

# **Computer aided classification of histopathological damage in images of haematoxylin and eosin stained human skin**

A thesis submitted by

Juliana Haggerty

for the award of Engineering Doctorate in Biopharmaceutical Process  
Development



Biopharmaceutical and Bioprocessing Technology Centre

School of Chemical Engineering & Advanced Materials

October 2015

---

## **Abstract**

Excised human skin can be used as a model to assess the potency, immunogenicity and contact sensitivity of potential therapeutics or cosmetics via the assessment of histological damage. The current method of assessing the damage uses traditional manual histological assessment, which is inherently subjective, time consuming and prone to intra-observer variability.

Computer aided analysis has the potential to address issues surrounding traditional histological techniques through the application of quantitative analysis. This thesis describes the development of a computer aided process to assess the immune-mediated structural breakdown of human skin tissue. Research presented includes assessment and optimisation of image acquisition methodologies, development of an image processing and segmentation algorithm, identification and extraction of a novel set of descriptive image features and the evaluation of a selected subset of these features in a classification model.

A new segmentation method is presented to identify epidermis tissue from skin with varying degrees of histopathological damage. Combining enhanced colour information with general image intensity information, the fully automated methodology segments the epidermis with a mean specificity of 97.7%, a mean sensitivity of 89.4% and a mean accuracy of 96.5% and segments effectively for different severities of tissue damage.

A set of 140 feature measurements containing information about the tissue changes associated with different grades of histopathological skin damage were identified and a wrapper algorithm employed to select a subset of the extracted features, evaluating feature subsets based their prediction error for an independent test set in a Naïve Bayes Classifier. The final classification algorithm classified a 169 image set with an accuracy of 94.1%, of these images 20 were an unseen validation set for which the accuracy was 85.0%. The final classification method has a comparable accuracy to the existing manual method, improved repeatability and reproducibility and does not require an experienced histopathologist.

## **Acknowledgements**

Thanks to my supervisors at Alcyomics, Prof. Anne Dickinson and Dr. Xiao Nong-Wang, for spending time to train me in manual histopathology, manually grade images and provide valuable insight and advice in addition to their sponsorship. I acknowledge support from my original academic supervisors at Newcastle Prof. Elaine Martin and Dr. Ming Tham but particular thanks go to Dr. Chris O'Malley who stepped in to provide support, feedback and advice when I needed it. I'd also like to thank Prof. Gary Montague for providing encouragement, support and not letting me give up. Thanks also to Dr. Trevor Booth in the Bioimaging Department who sat patiently with me whenever I forgot how to use the microscope.

My fellow EngD students have been a huge help, providing encouragement, advice, understanding and plenty of distraction when required. Meeting such a great bunch has been one of the best parts of the programme and will be an enduring network in my future career. I am also very grateful to my bosses at Centre for Process Innovation for allowing me time to get this thesis written up and pushing me to finish and submit.

I'd like to thank dad for encouraging to follow my dreams and fulfil my potential and mum for providing the initial inspiration for a career in science.

Last and most important is Sam, my partner, who has got me through all the hard times and given me the best support anyone could ask for.

## **Dedication**

To Dad, sorry you couldn't hold on long enough to see me finish this. Your unwavering faith and pride that I would gave me the push I needed.

## Table of Contents

Abstract.....	i
Acknowledgements.....	ii
Dedication.....	iii
List of Figures.....	1
List of Tables.....	9
Glossary.....	11
Chapter 1 Introduction.....	16
1.1 Academic Drivers for Research Project.....	16
1.2 Industrial Partner – Alcyomics Ltd .....	19
1.3 Research Problem and Commercial Motivation for Alcyomics Ltd.....	19
1.4 Project Aim and Objectives .....	20
1.5 Industry Drivers for Research Project.....	21
1.5.1 Pharmaceutical Industry Drivers .....	21
1.5.2 Cosmetics Industry Drivers.....	22
1.5.3 Chemical Industry Drivers .....	23
1.6 Research Field .....	23
1.7 Research Methodology and Approach.....	24
1.8 Research Significance and Contribution .....	26
1.9 Ethics.....	27
1.10 Organisation of Thesis .....	27
Chapter 2 Background and Theory .....	29
2.1 Histology and Histopathology.....	29
2.1.1 Sample Preparation .....	29
2.1.2 Colourimetric Staining and Brightfield Microscopy .....	30
2.1.3 Grading in Histopathology.....	31
2.2 Structure of human skin .....	32
2.2.1 The Epidermis.....	33
2.2.2 The Dermal-Epidermal Junction.....	34
2.2.3 The Dermis.....	35
2.2.4 The Hypodermis .....	36
2.2.5 Histopathology of the Skin.....	36
2.3 Graft versus Host Reactions .....	37

2.3.1	Skimmune skin explant assay.....	38
2.4	Digital Image Processing and Analysis Theory.....	43
2.4.1	Digital Image Representation .....	44
2.4.2	Image Types.....	45
2.4.3	Image Resolution: Spatial and Intensity.....	46
2.4.4	Colour Image Processing.....	47
2.4.5	Colourspace Theory.....	48
2.4.6	Image Transforms.....	55
2.4.7	Pixel Neighbourhoods .....	56
2.4.8	Image Mean Filtering.....	56
2.4.9	Contrast enhancement methods .....	57
2.4.10	Set and logical operations .....	61
2.4.11	Segmentation .....	63
2.4.12	Thresholding – A region based approach to segmentation .....	64
2.4.13	Connectivity and mathematical morphology .....	66
2.5	Machine Learning, Feature Selection and Classification Theory.....	71
2.5.1	Feature Extraction .....	71
2.5.2	Classification .....	77
2.5.3	Naïve Bayes Classification.....	79
2.5.4	Considerations in Statistical Classification.....	80
2.5.5	Feature Selection .....	81
2.5.6	Dimensionality Reduction – an Alternative Approach to Feature Selection .....	82
2.5.7	Classifier Evaluation.....	83
Chapter 3	Literature Review.....	86
3.1	Grading Variability in Manual Histopathology .....	86
3.1.1	Skin Explant Assay .....	89
3.1.2	Alternative Methods to test for Toxicity, Allergenicity and Immunogenicity .....	90
3.2	Digital Histopathology.....	91
3.2.1	Challenges of applying Computer Analysis in to Histopathology.....	91
3.2.2	Colour Normalisation in Digital Histopathology.....	94
3.2.3	Segmentation .....	96

3.2.4	Segmentation of Morphological Structures.....	98
3.2.5	Segmentation of Tissue .....	99
3.2.6	Feature Extraction .....	101
3.2.7	Feature Selection .....	104
3.2.8	Classification .....	104
3.2.9	Ground Truth .....	108
3.3	Imaging and automated analysis of skin.....	108
3.3.1	Alternative Imaging Modalities.....	108
3.3.2	Application of Image analysis techniques for skin histopathology...	110
Chapter 4	Data Generation and Image Acquisition.....	113
4.1	Data Source: Skin Explant Assay .....	113
4.2	Image variation in skin explant data set .....	113
4.3	Manual Slide Examination .....	117
4.4	Digital Representation Framework .....	117
4.4.1	Sampling .....	117
4.4.2	Multiresolution Image Acquisition .....	119
4.4.3	Whole Slide Imaging.....	119
4.5	Assessment of Leitz Wetzlar Microscope/ Canon Digital Camera .....	121
4.5.1	Field of View .....	122
4.5.2	Magnification and Image Resolution .....	122
4.6	Assessment of Zeiss Axio Imager A2 System.....	124
4.6.1	Magnification and Image Resolution .....	124
4.6.2	Image Tiling, White Balance Correction and Background Correction	126
4.7	Image Grading using Manual Approach .....	129
4.8	Discussion of Data Generation and Image Acquisition.....	131
Chapter 5	Image Processing and Segmentation .....	134
5.1	Sample and Epidermal Segmentation: Method Development.....	135
5.1.1	Sample segmentation and Image Cropping.....	137
5.1.2	Colour Normalisation.....	143
5.1.3	Colourspace Conversion .....	145
5.1.4	Contrast Enhancement.....	147
5.1.5	Linear Combination .....	148

5.1.6	Thresholding.....	150
5.1.7	Morphological Processing.....	151
5.1.8	Object Classification.....	153
5.1.9	User Interaction.....	154
5.2	Epidermal Segmentation Optimisation and Evaluation.....	154
5.2.1	Generation of a Ground Truth Data set.....	155
5.2.2	Performance Metrics.....	155
5.2.3	Optimisation of Algorithm Parameters.....	157
5.2.4	Optimisation of Object Classification Rules.....	167
5.2.5	Final Performance Evaluation for Epidermal Segmentation.....	169
5.3	Dermal Segmentation.....	176
5.3.1	Sample Perimeter Masking.....	178
5.3.2	Dermal Segmentation: Method.....	180
5.3.3	Dermal Segmentation: Results.....	183
5.4	Cleft Segmentation.....	185
5.4.1	Cleft Segmentation: Method.....	186
5.4.2	Cleft Segmentation: Results.....	190
5.5	Vacuole Segmentation.....	192
5.5.1	Vacuole Segmentation: Method.....	193
5.5.2	Vacuole Segmentation: Results.....	194
5.6	Size-based Classification of Vacuoles and Clefts.....	195
5.7	Discussion of Image Processing and Segmentation.....	197
Chapter 6	Feature Extraction, Selection and Classification.....	204
6.1	Feature Extraction and Selection.....	204
6.1.1	Morphological Features.....	204
6.1.2	Texture Features.....	207
6.1.3	Feature Selection.....	210
6.1.4	Data Preparation and Use in the Classification Task.....	211
6.1.5	Methodology for Selecting Features.....	212
6.1.6	Model Estimation of the Naïve Bayes Classifier.....	214
6.1.7	Testing of Predictive Accuracy.....	214
6.1.8	Results for Forwards Feature Selection using Cross Validation.....	214



6.1.9	Wrapper-based Backwards Feature Selection using Cross Validation	216
6.1.10	Analysis of Feature Subset.....	219
6.1.11	Wrapper-based Backwards Feature Selection using 10 fold Cross Validation.....	221
6.1.12	Final Feature List.....	222
6.2	Final Model Training.....	222
6.3	Final Model Validation.....	224
6.3.1	Investigation into Misclassified Images.....	224
6.3.2	Investigation into Effect of Prior Probabilities on Classifier Performance.....	226
6.3.3	Final Industrial Application of Image Processing, Feature Selection and Classification Algorithm .....	227
6.4	Discussion of Feature Extraction, Selection and Classification.....	228
Chapter 7	Final Discussion and Conclusions.....	232
7.1	Discussion of Research Approach.....	232
7.2	Assessment of Performance against Industrial Research Objectives.....	234
7.2.1	Automation .....	234
7.2.2	Non-expert user .....	234
7.2.3	Accuracy.....	235
7.2.4	Repeatability and Reproducibility.....	236
7.2.5	Robustness.....	236
7.3	Discussion of Academic Research Contributions .....	237
7.4	Future Work.....	238
Appendix A.....		240
Appendix B.....		242
Appendix C.....		245
References.....		247

## List of Figures

Figure 1.1 Diagram showing the key areas in which knowledge is required for the project and the specific details of knowledge required.....	24
Figure 2.1 H&E stained section of normal skin at x40 magnification showing the typical layers within the epidermis: the basal cell layer (stratum basale); the spinous layer (stratum spinosum); the granular layer (stratum granulosum); the horny layer (stratum corneum). Image credit: Lutz Slomianka 1998-2009, Blue Histology ( <a href="http://www.lab.anhb.uwa.edu.au/mb140/">http://www.lab.anhb.uwa.edu.au/mb140/</a> ) .....	34
Figure 2.2 H&E stained section of normal skin at x10 magnification showing the epidermis and dermis. The dermal-epidermal junction is highlighted and has clear projections caused by folding of the layers. ....	35
Figure 2.3 Section of H&E stained skin tissue showing changes associated with grade I damage. The whole epidermis is outlined in blue; the stratum corneum layer of the epidermis is outlined in yellow .....	41
Figure 2.4 Section of H&E stained skin tissue showing changes associated with grade II damage. Arrows indicate some of the sites of vacuolisation. ....	41
Figure 2.5 Section of H&E stained skin tissue showing grade III changes, typified by extensive cleft formation at the dermal-epidermal junction. Two such clefts have been outlined in blue to highlight. ....	42
Figure 2.6 Section of H&E stained skin tissue showing grade IV changes, with complete separation of the epidermis from the dermis.....	43
Figure 2.7 A representation of an RGB image of order, $m \times n$ . A pixel at spatial coordinates $(x,y)$ will be represented by the column vector, $\mathbf{x}$ , shown in the figure. ....	46
Figure 2.8 An RGB image displayed at four spatial resolutions, $412 \times 550$ , $206 \times 275$ , $41 \times 55$ and $21 \times 28$ .....	47

Figure 2.9 Representation of the RGB colourspace cube and representation in RGB colour model of three colours.....49

Figure 2.10 RGB image and the three colour channel intensity images of the RGB colourspace .....49

Figure 2.11 RGB image and the three colour channel intensity images of the HSV colourspace .....51

Figure 2.12 RGB image and the three colour channel intensity images of the YCbCr colourspace .....52

Figure 2.13 Representation of the L\*a\*b\* colourspace. (The Mathworks, 2010). ...53

Figure 2.14 RGB image and the three colour channel intensity images of the L\*a\*b\* colourspace .....53

Figure 2.15 Colour RGB image and the same image converted to greyscale .....55

Figure 2.16 Diagram showing how pixel neighbourhoods relate to point, local and global image operations.....56

Figure 2.17 Square 3x3 Kernel for Mean Filtering, with the origin placed in the centre.....57

Figure 2.18 Remapping grey-levels using a transform function .....58

Figure 2.19 Diagram illustrating the effect of contrast stretching the intensity histogram, using a linear transform. Two sets of potential high and low values for the remapping are show. ....59

Figure 2.20 Diagram illustrating how penetration points can be selected using a cumulative percentage histogram. ....59

Figure 2.21 Diagram illustrating the effect of selecting a set band of intensities for a linear remapping. ....60

Figure 2.22 The effect of normal and adaptive histogram equalisation on a greyscale image and its intensity histogram.....	61
Figure 2.23 Diagrammatic representation of set theory, showing union (b), intersection (c), complement (d), reflection (e) and translation (f).....	62
Figure 2.24 Representation of typical intensity histogram distributions and suitable threshold or threshold bands .....	65
Figure 2.25 Structuring element matrices which form the basis of disk and diamond shaped structuring elements with a size (radius) of 3. ....	67
Figure 2.26 Diagram showing the effect of erosion and dilation with a 3x3 structuring element on a simple binary image. ....	68
Figure 2.27 An example of a bounding box (in red) for a specific object .....	75
Figure 2.28 Examples of ellipses with the same second moment as specific objects in the image (marked with red ellipse boundaries). The major and minor axis lengths of one of the ellipses are shown by blue and green arrows respectively. ....	75
Figure 4.1 Four examples of H&E stained skin images, showing variation in shape, structure and orientation .....	115
Figure 4.2 Section of H&E stained skin section showing a tear at the dermal epidermal junction, indicated by the arrow.....	116
Figure 4.3 H&E stained skin image with grade I damage. Unusual break down in cell and tissue structure at cut sample edges is circled in blue.....	116
Figure 4.4 H&E stained skin section showing grade I changes with some focal grade III changes, circled in green.....	118
Figure 4.5 Schematic representation of the Leitz Wetzlar microscope and Canon camera image acquisition system. ....	121

Figure 4.6 Diagram indicating the portion of the microscope field of view that is captured by the camera sensor. Note the image used was not taken with this system. ....	122
Figure 4.7 Comparison of image resolution with four different microscope objective lenses. ....	123
Figure 4.8 A 1388 x 1040 pixel image captured using a x10 microscope objective lens. The image on the right is an enlarged section to show cellular detail .....	125
Figure 4.9 A 1388 x 1040 pixel image captured using a x20 microscope objective lens. The image on the right is an enlarged section to show cellular detail .....	125
Figure 4.10 Skin sample image, requiring 56 tiles at x20 magnification (objective). White balance, background correction and image stitching NOT applied. ....	127
Figure 4.11 Skin sample image, requiring 56 tiles at x20 magnification (objective). White balance applied, background correction and image stitching NOT applied. ....	128
Figure 4.12 Skin sample image, requiring 16 tiles at x10 magnification (objective). White balance, background correction and image stitching applied .....	128
Figure 5.1 Hierarchical structure of segmentation process, starting with the whole image and resulting in the segmentation of the critical histological features, vacuoles and clefts. ....	134
Figure 5.2 Main processing steps in the algorithm to segment the epidermis from a digital image of an H&E stained skin section. The text boxes on the right describe the function of the processing steps throughout the algorithm. ....	137
Figure 5.3 Bar chart showing the mode composite intensity, $bg_{thresh}$ , values for 50 images. ....	139
Figure 5.4 Effect of the automated image cropping procedure on an image which includes a number of small tissue fragments. ....	140

Figure 5.5 Effect of a pre-thresholding smoothing step on the subsequent thresholding operation. The figure shows the binary mask created by the thresholding operation without smoothing, and when the smoothing step is performed using a 9x9, 29x29 and 49x49 sized filter..... 141

Figure 5.6 The effect of a mean filtering step using a 29x29 filter on an RGB image. .... 143

Figure 5.7 Effect of colour normalisation on RGB skin images showing two RGB skin images before and after colour normalisation with different staining contrast, lighting during acquisition, overall colour hues, and proportions of epidermis and dermis tissue. The non-sample pixels have been changed to white in the normalised images. .... 145

Figure 5.8 Effect of linear combination of three sets of greyscale and b\* images. 150

Figure 5.9 Histogram of enhanced additive image showing Otsu threshold..... 151

Figure 5.10 Enlarged sections of images showing H&E stained epidermal cells. The arrows highlight the diameter of normal and vacuolised cells ..... 158

Figure 5.11 Representation of box, centre and axial points in a Central Composite Design..... 161

Figure 5.12 Residuals plots for model of key factor effect on mean sensitivity..... 164

Figure 5.13 Residuals plots for model of key factor effect on mean specificity..... 164

Figure 5.14 Optimisation plot for key parameters to maximise sensitivity ..... 166

Figure 5.15 Contour plot of the effect of area and extent object classification thresholds on mean sensitivity and specificity for segmentation of epidermis. ... 169

Figure 5.16 Four H&E stained images showing varying staining and lighting. A, B and C had sensitivities of < 60% and D had a sensitivity of 81%. .... 170

Figure 5.17 Boxplot of specificity, sensitivity and accuracy for epidermal segmentation in training and test sets– with and without user interaction .....	173
Figure 5.18 Boxplots showing effect of damage grade on specificity, sensitivity and accuracy of epidermal segmentation.....	174
Figure 5.19 A selection of images highlighting differences in the stratum corneum and areas of necrotic tissue.....	178
Figure 5.20 Diagram illustrating the effect of smoothing on the perimeter masking step.....	179
Figure 5.21 RGB image of H&E stained skin sample overlaid with mask of thickened perimeter.....	180
Figure 5.22 Histogram of dimension measurements for dermis and non-dermis objects.....	182
Figure 5.23 Histogram of area measurements for dermis and non-dermis objects.....	182
Figure 5.24 Subtraction of epidermis mask from sample mask.....	183
Figure 5.25 Subtraction of thickened sample perimeter mask .....	184
Figure 5.26 Removal of non-dermis objects with a classification rule.....	184
Figure 5.27 User-interactive removal of any remaining misclassified objects .....	185
Figure 5.28 A sub-epidermal cleft in an original RGB image and an image which has been normalised using histogram matching.....	186
Figure 5.29 Dermal cleft objects adjacent to the epidermis (Figure 5.29a ), the epidermal cleft objects adjacent to the dermis (Figure 5.29b), and the combination of both sets of cleft objects (Figure 5.29c).....	189

Figure 5.30 A skin sample showing grade III damage, with clefts at the DEJ. The thickened sample perimeter mask is shown masking a tear at one cut edge of the tissue.....	190
Figure 5.31 The effect of different thresholds on the binary cleft mask created during thresholding of two luminance images, one containing clefts and one with no clefts. ....	191
Figure 5.32 A section of the original RGB image, with the cleft boundaries identified using the cleft segmentation procedure plotted in green over the RGB image. ....	192
Figure 5.33 Sections of two RGB images, which have been through the whole vacuole segmentation procedure using thresholds of mode luminance, mode luminance - 80 and mode luminance -100. The final vacuole boundaries are plotted in blue over the RGB image. Yellow arrows indicate regions misclassified as vacuoles when using the lower threshold. ....	195
Figure 6.1 Four directions and sets of pixels pairs used to calculate texture features .....	209
Figure 6.2 Change in misclassification rate as the 25 best features are added sequentially.....	215
Figure 6.3 Change in misclassification rate as the 50 best features are added sequentially.....	216
Figure 6.4 Change in misclassification rate as features are removed from the 31feature subset .....	217
Figure 6.5 Change in standard error of the mean as features are removed from the 31 feature subset .....	218
Figure 6.6 Change in misclassification rate as features are removed from the 22 feature subset.....	221



Figure 6.7 The effect of changing the prior probabilities the performance of the  
final classifier. .... 226

## List of Tables

Table 1.1 Summary of limitations and issues with manual grading in histopathology and potential for computer-aided grading to solve these issues. ...	20
Table 2.1 Description of the main features, cell and tissue types in the three main skin layers (Cox and Coulson, 2010).....	33
Table 2.2 Lerner’s histological criteria for grading GVHR.....	37
Table 2.3 Histological criteria for grading GVHR in the Skimune assay .....	39
Table 2.4 Comparison of unsupervised and supervised approaches to classification tasks.....	78
Table 3.1 Summary of literature in digital histopathology grouped according to feature type used .....	103
Table 3.2 Classification Approaches used in Histopathology and Published Performance .....	105
Table 4.1 Images excluded from further analysis and reason for exclusion .....	129
Table 4.2 Grading agreement of skin explant samples by two expert operators ..	130
Table 5.1 Visual analysis of dermal epidermal contrast in 5 colourspaces. ....	146
Table 5.2 Effects and Coefficients for Factors and Interactions in the Sensitivity and Specificity Screening .....	159
Table 5.3 Effect of each term in the final model on sensitivity and specificity.....	163
Table 5.4 Range tested and optimised values for the four factors included in the model.....	165
Table 5.5 Effect of the tested area and extent thresholds on algorithm sensitivity and specificity .....	168

Table 5.6 Summary of statistics for accuracy, sensitivity and specificity performance metrics .....	171
Table 6.1 Summary of features retained in the 22 feature subset .....	220
Table 6.2 Final feature subset used in classification model.....	222
Table 6.3 Misclassification error estimated using 10 fold cross validation.....	223
Table 6.4 Data on manual grading of the 10 misclassified images in the 169 image dataset.....	225

## Glossary

<b>Term (and acronym)</b>	<b>Definition</b>
<b>Acanthocyte</b>	An red blood cell characterized by multiple spiny cytoplasmic projections.
<b>Acantholysis</b>	A loss of intercellular connections (desmosomes) between keratinocytes; causes change in cell shape from polygonal to round.
<b>Algorithm</b>	A formula or set of steps with unambiguous rules for solving a particular problem.
<b>Allergenicity</b>	The degree to which a substance can cause allergic sensitisation.
<b>Allogeneic</b>	Cells or tissues taken individuals of the same species, which are genetically different to each other because they are derived from separate individuals.
<b>Alloreactivity</b>	The reaction of lymphocytes or antibodies with alloantigens.
<b>Area of Interest (AOI)</b>	The area containing the features of interest, to be used in subsequent analysis.
<b>Assay</b>	A procedure in molecular biology for testing or measuring the activity of a drug or biochemical in an organism or organic sample.
<b>Basal (relating to epidermis)</b>	The deepest layer of the epidermis, located next to the dermal-epidermal junction.
<b>Benign</b>	A condition which will not metastasize and is not harmful in and of itself. Treatment/removal can alleviate symptoms (e.g., pressure on surrounding organs), and treatment/removal is considered sufficient for complete recovery.
<b>Bone Marrow Transplant (BMT)</b>	Delivers healthy bone marrow stem cells into the patient to replace damaged or defective bone marrow.
<b>Brightfield microscopy</b>	Microscopy techniques using a broad spectrum light source to visualize the specimen, where light passing through sample is differentially absorbed, creating contrast.
<b>Carcinoma</b>	A cancer of the epithelium.
<b>Chromacity</b>	An objective specification of the quality of a colour determined by hue and colourfulness (not luminance).

