following would be used:

```
C:\CDROM\CVCODE\javac -classpath ..\ Boundary.java
```

To run this code, a similar method is used:

```
C:\CDROM\CVCODE\java -cp ..\;. Boundary
```

Note that the images cannot be used 'as is' from the CDROM. They must first be converted to 24 bit uncompressed Targa files. In general, the code is intended to be used in 'iteration' mode. That is, repeatedly applied to a number of image files. These image files are specified in a *script file*, which lists the number of files, relevant directories and each individual file name. Below is an example of the first part of the UHWIS Clinical-view asymmetry script file.

```
73
"d:\\lesions\\"
"d:\\lesions\\masks\\"
1
2
```

The first line states how many lesions are to be processed (73). The second and third lines state directory information, firstly where the image files are to be found, and secondly where the image masks are located. The remaining lines lists the image files to be processed. The border irregularity and diameter algorithms only require boundary files, while the other image analysis algorithms generally require both image files and either mask or boundary data. Example script files for each of the algorithms is included in the code directories. The following example assumes that the image files contained in `cvimages` have all been converted to 24-bit uncompressed Targa files, and that the `Boundary.java` file has been successfully compiled.

```
C:\CDROM\CVCODE\java -cp ..\;. Boundary iterate bound.scr output.txt
```

This example will run the Irregularity index algorithm over all 73 images listed in `bound.scr` and will output the results in the `output.txt` file. Conversely,

```
C:\CDROM\CVCODE\java -cp ..\;. Boundary ..\CVLESIONS\cvbound\1.bnd
```

will calculate the Irregularity index of the boundary contained in the `1.bnd` file. A similar structure is used for the other algorithms although in most cases both image and mask data is required (see below).

C.3 Images

The Clinical-view images are contained in the `'cvimages'` directory, along with binary mask files (`'cvmasks'`) and boundary data (`'cvbound'`). Mask files are simple

files which store the mask of the lesion, and are the output of `PSPMaskToMask.java`. In general, the process of obtaining a mask file is to first outline the lesion in an image tool (such as Paint Shop Pro 5.01). By using cut, the 'skin' portion is eliminated. The resulting lesion image (with black background) is saved as a Targa file. `PSPMaskToMask` is then run on this file, which results in a mask file, containing binary data (true for lesion, false for skin). This data is simply written straight to the file, and therefore it is necessary when reading a mask file to open the corresponding image file to obtain the relevant width and height. It is not necessary to read the image file however.

Boundary data is stored in a text file. The first number in the file is the total number of boundary points. The boundary points are then listed. There is no supplied method for reading these points into a specified data structure. However, the following code will read these points into a Vector:

```
Vector boundaryPoints;
fr = new FileReader(dir1+((Integer)v.elementAt(i)).toString()+".bnd");
st = new StreamTokenizer(fr) ;
st.nextToken() ;
int numberOfPoints = (int)st.nval ;
for (int j = 0; j < numberOfPoints; j++) {
  st.nextToken() ;
  x = (int)st.nval;
  st.nextToken() ;
  y = (int)st.nval;
  boundaryPoints.addElement(new Point(x,y)) ;
}
```

Boundary data is usually obtained by running `TraceContour.java` on the mask file. `TraceContour.java` can also be run on a directory which contains the original image files, with a 'masks' subdirectory containing the mask (`.msk`) files.

The ELM images and related data is stored in `'smimages'` directory. The directory structure is identical to the Clinical-view images. All image files are stored in compressed TIF format. For use with the algorithms, the image files must first be converted to uncompressed, 24-bit Targa format.

References

Andreassi, L., Perotti, R., Burroni, M., Dell'eva, G. & Biagioli, M. (1995), 'Computerized image analysis of pigmented lesions', *Chronicals of Dermatology* **V**(1), 11–24.

Argenyi, Z. B. (1997), 'Dermoscopy (epiluminescence microscopy) of pigmented skin lesions: Current status and evolving trends', *Dermatologic Clinics* **15**(1), 79–95.

Arvo, J., ed. (1991), *Graphics Gems II*, Academic Press, Inc., San Diego, chapter 1.1 The area of a simple polygon.

Binder, M., Kittler, H., Seeber, A., Steiner, A., Pehamberger, H. & Wolff, K. (1998), 'Epiluminescence microscopy-based classification of pigmented skin lesions using computerized image analysis and an artificial neural network', *Melanoma Research* **8**, 261–6.

Binder, M., Kittler, H., Steiner, A., Dawid, M., Pehamberger, H. & Wolff, K. (1999), 'Reevaluation of the abcd rule for epiluminescence microscopy', *Journal of the American Academy of Dermatology* **40**(2, part 1), 171–6.

Binder, M., Schwarz, M., Winkler, A., Steiner, A., Kaider, A., Wolff, K. & Pehamberger, H. (1995), 'Epiluminescence microscopy: A useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists', *Archives of Dermatology* **131**, 286–291.

Bischof, L., Talbot, H., Breen, E., Lovell, D., Chan, D., Stone, G., Menzies, S., Gutenev, A. & Caffin, R. (1998), An automated melanoma diagnosis system, Technical report, University of Ballarat/University of Technology, Sydney. Presented at the Research Workshop on Automated Medical Image Analysis.

Borland, R., Marks, R. & Noy, S. (1992), 'Public knowledge about characteristics of moles and melanomas', *Australian Journal of Public Health* **16**(2), 370–5.

Breiman, L. (1996), 'Bagging predictors', *Machine Learning* **24**(2), 123–140.

Cascinelli, N., Ferrario, M., Bufalino, R., Zurrida, S., Galimberti, V., Mascheroni, L., Bartoli, C. & Clemente, C. (1992), 'Results obtained by using a computerized image analysis system designed as an aid to diagnosis of cutaneous melanoma', *Melanoma Research* **2**, 163–170.

Castleman, K. (1979), *Digital Image Processing*, Prentice Hall Signal Processing Series, Prentice-Hall Inc., New Jersey.

Champion, R. H., Burton, J. L., Burns, D. A. & Breathnach, S. M., eds (1998), *Rook/Wilkinson/Ebling Textbook of Dermatology*, 6th edn, Blackwell Science Ltd.

Claridge, E., Hall, P. N., Keefe, M. & Allen, J. P. (1992), 'Shape analysis for classification of malignant melanoma', *Journal of Biomedical Engineering* **14**, 229–234.

Cormen, T., Leiserson, C. & Rivest, R. (1990), *Introduction to Algorithms*, MIT Press, Cambridge, Massachusetts.

Cross, S. S., McDonagh, A. J., Stephenson, T. J., Cotton, D. W. & Underwood, J. C. (1995), 'Fractal and integer-dimensional geometric analysis of pigmented skin lesions', *The American Journal of Dermatopathology* **17**(4), 374–378.

Day, G. R. (In press), 'How blurry is that border? an investigation into algorithmic reproduction of skin lesion border cutoff', *Computerized Medical Imaging and Graphics* . Accepted for publication 23rd December 1999.

Day, G. R. & Barbour, R. H. (1996), Approximate algorithmic boundary detection of skin lesions, *in* 'Proceedings of Image and Vision Computing New Zealand', pp. 7–12. 29-30 August, Lower Hutt, New Zealand.

Day, G. R. & Barbour, R. H. (2000), 'Automated melanoma diagnosis: Where are we at?', *Skin Research and Technology* **6**, 1–5.

Del-Mar, C., Green, A., Cooney, T., Cutbush, K., Lawrie, S. & Adkins, G. (1994), 'Melanocytic lesions excised from the skin: what percentage are malignant?', *Australian Journal of Public Health* **18**(2), 221–223.

Dhawan, A. P. (1988), 'An expert system for the early detection of melanoma using knowledge-based image analysis', *Analytical and Quantitative Cytology and Histology* **10**(6), 405–416.

DiGuiseppi, C., Atkins, D. & Woolf, S. (1997), Screening for skin cancer - including counselling to prevent skin cancer, *in* D. e. Kamerow, ed., 'Report of the United States Preventive Services Task Force', second edn, US Preventive Services Task Force.

Dummer, W., Blaheta, H.-J., Bastian, B. C., Schenk, T., Bröcker, E.-B. & Remy, W. (1995), 'Preoperative characterization of pigmented skin lesions by epiluminescence microscopy and high-frequency ultrasound', *Archives of Dermatology* **131**, 279–285.

Elwood, M. & Glasgow, H. (1993), *Melanoma: The Prevention and Early Detection of Melanoma in New Zealand*, Cancer Society of New Zealand/New Zealand Department of Health.

Ercal, F., Chawla, A., Stoecker, W., Lee, H. C. & Moss, R. (1994), 'Neural network diagnosis of malignant melanoma from colour images', *IEEE Transactions on Biomedical Engineering* **41**, 837–845.