<b>Chromacity</b>	An objective specification of the quality of a color regardless of its luminance determined by its hue and colorfulness
<b>Chromatin</b>	Nuclear material that is readily stained, consisting of the nucleic acids and associated proteins.
<b>Cleft</b>	A space or opening, made as if by splitting.
<b>Colourimetric stains</b>	Coloured stain that binds specifically to certain chemical/biological constituents in the body.
<b>Counterstain</b>	A stain used as contrast to another, generally more specific, stain.
<b>Cytology</b>	The study of cells at a microscopic level, generally via a light microscopy technique.
<b>Cytoplasm</b>	All cell contents outside of the nucleus and enclosed within the cell membrane.
<b>Cytotoxic</b>	Toxic to cells.
<b>Densitometry</b>	Measurements related to the optical density of a sample.
<b>Desmosomes</b>	Specialised cell junctions characteristic of epithelia, especially obvious in skin.
<b>Dyskeratosis</b>	Abnormal, premature, or imperfect keratinisation of the keratinocytes below granular cell layer; often have brightly eosinophilic (pink-staining) cytoplasm.
<b>Epithelium</b>	The internal and external lining of cavities within the body; also the external covering (skin).
<b>Field of View (FOV)</b>	The diameter of the image that can be viewed at the microscope eyepiece. Also refers to the rectangular area that is captured by the camera sensor.
<b>Gleason grading</b>	A grading for prostate cancer, characterizing the tumor into one of 5 categories based on tumour differentiation.
<b>Graft versus Host Disease (GVHD)</b>	A complication following bone marrow/ stem cell transplants in which the transplanted material attacks the transplant patient's body.
<b>Graft versus Host Reaction (GVHR)</b>	In this thesis this will be used to refer to the skin reactions created in the skin explant assay which mimic cutaneous GVHD.
<b>Grey-level co-occurrence matrix (GLCM)</b>	A matrix representing the spatial relationship between grey levels of neighbouring pixels.

<b>Ground truth</b>	The correct/desired output (i.e., truth) for an image analysis algorithm. Originally a term from the remote sensing community, reflecting the truth obtained from a ground-based survey.
<b>Haematopoietic</b>	Associated with the formation of blood or blood cells in the body.
<b>Haematopoietic stem cell</b>	A stem cell from which red and white blood cells evolve.
<b>Haematoxylin and Eosin (H&amp;E)</b>	The popular staining method in histology, widely used stain in medical diagnosis.
<b>Histology</b>	The anatomical study of microscopic structure of plant and animal cells and tissues. In this thesis it will only be used to refer to animal tissue.
<b>Histopathology</b>	The application of histology for disease diagnosis.
<b>Human Leukocyte Antigen (HLA)</b>	Any of the numerous antigens (substances capable of stimulating an immune response) involved in the major histocompatibility complex in humans.
<b>Immunogenicity</b>	The ability of a particular substance, such as an antigen or epitope, to provoke an immune response in the body of a human or animal.
<b>Immunohistochemistry (IHC)</b>	The process of localizing antigens (e.g. proteins) in cells of a tissue section exploiting the principle of antibodies binding specifically to antigens in biological tissues.
<b>Immunomodulatory</b>	Capable of modifying or regulating immune functions.
<b>in vitro</b>	Experimentation performed not in a living organism but in a controlled environment such as a test tube.
<b>in vivo</b>	Biological process or experiment within a living organism.
<b>Keratinocyte</b>	The predominant cell type in the epidermis specialised to synthesise keratin.
<b>k-means</b>	A widely known and used unsupervised classification algorithm which clusters data into k clusters while minimising the intra-cluster variance.
<b><math>L^*a^*b^*</math> Colourspace</b>	A colourspace made up of luminance ( $L^*$ ), red-green chromacity ( $a^*$ ) and yellow-blue chromacity ( $b^*$ )
<b>Lymphocyte</b>	A type of leukocyte (white blood cell) that is of fundamental importance in the immune system.

<b>Maculopapular rash</b>	A rash characteristic of GVHD which contains both <i>macules</i> and <i>papules</i> , a <i>macule</i> being a flat discoloured area of the skin, and a <i>papule</i> a small raised bump.
<b>Major Histocompatibility Complex (MHC)</b>	A set of molecules displayed on cell surfaces that are responsible for lymphocyte recognition and "antigen presentation".
<b>Malignant</b>	A condition which will eventually lead to death if untreated. Malignant conditions tend to metastasize, grow uncontrollably, and lack proper tissue differentiation.
<b>Markup</b>	The specification of ground truth, often obtained from an expert by the physical marking of an image for regions of interest, etc.
<b>Mathematical Morphology (MM)</b>	A theory and technique for the analysis and processing of geometrical structures, based on set theory, lattice theory, topology, and random functions.
<b>Mononuclear cell</b>	A collective term for certain leukocytes and phagocytes cells in the haematopoietic system.
<b>Morphology</b>	The form or structure of an organism or one of its constituent parts, the term does not include the function.
<b>Necrosis</b>	The death of most or all cells in tissue (or an organ due to disease or injury).
<b>Nucleus</b>	Membrane bound structure inside cell containing hereditary information.
<b>Optical Density (OD)</b>	Provides a linear relationship between image intensity and stain density, based on Lambert-Beer's law describing the intensity of light transmitted through a specimen.
<b>Pathophysiology</b>	The functional changes associated with a disease or an injury.
<b>Posterior Probability</b>	The probability of assigning observations to groups given the data.
<b>Prior Probability</b>	The probability that an observation will fall into a group before you collect the data.
<b>Scalar</b>	In linear algebra, real numbers are called scalars and relate to vectors in a vector space through the operation of scalar multiplication.
<b>Segmentation</b>	The process of delineating an image object.

<b>Sensitivity</b>	A measure of the proportion of positives that are correctly identified as such.
<b>Skin Explant Assay</b>	An <i>in vitro</i> test producing GVHR using a skin biopsy.
<b>Specificity</b>	A measure of the proportion of negatives that are correctly identified as such.
<b>Spongiosis</b>	An increase of fluid between the epidermal cells, causing the cells to splay apart in the upper epidermal layers (resembling a sponge).
<b>Stroma</b>	Connective tissue.
<b>Supervised Learning</b>	A machine learning technique for deducing a function from training data.
<b>T cell</b>	An important type of white blood cell in the immune system.
<b>Thresholding</b>	A simple procedure to segment an image by setting all pixels whose intensity values are above a threshold to a foreground value and all the remaining pixels to a background value.
<b>Toxicity</b>	The degree to which a substance can damage an organism.
<b>Transform</b>	A procedure that changes one function into another.
<b>Vacuole (vacuolisation)</b>	A small cavity in the cytoplasm of a cell, bound by a single membrane and containing water, food, or metabolic waste. Vacuolisation is the state of having become filled with vacuoles.
<b>Vector</b>	A one dimensional array.
<b>Wrapper methodology</b>	A method for feature selection which searches for the optimal feature subset for a particular classifier and domain.



## Chapter 1 Introduction

This thesis describes the development of a digital image analysis process to assess the immune-mediated structural breakdown of human skin tissue. While optimised and modified versions of the developed process may be applicable in a number of other applications in digital histopathology, particularly those concerned with analysis of epithelial tissue, this research looks specifically at the assessment of histological damage in haematoxylin and eosin stained human skin samples. In this chapter the motivation for the research is introduced, some of the issues associated with traditional histopathology approaches are highlighted and the potential benefits of applying digital image analysis to this field are presented.

This thesis is being submitted as the research element for an Engineering Doctorate, which by its nature addresses an industrial engineering challenge. The industrial context, including background on the industrial partner, Alcyomics Ltd, is given in this introductory chapter, along with the commercial and industrial drivers for the research. An overview of the multi-disciplinary nature of this project and the general research approach taken is provided, and the main research contributions summarised. Finally, a summary of the organisation of this thesis is provided.

### 1.1 Academic Drivers for Research Project

Histopathology refers to the microscopic examination of human cell tissue for the study and diagnosis of disease through expert medical interpretation. Traditionally in histopathology the patient diagnosis or prognosis is made based on the appearance of specific features within stained tissue biopsies viewed through a microscope. Manual grading methods are time and labour-intensive, and the lack of quantitative characterisation can lead to issues relating to subjectivity and inter and intra-observer variability (Farmer *et al.*, 1996; Standish *et al.*, 2006; Van Putten *et al.*, 2011).

Digital image analysis has the potential to address some of the issues associated with traditional histopathological methods. Automated software to analyse characteristic features of damage quantitatively and classify the grade of damage,

provides the opportunity to reduce the analysis time required from the histopathologist, and in some cases allow a less experienced operator to carry out the analysis. A major benefit of using image analysis for histopathological analysis and grading is the opportunity to capture and use quantitative information. This could replace the histopathologist by providing a grading decision; however a more likely use would be to aid the decision-making process of the histopathologist, researcher or clinician. Although semi-quantitative grading criteria incorporating procedures such as cell counting are sometimes used in manual histopathology, image analysis offers the opportunity to significantly increase the amount of quantitative data which can be extracted from the image and analysed, with the ultimate aim being to improve the repeatability, reproducibility, objectivity and accuracy of the process.

Automated image processing and analysis have been used in other medical disciplines including cytology and radiology for a number of years, however the general uptake of these methods in histopathology has only occurred recently as a consequence of the development of high quality digital slide scanning technology, the increased level and availability of computing power and the development of new image analysis algorithms that are able to handle the inherent complexity in tissue images. Following the widespread adoption of digital scanning systems in pathology, the development of automated image analysis procedures for scanning, segmentation and ultimately diagnosis has become the focus of significant research in recent years (Gurcan *et al.*, 2009). Theoretically it is now possible for the entire histopathology assessment process to be automated, including slide digitisation, image processing and enhancement, identification of key features and final diagnosis. In practice, individual parts of the process tend to be automated and combined with some manual intervention, often identifying non-routine or complex cases for manual assessment. There are a number of examples in the literature which propose automated image analysis as a useful aid to histopathological and clinical diagnosis. They include a diagnostic tool for HER-2 status used to support treatment decisions in breast cancer (Dobson *et al.*, 2010) and an automated workflow for staining, slide-scanning, and image analysis to

explore cancer biomarkers which can be used in conjunction with traditional pathology to support drug discovery and development (Shinde *et al.*, 2014).

Although there is a growing body of work relating to image segmentation, feature extraction and classification in digital histopathology, the majority of research publications in the area relate to cancer pathology and have been developed for specific applications. In particular there are very few papers describing segmentation (identification) techniques for application to epithelial tissue such as human skin, and none specifically for epidermis tissue. The challenge of applying image analysis to histopathology relates to the high structural complexity of the images, the biological variation and the complex sample preparation procedures. An additional challenge of this research is that the skin tissue must be identified in various states of structural breakdown.

In histopathology mathematically driven feature extraction, based for example on texture or wavelets, can be used to provide a representation of spatial information for use in a classification algorithm. Such mathematically driven features can be difficult for a histopathologist or clinician to interpret as they do not relate to traditional histological criteria, which can in turn lead to mistrust and a subsequent slow uptake of the developed technologies. An alternative approach is to use object level features based on shape and size which have a stronger association with traditional histopathological grading criteria. Quantitative object based features offer an opportunity to build on traditional histological criteria and domain knowledge by analysing shape and size more accurately and in greater detail. In addition, using a computer based system enables existing methods such as counts of abnormal cells to be carried out much more quickly and accurately.

There is a significant amount of published research presenting classification algorithms for use in the histological diagnosis and grading of cancer. Grade based classification accuracies in the literature vary significantly, and often accuracy for a cancer/ non-cancer decision is much higher than the accuracy when discriminating between different grades. For instance, Keenan *et al* (2000), reported accuracies of between 62.3%-98.7% for discrimination of different grades in haematoxylin and eosin (H&E) stained cervical tissue and whilst Tabesh *et al* (2007) could

discriminate between cancer and non-cancer in 96.7% of prostate cancer tissue slides, discrimination between low and high cancer grades was much lower at 81%. Consistent discrimination between different grades of cancer is a challenge in histopathology, which researchers are attempting to tackle using ever more complex ensemble methods consisting of multiple classifier types. A different approach has been adopted in this research, with the focus being to extract relevant features, thus enabling a simpler classification method to be used successfully. The use of these features to train a Naïve Bayes classification algorithm, and the optimisation, testing and validation of that system formed the final part of the research.

## **1.2 Industrial Partner – Alcyomics Ltd**

Alcyomics Ltd provides screening services and novel solutions for product safety, potency, toxicity and efficacy testing in the pharmaceutical, chemical and cosmetics industries. The core technology is Skimune™, a laboratory test (or assay) which uses skin samples from healthy volunteers to carry out safety and efficacy assessments of novel compounds and drugs. These compounds may cause allergic or immunogenic responses, contact sensitivity or inflammatory damage in the tissues of the body, which is mimicked in the skin response. The Skimune assay is based on an approach originally used to predict the occurrence and study the pathophysiology of graft versus host disease (GVHD), a common complication following bone marrow transplants (BMT) (Vogelsang *et al.*, 1985; Sviland *et al.*, 1990). Alcyomics was spun out from Newcastle University in 2007 to exploit the commercial potential of the assay in the pharmaceutical, chemical and cosmetics industries.

## **1.3 Research Problem and Commercial Motivation for Alcyomics Ltd**

In the Skimune assay, a sample of human skin is removed and co-cultured with immune cells in the presence of the test compound. Any immune response caused by the test substance creates an immune reaction in the skin, which is assessed and graded for severity using histopathology. Traditionally in histopathology, a trained specialist examines a small sample of sectioned and stained tissue under the microscope, and uses qualitative analysis guided by a descriptive grading scale

to make a diagnosis. The requirement for a specialist limits the reach of technologies which use histopathology, as not all potential customers have access to a histopathologist with the relevant experience. For Alcyomics to expand their business beyond a service based model to a product, technology or platform model, the expert knowledge of the in-house team must be captured and made available for use by others in an accurate, reliable and reproducible way. The highly regulated industries which Alcyomics are targeting demand that any assay used in their research, development and product testing is reliable, repeatable, objective and validated. Moving to an automated computer-based grading procedure will improve the repeatability and objectivity of the assay output by removing the subjectivity associated with human interpretation of the qualitative grading criteria. Some of the issues associated with the current manual grading system, and the potential benefits offered by changing to an automated, computer-aided system are summarised in Table 1.1.

*Table 1.1 Summary of limitations and issues with manual grading in histopathology and potential for computer-aided grading to solve these issues.*

<b>Manual Grading Limitations</b>	<b>Potential Improvement offered by Computer Aided Grading</b>
Required operator training	Can be run by non-expert
Labour intensive	Potential for high throughput
Qualitative	Quantitative
Subjective	Objective
High inter/ intra operator variability	Improved repeatability

#### **1.4 Project Aim and Objectives**

The overall project aim was to develop an automated system to enable non-expert users to grade histological skin damage in the Skimune assay with a comparable level of accuracy, repeatability and reproducibility to that achieved through manual grading.

More specifically, the *project objectives* were defined as:

- Compare different image acquisition methods and determine the most appropriate in terms of resolution and practicality.
- Identify the specific histological features associated with damage in the Skimune assay and extract quantitative measurements associated with those features.
- Train a classification algorithm using the extracted features so that it is capable of identifying positive examples of histological skin damage with high sensitivity and specificity and which is capable of handling the normal variation within the skin images.

## 1.5 Industry Drivers for Research Project

### 1.5.1 Pharmaceutical Industry Drivers

The pharmaceutical industry is facing considerable challenges, including low revenue growth, the arrival of the patent cliff, a faltering product pipeline, poor stock performance, public concerns over transparency and ethics, and high attrition rates during development (DiMasi *et al.*, 2010). Drug development is an expensive process, with the main factors affecting cost being time and risk. High failure rates are a major source of risk in the pharmaceutical industry, with many drug candidates rejected at late stages of development due to issues of toxicity and lack of efficacy (Kola, 2008). As a result of these issues, there is a growing drive in the industry to identify promising candidates early, and to ensure toxic or non-efficacious candidates fail early in the process. The identification of products with a high chance of failure may use *in silico* prediction, pre-clinical laboratory screening or animal models.

Within the pharmaceutical industry, skin explant assays can be used to provide information on mode of action, toxicity, adverse reactions, safety and efficacy of therapeutics, prior to first in man clinical trials. The data generated can be analysed to investigate dose response and it also offers the potential to implement a stratified medicine approach to patient selection in clinical trials by identifying patient groups who respond differently to particular drugs. Through such *in vitro*

technology, it may be possible to minimise the risks to patient safety, reduce the financial risk associated with drug development, shorten development times and reduce costs due to costly animal trials.

The Skimune skin explant assay developed by Alcyomics is an alternative to the animal models traditionally used for pre-clinical safety testing. Alternative procedures are required as animal models are expensive and do not always predict human response accurately (Hackam and Redelmeier, 2006; Perel *et al.*, 2007). One reason proposed by Sena *et al* (2010) for the poor conversion of successful animal studies to drug approvals is bias against publishing negative results in the pharmaceutical industry.

One example of the potential impact of the skin explant test is provided by the TGN1412 phase I trial in 2006, where six human volunteers suffered multi-organ failure after being given a dose of CD28 superagonist antibody five hundred times smaller than that found safe in animal studies (Attarwala, 2010). The TGN1412 drug, when tested using the skin explant assay by Alcyomics, gave a positive result indicating the likely immunogenicity in humans.

### **1.5.2 Cosmetics Industry Drivers**

In the cosmetics industry, Alcyomics are taking advantage of a move to replace and reduce the use of animal models, as typified by the 7th Amendment to the Cosmetics Directive (Directive 76/768/EEC2), which prohibits all animal testing for cosmetics and toiletries. This Directive requires that alternative methods are introduced for a range of toxicological end points, including testing for skin sensitisation. The skin explant assay offers an alternative to animal testing for cosmetics that require human safety testing and could therefore be described as part of the 3R movement. The 3R movement, first described by Russell and Burch (1959), aims to refine, reduce and replace the use of animal models in scientific research. The market for 3R tests is likely to increase significantly in the next 5-10 years, due to a combination of macro-environment forces including EU legislation, UK government programmes, high costs of animal testing and strong public opinion (particularly regarding cosmetics testing).

The current procedure has limited throughput due to skin supply and manual measurement, and the subjectivity of assessment may concern industry regulators. Automating the assay read out, reducing bias, and improving the objectivity of the assay will allow Alcyomics to compete with some of the other alternatives to animal testing being used in the cosmetics industry such as artificial skin models and *in silico* prediction methods.

### **1.5.3 Chemical Industry Drivers**

There is a drive in the chemical industry, typified by the REACH system for controlling chemicals in Europe, to understand the risks associated with new and existing chemicals on the market. REACH is a European Union regulation concerning the **R**egistration, **E**valuation and **A**uthorisation of **C**hemicals. The REACH proposal requires industry to register all existing and future new substances with a new European Chemicals Agency. Part of the registration process requires the submission of data on toxicological properties of the chemical and any risks it may pose to human health. One of the aims of REACH is to promote the use of alternative methods in the chemical industry to assess the hazardous properties of substances.

In the chemical industry, Skimune could be used to test the immunogenicity, potency and toxicity of novel chemicals and identify those likely to create hypersensitivity responses. An automated, objective skin explant assay could serve as a useful, high throughput tool in this industry.

## **1.6 Research Field**

This EngD research project has involved the application of image analysis and classification methods from the fields of mathematics, engineering and computing, to a biological process. The project has been approached from an interdisciplinary perspective, combining knowledge and experience from a range of fields to develop the best solution to this industrial problem. It was necessary to gain a full understanding of the biology of skin reactions in the Skimune assay, including the nature of the different tissue types and cells within the skin, typical biological variation, and the appearance and typical features of a Graft versus Host Reaction



(GVHR). An understanding of the traditional methods used in histopathology and their limitations was also essential when designing a new approach to the grading of the skin explant assay. Having understood the problem, knowledge of the mathematical tools necessary to automate a complex human process was required; these tools can be split into data acquisition, image processing, and feature selection and classification. Figure 1.1 summarises the main areas of importance to the research and how each area is linked to the research project. Areas relating to biology are shown in red and those relating to mathematics or engineering are shown in blue.

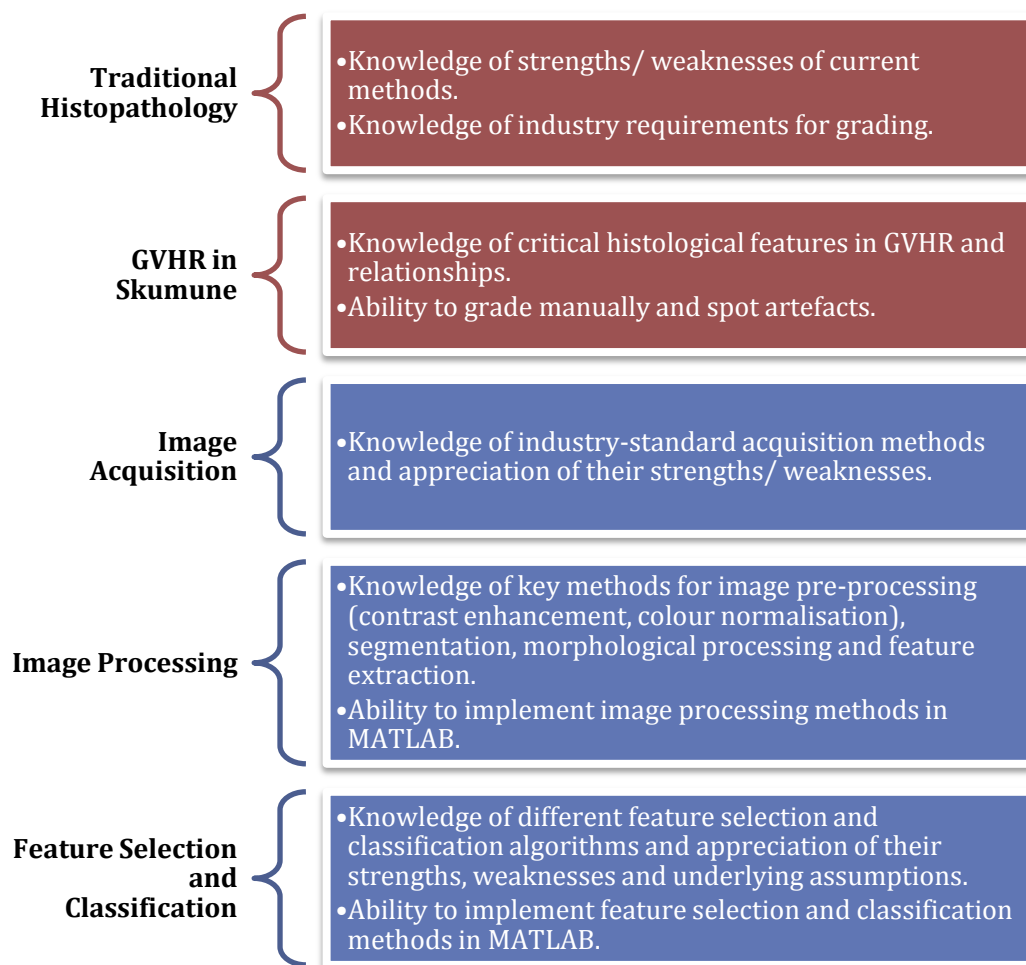


Figure 1.1 Diagram showing the key areas in which knowledge is required for the project and the specific details of knowledge required.

## 1.7 Research Methodology and Approach

The particular dataset of images used in this study and the needs of the industrial partner guided the choice of research methodology and approach. The large degree

of variation within the images which is not related to the changes associated with histological damage presented a major challenge in this research. Although the current trend in machine learning and image analysis is moving away from mimicking expert systems towards a more data driven approach (Bishop, 2010), this research attempts to show that there is value in combining expert knowledge with computational numerical capability and accuracy.

The human method for carrying out the assessment of histological grading is hierarchical. It proceeds by first locating the specific regions of interest within the image, then identifying the specific features before making an estimate of their extent. The human approach has been used as a guide to locate the regions of interest and identify the features of interest. It is after this point that the strengths of the computer system are utilised, extracting a large number of size and shape based statistics describing the features which a human would not be capable of doing. Wrapper based feature selection and finally the automated training of a classification model were used to identify the most useful features without any human bias to create the most accurate classifier possible.

The thesis does not attempt to present new individual image processing or classification methods, but rather describes the novel application of a variety of image processing, segmentation, feature extraction and classification techniques to a new and challenging industrial problem. The research is not presented as a computer sciences thesis as are many in the field of image analysis, instead the general approach has been to design an industrially useful solution based on a deep understanding of the biological system and process. Computing is used to enhance the parts of the process in which humans are weakest, but the new process learns from the parts of the human approach which are well adapted. In summary, the research has aimed to design of a system to capture expert histological knowledge and apply it in a reproducible manner through a computer-based algorithm.