Ercal, F., Lee, H., Stoecker, W. & Moss, R. (1994), Skin cancer diagnosis using heirarchical neural networks and fuzzy systems, *in* 'Intelligent Engineering Systems Through Artificial Neural Networks', Vol. 4, ASME Press, New York, pp. 613–618.

Ercal, F., Moganti, M., Stoecker, W. & Moss, R. (1993), 'Detection of skin tumor boundaries in color images', *IEEE Transactions on Medical Imaging* **12**(3), 624–627.

Fitzpatrick, T. B., Eisen, A. Z., Wolff, K., Freedberg, I. M. & Austen, K. F., eds (1993), *Dermatology in General Medicine*, McGraw-Hill Inc.

Friedman, R., Rigel, D. & Kopf, A. (1985), 'Early detection of malignant melanoma: The role of physician examination and self examination of the skin', *Ca- A Cancer Journal for Clinicians* **35**, 130–151.

Friedman, R., Rigel, D., Silverman, M., Kopf, A. & Vossaert, K. (1991), 'Malignant melanoma in the 1990s: The continued importance of early detection and the role of physician examination and self-examination of the skin', *Ca-A Cancer Journal for Clinicians* **41**(4), 200–226.

Giles, G. G., Armstrong, B. K., Burton, R. C., Staples, M. P. & Thursfield, V. J. (1996), 'Has mortality from melanoma stopped rising in australia? analysis of trends between 1931 and 1994', *British Medical Journal* **312**, 1121 1125.

Giles, G. & Thursfield, V. (1997), *Canstat-Trends in Cancer Mortality, Australia 1910-1994*, number 24, Cancer Epidemiology Centre, Anti-Cancer Council of Victoria.

Glasbey, C. (1993), 'An analysis of histogram-based thresholding algorithms', *Computer Vision and Image Processing: Graphical Models and Image Processing* **55**(6), 532–537.

Golston, J. E., Stoecker, W. V., Moss, R. H. & Dhillon, I. P. S. (1992), 'Automatic detection of irregular borders in melanoma and other skin tumors', *Computerized Medical Imaging and Graphics* **16**(3), 199–203.

Golston, J., Moss, R. & Stoecker, W. (1990), 'Boundary detection in skin tumor images: An overall approach and a radial search algorithm', *Pattern Recognition* **23**, 1235–1247.

Gonzalez, R. & Wintz, P. (1987), *Digital Image Processing*, Addison-Wesley Publishing Company.

Green, A., Martin, N., McKenzie, G., Pfitzner, J., Quintarelli, F., Thomas, B., O'Rourke, M. & Knight, N. (1991), 'Computer image analysis of pigmented skin lesions', *Melanoma Research* **1**, 231–236.

Green, A., Martin, N., Pfitzner, J., O'Rourke, M. & Knight, N. (1994), 'Computer image analysis in the diagnosis of melanoma', *Journal of the American Academy of Dermatology* **31**, 958–964.

Grin, C., Kopf, A., Welkovich, B., Bart, R. & Levenstein, M. (1990), 'Accuracy in the clinical diagnosis of malignant melanoma', *Archives of Dermatology* **126**, 763–766.

Grob, P. J. J. (1997), Cost effectiveness in skin cancers prevention, *in* P. Altmeyer, K. Hoffman, M. Stücker, H. P. Schwarze & M. Freitag, eds, 'Skin Cancer and UV Radiation', Springer-Verlag, Berlin, pp. 902–908.

Gutkowicz-Krusin, D., Elbaum, M., Szwaykowski, P. & Kopf, A. W. (1997), 'Can early malignant melanoma be differentiated from atypical melanocytic nevus by in vivo techniques? part ii. automatic machine vision classification', *Skin Research and Technology* pp. 15–22.

Habif, T. (1996), *Clinical Dermatology: A Color Guide to Diagnosis and Therapy*, C. V. Mosby Company, StLouis Missouri.

Hall, P., Claridge, E. & Morris Smith, J. (1995), 'Computer screening for early detection of melanoma-is there a future?', *British Journal of Dermatology* **132**, 325–338.

Hanley, J. A. & McNeil, B. J. (1982), 'The meaning and user of the area under a receiver operating characteristic (roc) curve', *Radiology* **143**, 29–36.

Health-Waikato (1995), 'Regional tumour registry annual report 1995', Health Waikato Ltd.

Herbin, M., Venot, A., Devaux, J. Y. & Piette, C. (1990), 'Color quantitation through image processing in dermatology', *IEEE Transactions on Medical Imaging* **9**(3), 262–269.

Hintz-Madsen, M., Hansen, L. K. & Larsen, J. (1995), Design and evaluation of neural classifiers application to skin lesion classification, *in* 'IEEE Workshop on Neural Networks for Signal Processing 1995 (NNSP'95)'.

Hintz-Madsen, M., Hansen, L. K., Larsen, J., Olesen, E. & Drzewiecki, K. (1996), Detection of malignant melanoma using neural classifiers, *in* A. Bulsari, S. Kallio & D. Tsaptsinos, eds, 'Solving Engineering Problems with Neural Networks (Proceedings of the International Conference EANN'96)', Systeemiteknikan Seuru Ry, Finland, London, England, pp. 395–398.

Horsch, A., Stolz, W., Neiβ, A., Abmayr, W., Pompl, R., Bernklau, A., Bunk, W., Dersch, D. R., Glāβl, A., Schiffner, R. & Morfill, G. (1997), Improving early recognition of malignant melanomas by digital image analysis in dermatoscopy, *in* C. P. et al., ed., 'Medical Informatics Europe 1997', IOS Press, pp. 531–5.

Hosmer, D. W. & Lemeshow, S. (1989), *Applied Logistic Regression*, Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons, Inc.

Huang, C. L., Wasti, Q., Marghoob, A. A., Kopf, A. W., de David, M., Rao, B. K. & Bart, R. S. (1996), 'Border irregularity: atypical moles versus melanoma', *European Journal of Dermatology* **6**(4), 270–3.

Jackson, A., Wilkinson, C. & Pill, R. (1999), 'Moles and melanomas - who's at risk, who knows, and who cares? a strategy to inform those at risk', *British Journal of General Practice* **49**, 199–203.

Jackson, A., Wilkinson, C., Ranger, M., Pill, R. & August, P. (1998), 'Can primary prevention or selective screening for melanoma be more precisely targeted through general practice? a prospective study to validate a self administered risk score', *British Medical Journal* **316**, 34–38.

Kenet, R. O., Kang, S., Kenet, B. J., Fitzpatrick, T. B., Sober, A. J. & Barnhill, R. L. (1993), 'Clinical diagnosis of pigmented lesions using digital epiluminescent microscopy', *Archives of Dermatolgy* **129**, 157–174.

Kittler, J. & Illingworth, J. (1986), 'Minimum error thresholding', *Pattern Recognition* **19**, 41–47.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E. & Nizam, A. (1998), *Applied Regression Analysis and Other Multivariable Methods*, 3rd edn, Brooks/Cole Publishing Company, United States.

Kopf, A. W., Elbaym, M. & Provost, N. (1997), 'Editorial: The use of dermoscopy and digital imaging in the diagnosis of cutaneous malignant melanoma', *Skin Research and Technology* **3**, 1–7.

Landau, M., Matz, H., Tur, E., Dvir, M. & Brenner, S. (1999), 'Computerized system to enhance the clinical diagnosis of pigmented cutaneous malignancies', *International Journal of Dermatology* **38**, 443–446.

Lee, H. (1994), Skin cancer diagnosis using heirarchical neural networks and fuzzy logic, Master's thesis, Computer Science, University of Missouri-Rolla, Missouri, USA.

MacKenzie-Wood, A. R., Milton, G. W. & de Launey, J. W. (1998), 'Melanoma: Accuracy of clinical diagnosis', *Australasian Journal of Dermatology* **39**, 31–33.

MacKie, R. (1989), *Skin Cancer*, Martin Dunitz Ltd., London.

MacKie, R. & Doherty, V. (1988), 'Educational activities aimed at earlier detection and treatment of malignant melanoma in a moderate risk area', *Pigment Cell* **9**, 140–152.

MacKie, R. M. (1985), An illustrated guide to the recognition of early malignant melanoma, Technical report, Department of Dermatology, University of Glasgow.

MacKie, R. M. & Hole, D. J. (1996), 'Incidence and thickness of primary malignant tumours and survival of patients with cutaneous malignant melanoma in relation to socioeconomic status', *British Medical Journal* **312**, 1125–1128.

Mandelbrot, B. (1983), *The fractal geometry of nature*, revised edn, W. H. Freeman and Company.

Marks, R. (1994), 'Skin cancer control in australia: have we made any difference?', *Australian Journal of Public Health* **18**(2), 127–8.

Martin, R. H. (1995), 'Relationship between risk factors, knowledge and preventive behaviour relevant to skin cancer knowledge in general practice patients in south australia', *British Journal of General Practice* **45**, 365–7.