## 1.8 Research Significance and Contribution

In this thesis, an image analysis process for the automated grading of histological skin damage is described. The overall industrial goal is to create a solution for Alcyomics which enables the output in the Skimmune assay to be assessed in an objective and repeatable manner without an expert histopathologist. The main academic contributions of this work are three-fold:

- First, a method has been developed which is able to identify epidermis tissue from H&E stained skin sections showing varying degrees of histopathological damage. Although many methods have been described for segmentation of histology images, most are for cell, gland or nuclear segmentation rather than tissue segmentation. The epidermis segmentation algorithm is a useful addition to this small but growing area of research and could provide a framework for segmentation of other epithelial tissues. The author is unaware of any segmentation methods that have been applied to images showing severe histological damage such as graft versus host type reactions. This part of the work has been published in the peer reviewed open access academic journal, BMC Medical Imaging, where it has been classified as highly accessed. The paper is available online at <http://www.biomedcentral.com/content/pdf/1471-2342-14-7.pdf>.
- Second, a novel set of object and spatial level quantitative features have been defined and a method for their extraction created. The extracted feature measurements are relevant to the expert grading criteria for histological damage but add a quantitative dimension. While this has direct application to the grading the Skimmune assay, this set of feature measurements could also be applied in an automated version of the Lerner grading used in the diagnosis and prediction of GVHD.
- An approach to histopathological tissue classification, which combines expert domain knowledge in the design of potential features, with a fully objective feature selection and classification approach.

## 1.9 Ethics

All postgraduate research projects require ethical review by both the university, as part of the project approval process, and by funding bodies such as the EPSRC. Newcastle University Ethics Committee requires ethical approval in cases where research involves human subjects (including the use of organs, tissue or information) or certain live animal subjects. The Alcyomics research project uses skin tissue samples from NHS patients for research. It was therefore necessary to complete an ethics approval form at the start of the project, so that it could be reviewed by the Ethics Committee.

The 2004 Human Tissue Act provides a legislative framework around the removal, use, storage and disposal of human tissue, sets out the requirements for participant consent, and the mechanism of regulation by the Human Tissue Authority. Alcyomics Ltd had external approval from the local ethics research committee for the use of human bodily samples in research and development in place prior to the start of the EngD project. This approval was for the use of skin and blood samples in the development of an *in vitro* skin safety assay for the detection of immunogenicity and hypersensitivity reactions to novel compounds and drugs. The approval was granted for five years from November 2010, on the basis that all patients gave informed consent for their tissue to be stored and used for commercial research. The tissue is stored in a HTA licensed tissue bank (Newcastle upon Tyne Research Tissue Bank, Licence No. 12048).

## 1.10 Organisation of Thesis

Chapter 1 introduces the industry partner, research problem, industry drivers for the research and highlights the research significance and general approach.

Chapter 2 presents the background theory required for full understanding of the thesis in terms of: human skin histopathology; microscopy and grading in histopathology; the Skimune assay; digital image processing and analysis theory; and theory of machine learning, feature selection and classification.

Chapter 3 examines the relevant published literature and current state of the art in the research area. The challenges of grading in histopathology and potential solutions are examined, and some of the technologies competing with the Skimune assay are described. The application of image processing and analysis techniques in histopathology is surveyed, focusing on the specific areas of colour normalisation, segmentation (with particular emphasis on the segmentation of epithelial tissues), features for histological assessment, automated grading, and assessment of performance through the use of a ground truth.

Chapter 4 is focussed on Image Acquisition, and gives technical background information, a description of the systems tested and an analysis of the suitability of each system for the project.

Chapter 5 describes the development of new processes for the hierarchical segmentation of the sample, the relevant tissue types and finally the features of damage. The final segmentation algorithms are described fully in this chapter. An analysis of the key epidermal segmentation stage is presented, using a manual segmentation to determine the sensitivity, selectivity and accuracy of the algorithm.

Chapter 6 describes the extraction of a set of feature measurements, the pre-processing of the feature vector dataset, the selection of an appropriate feature subset and the training and testing of the classification algorithm.

Chapter 7 includes a general discussion of the research contributions, and a set of suggestions for future work to improve and extend the current research.

## Chapter 2 Background and Theory

This chapter presents the background and theoretical knowledge that underpins the development of an automated image analysis and classification procedure for the grading of histological reactions in human skin. Background theory in a number of areas is introduced in this chapter to explain both the nature of the histological reactions and the potential approaches for image analysis and classification that can be applied. A general introduction to the histopathology and the structure of human skin is provided in sections 2.1 and 2.2, prior to explaining the skin explant assay in section 2.3. Section 2.4 describes typical image acquisition methods used in digital histopathology and the background and mathematical basis to the relevant image processing and analysis techniques used in the research. Section 2.5 gives a general introduction to the field of machine learning, with particular emphasis on feature selection, supervised classification techniques, the training and evaluation of classification models and the classification algorithm used in this research.

### 2.1 Histology and Histopathology

Histology is the microscopic study of plant and animal cell tissue and histopathology refers to the use of histology for the study and diagnosis of disease through expert medical interpretation. Histopathology can be used in the diagnosis of disease and is the “gold standard” in cancer diagnosis (Rubin *et al.*, 2007). It is also used to assess the severity and progression of a disease or to research disease mechanisms. Traditional histopathology relies on the examination of tissue samples obtained through biopsy by an expert histopathologist. A biopsy is performed by removing a small piece of tissue (often from a lesion or tumour) surgically. Commonly used procedures for biopsy include excision, punch biopsy, shave biopsy and curettage biopsy (Kempf *et al.*, 2008). The punch biopsy used in the skin explant assay involves the removal of a 3-4mm diameter cylindrical plug of skin tissue.

#### 2.1.1 Sample Preparation

The tissue from the biopsy must be prepared before it can be assessed by a histopathologist. This preparation stage ensures the sample is stable, thin enough

to be examined under the microscope and has sufficient contrast between the structures of interest. Although there are a variety of methods used in histopathology, the following method is typical of the preparation used in the skin explant assay. The tissue is first cut into small pieces, and then chemically fixed using a formaldehyde solution to preserve the sample while retaining the original tissue morphology. The fixed sample is then embedded in paraffin to harden; this enables it to be sectioned into slices 3-10 microns thick using a specialised automated instrument called a microtome. The tissue cross sections can then be mounted on to glass slides before removing the paraffin from the tissue so it is ready for staining.

### **2.1.2 Colourimetric Staining and Brightfield Microscopy**

Many biological specimens do not have sufficient contrast for structural details to be easily visible under a microscope. However, the tissue can be stained with colourimetric stains to improve the contrast. A variety of light-absorbing stains can be used to visualise cells and cell constituents, and they are essential for the recognition of tissue types and morphological features. The method of haematoxylin and eosin (H&E) staining has been used for over a century and remains one of the most widely applied in histology for reasons of cost, availability, simplicity and historical precedent (Gartner *et al.*, 2007). Haematoxylin stains the chromatin in the nuclei of cells a blue/purple shade. Eosin is a counterstain that binds non-specifically to proteins in the cytoplasm, connective tissue and other extracellular substances, staining them various shades of pink. Alternative stains can be used to stain the tissue selectively, for instance Oil Red O is used to stain lipids red and nuclei blue/ black, and Prussian Blue is used to stain iron bright blue.

Once the tissue samples have been sectioned and stained they are ready to be examined under a microscope. Although newer techniques such as digital slide scanning are becoming more common in histology to visualise tissue, the more traditional technique of brightfield microscopy is still commonly used for the visualisation of H&E stained tissue and is the current method of assessment used in the Skimune assay.

In brightfield microscopy the sample is placed on a stage above the light source and light is focussed onto the sample by a condenser lens placed between the light source and the sample. Light which is not absorbed by the sample is captured by an objective lens above the stage, which magnifies the light before transmitting it through an eyepiece into either the operator's eyes or onto an optical sensor. The sample appears dark in contrast to the bright viewing field, hence the name "brightfield". Brightfield microscopy images are generally acquired using either a digital camera attached to a microscope, a specialised microscope imaging system, or a digital slide scanner.

The simplest systems consist of a digital camera mounted on a standard microscope. This was the first digital set-up adopted in many histology labs, due to the low cost and simplicity. In addition to the quality of camera, the accuracy of the system set-up can also have a large impact on the image resolution in these systems. The main issues with such systems are maintaining consistency in terms of lighting and focusing, and the challenge of creating a whole slide image. To address the issues with the simple camera/ microscope systems, more complex specialised microscope-based systems have been designed. Motorised stages on which the sample is mounted can be linked with software to allow adjacent fields of view to be captured by automatically moving the microscope stage to a new position beneath the fixed optics. The stored stage positions can then be used to join together individual field of view images to create a single whole slide image. These systems often come with additional functionality and software to allow auto-focussing, background correction, and reproducible illumination and contrast setting. Digital slide scanners convert glass microscope slides to high resolution whole slide digital images using a completely automated process.

### **2.1.3 Grading in Histopathology**

In manual histopathology a medical specialist known as a histopathologist will examine the tissue and look for the presence of features associated with a particular disease. Scoring and grading systems with detailed criteria are then used to determine the severity of the disease. For example, grading systems in cancer histopathology are used to assess the extent of the disease, estimate patient



prognosis and determine the optimal treatment. Grading subdivides a diagnostic category to assist clinicians in making decisions about treatment, and is commonly used when assessing tumours and inflammatory conditions. The primary purpose of most grading systems is to help predict the biological behaviour of a disease and direct clinicians to the appropriate treatment.

## **2.2 Structure of human skin**

The skin is the largest organ in the human body and is the primary interface between the body and its environment. It has multiple functions including protection, sensation, thermoregulation, synthesis, storage, excretion and absorption (Gartner *et al.*, 2007). The anatomy of the skin reflects this functional complexity and comprises many different cell types, extracellular structures, and specialised appendages such as hair follicles and sweat glands. There is significant regional variation in the skin in terms of skin thickness, composition and appendage density; the appearance and structure of normal skin also varies according to the age, sex and ethnicity of the subject (Freinkel and Woodley, 2001). The skin comprises three separate layers, the epidermis, dermis and hypodermis, with the epidermis being the layer closest to the surface. The main features of the three layers are described in Table 2.1. As this research involves images showing only the epidermis and dermis, and is primarily concerned with changes in the epidermis and at the dermal–epidermal junction (DEJ), these areas will be described in more detail.

Table 2.1 Description of the main features, cell and tissue types in the three main skin layers (Cox and Coulson, 2010)

Skin Layer	Description	Main cell and tissue types
<i>Epidermis</i>	A keratinised, stratified, continually renewing epithelium.	Mainly keratinocytes, some melanocytes, Langerhans cells, and Merkel cells
<i>Dermis</i>	A dense, fibrous connective tissue consisting of a thin papillary and thicker reticular layer.	Extracellular collagen fibers, ground substance, and fibroblasts. Mast cells, lymphocytes and macrophages
<i>Hypodermis</i>	Adipose/fibrous connective tissue	Adipocytes (fat cells), fibrous tissue.

### 2.2.1 The Epidermis

The thin outer layer of the skin, known as the epidermis, is a type of stratified squamous epithelium. Epithelium is a type of animal tissue used to line cavities and surfaces in the body; stratified squamous epithelium comprises multiple layers of flat, plate-like epithelial cells resting on a basement membrane. The epidermis consists predominantly of structural cells known as keratinocytes, which synthesise a protein called keratin. The keratinocytes are organised into several epidermal layers according to their state of maturity. Langerhans cells are present in the epidermis, and are a type of dendritic cell which have a role in the skin's immune system. Merkel cells in the epidermis are thought to have a sensory touch function, and are difficult to visualise, requiring a specific immunohistochemical stain. Migrating cells such as lymphocytes, which are an essential component of the human immune system, may also be present in the epidermis transiently in diseased states. The epidermis regenerates in an orderly fashion starting with cell division of keratinocytes in the basal layer (*stratum basale*) of the epidermis. In the basal layer, keratinocytes are columnar in shape and attached to surrounding cells by structures known as desmosomes. Melanocytes are also present in the basal cell layer and are the cells responsible for the production of melanin, which protects skin from ultraviolet radiation.

The next layer is the spinous layer (*stratum spinosum*), so-called because of the 'spiny' appearance of the desmosome connections between keratinocytes. The

spinous cells change from a polyhedral shape near the basal layer to a larger, flatter shape near the granular layer. The granular layer (*stratum granulosum*) is characterised by cells with visible granules in the cytoplasm and once these cells die, they become keratinised and form the tough, outermost layer of skin, the horny layer (*stratum corneum*). The four layers of the epidermis and the changing shape of the cells throughout the layers can be observed in Figure 2.1. (Gartner *et al.*, 2007)

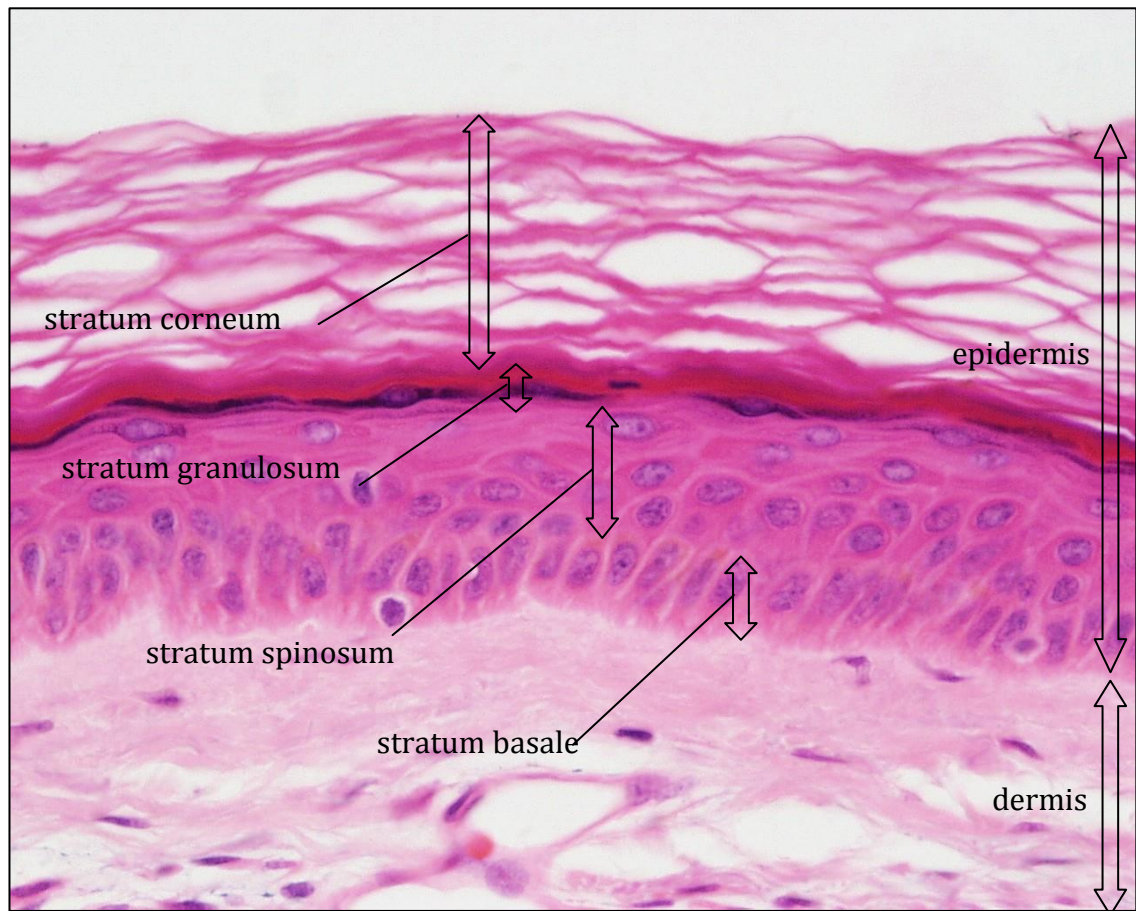


Figure 2.1 H&E stained section of normal skin at x40 magnification showing the typical layers within the epidermis: the basal cell layer (*stratum basale*); the spinous layer (*stratum spinosum*); the granular layer (*stratum granulosum*); the horny layer (*stratum corneum*). Image credit: Lutz Slomianka 1998-2009, Blue Histology (<http://www.lab.anhb.uwa.edu.au/mb140/>)

### 2.2.2 The Dermal-Epidermal Junction.

The junction of the epidermis and dermis, known as the dermal-epidermal junction (DEJ) is a type of basement membrane. Basement membranes are complex multi-layered structures found at the interface between cell sheets or between cells and connective tissue. The DEJ is important for the mechanical support of the epidermis, the anchoring and adhesion of the two layers, and the transport to and

from the epidermis (Gartner *et al.*, 2007). The complex variety of proteins present in this layer can affect proliferation, migration and differentiation of keratinocytes. The basal layers of the epithelium are folded to form dermal papillae where the dermis tissue forms projections into the epidermis, giving a characteristic wave-like appearance to the DEJ which can be seen in Figure 2.2.

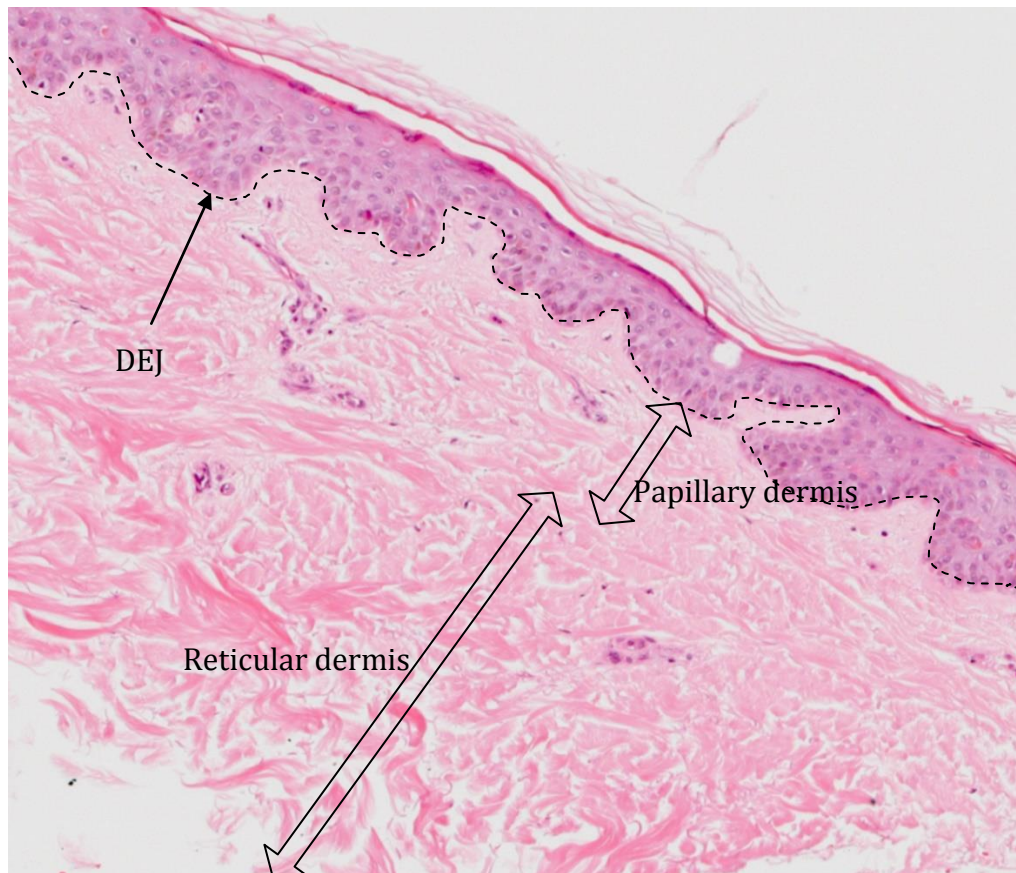


Figure 2.2 H&E stained section of normal skin at x10 magnification showing the epidermis and dermis. The dermal-epidermal junction is highlighted and has clear projections caused by folding of the layers.

### 2.2.3 The Dermis

The dermis is a tough, supportive cell matrix comprising a thin papillary and thick reticular layer, as shown in Figure 2.2. The papillary dermis connects with the epidermis and consists of thin, collagen fibres which stain light pink in traditional H&E staining. The collagen fibres, which make up 70-80% of the dermis, are loosely arranged in the papillary layer, but form a dense network in the reticular layer (Fraga, 2012). This gives the two dermal layers different textures within the image. In addition to the collagen fibres, the dermis is made up of collagen-producing fibroblasts, elastin, structural proteoglycans, immunocompetent mast

cells and macrophages (Gartner *et al.*, 2007). Groups of cells clustered throughout the dermis stain in a similar manner to the keratinocytes in the epidermis with H&E stain, and can therefore appear similar to a small patch of epidermis tissue.

#### **2.2.4 The Hypodermis**

The hypodermis lies immediately below the dermis and mainly consists of fat cells (called adipocytes), nerves and blood vessels. The fat cells are organised into lobules, which are separated by structures called septae containing nerves, larger blood vessels, fibrous tissue and fibroblasts. This layer will not be discussed again, as the specific skin damage that is the subject of this thesis does not cause changes in this layer. The hypodermis tissue layer is not captured in the biopsy samples being used in this research, and so there is no requirement to eliminate it from the analysis.

#### **2.2.5 Histopathology of the Skin**

Skin diseases tend not to have a single cause, unique to the particular disorder, and as a consequence skin disease definitions often rely on a complex combination of clinical, histopathological, immunopathological and genetic features (Cox and Coulson, 2010). However, four broad groupings can be defined (Fraga, 2012):

- *Genodermatoses* – skin changes associated with abnormal development of the epidermis.
- *Inflammatory dermatoses* – a broad category of skin diseases associated with inflammation.
- *Infections and infestations* – skin diseases caused by external organisms such as bacteria, viruses or fungi (infection) or parasites (infestation).
- *Neoplasms* – abnormal tissue mass due to abnormal cell proliferation which can be benign, premalignant (pre-cancer), or malignant (cancer).

The GVHRs which form the basis of this research are a form of *interface dermatitis*, where damage to the epidermis is caused by a T-cell mediated immune reaction. This type of reaction can be classified as an *inflammatory dermatosis*.

### 2.3 Graft versus Host Reactions

In transplant operations, blood-forming stem cells are given to a patient intravenously to restore hematopoietic function. GVHD occurs when immune cells in the transplanted material mount an immunologic attack against the patient's own tissues. The skin is usually the first and most commonly affected organ in GVHD. The typical histological reactions found in GVHD are known as graft versus host reactions (GVHR). Lerner *et al* (1974) described the histopathology of GVHR in detail following the investigation of a large number of marrow graft recipients and investigating typical histological reactions. The study resulted in a grading system being established to assess the severity of the reactions and predict clinical outcome based on a series of histological criteria. The grading system described by Lerner continues to be used to assess GVHR today. The criteria utilised in the Lerner grading system are given in Table 2.2 as a baseline for comparison with the reduced set of criteria used in the skin explant assay (Table 2.3 in the following section).

Table 2.2 Lerner's histological criteria for grading GVHR

Grade	Description by Lerner 1974
<i>0</i>	Normal skin
<i>I</i>	Focal or diffuse vacuolar degeneration of epidermal basal cells and acanthocytes. Lesion varies from vacuolization of basal cell cytoplasm to frank necrosis in the basal and suprabasal layers.
<i>II</i>	In addition to Grade I changes, focal and diffuse spongiosis (separation and intracellular edema of basal cells and acanthocytes), dyskeratosis or eosinophilic degeneration of epidermal cells, tending to occur in scattered individual cells.
<i>III</i>	In addition to Grade II changes, occurrence of clefts/spaces (acantholysis, epidermolysis) after necrosis of basal cells and acanthocytes in the basal cells and more superficial layers, resulting in separation of the dermal-epidermal junction.
<i>IV</i>	In addition to Grade III changes, frank loss of epidermis.

The main changes described by Lerner relate to the appearance of vacuoles, clefts and changes at the basement membrane. Vacuoles are membrane bound cavities in the cytoplasm or between cells, which appear as white regions in H&E stained images as they are not stained by either haematoxylin or eosin. Vacuolisation describes a state where the tissue becomes filled with vacuoles either within or adjacent to the cells, and is often seen at the base of the epidermis in skin histopathology. Clefts are a progression of vacuolisation at the basement membrane, i.e., as vacuolisation becomes more severe the vacuoles begin to fuse together and clefts occur at the DEJ. As the severity of the immune reaction increases, the clefts cover an increasing proportion of the DEJ until there is complete separation of the epidermis and dermis.

### **2.3.1 Skimune skin explant assay**

The Skimune assay is based on an *in vitro* skin explant assay that uses a surgically excised section of skin tissue (Vogelsang *et al.*, 1985). The assay has been used to predict both the occurrence of GVHD in human leukocyte antigen matched bone marrow transplant patients (Sviland *et al.*, 1990) and to study the pathophysiology of the disease (Dickinson *et al.*, 1994; Jarvis *et al.*, 2002). The commercialised assay applies the technology in such a way that the immunological, allergenic and toxic effects of a particular compound can be tested in a practical and safe *in vitro* environment, while retaining an authentic human immune response. Although the assay in its commercial form is not being used to classify GVHD, the immune reactions and histological changes being measured in the assay are assessed using criteria based on those used to classify GVHD. The original skin explant assay protocol is described in detail in Sviland and Dickinson (1999) the modified assay used for commercial applications is now discussed.