McGee, R., Elwood, M., Adam, H., Sneyd, M., Williams, S. & Tilyard, M. (1994), 'The recognition and management of melanoma and other skin lesions by general practitioners in new zealand', *New Zealand Medical Journal* **107**, 287–290.

Menzies, S., Crotty, K., Ingvar, C. & McCarthy, W. (1996), *An Atlas of Surface Microscopy of Pigmented Skin Lesions*, McGraw Hill Book Company Australia Pty Limited, Sydney, Australia.

Menzies, S. W., Bischof, L. M., Peden, G., Talbot, H. G., Gutenev, A., Thompson, R. L., McNamara, K. W., Burlutski, G., McCarthy, W. H. & Skladnev, V. N. (1997), Automated instrumentation for the diagnosis of invasive melanoma: Image analysis of epiluminescence microscopy, *in* P. Altmeyer, K. Hoffman, M. Stücker, H. P. Schwarze & M. Freitag, eds, 'Skin Cancer and UV Radiation', Springer-Verlag, Berlin Heidleberg, pp. 1064–1070.

Menzies, S. W., Crotty, K. A. & McCarthy, W. H. (1995), 'The morphological criteria of the pseudopod in surface microscopy', *Archives of Dermatology* **131**.

Miller, D. R., Geller, A. C., Wyatt, S. W., Halpern, A., Howell, J. B., Cocerell, C. & Riley, B. A. (1996), 'Melanoma awareness and self-examination practices: Results of a united states survey', *Journal of the American Academy of Dermatology* **34**(6), 962–970.

Morton, C. A. & MacKie, R. M. (1998), 'Clinical accuracy of the diagnosis of cutaneous malignant melanoma', *British Journal of Dermatology* **138**, 283–287.

Nachbar, F., Stolz, W., Merkle, T., Cognetta, A. B., Vogt, T., Landthaler, M., Bilek, P., Braun-Falco, O. & Plewig, G. (1994), 'The abcd rule of dermatoscopy', *Journal of the American Academy of Dermatology* **30**(4), 551–9.

Ng, V. & Lee, T. (1996), Measuring border irregularities of skin lesions using fractal dimensions, *in* C.-S. Li, R. L. Stevenson & L. Zhou, eds, 'Electronic Imaging and Multimedia Systems', Vol. 2898, SPIE-The International Society for Optical Engineering, pp. 63–72.

Ngan, P. & Coombs, B. (1994), 'Segmentation of intensity basins in gray-scale images', *Computers and Biomedical Research* **27**, 39–44.

Nilles, M., Boedeker, R.-H. & Schill, W.-B. (1994), 'Surface microscopy of naevi and melanomas - clues to melanoma', *British Journal of Dermatology* pp. 349–355.

Pal, N. & Kak, S. (1993), 'A review on image segmentation techniques', *Pattern Recognition* **26**, 1277–1294.

Pehamberger, H., Binder, M., Steiner, A. & Wolff, K. (1993), 'In vivo epiluminescence microscopy: improvement of early diagnosis of melanoma', *Journal of Investigative Dermatology* **100**, 356S–362S.

Pehamberger, H., Steiner, A. & Wolff, K. (1987), 'In vivo epiluminescence microscopy of pigmented skin lesions. i. pattern analysis of pigmented skin lesions', *Journal of the American Academy of Dermatology* **17**(4), 571–583.

Rademaker, M. & Zainal, Z. (1997), Melanoma blackspot of the world? Department of Dermatology, Health Waikato.

Ramsay, D. & Weary, P. (1996), 'Primary care in dermatology: Whose role should it be?', *Journal of the American Academy of Dermatology* **35**, 1005–1008.

Rigel, D. S., Friedman, R. J. & Kopf, A. W. (1996), 'The incidence of malignant melanoma in the united states: Issues as we approach the 21st century', *Journal of the American Academy of Dermatology* **34**(5, Part 1), 839–47.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press.

Rosenfeld, A. & Kak, A. (1982), *Digital Picture Processing*, Vol. 1 of *Computer Science and Applied Mathematics*, Academic Press, Inc., Orlando.

Rosenfeld, A. & Kak, A. C. (1976), *Digital Picture Processing*, Academic Press, Inc.

Russ, J. C. (1990), *Computer Assisted Microscopy: The Measurement and Analysis of Images*, John Wiley and Sons, Inc.

Russ, J. C. (1999), *The Image Processing Handbook*, CRC Press.