The modifications include using blood and skin tissue from single healthy volunteers rather than combining cells and tissue from patients and potential donors. There are also some changes to the laboratory procedures which have been made to improve the reproducibility of the assay. For the assay, blood samples and punch skin biopsies are taken from healthy volunteers after informed consent is attained. The 4mm<sup>2</sup> punch biopsies are taken from the back below the

waist-line and each biopsy is dissected into four equally sized sections. In the culture phase, primary dendritic and T-cells are extracted from the blood sample and then cultured with the test compound or drug. These pre-primed immune cells are then co-cultured with the skin samples to induce tissue damage in cases where there has been an immune response. Skin sections cultured in medium alone are used as controls. After 72hr incubation at 37°C, skin sections are fixed in formalin, sectioned and stained with H&E. The manual histopathological grading is then assessed and confirmed independently by two experts.

The Skimune assay is assessed using a modified version of Lerner's original grading criteria. Table 2.3 shows the simplified criteria adopted by Alcyomics during the development of the commercial assay. These criteria focus on the major morphological changes in the tissue such as vacuolisation, cleft formation and the appearance of dyskeratotic bodies, disregarding some other features such as necrosis and spongiosis which were determined not to be critical in the commercial application.

Table 2.3 Histological criteria for grading GVHR in the Skimune assay

Grade	Skimune Histological Criteria
0	Normal skin
I	Vacuolisation of epidermal basal cells.
II	Diffuse vacuolisation of basal cells with dyskeratotic bodies.
III	Sub-epidermal cleft formation
IV	Complete epidermal separation

Controls are included in every assay and include skin biopsies cultured in medium only. The *in vitro* culture process creates some baseline histopathological changes even when there is no immune response, and so grade 0 tissue with no damage at all is not usually observed in the assay. The control samples usually show grade I changes and so a grade I reaction is counted as a negative result. Reactions of grade II or above are considered a positive result in this assay, however if the control sample has grade II, III or IV changes the assay is repeated. In addition to vacuolisation, cleft formation and the presence of abnormal cells with a high



keratin content called dyskeratotic bodies, some of the samples generated in the assay include regions of necrotic (dead) tissue.

Figure 2.3 is an image of an H&E stained skin section from the Skimune assay showing grade I changes. The epidermis tissue boundary is shown in blue and the rest of the tissue shown is part of the dermis. The part of the epidermis outlined in yellow is the *stratum corneum*, the uppermost layer of the epidermis consisting of dead, keratinised cells (first described in section 2.2.1). This layer is highlighted as it is the only part of the epidermis not of interest when assessing GVHR and other immunological reactions. Figure 2.3 is characteristic of a grade I reaction, in that the cells of the epidermis are tightly packed, with very few vacuoles within the cells and no clefts at the DEJ.

A tissue sample with classic grade II changes is shown in Figure 2.4. There is extensive vacuolisation throughout the epidermis, causing the structure of the tissue to break down. Dyskeratotic bodies, which while not always present at grade II, are indicative of at least a grade II reaction when present. Characterised by highly stained pink cytoplasm and a condensed, small, dark nucleus, dyskeratotic bodies are difficult to identify for an inexperienced operator.

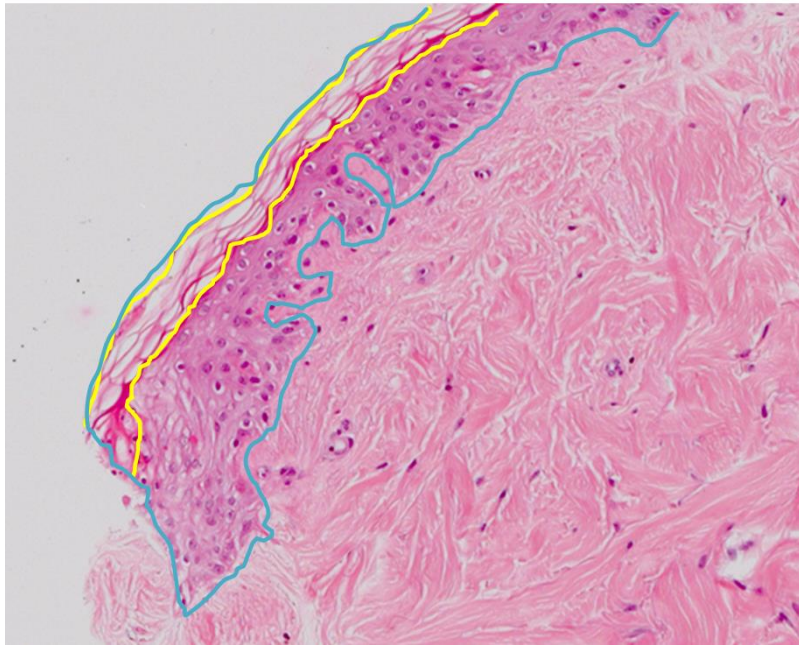


Figure 2.3 Section of H&E stained skin tissue showing changes associated with grade I damage. The whole epidermis is outlined in blue; the stratum corneum layer of the epidermis is outlined in yellow

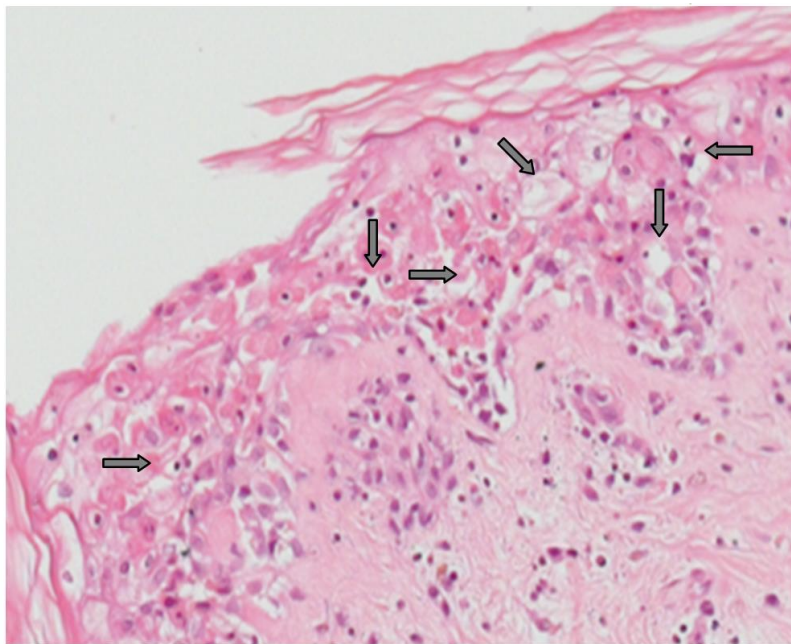
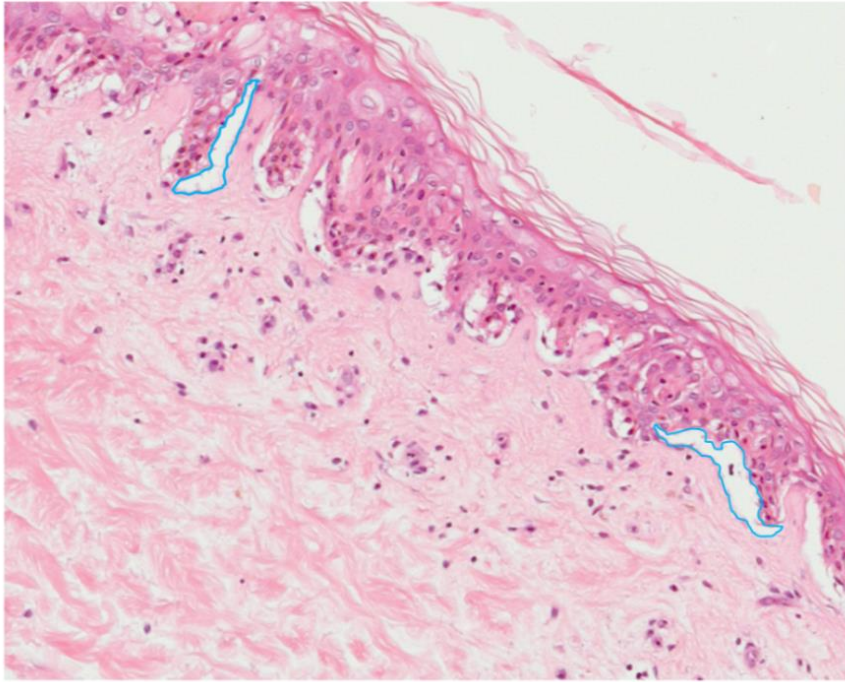


Figure 2.4 Section of H&E stained skin tissue showing changes associated with grade II damage. Arrows indicate some of the sites of vacuolisation.

Figure 2.5 is an image of an H&E stained tissue section showing grade III changes, typified by extensive cleft formation at the DEJ. Two of the clefts are outlined in blue.



*Figure 2.5 Section of H&E stained skin tissue showing grade III changes, typified by extensive cleft formation at the dermal-epidermal junction. Two such clefts have been outlined in blue to highlight.*

An H&E stained skin section showing grade IV changes is shown in Figure 2.6. The epidermis has completely separated from the dermis tissue in this skin section. It is of interest to note that the vacuolisation within the epidermis in the grade III and grade IV images does not always appear as severe as in some of the images of grade II reactions. For grade III and grade IV reactions, the presence of clefts at the DEJ takes precedence over the amount of vacuolisation when determining the final grading.

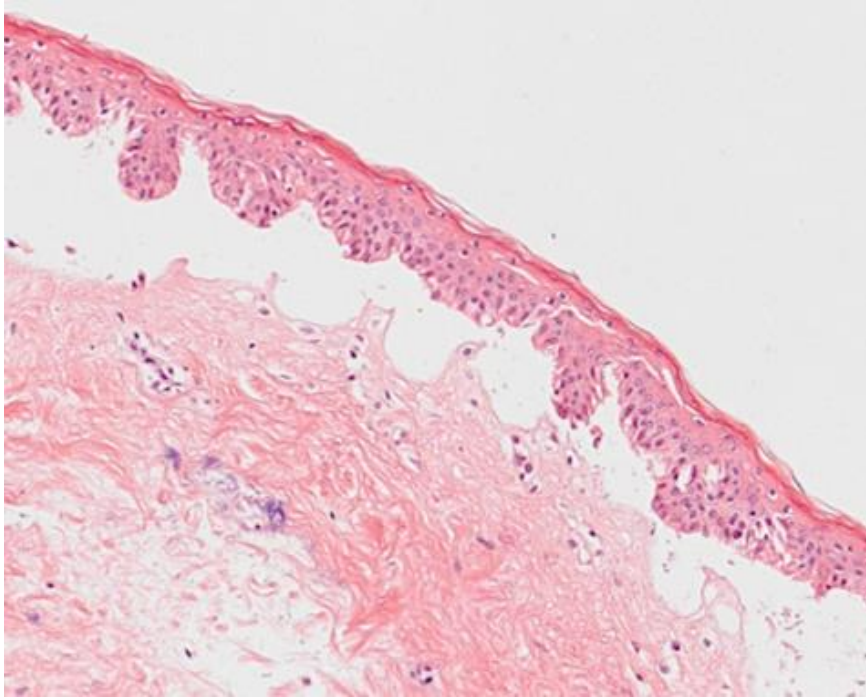


Figure 2.6 Section of H&E stained skin tissue showing grade IV changes, with complete separation of the epidermis from the dermis.

## 2.4 Digital Image Processing and Analysis Theory

Digital image processing and analysis involves the use of computer algorithms to create, process, communicate, display, analyse and extract information from images. Image processing and analysis involves many different processes, and they can be classified and described as low, mid and high level processes. Low-level processes are used to manipulate and process information within the image and include image transformations, noise reduction, and contrast enhancement; both the input and the output of these steps are images. The overall aim of low level processing in the context of an image classification process is to suppress image characteristics and features which are not relevant to the image classification, and enhance those features which aid discrimination between classes. Mid-level processes extract information from the image, using tasks such as segmentation and edge detection to partition the image into regions of interest, and feature extraction to isolate quantitative measurements that represent important image characteristics. High-level processes interpret the information extracted from images, and include image recognition and classification through to systems mimicking human vision and cognition. Low, mid and high level processes are all used within this research, and the rest of this chapter will give background theory

on the processes and operations used in the final algorithm. First an introduction to the terminology and conventions used when discussing images is provided to aid understanding of the image processing operations that follow.

### 2.4.1 Digital Image Representation

Images are representations of an analogue world and to enable computer processing they must be digitised. The image acquisition techniques described in Section 2.1.2 give an output which is a quantised representation of the sample that has been converted from an analogue to a digital signal. The real world image scene can be represented by a continuous 2-dimensional function  $f(x, y)$ , where  $x$  and  $y$  are coordinates and the amplitude of  $f$  is the intensity or grey level. The information in the analogue signal is digitised by the two operations of sampling and quantisation. Sampling is the discretisation of space, while quantisation is the discretisation of intensity and once the image is digitised,  $f$ ,  $x$  and  $y$  are all discrete quantities. Each finite element within an image is called a picture element or pixel. If an image has  $m$  pixels in the vertical direction and  $n$  pixels in the horizontal direction, the image can be described as a matrix,  $\mathbf{A}$ , of order  $m \times n$  with the individual matrix elements given by  $a_{ij}$ .

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \dots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

The density of the sampling grid with regards to spatial resolution, the number of different intensity levels chosen for the discretisation and the level of noise generated in the system are the most important factors in terms of defining the quality of the digital representation.

Once the digital image has been captured it must be imported into a suitable environment capable of performing the different image processing operations. A range of image processing systems are available:

- Adobe Photoshop is a graphics editing program which is easy to use and has many existing functions and processing operations accessed through a graphic

user interface. While there are many useful features, the program does not offer advanced feature selection or classification functions.

- ImageJ: This freeware has powerful image processing functionality to edit, process and enhance images, but like Adobe Photoshop, it is lacking the functionality for higher level processing such image classification.
- Specialised operating systems set up to implement algorithms written in C++, Python and Java programming languages. These systems offer maximum flexibility and functionality, but require advanced computer programming skills.

In this research project, MATLAB® (The Mathworks®, Nantick, Mass), a technical computing and programming language and environment has been used. MATLAB can be used for algorithm development, data analysis, visualisation, and numerical computation, and is particularly appropriate for image processing as it is designed around matrix manipulation. A number of toolboxes are available which offer tailored algorithms and tools in different application areas. Those used in the course of this research include the Image Processing Toolbox™, which provides algorithms and graphical tools for image processing, visualisation, analysis and algorithm development and the Statistics Toolbox™, which provides statistical and machine learning algorithms and tools to organise, analyse and model the data. This environment was chosen as it is flexible enough to allow tailored development of low, mid and high levels processes, offers a good range of pre-existing functions, and can be used to develop new algorithms. In addition, the programming language was more easily mastered within the time frame of a doctorate than other options such as C++.

### **2.4.2 Image Types**

A greyscale image measures light intensity, and can be represented by a two dimensional matrix with each pixel value proportional to the brightness. The least bright areas are represented by black, and the brightest, by white. In addition to brightness information, a colour image also contains information about colour. For the most common colour image representation, the RGB representation, the intensity matrix is three dimensional ( $x, y, z$ ) with three separate matrices (or



planes) representing the intensity of red, blue and green light. A representation of the RGB image is given in Figure 2.7, which shows the three 2D planes which contain information about the intensity of each colour at each location in the image.

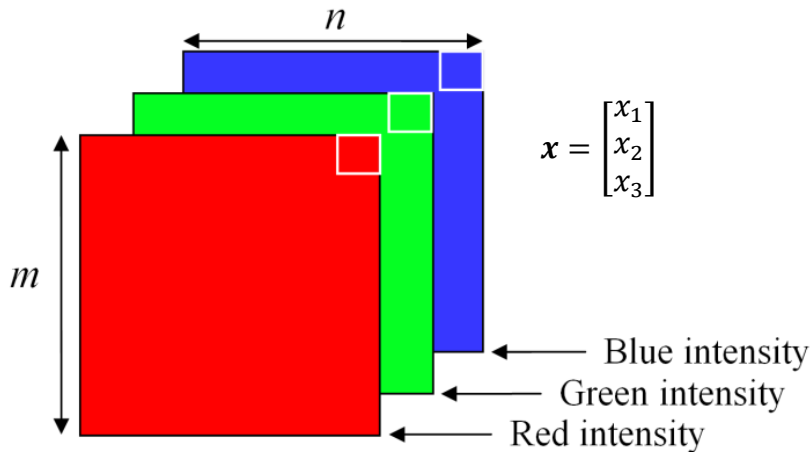


Figure 2.7 A representation of an RGB image of order,  $m \times n$ . A pixel at spatial coordinates  $(x, y)$  will be represented by the column vector,  $\mathbf{x}$ , shown in the figure.

### 2.4.3 Image Resolution: Spatial and Intensity

The spatial resolution determines the level of detail which can be attained from an image. When considering an  $m \times n$  matrix representing an image, the values of  $m$  and  $n$  tend to be powers of 2 (e.g., 128, 256, 512) to align with standard computer architecture. The resolution of an image is dependent on the magnitude of  $m$  and  $n$ ; if the magnitude is too small (e.g.,  $<32$ ) then the image appears to be a collection of squares and interpretation is lost. In Figure 2.15 the shape of the bike can be seen clearly in the first two images which have a spatial resolution  $412 \times 550$  and  $206 \times 275$ , it is just discernible in the  $41 \times 55$  image, but cannot be identified in the  $21 \times 28$  resolution image. The exact pixel number required is dependent on the complexity of the image, the image contrast, and the information required, but it should be a high enough resolution to observe the required detail but not so high that unnecessary computer power and storage space are required.



Figure 2.8 An RGB image displayed at four spatial resolutions, 412 x 550, 206 x 275, 41 x 55 and 21 x 28.

As with spatial resolution in an image, intensity resolution is usually measured in powers of 2 with an 8-bit greyscale image having 256 ( $2^8$ ) possible intensity levels (or grey-levels). Measurement noise or the display system used can limit resolution; computer screens are typically limited to the values of 0-255. The most common image representation is 1 byte per pixel (8-bit), which has intensity levels over the range 0-255. Images where intensity is limited to two values, 0 or 1, are known as binary images. These images are important in segmentation, morphological processing and classification of objects within an image as will be described later in this chapter.

#### **2.4.4 Colour Image Processing**

Visible light is made up of electromagnetic radiation within the 380nm – 780nm band of frequencies. Different coloured light corresponds to set wavelengths, and the colour of an object is the product of the wavelength spectrum of the incident light and the absorption and reflection properties of the object.



Colour theory is made more complex due to the fact that human colour perception and vision is limited by our optical system. The tristimulus theory of colour perception states human colour vision and perception is based on three types of colour receptors (called cones), which integrate over the red, green and blue parts of the spectrum. Humans see colour as variable combinations of the primary colours, red, green and blue. Combining two primary colours of light produces the secondary colours magenta (red and blue), cyan (green and blue) and yellow (red and green) and combining all three primaries produces white light. In contrast, coloured pigments work using a subtractive colour model, where a particular coloured pigment will absorb one primary colour of light and reflect the other two (e.g., absorb magenta, and transmit cyan and yellow). While the red, blue, green colour representation is the most familiar, there are a variety of other representations grouped under the term colourspaces.

#### **2.4.5 Colourspace Theory**

Colourspaces are mathematical models which describe the representation of colour using colour components. More specifically, they aid the description, specification, visualisation and transfer of information about colour, between people and machines, or between different machines. Colourspaces can exist in 2D, 3D or 4D, so a set of 2, 3 or 4 coordinates can be used to express any colour and its position within the colourspace. Most models are based on a three components system similar to the human colour system. There are a variety of colourspaces and the application under study and equipment being used often determines which is the most appropriate. Some of the colourspaces investigated in this research are described in the following section.

**Red, Green, Blue (RGB) Colourspace:** The RGB colourspace is based on human visual trichromatic theory and is an additive colourspace, meaning that from a start point of black (or darkness), colours are created by the addition of different coloured light. The colourspace uses a Cartesian coordinate system and can be represented as a cube, Figure 2.9. Images from colour cameras are RGB images, and consist of three  $m \times n$ , 1 byte-per-pixel images, each representing the intensity of red, green or blue. For an 8-bit image the coordinates (0, 0, 0) represent black,

(255, 255, 255) white and (0, 255, 255) cyan. The  $1 \times 3$  colour vector used to represent some of the colours commonly found in the research images are given in Figure 2.9.

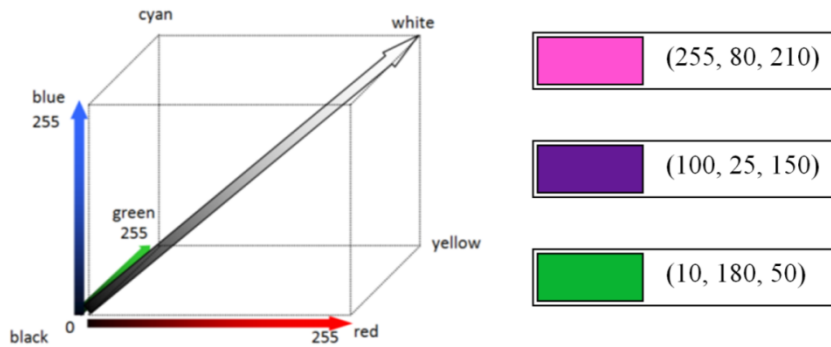


Figure 2.9 Representation of the RGB colourspace cube and representation in RGB colour model of three colours.

The RGB colourspace is not particularly intuitive or easy to interpret. Figure 2.10 shows the different colour channels of an RGB image; counter-intuitively the oranges and carrots have higher red intensity than objects which appear a pure red to the human eye (e.g., the red pepper) and the yellow objects also appear to have a higher green intensity than the more obviously green objects. The reasoning behind this is that orange colour contains a high intensity of red light, but the high intensity of green light combines with the red to produce orange.



Figure 2.10 RGB image and the three colour channel intensity images of the RGB colourspace

A number of colourspaces separate the lightness/ brightness component from the colour components. This enables the colour and greyscale information to be stored and transmitted at different resolutions and bandwidths, and in microscopy applications isolates staining and lighting differences from the lightness component. Many colourspaces use hue and saturation as the colour components and a third descriptor such as value, brightness, intensity or luminance. Examples include hue saturation luminance (HSL) model, where the RGB data is transformed to give an achromatic additive RGB signal and two differential chromatic signals (Garbay *et al.*, 1981) HSV (hue, saturation and value) which is described in more detail in the following section, HSL (hue, saturation, lightness) and HSI (hue, saturation, intensity) (Geladi and Grahn, 1996; The Mathworks, 2010).

**HSV Colourspace:** HSV represents colour in an intuitive manner, a hue can be selected and then the saturation and intensity modified. *Hue* represents the wavelength of the colour, *saturation* is the dominance/ purity of a hue in the final colour (0% being grey and 100% the pure colour) and *value* relates to the lightness of the colour. Figure 2.11 shows the HSV colour channels. While the yellow, orange and red objects are difficult to differentiate using hue, the purple cabbage and green vegetables exhibit greater contrast. Some of the different shades of green, red and orange can be identified using the saturation and value channels; however similar colours (e.g., carrots and oranges) are difficult to differentiate by eye. It may be possible to quantify these finer differences when examining the numbers, which is a benefit of computer based image analysis.



Figure 2.11 RGB image and the three colour channel intensity images of the HSV colourspace

Other colourspaces separate out the colour component using two or more measures of chromaticity (an objective colour specification independent of luminance which combine hue and saturation information), and a measure of brightness.

**YCbCr Colourspace:** Widely in digital video and photography systems, this is not a true colourspace, but rather a way of encoding RGB information differently using a linear transformation. The transform rotates the RGB reference axis so the diagonal of the cube (from the black to white corners in Figure 2.9) forms the main  $x$  axis, representing luminance ( $Y$ ). The two remaining axes ( $y$  and  $z$ ) contain the colour information. This approach is taken because humans are more sensitive to luminance than colour, and splitting the information in this way allows a greater emphasis to be given to the luminance component and bandwidth compression to be performed. The two colour channels are named blue-difference (Cb) and red difference (Cr). Figure 2.12 shows the YCbCr colour channels. The Cb channel highlights the contrast between the orange objects and the red and green objects, while the Cr enhances the contrast between the red/ orange objects and the green objects.



Figure 2.12 RGB image and the three colour channel intensity images of the YCbCr colour space

**$L^*a^*b^*$  Colourspace:** While the tristimulus theory implies that any colour can be created using a combination of red, green and blue, this is not the case for all visible colours. This issue was addressed by the Commission Internationale de l'Éclairage (CIE), who defined three standard primaries in 1931,  $X$ ,  $Y$  and  $Z$ , which can be combined to make any visible colour. The CIE has developed a number of additional colourspaces which aim to improve on the original CIE  $XYZ$ , including  $L^*a^*b^*$  which aims to provide a perceptually uniform colourspace. In a uniform colour scale, differences between points defined in the colourspace correspond to visual differences as perceived by a human. The colourspace is a non-linear system in contrast to those already discussed, and these non-linear relationships of the components are based on the logarithmic response in the human visual system.

$L^*a^*b^*$  has a luminance channel ( $L^*$ ) and two chromaticity channels for red-greenness ( $a^*$ ) and yellow-blueness ( $b^*$ ). The representation in Figure 2.13 shows a vertical axis which contains the luminance information, and two perpendicular axes to show the two colour channels.

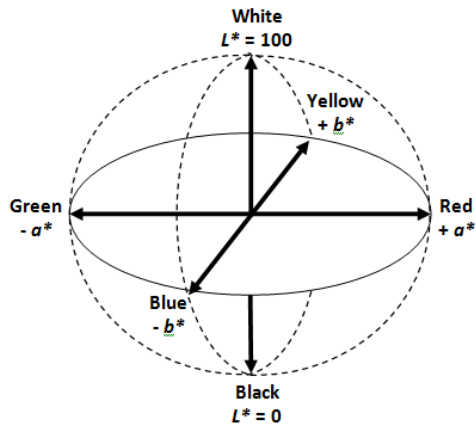


Figure 2.13 Representation of the  $L^*a^*b^*$  colour space. (The Mathworks, 2010).

A similar colour space to the  $L^*a^*b^*$  is the  $L\alpha\beta$ , obtained by applying principal components analysis to a set of natural images (Ruderman *et al.*, 1998). The first three principal components were found to represent luminance, yellow-blue and red-green. This finding demonstrates the usefulness of these two colour spaces when processing images from nature. Figure 2.14 shows the  $L^*a^*b^*$  colour channels. The  $a^*$  (red-green) channel enhances the contrast between red and green objects very effectively, while the  $b^*$  (yellow-blue) channel enhances the contrast between the purple cabbage and the yellow/orange objects. The  $L^*$  channel allows similar hues of differing luminance to be differentiated, e.g., green lettuce, broccoli and pepper.



Figure 2.14 RGB image and the three colour channel intensity images of the  $L^*a^*b^*$  colour space

It is sometimes desirable to convert colour images to greyscale, so that an overall measure of intensity can be made. An image can be converted from colour to greyscale in a number of ways, including:

- The lightness method, which averages the most prominent and least prominent colours:  $(\max(R, G, B) + \min(R, G, B)) / 2$ .
- The average method, which averages the values:  $(R + G + B) / 3$ .
- The luminosity method, which forms a weighted average to account for the fact that humans are more sensitive to green than other colours.

One approach for converting from RGB images (the effect of which is shown in Figure 2.15) is to apply the MATLAB function `rgb2gray`, which extracts the luminance information based on conversion from the RGB colour space to the National Television System Committee's YIQ colour space. The luminance (Y) is the greyscale signal used to display pictures on monochrome (black and white) televisions, and the other components carry the hue and saturation information. The underlying calculation (Eq.2.1) uses the weightings that define the luminance when converting from RGB to YIQ:

$$Y_{(i,j)} = (0.2989 \times R_{(i,j)}) + (0.5979 \times G_{(i,j)}) + (0.1140 \times B_{(i,j)}) \quad \text{Equation 2.1}$$

where **R** is the red colour channel image, **G** is the green colour channel image and **B** is the blue colour channel image. The weighting Figure 2.15 compares a RGB image and the equivalent greyscale image produced using `rgb2gray`. The figure shows how much information is lost during this process, the red pepper, lettuce and grapes which were easily distinguishable by colour contrast in the colour image all have a similar intensity in the greyscale image. To prevent this loss of information, it is sometimes better to consider the intensity of the colour planes separately.





Figure 2.15 Colour RGB image and the same image converted to greyscale

#### 2.4.6 Image Transforms

Image transforms are transformative operations performed on an image to change its representation. These may involve mathematical operations (simple image arithmetic, Fourier transform, Hough transform), histogram modification (equalisation and adaptive equalisation), or geometric operations (rotation, scaling). Simple image arithmetic operations involve the point wise combination of two images using basic arithmetic or logical operators including addition, subtraction, multiplication, division, logical AND and NOT. For the addition of two images, **A** and **B**, the  $(i,j)$ th element of the output image (**C**) is given by

$$C_{(i,j)} = A_{(i,j)} + B_{(i,j)} \quad \text{Equation 2.2}$$

If an operation, **H**, is carried out on an image matrix, **C**, the result can be described using standard matrix notation. If the operation produces a new image **D**, then  $D = H(C)$ . Typical low level operations of this form include shading correction, contrast enhancement, binarisation and noise reduction, and geometric transforms such as rotations, stretching and shrinking. If the operation produces a vector **d**, then  $d = H(C)$ . This type of mid-level operation is usually a data reduction and the vector, **d**, may be the grey-level histogram of the original image. If the operation produces a scalar *d*, then  $d = H(C)$ . This mid or high level operation is always a data reduction operation and the scalar output might be a key piece of information such as number of cells present in the image. It is likely that such an operation would be complex and involve a number of steps to move from the original image to the final scalar value, (Geladi and Grahn, 1996).



### 2.4.7 Pixel Neighbourhoods

The pixels surrounding a particular pixel define the pixel neighbourhood. This neighbourhood might consist of the pixels above, below and at either side of the pixel (4-connected) or it may also include the diagonal pixels (8-connected). In image to image operations, the pixel neighbourhood is an important difference between global, local and point operations. Point, local and global operations are summarised in Figure 2.16, for an input image **A** and an output image **B**, the pixel output at  $B_{(i,j)}$  is dependent only on the pixel at  $A_{(i,j)}$  for point operations, while for local operations the pixel output also depends on a neighbourhood of pixels around  $A_{(i,j)}$ . For global operations  $B_{(i,j)}$  is dependent on all pixels in image **A**, (Gonzalez and Woods, 2008).

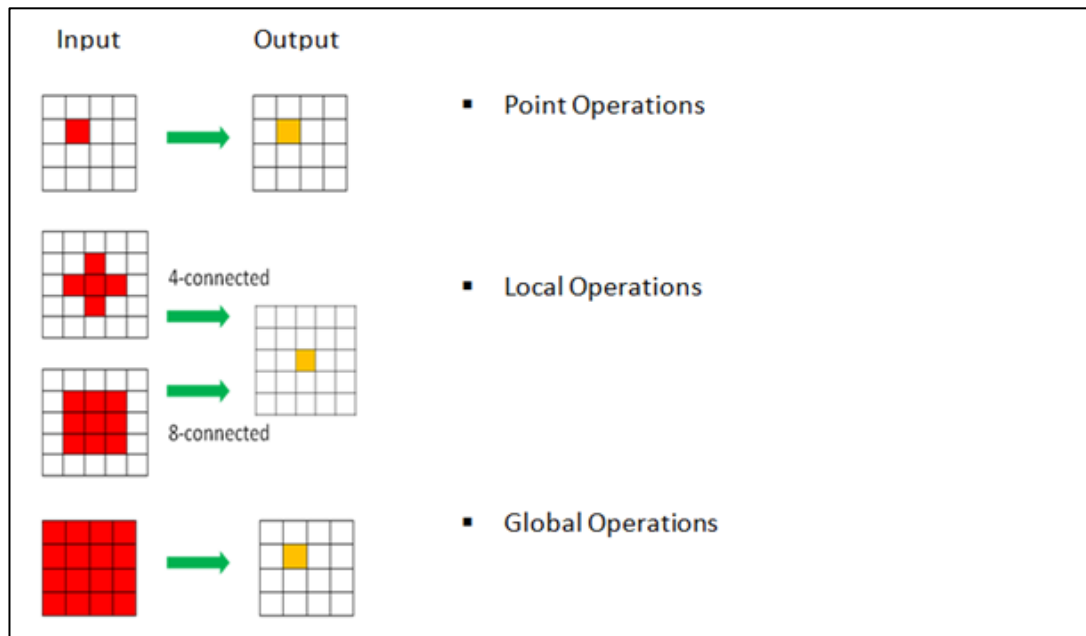


Figure 2.16 Diagram showing how pixel neighbourhoods relate to point, local and global image operations

### 2.4.8 Image Mean Filtering

Mean filtering can be used to smooth images by reducing the intensity variation within defined pixel neighbourhoods. When mean filtering is applied, each pixel in the input image is replaced with the average value of its neighbours. The technique uses a matrix of numbers of a smaller order than the input image, which is known as a kernel. The kernel defines the shape and size of the neighbourhood over which the average is calculated. An example of a square 3x3 kernel is shown in Figure 2.17. For a kernel of size  $m \times n$ , each element is given a value of  $1 / (m \times n)$  which is then

used to perform a weighted multiplication. The origin pixel, shaded in the figure, defines the position of the pixel being affected in the output image and can be any position within the kernel, but is usually placed at the centre.

1/9	1/9	1/9
1/9	1/9	1/9
1/9	1/9	1/9

Figure 2.17 Square 3x3 Kernel for Mean Filtering, with the origin placed in the centre.

Convolution is the process by which the two matrices (the input image and kernel) are multiplied together. The kernel slides over the input image into each position in which the kernel fits in its entirety. In each position, each kernel pixel (or weighting) is multiplied with the underlying input image pixel, then the sum of all these pixel products are added together and used to define the pixel value in an output image at the origin position of the kernel.

#### 2.4.9 Contrast enhancement methods

Formally, the contrast  $c$  between two intensity values  $x_1$  and  $x_2$  can be defined as the absolute value of their difference:  $c(x_1, x_2) = |x_1 - x_2|$ . Low contrast images have intensity values across a narrow distribution (i.e., mainly bright, mainly mid-tone, or mainly dark). Contrast enhancement adjusts the relative brightness and darkness in an image to improve the visibility of certain objects or features. Grey level histograms can be used to illustrate image contrast as they show the frequency distribution of pixels with regards to pixel intensity. Contrast enhancement can be achieved using remapping, a process by which the grey levels in the original image are mapped onto new values using a transform mapping function. A function  $g$  can be used to generate a contrast enhanced image,  $\mathbf{B}$ , from image,  $\mathbf{A}$ :

$$\mathbf{B}_{(i,j)} = g(\mathbf{A}_{(i,j)}),$$

Equation 2.3

for  $i = 0, \dots, n - 1, j = 0, \dots, m - 1$

The transform determines how the intensity distribution is remapped, as shown in Figure 2.18. In the figure a sigmoidal transfer function is shown which can be used to map any input grey level (on the x axis) to a new grey level on the y axis. A variety of linear and nonlinear fixed functional forms including log,  $n^{\text{th}}$  root, linear,  $n^{\text{th}}$  power, inverse log and gamma correction can be used to transform the original grey levels to their new values. Alternatively adaptive transforms including histogram equalisation or histogram matching can be used.

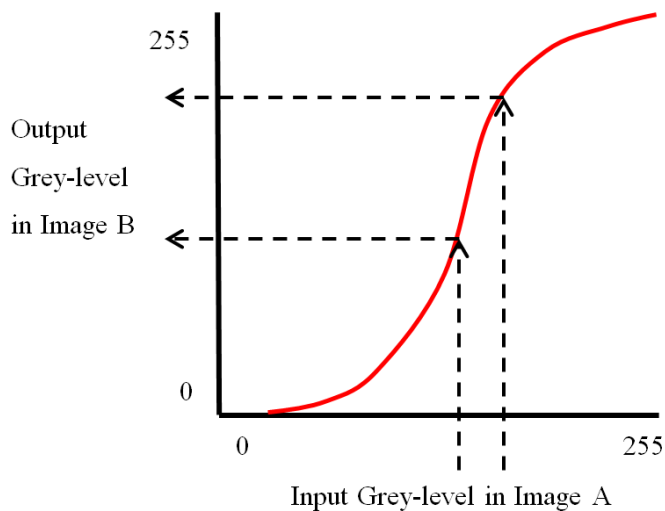


Figure 2.18 Remapping grey-levels using a transform function

One of the simplest transforms used to enhance contrast is a linear transform which stretches the grey-level values to create a histogram spanning the full dynamic range (0 255). A low contrast image,  $\mathbf{G}$ , is transformed to a high contrast image,  $\mathbf{G}'$ , by remapping the grey levels. The lowest grey level in  $\mathbf{G}$ ,  $GL_{\min}$ , is mapped to a new minimum grey level  $GL'_{\min}$ , and the highest grey level in  $\mathbf{G}$ ,  $GL_{\max}$ , to a new maximum grey level  $GL'_{\max}$ . The linear transform is given by:

$$\mathbf{G}'_{i,j} = INT \left\{ \frac{GL'_{\max} - GL'_{\min}}{GL_{\max} - GL_{\min}} [\mathbf{G}_{i,j} - GL_{\min}] + GL'_{\min} \right\} \quad \text{Equation 2.4}$$

where the  $INT$  function returns the integer value.  $GL_{\min}$  and  $GL_{\max}$  can be replaced with points  $P_{\min}$  and  $P_{\max}$  which lie within the grey level histogram. Figure 2.19 illustrates the effect of the linear transform on a grey level histogram and shows that the original intensity distribution is approximately preserved in this type of contrast enhancement.

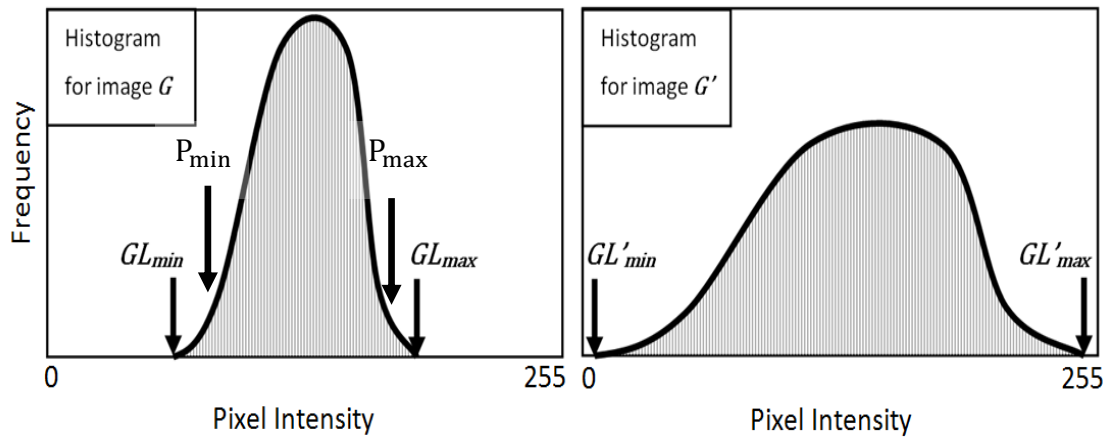


Figure 2.19 Diagram illustrating the effect of contrast stretching the intensity histogram, using a linear transform. Two sets of potential high and low values for the remapping are show.

The penetration points can be determined using a cumulative percentage histogram which shows the percentage of pixels between zero and each grey-level. Figure 2.20 shows the selection of  $P_{min}$  and  $P_{max}$  from a cumulative percentage histogram. The proportion of pixels excluded can be chosen based on the application, for example the lowest and highest 1% of intensities could be mapped to 0 or 255 respectively, meaning resolution at the extremes of intensity is lost. This is a useful technique if the intensity band of interest is at the mid grey level as opposed to the extreme grey levels.

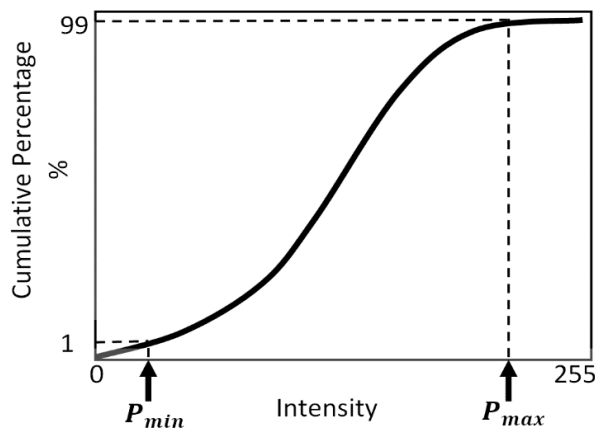


Figure 2.20 Diagram illustrating how penetration points can be selected using a cumulative percentage histogram.

An alternative method to enhance resolution in a set band of intensities is to select penetration points bordering a particular band of intensities, as shown in Figure 2.21. This has the effect of removing the high intensity “shoulder” from the

distribution (which may represent a particular region of the image which is not of interest, such as background), and allowing the rest of the distribution to be contrast enhanced by stretching the remaining pixels out to intensities of 0-255.

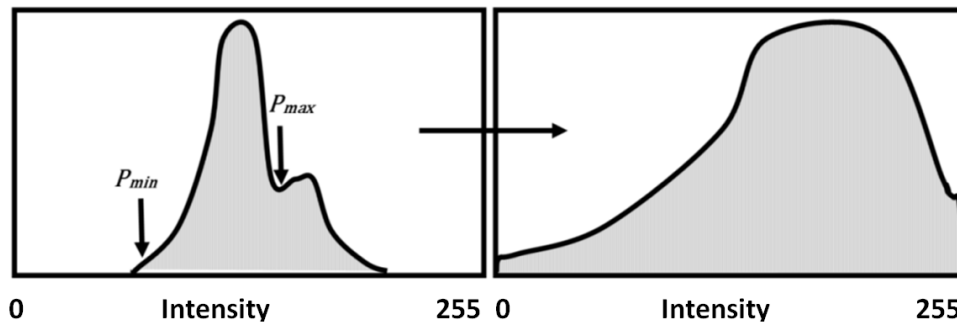


Figure 2.21 Diagram illustrating the effect of selecting a set band of intensities for a linear remapping.

More complex histogram equalisation methods are available which modify the dynamic range and enhance the contrast in images by adjusting the shape of the intensity histogram. In some forms of histogram equalisation (e.g., the [histeq](#) function in MATLAB), the pixel intensities from an image histogram are mapped to new values using a non-linear transfer function, such that the resulting new image has a uniform distribution of intensities and a flat intensity histogram. In contrast, adaptive histogram equalisation defines a pixel neighbourhood then derives a transfer function which will ensure each pixel is mapped to any new distribution specified. The [adapthisteq](#) function in MATLAB uses this type of process. The process can be limited by specifying a certain range for the final mapping which may help to avoid amplifying noise. Figure 2.22 shows the effect of the [histeq](#) and [adapthisteq](#) functions on a greyscale image and its histogram. The adaptive histogram equalisation preserves the “sense” in the original image better, with the background remaining white and the contrast enhancement showing the tissue in more detail. Performing the contrast enhancement on separate tiles means that variations in intensity across the image (e.g., due to lighting inconsistencies) do not create problems. In some cases the non-adaptive mapping to a flat intensity histogram can reveal hidden features that were not obvious in the original image. The selection of an appropriate method for contrast enhancement is dependent on the shape and dynamic range of the original image, the regions of interest, and the variation in contrast across the image.

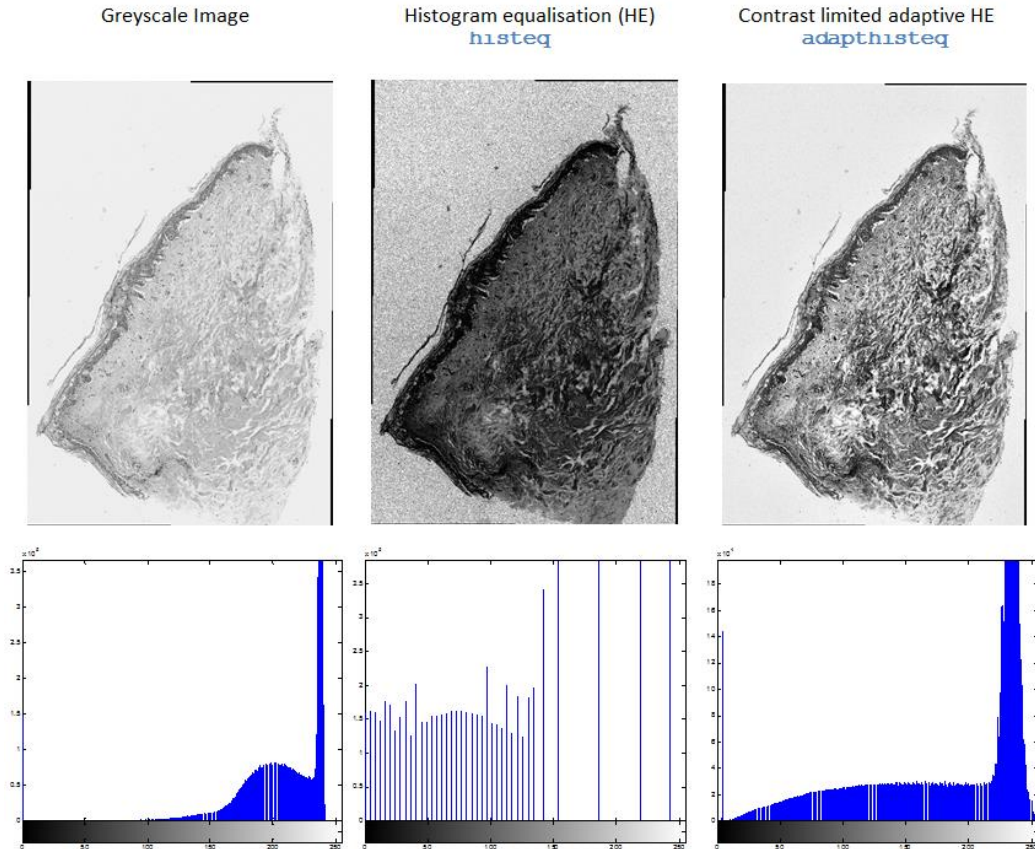


Figure 2.22 The effect of normal and adaptive histogram equalisation on a greyscale image and its intensity histogram.

#### 2.4.10 Set and logical operations

Set theory is a mathematical field that studies sets of objects, and it is an essential tool when working with binary images and identifying objects in an image. Each member of a set is referred to as an element of the set. In image processing, the set can represent the whole image, or alternatively an object or feature in the image, in which case the pixel coordinates of all pixels in the object will be the elements of the set. Typically, sets are represented by uppercase letters, such as  $A$ ,  $B$ , and  $C$ , and set members by equivalent lowercase letters, such as  $a$ ,  $b$ , and  $c$ . If an element  $a$  is a member of a set  $A$ , then  $a \in A$ , and if it is not a member of  $A$ , then  $a \notin A$ . For example, if  $E$  is set comprising all even numbers, the set,  $F$ , of all even numbers less than 100 can be denoted as:

$$F = \{f \in E \mid f < 100\} \quad \text{Equation 2.5}$$

If all the members of set  $A$  are also members of set  $B$ , then  $A$  is a subset of  $B$ , denoted as  $A \subseteq B$ . A set with no elements or members is known as an empty or null set and is referred to using the symbol  $\emptyset$ . Binary operations can be carried out on sets and those that are useful in image analysis were defined by Gonzalez and Woods (2008):

- **Union** of sets, e.g.,  $A \cup B$ , is the set of all objects which are members of  $A$ , or  $B$  or both. Sets  $A$  and  $B$  are shown in Figure 2.23a, the union is labelled  $C$  in Figure 2.23b.
- **Intersection** of sets,  $A \cap B$ , is the set of all objects which are members of both  $A$  and  $B$ . The intersection of sets  $A$  and  $B$  is labelled  $D$  in Figure 2.23c.
- **Complement** of a set,  $A$ , is all the elements ( $a$ ) in a given object universe (e.g., the whole image) that are not in set  $A$ . It is defined as  $A^c = \{a|a \notin A\}$ , and the complement of set  $A$  is labelled  $A^c$  in Figure 2.23d.
- **Reflection (transposition)** of set  $A$ ,  $\hat{A}$  is the reflection of all elements of  $B$  about the origin. If  $A = \hat{A}$ , the set is symmetric. It is denoted  $\hat{A} = \{-a|a \in A\}$  and is labelled  $E$  in Figure 2.23e.
- **Translation** of set  $A$ ,  $(A)_z$  is the translation of the origin of  $A$  to point  $z$ . It is defined as  $(A)_z = \{a|a = a + z, \text{ for } a \in A\}$  and is labelled  $F$  in Figure 2.23.