Saxe, N., Hoffman, M., Krige, J., Sayed, R., King, H. & Hounsell, K. (1998), 'Malignant melanoma in cape town, south africa', *British Journal of Dermatology* **138**, 998–1002.

Schapire, R. E. (1999), A brief introduction to boosting, *in* T. Dean, ed., 'Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI'99) Stockholm, Sweden'.

Schindewolf, T., Schiffner, R., Stolz, W., Albert, R., Abmayr, W. & Harms, H. (1994), 'Evaluation of different image acquisition techniques for a computer vision system in the diagnosis of malignant melanoma', *Journal of the American Academy of Dermatology* **31**(1), 33–41.

Schindewolf, T., Stolz, W., Albert, R., Abmayr, W. & Harms, H. (1993*a*), 'Classification of melanocytic lesions with color and texture analysis using digital image processing', *Analytical and Quantitative Cytology and Histology* **15**(1), 1–11.

Schindewolf, T., Stolz, W., Albert, R., Abmayr, W. & Harms, H. (1993*b*), 'Comparison of classification rates for conventional and dermatoscopic images of malignant and benign melanocytic lesions using computerized colour image analysis', *European Journal of Dermatology* **3**(4), 299–303.

Sedgewick, R. (1992), *Algorithms in C++*, Addison-Wesley Publishing Company Inc.

Seidenari, S., Pellacani, G. & Giannetti, A. (1999), 'Digital videomicroscopy and image analysis with automatic classification for detection of thin melanomas', *Melanoma Research* **9**, 163–171.

Seidenari, S., Pellacani, G. & Pepe, P. (1998), 'Digital videomicroscopy improves diagnostic accuracy for melanoma', *Journal of the American Academy of Dermatology* **39**(2), 175–181.

Sinclair, R. (1998), 'Commentary: Start with the kiss principle', *British Medical Journal* **316**, 39–89.

Skegg, D. (1994), 'Melanoma: The public health commission's advice to the minister of health 1993-1994', Published by: Public Health Commission, Wellington.

Sober, A. J. & Burstein, J. M. (1994), 'Computerized digital image analysis: An aid for melanoma diagnosis', *The Journal of Dermatolgy* **21**, 885–890.

Steiner, A., Pehamberger, H. & Wolff, K. (1987), 'In vivo epiluminescence microscopy of pigmented skin lesions. ii. diagnosis of small pigmented skin lesions and early detection of malignant melanoma', *Journal of the American Academy of Dermatology* **17**(4), 584–591.

Stoecker, W., Moss, R., Ercal, F. & Umbaugh, S. (1995), 'Nondermatoscopic digital imaging of pigmented lesions', *Skin Research and Technology* **1**, 7–16.

Stoecker, W. V., Li, W. W. & Moss, R. H. (1992), 'Automatic detection of asymmetry and skin tumors', *Computerized Medical Imaging and Graphics* **16**(3), 191–197.

Stoecker, W. V. & Moss, R. H. (1992), 'Editorial: Digital imaging in dermatology', *Computerized Medical Imaging and Graphics* **16**(3), 145–150.

Stolz, W., Braun-Falco, O., Bilek, P., Landthaler, M. & Cognetta, A. B. (1994), *Color Atlas of Dermatoscopy*, Blackwell Science Ltd.

Swerlick, R. & Chen, S. (1996), 'The melanoma epidemic: is increased surveillance the solution or the problem', *Archives of Dermatology* **132**, 881–884.

Tabachnick, B. G. & Fidell, L. S. (1996), *Using Multivariate Statistics*, 3rd. edn, HarperCollins Publishers Inc.

Thursfield, V., Giles, G. & Staples, M. (1995), *Canstat-Skin Cancer*, Anti-Cancer Council of Victoria.

Tomatis, S., Bono, A., Bartoli, C., Tragni, G., Farina, B. & Marchesini, R. (1998), 'Image analysis in the rgb and hs colour planes for a computer-assisted diagnosis of cutaneous pigmented lesions', *Tumori* **84**, 29–32.

Umbaugh, S. E. (1990), Computer Vision in Medicine: Color Metrics and Image Segmentation Methods for Skin Cancer Diagnosis, PhD thesis, Electrical Engineering, University of Missouri-Rolla, United States.

Wolpert, D. H. (1992), 'Stacked generalization', *Neural Networks* **5**, 241–259.

Young, I., Walker, J. & Bowie, J. (1974), 'An analysis technique for biological shape i', *Information and Control* **25**(371), 357–370.