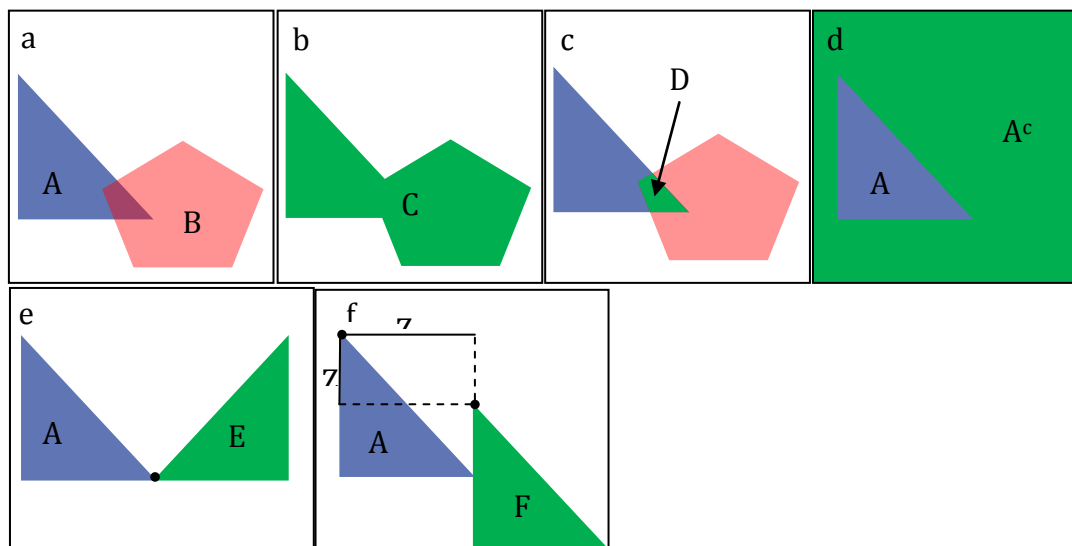


Figure 2.23 Diagrammatic representation of set theory, showing union (b), intersection (c), complement (d), reflection (e) and translation (f).

The low level processes described in the previous sections include image transforms, image filtering, contrast enhancement methods and logical operations. These processes can either be used to pre-process an image before segmentation or as part of the segmentation step itself. The process of segmentation is described in the following section.

#### **2.4.11 Segmentation**

After image pre-processing, the first step in image analysis is typically segmentation, a process in which an image is partitioned into constituent parts or objects, comprising sets of pixels. This key step marks the transition from analysing the image at pixel level to analysing the objects comprising sets of contiguous pixels. Identifying objects in a complex image can be made easier by converting the image to a logical or binary form where the pixels of the objects of interest are labelled with a 1, and the remaining pixels are labelled with a 0. Once the objects have been identified, a variety of measurements including area, position or texture can be extracted and the resulting data analysed statistically and used for image classification.

Segmentation approaches are generally sub-divided into region and contour based approaches. Region-based methods create sets using pixel or neighbourhood properties such as colour, intensity, location or texture. Examples of region based methods are thresholding, region growing, and region splitting/ merging. Contour based approaches look for discontinuities in an image, using edge or boundary detection with local processing techniques, global approaches or more complex methods such as active contours. Segmentation is simplest either when pixels in a particular object or regions of interest have similar greyscale values (in which case a region based approach is most appropriate), or when neighbouring pixels in different objects have dissimilar values (in which case a contour based method will be most suitable).

In histopathology, segmentation is usually used to identify the presence, number, distribution, size and morphology of diagnostic features including tumours, specific cells, nuclei and glands. The accurate identification of these structures is an



essential first step in the diagnosis, staging and grading of disease using image analysis.

#### 2.4.12 Thresholding – A region based approach to segmentation

Thresholding provides one approach to identifying regions of interest in an image. In the underlying process, the intensity or colour characteristics of pixels are used to classify them as either background or foreground, although multiple thresholds can be used to create more complex segmentations. The input for thresholding is typically a colour or greyscale image with the output being a binary image in which pixel intensities are assigned as 0 (background) or 1 (foreground). In the simplest form of thresholding a single intensity threshold is set, with pixels above the threshold in the input image assigned a 1 and displayed as white and those below assigned a 0 and displayed as black in the output image. For the conversion of an  $m \times n$  8-bit image  $A$  with values from 0-255, to a  $m \times n$  binary image,  $BW$ , using a threshold value of 100 then:

$$BW_{(i,j)} = \begin{cases} 1 & \text{if } A_{(i,j)} > 100 \\ 0 & \text{else } A_{(i,j)} \leq 100 \end{cases} \quad \text{Equation 2.6}$$

More complex thresholding techniques specify pixels within a certain intensity band, specify multiple thresholds or bands for different colour channels, or retain colour information in the feature regions rather than changing them to black or white. Figure 2.24 is an illustration of how thresholds might be selected using intensity distribution histograms, the arrows represent potential points for thresholds or thresholding bands to be set to create a segmentation (Gonzalez and Woods, 2008). In Figure 2.24, a simple selection method has been illustrated, whereby each separate peak in the intensity histogram is assumed to represent a region of interest that should be segmented. However this may not be the case and so often more sophisticated methods are required.

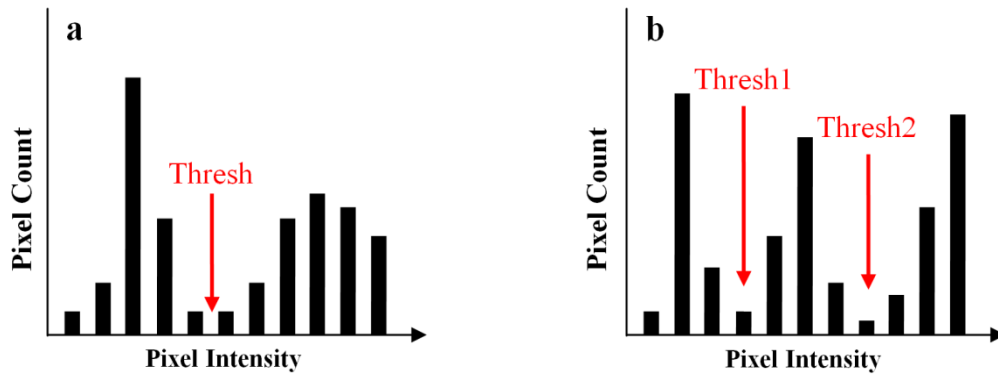


Figure 2.24 Representation of typical intensity histogram distributions and suitable threshold or threshold bands

While thresholds can be set manually using application knowledge, automating the process is quicker and less subjective. In their survey of image thresholding methods, Sezgin and Sankur (2004) defined six algorithm categories based on the information used to create the threshold: histogram shape, measurement space clustering, histogram entropy, image attribute detail, spatial data and local characteristics. Methods for selecting the threshold automatically include:

- The **mean** method: The mean grey level in the image is used as a threshold. This method is most useful as a first guess threshold used as the starting point for some of the other methods.
- The **intermeans algorithm (or IsoData)**: Proposed by Ridler and Calvard (1978) and Trussell (1979), this method starts with an estimate of the threshold value from which the mean values of pixels in each set (foreground and background) are made. In an iterative process of mean determination and incremental threshold change, the threshold is repositioned to lie exactly half way between the two means
- The **intermodes algorithm**: This approach assumes a bimodal distribution. The histogram is iteratively smoothed until two local maxima remain,  $j$  and  $k$ . The threshold is then calculated as  $(j + k)/2$  (Prewitt and Mendelsohn, 1966).
- The **Otsu thresholding** method: Proposed by Otsu (1979), the method is a point-dependent global thresholding technique that can be applied to bimodal grey-level histograms. The clustering algorithm maximises the separation of the foreground and background pixel sets by searching for a threshold which minimises the intra-class variance. To do this, the variance of the foreground

set (weighted according to the number of pixels in this set) is added to the variance of the background set (weighted according to pixel number). While this approach can be used to screen all possible thresholds until the weighted inter class variance is minimised, a faster approach exploits the fact that the threshold with the minimum intra-class variance also has the maximum inter-class variance. As the inter-class variance is much quicker to calculate than the intra-class variance this approach is most commonly used.

The relative merits and the reasons for the choice of method used in this research are discussed in Chapter 3. The final image processing technique to be introduced is mathematical morphology, which can be performed after segmentation to modify the segmented regions within the image and extract information from them.

#### **2.4.13 Connectivity and mathematical morphology**

First described by Georges Matheron (1975), Mathematical Morphology (MM) is a theoretical approach to the analysis of geometric structures and encompasses a range of operations utilising set theory. MM processing can be applied to greyscale or binary images, but discussion of it in this thesis focusses on its use for processing binary images. MM operations are particularly useful for object recognition in image analysis, as the operations can preserve the key shape characteristics of an image while removing uninformative variations in intensity (Haralick *et al.*, 1987). By distinguishing between meaningful shape information and irrelevant shape information, this approach mimics human visual perception.

Morphological processing operations require the interaction of an input image pixel set with an external pixel set in the form of a structuring element (SE). A SE is a matrix of 0's and 1's, generally much smaller than the image being processed. The 1's define the shape and size of a pixel neighbourhood. An example of a disk and diamond shaped SE is shown in Figure 2.25. In MM and binary processing, objects are contiguous regions of foreground pixels with a value of one. Objects within the input image are analysed using an appropriately shaped SE, and the output image pixels are based on a comparison of the corresponding pixel in the input image with its neighbourhood, as defined by the SE. By varying the size and

shape of the SE, it is possible to extract shape information for different parts of the image. The shape of the SE is chosen based on the shape of the regions or features of interest. For example, when MM is applied to images of circuit boards, horizontal and vertical linear SEs tend to be used to help identify the circuits, while biological applications attempting to identify cells are more likely to use a disk shaped SE.

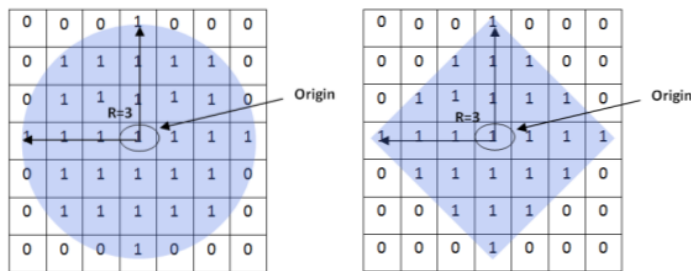


Figure 2.25 Structuring element matrices which form the basis of disk and diamond shaped structuring elements with a size (radius) of 3.

The most basic MM operations are dilation and erosion; these two operations are used both on their own and as the basis of more complex operations including opening and closing. An example of the effect erosion and dilation have on binary objects is shown in Figure 2.26. The 3 x 3 SE is moved sequentially across the original image, and the pixels in the 3 x 3 neighbourhood are averaged to determine the output pixel in the new image. Dilation tends to make objects bigger, smooth uneven edges and bridge gaps whereas erosion tends to make objects smaller, remove protuberances and break bridges. The definitions and mathematical notation in this thesis are based on those used by Gonzalez and Woods (2008).

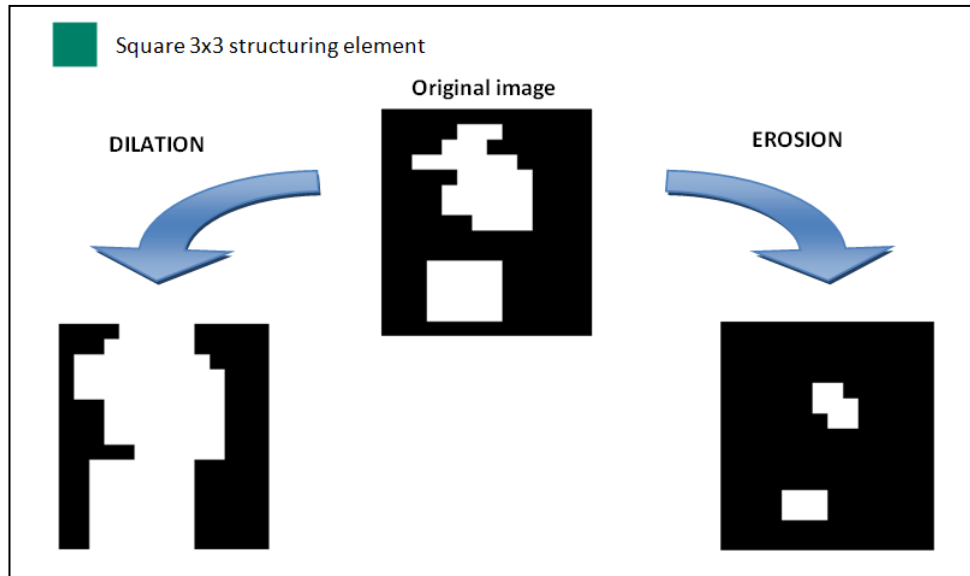


Figure 2.26 Diagram showing the effect of erosion and dilation with a 3x3 structuring element on a simple binary image.

### Dilation and Erosion

In *dilation*, pixels are added to the boundaries of objects using a rule stating that the output pixel is set to 1 if any pixels in the input neighbourhood are 1. In mathematical terms, if  $A$  is a set of pixels in the input image and  $B$  a SE, then  $\hat{B}$  is a reflection of  $B$  about its origin and dilation,  $A \oplus B$ , is the set of all pixel locations  $z$ , where the reflected SE overlaps with pixels in  $A$  with a value of 1 when translated to  $z$ :

$$A \oplus B = \{z | (\hat{B})_z \cap A \neq \emptyset\} \quad \text{Equation 2.7}$$

*Erosion* is the opposite of dilation; pixels are removed from the outside of objects based on a rule stating the output pixel is set to 0 if any pixels in the input neighbourhood are 0. The erosion,  $A \ominus B$ , of set  $A$  by structuring element  $B$ , is the set of all pixel locations  $z$  such that  $B$  overlaps with pixels in  $A$  with a value of 1 when translated to  $z$ :

$$A \ominus B = \{a | (B)_a \subset A\} \quad \text{Equation 2.8}$$

Erosion and dilation are dual with respect to reflection and complementation, more specifically, erosion of the image background is equivalent to dilation of the image foreground, i.e.,  $(A \ominus B)^c = A^c \oplus \hat{B}$ .

### Opening and Closing

Morphological opening can be used to remove small objects, smooth object contours, remove thin protrusions from and bridges between objects in a similar manner to erosion, however it also approximately preserves object size. The opening operation is erosion of a set  $A$  by a structuring element  $B$  followed by dilation of the resulting set by  $B$ . Opening, denoted  $A \circ B$  is given as:

$$A \circ B = (A \ominus B) \oplus B \quad \text{Equation 2.9}$$

Morphological closing can be used to smooth object contours, fuse narrow breaks and thin gulfs, and remove small holes in objects like dilation, while approximately preserving object size. The closing operation is dilation of a set  $A$  by a structuring element  $B$ , followed by erosion of the resulting set by  $B$ . Closing, denoted  $A \cdot B$  can be given as:

$$A \cdot B = (A \oplus B) \ominus B \quad \text{Equation 2.10}$$

### Hit-or-Miss Transform

This transform for shape detection aims to identify the location of a shape,  $X$  in a set  $A$ . It can be used to locate isolated foreground pixels, or endpoints and contour points of foreground objects. If  $B$  is the set including  $X$  and its background, the match of  $B$  in  $A$ , denoted  $X \odot B$  is given as:

$$A \odot B = (A \ominus B_1) - (A \oplus \widehat{B_2}) \quad \text{Equation 2.11}$$

where  $B_1$  is the set of elements of  $B$  associated with the object,  $X$ , and  $B_2$  is the set of elements of  $B$  associated with the corresponding background.

### Locating Boundaries/ Object Perimeters

The perimeter or boundary of a set of connected pixels  $A$ , denoted  $\beta(A)$  can be found by eroding  $A$  by  $B$ , then calculating the set difference of the original and eroded  $A$ . This is given as:

$$\beta(A) = A - (A \ominus B) \quad \text{Equation 2.12}$$

### Locating objects/ region filling

This process takes known point inside the boundary of an object, and then fills the object region with 1's. Let  $Y$  represent a connected component (object) in set  $A$ . The process of selecting the object begins with a zero array  $X_0$  the same order as  $A$ , with a point  $p$  at some location inside the boundary of the object. An iterative procedure can then be applied to grow the region until all the elements of  $Y$  have been found. The iteration applied is given as:

$$X_k = (X_{k-1} \oplus B) \cap A, \quad k = 1, 2, 3, \dots, \quad \text{Equation 2.13}$$

where  $B$  is the SE. The intersection with  $A$  limits the dilation so it does not extend beyond the region of interest. The iteration process continues until  $X_k = X_{k-1}$  at which point  $Y = X_k$ . The process can be applied to multiple objects, as long as a pixel location within each object is known.

### Filling Holes

A hole can be defined as a connected region of background pixels surrounded by foreground pixels. If  $A$  is the set of pixels surrounding a hole, the task is to fill the holes with foreground pixels. The process of filling a hole begins with a zero array  $X_0$  the same size as  $A$ , with a point  $p$  at some location inside the boundary of the hole. The following procedure is then used to fill the hole with foreground pixels:

$$X_k = (X_{k-1} \oplus B) \cap A^c, \quad k = 1, 2, 3, \dots, \quad \text{Equation 2.14}$$

where  $B$  is the SE. The intersection with the complement of  $A$ , limits the dilation so it is always inside the region of interest. The iteration process continues until  $X_k = X_{k-1}$ . To fill a hole within an object already identified,  $A$  is replaced with  $Y$  determined as described in the previous section.

### Thickening

Thickening is used to thicken and grow concavities in objects without them merging completely. The thickening of set  $A$  using SE  $B$  can be defined as a hit-or-miss transform and union:

$$A \odot B = A \cup (A \otimes B) \quad \text{Equation 2.15}$$

The additional pixels identified by the hit-or-miss operation are added to the original pixels in the object. The operation can also be defined as a sequential operation:

$$A \odot \{B\} = (((A \otimes B^1) \otimes B^2) \dots) \otimes B^n \quad \text{Equation 2.16}$$

where a series of rotated SEs ( $B^1, B^2, \dots, B^n$ ) is used to carry out the procedure.

## 2.5 Machine Learning, Feature Selection and Classification Theory

In machine learning a computer or automated system learns to carry out a task or solve a problem from a series of data based examples, as opposed to from a set of programmed rules. It is generally used to describe processes which aim to reproduce human learning. The data examples (known as *observations*), are analysed and a set of properties extracted and used as inputs for the learning machine. These inputs are known called *features* (or explanatory variables), and can be binary (e.g., true, false), categorical (e.g., blonde, brunette and red), ordinal (e.g., low, medium and high), integer-valued (e.g., number of words in an email) or real-valued (e.g., cholesterol level in the blood). Once the features are extracted they are used to adjust internal parameters of a predictive model so that the model captures the underlying patterns and can begin to make accurate predictions.

Further background theory on classification is provided in section 2.5.2. Prior to this some background is given on methods for extracting the features used in the classification algorithm.

### 2.5.1 Feature Extraction

Features are measurements or attributes which capture important information representing the differences and similarities between input observations in a classification system. A set of feature measurements can be stored in a feature vector, which represents the information in a 3 dimensional colour image (e.g., a RGB image with 3 colour channels) in a 1 dimensional list of numbers. Feature extraction aims to capture the information required to develop an accurate classifier within a significantly smaller dataset than the original image. For



example, a typical RGB image used as an input in this research project is of the order of 2666 x 3863 x 3 pixels, and hence will be represented by 30,896,274 values. The extraction of pertinent high level features could allow the image to be represented by as few as five numbers in the final classification model. Feature extraction is one form of dimensionality reduction used in image analysis, other forms will be discussed in section 2.5.6.

To be useful for image classification, the features should vary between classes/ categories, remain as resistant as possible to other variation in the image including lighting, rotation and staining and also be detectable using an automated process.

Features used in histological image analysis can be categorised as:

- Morphometric features, e.g., size, shape
- Intensity/ colour features, e.g., hue, intensity, saturation, optical density
- Texture, e.g., co-occurrence matrix, Gabor, energy, fractal and wavelet.
- Architectural or graph based spatial features, e.g., node number, clustering coefficient, spectral radius

In histopathology, morphometric features are often based on visual attributes used by clinicians and histopathologists to grade or classify disease. These features are usually object based and associated with the shape and size of tissue structures such as glands, tumours, whole cells or cellular components including nuclei or cytoplasm. Typically these structures or cellular components have been segmented in the previous stage of image processing. In addition to size and shape based features, measurements such as intensity, optical density or hue can be used to determine and quantify specific colourimetric or immunohistochemical stains which highlight biochemical and structural changes in the cell and tissue. Features such as texture, intensity and colour can also be extracted from a limited set of pixels representing an object (such as a cell or nucleus) in which case they can also be defined as object level features.

Features can also be extracted which attempt to represent or describe the global image texture. These features capture repeating patterns of variation in image intensity, capturing information about the spatial distribution of intensity levels.

These statistical methods analyse the spatial distribution of grey values, by computing local features at each point in the image, and deriving statistics from the local feature distributions.

First order statistical features estimate properties (e.g., mean, median, standard deviation) of pixels in the region of interest and can be calculated using the image histogram. First-order statistics ignore the spatial interaction between image pixels whereas second- and higher-order statistics estimate properties of two or more pixel values occurring at specific locations relative to each other. Second order statistical features are calculated using a grey-level co-occurrence matrix (GLCM). This form of statistical texture description specifies the grey-level spatial dependencies within a texture, and quantifies the distribution of co-occurring grey-level values at various angles and distances (Haralick *et al.*, 1973).

For an image,  $I$ , the GLCM,  $\mathbf{P}_{(i,j)}$ , is defined by counting all pairs of pixels with grey-levels  $i$  and  $j$ , which are separated by a distance,  $k$ , in direction,  $d$ . The normalised GLCM,  $\mathbf{PN}_{(i,j)}$ , is created by dividing each element in  $\mathbf{P}$  by  $N$ , the total number of co-occurrence pairs in  $\mathbf{P}$ :

$$N = \sum_i \sum_j \mathbf{P}_{d(i,j)}$$

Equation 2.17

$$\mathbf{PN}_{d(i,j)} = \frac{1}{N} \mathbf{P}_{d(i,j)} \cdot \mathbf{PN}_d$$

The GLCM can be scaled to include different numbers of intensity levels by increasing  $N$ . The offset can also be changed, by altering  $k$  and  $d$ . The spatial statistics calculated from the GLCM can be used to compute various features which capture textural information about the image.

Graph based spatial and architectural methods for extracting features include Voronoi Tessellation (Toussaint, 1980) and Delauney Triangulation (S. Doyle *et al.*, 2007). These type of features are used to quantify the spatial arrangement of specific features, usually cell nuclei in histopathology. In Voronoi Tessellation (VT), a set of nodes is identified (e.g., centroids of nuclei) and the VT creates polygonal

cells around the nodes such that all pixels within a given cell are closer to the cell node than any other node in the image. The VT is often used in combination with Delaunay Triangulation, which is a commonly used triangulation algorithm. From the set of all possible triangles, a triangle is accepted if its circumcircle contains no other nodes besides the triangle vertices.

The main feature types used in this research were object level features and texture features. The object level features provide information about the features of interest such as clefts and vacuoles, texture features give a more general measure of tissue structure in the epidermis and at the DEJ. The reasons for these choices are discussed in Chapter 3. The two feature types are described below in more detail.

### **Object level features**

Once histological objects such as clefts and vacuoles have been identified within the skin explant images, morphological features can be extracted to describe them. Many of these features can be applied to various regions of interest including the epidermis, individual clefts and vacuoles. The specific features investigated in this research are listed below:

- *Area* – The total number of pixels in the region of interest.
- *Bounding box* – The smallest rectangle which can contain the region of interest. The bounding box of an object is plotted over the binary mask of an object as illustrated in Figure 2.27.
- *Eccentricity* – The eccentricity of an ellipse with the same second moments as the region of interest. It is calculated as the ratio of the distance between the foci of the ellipse, and the ellipse's major axis length. An ellipse has two foci, and is the locus of points such that the sum of the distance to each focus is constant. Examples of the equivalent ellipses are shown in Figure 2.28.
- *Extent* – This is the *area* divided by the area of the bounding box
- *Major Axis Length* – The length (in pixels) of the major axis of the ellipse with the same normalised second central moment as the region of interest. Marked with the blue arrow in Figure 2.28.

- *Minor Axis Length* – The length (in pixels) of the minor axis of the ellipse with the same normalised second central moment as the region of interest. Marked with the green arrow in Figure 2.28.
- *Perimeter* – The distance around the outer boundary of the region of interest.



Figure 2.27 An example of a bounding box (in red) for a specific object

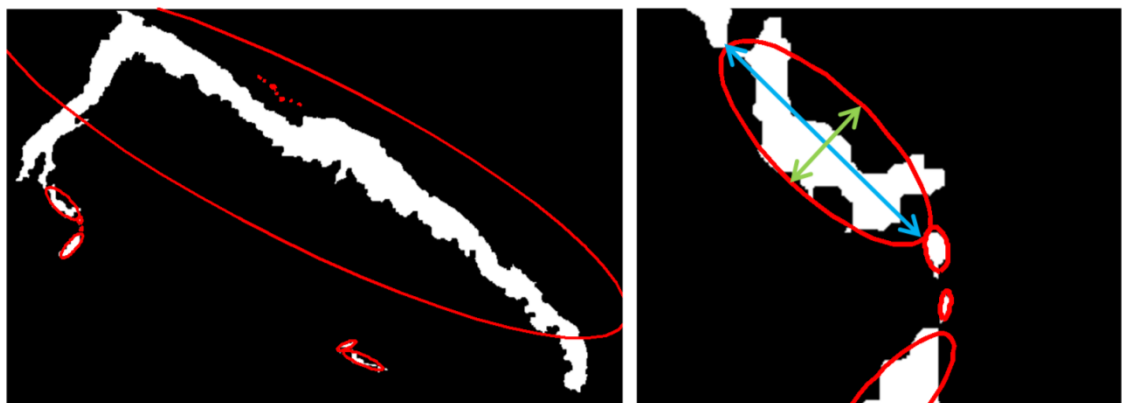


Figure 2.28 Examples of ellipses with the same second moment as specific objects in the image (marked with red ellipse boundaries). The major and minor axis lengths of one of the ellipses are shown by blue and green arrows respectively.

Statistical features describing the all examples of a particular shape within the image may also be informative; for instance, the median area of a vacuole in an image, or the mean inter-quartile range of vacuoles in a particular image. Object population statistics used in this research include count, mean, median, standard deviation, interquartile range, range, skewness and kurtosis.

### Texture Features

For this research the following features were investigated:

- *Contrast*: Measure of intensity variance or contrast between the pixel pairs  $i$  and  $j$  over a whole image.  $p_{(i,j)}$  is the element  $i, j$  of the normalised symmetrical GLCM. Contrast is zero for a constant image. Contrast is defined as:

$$\sum_{i,j} |i - j|^2 p_{(i,j)} \quad \text{Equation 2.18}$$

- *Correlation*: Measure of joint probability occurrence or correlation of pixel pairs  $i$  and  $j$  over the whole image.  $\mu$  is the GLCM mean,  $\sigma$  is the variance. Correlation is 1 for a completely positively correlated image, -1 for a perfectly negatively correlated image and NaN for a constant image. Correlation is defined as:

$$\sum_{i,j} \frac{(i - \mu_i)(j - \mu_j)p_{(i,j)}}{\sigma_i \sigma_j} \quad \text{Equation 2.19}$$

- *Energy*: The sum of squared elements in the GLCM, which is the angular second moment and is a measure of uniformity. The fewer grey level transition within an image, the higher the energy will be; energy is 1 for a constant image. It can be defined as:

$$\sum_{i,j} p_{(i,j)}^2 \quad \text{Equation 2.20}$$

- *Homogeneity*: Measure the closeness of the distribution of elements in the GLCM to the GLCM diagonal, and is a measure of uniformity in the image. Homogeneity can be defined as:

$$\sum_{i,j} \frac{p_{(i,j)}}{1 + |i - j|} \quad \text{Equation 2.21}$$

These features can be calculated using the [graycoprops](#) function in MATLAB for different directions and distances, and for different colour channels (e.g., R, G and B). It is important to calculate features for a variety of directions and distances to capture texture and pattern for different orientations and at different scales.

The features extracted from the images were used to represent the images during classification, which is described in the following section.

### 2.5.2 Classification

In this research, the technique of interest within machine learning is classification, which is a method for assigning class labels or identifying to which of a set of categories (or *classes*) a new observation belongs, based on a training set of observations. The classification process depends on the type of data used, for instance data vectors, lists, text strings and images all provide different challenges. Classification problems may be binary and require a simple yes/ no output, or multiclass, in which case multiple categories are possible.

In image classification the numerical properties of various image features are analysed and organised into categories. There are two main learning approaches used in classification, supervised and unsupervised. In supervised classification a set of training observations (e.g., a set of features representing each image) are accompanied by the correct output label (e.g., the grade of damage shown in the image). This enables the classifier to determine rules or patterns and predict the grade of a new image that has not previously been presented to the classifier. More specifically, the training phase uses a data set of inputs,  $\mathbf{X}$  and targets,  $\mathbf{y}$ , with each observation consisting of an input vector,  $\mathbf{x}_i$  and a class label,  $y_i$ . The input vector,  $\mathbf{x}_i$  can also be described as a *feature vector*. The relationship between the feature vectors and equivalent class labels is analysed and this information is used to build a mathematical model which contains a unique description of the features relevant to each training class. In the testing phase the model is used to predict output label of a new observation given the feature vector. The performance of the classifier in the testing phase is known as the generalisation ability of the classifier.

Unsupervised classification, often described as clustering, differs in that the training images are not labelled with a particular output; instead the observations are grouped or clustered into categories based on some underlying similarity or pattern present in the data. No information is provided on the number or type of classes during the learning process, instead decisions on the number and nature of categories are based purely on the input data. The important features of each class

are extracted, and this information is used to enable the classification of a new observation. Supervised and unsupervised approaches to classification are compared in Table 2.4.

Table 2.4 Comparison of unsupervised and supervised approaches to classification tasks

<b>Unsupervised</b>	<b>Supervised</b>
Number of classes unknown.	Number of classes known.
Allocates patterns to naturally occurring groups based on similarity/ cluster density.	Uses a training set of patterns to set up the internal parameters of model.
Number of classes and class structure must be learnt.	Uses a model to estimate class membership for an unknown observation/ image.
<b>Advantages</b>	
Fast and consistent for large data sets, as there is no requirement for a separately determined label set.	Utilises domain knowledge, but may introduce bias if class labelling is subject to bias.
No need to label observations.	Can learn complex patterns
<b>Disadvantages</b>	
Clusters may not correspond with desired groups if irrelevant features are included and may be difficult to interpret. This can be mitigated using careful feature selection.	Selection and preparation of training data can be expensive and time consuming as training data must be representative of the true distribution and labelled accurately.
<b>Examples</b>	
k-means, Mixture Models (e.g., Gaussian), Hierarchical Clustering, Self-Organising Maps	Support Vector Machines, Decision Tree, Naïve Bayes, Nearest Neighbours

A supervised approach was chosen for this research due to the availability of labels for the training set and the existing categories in place for grading the samples. Supervised learning approaches often use probability as a basis. Probability is used in different approaches, discriminative and generative, which build models in different ways.

Determining the posterior probability is a useful first step as this knowledge allows the image to be classified so as to minimise a particular loss function, for example, the misclassification rate. In the discriminative approach a parametric model is used to calculate the posterior probabilities, inferring the values of the parameters from a set of labelled training data. Posterior probabilities are estimated directly and there is no attempt to model underlying probability distributions. In the generative approach the joint distribution of images and labels is modelled. One way of doing this is to learn the class prior probabilities and the class-conditional densities separately using Bayesian classification, a generative supervised learning method. Based on a probabilistic model, the method captures uncertainty by determining probabilities of each possible output. The classification method is named after Thomas Bayes (1702-1761), who proposed Bayes Theorem (Equation 2.22). Bayes Theorem describes how the probability of a hypothesis ( $h$ ) being true is affected by new evidence and can be written:

$$p(h|D) = \frac{p(D|h)p(h)}{p(D)} \quad \text{Equation 2.22}$$

where  $p(h)$  is the prior probability of  $h$ ,  $p(D)$  is the prior probability of the training data,  $D$ ,  $p(h|D)$  is the probability of  $h$ , conditional on  $D$  and  $p(D|h)$  is the probability of  $D$ , conditional on  $h$ .

One implementation of the Bayesian classification is the naïve Bayes classifier which was used in this research project. A full evaluation of classification approaches used in the field and the rationale for the choice of the generative naïve Bayes classifier is given in Chapter 3, section 3.2.8.

### **2.5.3 Naïve Bayes Classification**

In the Naïve Bayes Classifier, the most likely class is assigned to a given image based on its feature vector. Learning in such classifiers can be simplified by assuming that features are independent given class, meaning the 1-dimension class conditional density can be determined for each feature individually:



$$P(\mathbf{x}|c) = \prod_{i=1}^n P(x_i|c) \quad \text{Equation 2.23}$$

where  $\mathbf{x} = (x_1, \dots, x_n)$  is a feature vector and  $c$  is a class.

The assumption of independence allows the method to estimate model parameters using less training data, so it is particularly useful where the number of features is high and the number of observations is low. The independence assumption greatly reduces the variance of this model; however the bias can be large.

For the Naïve Bayes classifier, each feature in each class is modelled using a distribution which is used in the prediction phase to determine the posterior probabilities of class membership. A Gaussian distribution can be used, in which case a normal distribution is estimated for each class based on the mean and standard deviation. Kernel density estimation (KDE) is a nonparametric technique for estimating the probability density function,  $f(x)$ , of a continuous variable  $X$ . KDE was first described for use with Naïve Bayes classifiers by John and Langley (1995) and the major benefit is that it does not assume normality for each class distribution. The method estimates the  $f(x)$  of class distribution by averaging a known density or weighting function (the kernel) over the observed data to create a smoothed approximation.

#### **2.5.4 Considerations in Statistical Classification**

**Prior Probabilities:** In many classification problems there is not an equal likelihood of a given observation belonging to each class. The performance of a classification algorithm can sometimes be improved if information on the typical proportions of known observations in each of the classes is included in the classification algorithm. This information can be incorporated using prior probabilities which increase the likelihood of predicting classes with higher priors. If the detection of a rare class is particularly important, it is advisable to over represent the class in the training set.

**Curse of Dimensionality:** In most decision making activities, having more information is considered to provide an advantage in arriving at the correct conclusion. In a classification task, it would be expected that including more

features in a classification model would result in a more accurate result. However this is not the case when working with high dimensional data, due to an effect referred to as the “curse of dimensionality”, first described by Bellman (1961). The effect occurs because observations become more sparsely spaced and dissimilar as the number of dimensions is increased, making the statistical grouping of observations difficult. The effect can be illustrated by considering a set of 10 training observations plotted across a 1 dimensional feature axis. Assuming the total range of feature measurements covers 5 unit intervals, the sample density would be  $10/5=2$ . If a second feature is added, the same 10 observations cover a feature space of  $5 \times 5=25$  unit squares, giving a sample density of  $10/25=0.4$  samples per interval. For 5 features, the sample density would be  $10/3125=0.0032$  and this exponential decrease in sample density continues as more features are added.

The sparsity of samples in high dimension feature space makes it easy for a classifier to obtain a hyperplane which separates two groups of observations. However, because the classifier has learned based on the appearance and position of specific instances and exceptions in the training set, it may be modelling random error or noise rather than the correct underlying relationship of features. This issue is referred to as overfitting, and is a direct effect of the curse of dimensionality. An overfitted model will perform poorly when presented with a new observation that does not adhere to the same exceptions, and is said to have poor generalisation ability. In order to avoid overfitting, either the number of dimensions must be kept low, or the training set must grow exponentially as features are added to maintain coverage of the feature space. Reducing the number of dimensions using feature selection is discussed next.

### **2.5.5 Feature Selection**

Feature selection involves selecting a subset of features (or variables) from the whole set, to use in the training and application of a classifier. It reduces the total number of features and avoids overfitting, facilitates data understanding, reduces measurement and storage requirements, and can reduce the time required for model training and utilisation. The general objective is to remove redundant

features (which provide no additional information to other features in the subset), or irrelevant features (which provide no useful information). Two approaches to feature selection which differ with respect to their treatment of redundant and irrelevant features are filter and wrapper methods.

Filter methods rank features using an evaluation criterion based on correlation coefficients or by other statistical tests such as t-tests or Chi-squared, to assess the relationship between individual features and the response or output of interest. Kohavi and John (1997) proposed an alternative methodology for feature selection with the objective of identifying the set of features most useful for building a classifier, rather than those most relevant to the output. Referred to as the wrapper methodology, this approach assesses subsets of features based on their ability to maximise the predictive performance of a classifier.

Feature selection methods commonly used to identify subsets in the wrapper methodology include the sequential search methods, sequential forward selection (SFS) and sequential backward selection (SBS) (Whitney, 1971). SFS starts with a single feature that is either chosen at random or selected as the most relevant to the classification task using a filter method. Features are then added sequentially, based on whether their addition improves classification performance. SBS starts with all features included and they are removed successively until any further removal results in a decrease in classifier performance. Floating search methods, including sequential floating forward search (SFFS) and sequential floating backward search (SFBS) have been proposed, which allow previously selected/discarded features to be re-evaluated at a later stage (Pudil *et al.*, 1994). Both sequential methods and their floating counterparts suffer from the “nesting” effect, where suboptimal subsets are possible due to the fact that previously selected features cannot be discarded and discarded features cannot be reselected. Two highly correlated variables might be included if it gives the best performance in the SFS evaluation.

### **2.5.6 Dimensionality Reduction – an Alternative Approach to Feature Selection**

An alternative method for dealing with high dimensional data is to carry out dimensionality reduction. Feature extraction is one approach to dimensionality

reduction used in image analysis already discussed in section 2.5.1, whereby a 1-dimensional vector of feature measurements is extracted from the original 3-dimensional matrix of the RGB image. The techniques more commonly implied by the term dimensionality reduction are typically based on data projection techniques. The main linear technique for dimensionality reduction is Principal Component Analysis which remaps the data onto a lower dimensional space in a way that maximises the variability in the lowest dimension (Kohavi, 1995). In each subsequent dimension less of variability is explained. The principal components can be used as an alternative to the untransformed features to capture more information in fewer features.

### 2.5.7 Classifier Evaluation

Classification can be evaluated in a number of ways, including accuracy, speed, robustness to noise, computational resource requirements, scalability, interpretability, ability to fully use the information content of the data, uniform applicability, and objectiveness. In reality, no classification algorithm can satisfy all these requirements nor be applicable to all studies, due to the complexity of histological classification. Classification accuracy assessment is, however, the most common approach for an evaluation of classification performance, which is described next.

- **Accuracy:** measures the probability of a correct classification. At pixel level (for segmentation) it refers to  $N_{cp}/N_p$ , where  $N_{cp}$  is the number of correctly classified pixels and  $N_p$  is the total number of pixels. For whole image classification it refers to  $N_{ci}/N_i$ , where  $N_{ci}$  is the number of correctly classified images and  $N_i$  is the total number of images.
- **Misclassification Rate:** measures the probability of an incorrect classification,  $N_i/N$ , where  $N_i$  is the number of incorrect cases (pixels or images), and  $N$  is the number of cases.
- **Sensitivity:** measures the probability that a known positive case will be classified correctly. This may be quoted as a probability or rate, or as a percentage.  $N_{tp}/(N_{tp}+N_{fn})$  where  $N_{tp}$  is the number of true positives and  $N_{fn}$  is the number of false negatives.

- **Specificity:** measures the probability that a known negative case will be classified correctly. This may be quoted as a probability or rate, or as a percentage.  $N_{tn}/(N_{tn}+N_{fp})$  where  $N_{tn}$  is the number of true negatives and  $N_{fp}$  is the number of false positives.
- **Robustness** measures the ability of the classifier to perform effectively while its variables or assumptions are altered, operating without failure under a variety of conditions. For instance, can the system grade a mild positive reaction accurately using feature measurements in the presence of non-relevant variability in the input images such as staining differences?
- **Repeatability** means that given the same inputs, the same result will consistently occur. For instance, will the system give the same grade given the same input image each time?
- **Reproducibility** is a measurement of consistency between systems. For example, when run on different computers by different users, does the system give the same output? Will images acquired from different acquisition systems be graded the same? How much variation does any user interaction introduce?

A commonly used method to evaluate the performance of a classifier is to estimate the misclassification rate for new unseen images using cross validation. This measurement is informative because it evaluates the generalisation ability of the model. One form of cross validation method is *k-fold* cross validation. In this method, a dataset of size  $N$  is first divided randomly into  $k$  mutually exclusive subsets. Each subset in turn is used as the test set for the classification model which has been trained using a training set made up the remaining  $k-1$  subsets. Once the model training and testing has been repeated  $k$  times, the average number of incorrectly and correctly classified observations across  $k$  tests is calculated. The size of  $k$  can vary from 2 to  $N$ , with the case of  $k=N$  known as *leaveout* or *leave-one-out* cross validation. As  $k$  is increased, the bias of the error estimate is reduced and the variance increases. Higher  $k$  values result in large training sets and small test sets and because larger training sets tend to result in similar performance on multiple iterations, higher  $k$  can lead to overfitting. It has been demonstrated that leaveout cross validation is a high variance estimator of generalisation error and gives overly optimistic results, the method often overfits

through the inclusion of too many features in the model, resulting in poor prediction performance for new data (Hastie *et al.*, 2011).

Monte-Carlo cross-validation can also be used as an alternative to *k-fold* cross-validation (Lu and Mandal, 2012). In this method the data is split randomly into training and test sets and observations can be included in multiple test or training sets, unlike *k-fold* cross-validation where the *k* subsets are mutually exclusive and no repeats are permitted. While *k-fold* cross-validation will give a very low bias, the estimation can have a high variance if the test set is small. For example, *10-fold* cross-validation for a sample of 100 observations will test prediction on 10 observations at a time. Monte Carlo cross-validation tends to result in low variance models as many more possible partitions of the dataset can be explored due to the fact that the subsets do not need to be mutually exclusive. However a consequence of this is that some observations may be used more than others resulting in a higher bias. One approach to balance bias and variance is to repeat *k-fold* several times, which retains low bias but reduces the variance of the estimate.

This chapter has introduced the theory and background of the biological and computational aspects of the research. Some of the key aspects will be considered in more detail in the next chapter, and the relevant published literature evaluated with a view to justify some of the research decisions made and qualify the need for this research.

## Chapter 3 Literature Review

This chapter provides a review of relevant literature, providing evidence of the need for an alternative to manual grading systems in histopathology both generally and in the case of the skin explant assay used by Alcyomics. A review of the competing technologies for toxicity, allergenicity and immunogenicity is also provided. Literature relating to the main aspects and challenges of the research is given next, specifically relating to colour normalisation, segmentation, feature extraction, classification and ground truth. Finally, literature on the application of image analysis to analysis of human skin is reviewed

### 3.1 Grading Variability in Manual Histopathology

Manual grading methods used in histopathology are time and labour-intensive, and the lack of quantitative characterisation can lead to issues relating to subjectivity and inter and intra-observer variability. Many grading scales used in histopathology are qualitative or semi-quantitative (Pilette *et al.*, 1998; Taylor and Levenson, 2006). A qualitative histological grading system uses information such as the presence/ absence, severity, distribution and morphology of particular histological features to determine the final grade. More advanced semi-quantitative systems, such as the Nottingham Grading System used for breast cancer staging and grading, combine quantitative information such as mitotic count and rate with qualitative information such as the 'degree of nuclear pleomorphism' (C. W. Elston and Ellis, 2002). A key barrier to objectivity for many current grading scales stems from the attempt to use a qualitative system to measure continuous variables; boundaries are not set clearly and instead ordinal variable language (low, medium, high) is used to guide decisions.

High signal to noise ratio, low inter and intra-observer variability and the placing of category boundaries to separate natural cluster of cases are all important considerations when designing grading systems (Morris, 1994). In both clinical and research situations it is important to consider the purpose of grading and decide whether multiple subdivisions provide useful information. When Morris (1994) discussed the theory of information transmission when applied to histopathology grading systems, he showed that reducing the number of categories

improved inter-observer agreement but reduced the amount of information transmitted. His suggestions of 100 point scales and quotation of confidence limits could provide a more rigorous but useful system in histopathology.

The qualitative visual assessment at the heart of histopathology is inherently subjective, resulting in a high degree of variation in analysis between different histopathologists. The inter-observer disagreement in manual histopathology diagnoses has been reported in numerous published studies, some of which are described below.

In a review of 500 neuropathology diagnoses of brain or spinal cord biopsies, Bruner *et al* (1997) reported some degree of disagreement in 42.8% of cases, with 8.8% classified as serious disagreements which were defined as having immediate significance for therapy or intervention. Reviewing inter-observer variation in pathological diagnosis of brain tumour patients, van de Bent (2010) stated that 'more objective, quantitative and reproducible criteria are urgently needed'. Van Putten also stated that poor performance in grading of non-advanced and advanced adenomas in colorectal cancer diagnosis suggested that more objective criteria were required (2011) A comparison of essential thrombocythemia in bone marrow biopsies in 370 patients by 3 experienced haematopathologists using World Health Organization classification criteria showed substantial inter-observer variability both for overall diagnosis and certain cellular characteristics (Wilkins *et al.*, 2008). Elsewhere there have been reports of poor agreement in diagnosis of melanomas and melanocytic nevi (Hastrup *et al.*, 1994; Farmer *et al.*, 1996), analysis of cervical biopsy specimens (Robertson *et al.*, 1989; de Vet *et al.*, 1990), scoring of chronic hepatitis in liver biopsy (Rousselet *et al.*, 2005) and grading of dysplasia in ulcerative colitis (Eaden *et al.*, 2001). Even a relatively simple task of classifying cell nuclei in renal cell carcinoma highlighted inter-observer variability; when classifying 180 nuclei, five experts agreed on 24 normal and 81 atypical nuclei, but there was an inter-observer classification error of 18-30% for the other 75 nuclei (Fuchs and Buhmann, 2011). Improvements in inter-observer agreement have been obtained by adding quantitative measures to previously qualitative grading systems. For example, the Nottingham modification



of the Scarff-Bloom-Richardson grading for breast carcinoma that added quantitative measures reported inter-observer agreement of 70-90% (C.W. Elston and Ellis, 1991; Frierson *et al.*, 1995; Dalton *et al.*, 2000).

While subjective language and qualitative grading criteria are responsible for some of the grading variability, fatigue, training level and sampling may be important factors. Manual grading is time consuming and the drive for more consistent grading through the collection of more detailed quantitative data is likely to increase analysis time unless automated processes are adopted. Fatigue and lapses in concentration can become significant when carrying out repetitive tasks and periods of intense concentration may result in changes of visual perception that reduce the chance of detecting unusual events, a state termed “inattentional blindness” by Mack and Rock (1998).

Tissue sections used in histopathology can be large (cm rather than mm in dimension) and high magnifications are often required to assess specific histological features at the microscope. The size of the sections means that a full assessment of the whole tissue sample at high magnification is often not practical, therefore sampling is often used to increase throughput. Unfortunately sampling of such large images may miss isolated or focal areas of change, increasing the chance of false negative results.

The requirement of an experienced histopathologist may exacerbate the issues mentioned above by increasing workload on key staff. The level of experience of the histopathologist is known to be an important factor in grading and was found to be the most important factor affecting scoring variability in a study investigating chronic viral hepatitis grading (Rousselet *et al.*, 2005).

The prevalence of variability in manual histopathology supports the development of new automated methods. Based on a review of the literature, new methods should increase quantitative measurement, reduce the workload of experienced histopathologists and incorporate automation in order to succeed.

### 3.1.1 Skin Explant Assay

In the case of both the Lerner scale and the modified Lerner scale used in the Skimune assay, the differentiation between grade I and grade II reactions is particularly challenging. The distinction is made by considering the presence of dyskeratotic bodies and the severity of vacuolisation in grade II when compared to grade I. Unfortunately dyskeratotic bodies are not always present in grade II samples, consequently sometimes the degree of vacuolisation alone must be used for grading. The difficulty of grade I and II differentiation is supported by the findings of Massi *et al* (1999) who found that while inter-observer agreement was almost perfect for grade III reactions, grades 0, I and II showed lower levels of agreement. Massi *et al* suggested this situation could be improved by the inclusion of an additional manual estimation of inflammatory infiltrate as an additional criterion for grade II damage. The appearance of mononuclear cell infiltrate into the tissue was also proposed as a better indicator of early GVHD than the presence of dyskeratotic bodies by Horn *et al* (1994), who proposed modifying Lerner's criteria to include dermal lymphocytic infiltrate at Grade II. In the *in vitro* skin explant assay there is no inflammatory infiltrate so this approach cannot be used.

The inter-observer variability for the original skin explant assay was assessed using three transplant centres across Europe (Sviland *et al.*, 2001), with 503 slides graded by each of the centres then reviewed and graded blindly by an experienced independent pathologist. Of the 503 slides, there was disagreement in 8% of cases across the four centres, with 14.5% disagreement found for samples with grades II- IV damage, compared to the 2.2% disagreement for samples with grades 0-I damage. Most of the differences were for cases at the boundary between grade I and II, which is the borderline between a normal and positive result.

The boundary between grade III and IV changes can also be open to interpretation in cases of very severe cleft formation. For instance, some operators would only classify a sample as grade IV if there is complete separation of the epidermis and dermis, while others may also grade a sample with a very small proportion of DEJ intact as a grade IV. Despite the challenges described, the simplified criteria used in the Skimune assay have been shown to retain good prediction of GVHD in bone

marrow transplant patients (Sviland *et al.*, 1990) and so have been retained for use by Alcyomics in their commercial assay.

### **3.1.2 Alternative Methods to test for Toxicity, Allergenicity and Immunogenicity**

In the commercial environment into which Alcyomics are launching the Skimune assay, there are a number of alternative tests and assays available. The Skimune assay can be used as an alternative for animal tests, which are established and accepted methods for measuring safety, toxicity and allergenicity in the pharmaceutical, chemical and cosmetics industries. *In vivo* experimental animal models have traditionally been used to provide information about the safety and toxicity of a range of products, including pharmaceuticals, industrial and household chemicals, cosmetics and agrochemicals. Testing of acute systemic toxicity to estimate the acute lethal dose (LD50) or concentration (LC50) and tests of skin and eye irritation account for many of the tests carried out on mammals. A combination of public opinion, new regulations, cost and logistics are making these industries look to alternative methods. For instance, the unprecedented numbers of new chemicals being introduced every year which REACH regulations now require to be tested make the sole use of animal tests logistically impossible, as well as extremely time and cost intensive (Frazier, 1992).

There are a number of 3R methods available, all of which aim to reduce, refine and replace the use of animal tests including: the use of human volunteers; artificial skin tests; molecular and cell culture methods; and *in silico* methods such as Quantitative Structure Activity Relationships. There are a number of companies offering *in vitro* skin models, which grow human skin cells into a life-like structure. Some models are long established, such as Epiderm (MatTek Corporation, US) which has been available for >15 years. Several of the models have already been through European validation, including EpiSkin™ (SkinEthic Laboratories, France), which the European Centre for the Validation of Alternative Methods (ECVAM) Scientific Advisory Committee recommended as a 'reliable and relevant method for predicting skin irritation' (SCCP (Scientific Committee on Consumer Products), 2007). There are also models which use *ex vivo* mouse skin, but the *ex vivo* human skin model from Alcyomics is unique because it mimics the autologous immune

system, matching immune cells and skin from specific individuals to give an accurate representation of the human immune response.

In the pharmaceutical market, immunogenicity assessment of biotherapeutics has received significant attention in recent years, with a number of industry white papers being published (Lu and Mandal, 2014; Xu and Mandal, 2015) alongside EMEA and FDA guidance on the clinical assessment of antidrug antibodies and the need for an immunogenicity screening framework. Immunogenicity in many cases leads to the loss of efficacy of a drug, but can also lead to the production of severe adverse side effects. New and improved methods to assess immunogenicity of potential drug candidates are therefore in demand by drug developers to help them reduce risk during drug development.

### **3.2 Digital Histopathology**

As increasing numbers of high resolution, high quality images are produced in pathology labs utilising the latest slide scanning technology, the analysis of these images becomes the bottleneck in the process. It is for this reason that there is currently intense focus in industry, academia and clinical environments on the development of useful and accurate image analysis algorithms.

#### **3.2.1 Challenges of applying Computer Analysis in to Histopathology**

Application of computer analysis to any biological sample is challenging due to the high degree of biological variability, complexity and problems of sampling bias (Paizs *et al.*, 2009). In histopathology, additional challenges are presented by the high data density of histopathology images, the complexity of the tissue structures, and the inconsistencies in tissue preparation (Gurcan *et al.*, 2009). McCann *et al.* (2014) describe the three main sources of variability in a histology-based diagnosis as:

- Biological variability, which encompasses the differences between people and also the variability of pathological process occurring in the tissue, meaning that slides of the same tissue from different people will look different.
- Inter-observer variability, which describes the impact of subjectivity and human judgement on histopathology analysis

- Technical variability, which is caused by differences in how the slide is prepared.

Biological variability can lead to significant variation within a single cell type, compounded by variation in the formation of tissue and the number, placement and morphology of tissue structures such as glands. This structural variation is often exacerbated by technical variability such as inconsistent sample preparation and staining procedure, differences in stain colour or reactivity between batches and the effect of section thickness on light transmission (Magee *et al.*, 2009). There are a broad range of histological patterns and features seen in different diseases and organs, with significant overlap in features both between different diseases and between different grades of the same disease. In addition to the general technical challenges, Levenson (2004) reported limited enthusiasm among pathologists for a switch from subjective to more quantitative scoring schemes due to difficulties of multi-centre implementation and a lack of recognised image analysis standards.

When replacing a manual process with a digital one, it is worth considering the strengths and weaknesses of the original system. The human visual perception system is particularly skilled and well adapted to interpreting visual scenes and this provides advantages when evaluating tissue sections. Human perception has the advantage of being reasonably resistant to image noise and contrast and invariant to changes in position, scale and orientation (Gonzalez and Woods, 2008). In histopathology, this means that humans can easily switch from low to high magnification, search for features of interest, and ignore artefacts and noise. The object oriented nature of our visual perception system is also well suited to identifying histological features such as cells, regions of tissue and structures such as glands. The analysis of cellular shape, size and organisation in histology uses pattern recognition, a process fundamental to human cognition which has been perfected over years of evolution. However, while certain facets of human perception bring significant advantages to histopathological analysis, the weaknesses associated with manual grading such as inter and intra-operator

variability, subjectivity, bias and fatigue that have already been discussed in detail are also related to the 'human factor'.

Replacing a manual process carried out by a human with an automated computer system provides significant challenges. It is challenging to try and replicate any human process of image understanding, interpretation and decision-making, as it can be difficult for the expert to explain exactly how certain decisions are made. In manual histopathology there are often significant levels of implicit knowledge required to make accurate decisions which are gained through experience and not always included in the traditional written grading criteria. One solution is to work more deeply with the experts to try and ascertain as much of this implicit knowledge as possible, and then codify this knowledge in the image analysis algorithm. Alternatively, researchers are increasingly extracting huge numbers of features in the hope that both explicit and implicit knowledge will be captured somewhere within the dataset (Bins and Draper, 2001)

Although the objective of both the human and computer process is the perception, understanding and interpretation of image information, the way in which this is achieved is fundamentally different. In contrast to the object-oriented world view that the human visual perception system uses, computer vision tends to represent images of the world at a pixel level. Although this approach is increasingly used (Bishop, 2010), and can be very successful, it can be difficult to separate the variation of interest to the other background variation already discussed. Computers are more suited than the human visual system for quantitative functions such as counting or area estimation, creating the potential for improvements in quantitation, throughput, objectivity, repeatability and reproducibility. Rather than examining a small percentage of the total cells as in manual histopathology, it is possible for a computer to analyse every pixel in every cell of the whole slide. Once algorithms and software programs are set up, computer-aided analysis has the potential to be much faster than manual analysis, although this is dependent on image file size, computer processing speed, and algorithm complexity. In this thesis, the research presented aims to mimic and capture human expertise and domain knowledge, but utilise the key strengths of

computational methods, i.e. quantitative measurement, objectivity, speed and reproducibility of analysis.

Although extremely common, claims that computer-aided image analysis is a completely objective solution have been disputed, with the argument that human devised and implemented algorithms are subject to human bias and judgement. Tadrous (2010) argued that image analysis methods simply implement the subjective decisions taken by the programmer throughout algorithm design in an objective manner, and the real benefits of the methods were speed, indefatigability and standardisation. In this research, the influence of domain knowledge and the bias that this may bring is not disputed, one of the main hypotheses of this research is that incorporating such knowledge into the early stages of the image processing and feature extraction will enable variation relevant to skin damage to be distinguished from non-relevant image variation.

In light of the challenges of histopathology, a number of commentators have suggested that image analysis tools in histopathology are not at present able to compete with the breadth and depth of expertise of a pathologist and their role should be to complement the role of the pathologist or histopathologist rather than to replace (Madabhushi, 2009). At the very least, a close relationship with the histopathologist is required to obtain feedback, and aid interpretation of results (Gurcan *et al.*, 2009).

### **3.2.2 Colour Normalisation in Digital Histopathology**

In histology, coloured chemical stains which bind specifically to proteins are used to aid identification of different tissue types. The final colour is affected by the quantity/ density of protein molecules in the stain, variability of the chemical stain colour or reactivity, variability of the staining procedure, tissue thickness (light transmission is a function of tissue thickness) and lighting during image capture and digitisation (Magee *et al.*, 2009). These differences, often referred to as batch effects, do not create insurmountable issues in manual analysis because in the human vision system colours can be perceived and identified easily under varying illumination conditions. However these batch effects can create bias in the performance of automated classification methods and so approaches are required

to normalise the colour distribution in an image and facilitate subsequent processing steps.

Much work in histopathology bypasses the issue of colour by converting to grey-scale, and while this can be successful in cases where there are clear intensity differences between features and non-features, there is a significant amount of information lost concerning which stains are present and in what proportions. It is often necessary to compensate for differences in staining intensity by normalising the image intensities. One method to allow for staining inconsistencies is presented by Paizs *et al* for the quantification of inflammation in murine spinal cord (Paizs *et al.*, 2009). An internal reference area unaffected by experimental treatment or disease condition was used to specify a staining baseline. However this technique requires operator input to identify areas of interest and so is not an ideal solution due to the impact on throughput, the requirement for operator intervention and the potential introduction of subjectivity. Reinhard *et al* (2001) described a method for colour normalisation which maps the pixels in an input image to the colour distribution of a target image by equalising the mean and standard deviation for each dimension of a  $l\alpha\beta$  colourspace. Reinhard's method has been applied to H&E stained histology images by Wang *et al* (2007b), and Magee *et al* (2009) used the method as a benchmark to test their novel colour normalisation methods against. This approach is simple and can be applied to multiple images, however it assumes that all areas of the image can be normalised with the same transform. In reality, the inter-image variation for different image regions (e.g. background, different tissue types), results from differing sources. The approach works well when a single stain (e.g. Eosin) dominates the image, however because the approach uses a single linear transform for all pixels it would result in the incorrect mapping of many typical histology images.

When multiple stains are used in the same slide, overlapping absorption spectra can create difficulties in identifying and quantifying features. This is particularly important in immunohistochemistry (IHC) where different stains are used to locate and quantify particular substances; however it is also important in H&E stained slides to identify structures. Narrow band filters have been used during



image acquisition to separate the stains (Zhou *et al.*, 1996), however overlapping spectrums are still a problem with this method. Methods based on colour space transforms have been described, for example Lehr *et al* (1999) applied Photoshop® tools to tissue images utilising their hue-saturation-luminance (HSL) characteristics and other used a stain-specific transform (Ruifrok, 1997). A further development by Ruifrok, allowing the determination of the relative contribution of each stain to a pixel's colour is a method known as colour deconvolution. This method uses the specific optical density (OD) for each of the RGB channels rather than the intensity to describe each stain. Each pure stain can be characterised by a specific OD for the light in each of the three RGB channels, which means it can be represented by a 3 x 1 OD vector describing the stain in the OD-converted RGB colour space.

### 3.2.3 Segmentation

Segmentation is a critical first step in many image analysis applications since by locating regions of interest early in the analysis subsequent steps can become more accurate and computationally efficient. A full review of segmentation approaches is beyond the scope of this thesis and thorough reviews have been published including one by Segzin and Sankur (2004). A brief discussion of the limitations of traditional segmentation techniques in histopathology follows, but a comprehensive discussion of segmentation approaches being used in histology for global scene segmentation and local structure, cell and nuclear segmentation can be found in the review paper by Gurcan *et al* (2009).

Traditionally, segmentation approaches can be sub-divided into contour or edge detection based methods and region or histogram based approaches. A widely used contour based approach for the segmentation of biomedical images is active contours (or *snakes*) which were first described by Kass *et al* as energy-minimising deformable splines “guided by external constraint forces and influenced by image forces” that localise towards edges and boundaries (Kass *et al.*, 1988). However, active contours can only be semi-automated; an initial curve must be defined by the user and this initialisation step influences processing time and result quality (Angenent *et al.*, 2006). The inherent structural complexity of histopathology

images and the frequent presence of overlapping objects make the application of contour based approaches in histopathology problematic. The skin images used in this research contain a number significant discontinuities that are likely to be identified by edge detection algorithms, including the loose, linear surface layers of the epidermis (the *stratum corneum*), the fibrous structures of connective tissue in the dermis tissue, the basement membrane at the junction of the epidermis and dermis, cleft boundaries at the DEJ and cell membrane boundaries. The large number of potential 'edges' in the images make the use of edge or contour based approaches to find the boundary of the epidermis tissue difficult and prone to error. Region-based methods create sets using pixel or neighbourhood properties such as colour, intensity, location or texture. Location cannot be used for the skin images used in this research, as the orientation and structure of the images varies significantly. Colour, intensity and texture are more applicable as the different tissue types (epidermis and dermis) stain differently and have different morphology (and therefore texture).

Thresholding is the simplest of the region-based methods. It involves selecting an intensity threshold to create a binary image with the two image states representing foreground and background. While the threshold can be selected manually, automating the process is quicker and more objective. The aim in this research is to threshold the epidermis as foreground, leaving the dermis tissue as background. To achieve this, a threshold must be chosen to separate the epidermis and dermis pixel sets. The relative proportion of epidermis varies between images and so algorithms based on the percentage of foreground pixels are not useful. Simpler methods for choosing the threshold automatically including the 'intermodes' algorithm (Prewitt and Mendelsohn, 1966) introduced in Section 2.4.12, which finds two local maxima and sets the threshold half way between them. This method does not work well when the grey level histogram has very unequal peaks, which can be the case for the images used in this research. The 'intermeans' algorithm proposed by Ridler and Calvard (1978) and Trussell (1979) and described in Section 2.4.12 iteratively adjusts the threshold so it lies half-way between the means of the background and foreground pixels sets. This algorithm tends to find a threshold which splits the pixels into two sets of approximately

equal number, and this would not be appropriate for the skin images used in this research, which have varying proportions of dermis and epidermis tissue.

One of the most popular approaches for automatic threshold selection, proposed by Otsu (1979), chooses a threshold to minimise intra-class variance in the foreground and background pixel sets. This approach of minimising variance within each set has potential as there is usually an intensity difference in the pixels in the dermis and epidermis. However, the variation within the epidermis and dermis pixel sets (the inter-class variance) due to the biological and technical variability described in section 3.2.1 causes problems when using all thresholding techniques, including Otsu's method, which work best when there is relatively little variation within a set of images (Gurcan *et al.*, 2009).

The use of hybrid segmentation methods is becoming more common as researchers find that a single technique is unable to segment all structures adequately; multi-resolution approaches, feature based classifiers and post-processing steps are all popular additions to the traditional segmentation approaches. For instance, the addition of binary morphology to adaptive thresholding resulted in correct segmentation of 89% of three nuclei types used in cancer grading, albeit with a limited sample size of 24 (S. Petushi *et al.*, 2004). Fuzzy c-means clustering and active contours were combined to segment prostate cancer tissue with an accuracy of 84% (Hafiane *et al.*, 2008). A Bayesian classifier used to inform level set and template matching algorithms identified nuclear and glandular structures in prostate and breast cancer with comparable accuracy to manual segmentation (Naik *et al.*, 2008). A modification to the EM algorithm, using Linear Discriminant Analysis in neuroblastic tumour segmentation, was deemed a success based on a faster convergence rate than k-means clustering, despite similar accuracy (Jun Kong *et al.*, 2007).

#### **3.2.4 Segmentation of Morphological Structures**

Automated detection of tissue structures is of particular interest in histopathology. Traditionally quantitative analysis of morphological structures, or morphometry, has involved superimposing grids over the sample to aid counting, however these methods are susceptible to human counting errors. Automated detection of

nuclear and glandular structures has been achieved by combining information from multiple scales; low level pixel information in a Bayesian classifier and high level extracted with level set and template matching algorithms (Naik *et al.*, 2008). Segmentation of nuclei or cells is a very common first step in the image analysis of biological tissue. One fully automated method proposed by di Cataldo *et al* (2009) used the morphological and chromatic characteristics of tissue to segment nuclei. The methodology incorporated Ruifrok and Johnson's (2001) colour deconvolution algorithm to separate an RGB image into two monochromatic images for the H and DAB stains prior to local adaptive thresholding and classification, resulting in a higher segmentation accuracy than either edge or region based snakes. Local adaptive thresholding was identified as the key factor in the success of the morphological approach; this approach has the ability to cope with inhomogeneous staining and illumination by taking into account the specific neighbourhood of the relevant pixels.

### **3.2.5 Segmentation of Tissue**

While there have been a number of methods proposed for nuclear and individual cell segmentation in H&E stained tissues and segmentation of structures such as glands, there are few which attempt to segment particular tissue types as a whole, a useful first step in identifying disease features if these features are known to occur within a particular tissue. The wide variety of tissues and their complexity of appearance make this a challenging problem. Chen *et al* (2011) segmented bone, cartilage, and fat tissue in teratoma tumour images using local pixel intensities as features with accuracies of 59.7%, 73.18%, 91.09% respectively which shows the difficulty of creating a generalised solution applicable to multiple tissues.

The following papers are those most closely related to segmentation of epidermal tissue in H&E stained samples.

Lu and Mandal (2012) used a multi-resolution approach to segment the epidermis in images of skin. The approach uses global thresholding and shape analysis on a monochromatic image to get a coarse segmentation, before generating high resolution image tiles for further manual or automated analysis. Results on 16 whole slide skin images resulted in a 92% sensitivity rate, 93% precision and 97%

specificity rate. The average processing time to segment the epidermis area for an image with 2800 by 3200 pixels was ~ 2.38 seconds, which is 3,764,705 pixels per second.

Mokhtari et al. (2014) proposed an epidermal segmentation procedure as part of a system for measuring melanoma depth of invasion in microscopic images. Using morphological closing and global thresholding, it assumes that the morphological closing operation will remove components such as cell nuclei from the dermis area. While it is of relevance to this research, there is no quantitative measure of performance by which to compare it with other techniques.

A relatively simple approach consisting of shading correction, a low pass filter and a threshold finding algorithm based on the grey-level histogram was also proposed to segment the epidermis (Smolle and Hofmann-Wellenhof, 1998), however no quantitative data was given relating to how accurate the segmentation is and the authors stated that this type of approach was dependent on good quality staining of the sample.

Wang *et al* (2007a) described an approach for the segmentation of squamous epithelium from cervical virtual slides using a multi-resolution approach. They used block based texture features in a support vector machine algorithm to create a rough segmentation at x2 magnification, then fine-tuned at x40 magnification. They reported excellent accuracies of 94.9 – 96.3%, however performance statistics were only reported for two of the 20 test images, in addition, sensitivity and specificity were not quoted and it is noted in the paper that the approach tended to misclassify red blood cells and columnar epithelium cells. The algorithm was reported to take 21 minutes to segment one image on a 120000 x 80000 pixel image on a Pentium 4 3.4GHz processor with 2GB RAM, which can be scaled to approximately 7,619,048 pixels per second. This speed raises questions over the suitability of the approach in its current form in any application with large images requiring high throughput.

Datar *et al* (2008) segmented prostate tissue microarrays into their constituent tissue types, using Hierarchical Self-Organizing Maps to classify pixels based on

colour and texture features, followed by unsupervised colour merging. While the segmented images appear to show good performance against a benchmark method, it is not possible to quantitatively compare the results with other methods as no accuracy metrics are quoted and no indication of computational efficiency or segmentation time is given.

Eramian *et al* (2011) presented a graph-cut method to segment epithelium in haematoxylin & eosin (H&E) stained samples of odontogenic cysts. They also included a luminance and chrominance standardisation procedure to reduce the variation resulting from sample preparation. For a set of 35 test images they reported mean sensitivity and specificities of  $91.5 \pm 14\%$  and  $85.1 \pm 19\%$  respectively, and a mean segmentation accuracy of  $85 \pm 16\%$ . The average run time of their method was 7.2s per image, which can be scaled to 189,583 pixels per second.

This area of research is relatively new with few groups working on the segmentation of epidermal tissue. The five papers presented provide the most useful benchmark by which to assess the success of the segmentation procedures developed in this research. A full comparison is included in Chapter 5 when discussing the results of the epidermal segmentation algorithm.

### **3.2.6 Feature Extraction**

Feature extraction and selection are essential components of many image processing and analysis applications, including image retrieval (Antani *et al.*, 2002), registration and matching (Zitová and Flusser, 2003) and pattern recognition (Gonzalez and Woods, 2008). Features used in the majority of histopathology classification systems presented in the literature tend to be inspired in some way by visual patterns or attributes used by clinicians for disease diagnosis in traditional histopathology. Often the features will relate to particular objects of importance such as cell nuclei, glands, or particular tissue types. An extensive review by Gurcan *et al* (2009) categorises the features used in histopathology into object level and graph based features. The object level features are split into four categories: size and shape; radiometric and densitometric; texture; and chromatin-specific. The graph based spatial features named include

Voronoi Tessellation, Delauney Triangulation and a variety of neighbourhood based graphs. One problem with this categorisation is the grouping of very different feature types under the label of “object based”, which have very little in common except that they can be applied to sets of pixels. The splitting out of chromatin based features is understandable as they are so commonly used, however the majority of the features named (e.g. area, optical density, number of regions) could be applied to any object and so are not strictly chromatin-specific. An alternative categorisation is proposed here which groups features in the following way: morphometric features (e.g. size, shape); intensity/ colour features (e.g. hue, intensity, saturation, optical density); texture (e.g. co-occurrence matrix, energy, fractal and wavelet), and graph based spatial features (e.g. node number, clustering co-efficient, spectral radius). Some of the published work using these types of features in histological image analysis is summarised in Table 3.1.

Morphometric and colour based features are popular with histopathologists (and those designing automated programs for histopathologists) as they are easier to interpret than some of the other types of features such as texture and they were some of the earliest features to be used in this field. The ease of interpretation of morphometric features can be attributed to the fact that human visual perception is object based. However because computer vision tends to be largely pixel based it can be challenging to try and replicate human visual analysis such as histopathology classification using a computer system. As is clear from the selection in Table 3.1, texture based features have been widely used by computer scientists developing automated methods for use in histopathology. They offer a way of extracting a quantitative measure to represent complex tissue architecture and staining patterns. However, the limited biological interpretability of some texture features has been highlighted as barrier to acceptance by clinicians and pathologists (Kothari *et al.*, 2013). Graph based features offer an alternative way of capturing and representing complex spatial architecture and structural information by defining a large set of topological features. Graphs have the ability to represent spatial arrangements and neighbourhood relationships of different tissue components. Very large numbers of graph based features are extracted to represent the structural and spatial information used by histopathologists to

classify disease states and they offer a new way of capturing the tacit knowledge a histopathologist uses when grading a tissue sample (Ghaznavi *et al.*, 2013).

*Table 3.1 Summary of literature in digital histopathology grouped according to feature type used*

Feature Classification	Feature Type	Appearance in published literature
Morphometric	Area, size	(Adiga <i>et al.</i> , 2006; Sokol Petushi <i>et al.</i> , 2006; Kothari <i>et al.</i> , 2011)
	Boundary (perimeter, perimeter curvature and fractal dimension)	(Naik <i>et al.</i> , 2008; Kothari <i>et al.</i> , 2011)
	Shape (eccentricity, sphericity, elongation, compactness, major/minor axis length)	(Price <i>et al.</i> , 2003; S. Doyle <i>et al.</i> , 2007; Naik <i>et al.</i> , 2008; Filipczuk <i>et al.</i> , 2011; Kothari <i>et al.</i> , 2011)
Intensity/ colour	Hue, saturation, optical density, intensity	(Jun Kong <i>et al.</i> , 2007; Tabesh <i>et al.</i> , 2007; Kothari <i>et al.</i> , 2011)
	Colour texture features	(J. Kong <i>et al.</i> , 2009; Sertel <i>et al.</i> , 2009; Kothari <i>et al.</i> , 2011)
Texture	Co-occurrence matrices (inertia, energy, entropy, homogeneity)	(Scott Doyle <i>et al.</i> , 2006; S. Doyle <i>et al.</i> , 2007; Al-Kadi, 2010)
	Haralick and Gabor filter features	(Diamond <i>et al.</i> , 2004; S. Doyle <i>et al.</i> , 2007; Kothari <i>et al.</i> , 2011)
	Discrete texture, Markovian texture, run length texture	(Al-Kadi, 2010)
	Wavelets	(Scott Doyle <i>et al.</i> , 2006; Kothari <i>et al.</i> , 2011)
Graph based / Spatial	Voronoi diagram,	(S. Doyle <i>et al.</i> , 2007; Basavanhally <i>et al.</i> , 2008; Jondet <i>et al.</i> , 2010; Kothari <i>et al.</i> , 2011)
	Delaunay triangulation	(S. Doyle <i>et al.</i> , 2007; Basavanhally <i>et al.</i> , 2008; Jondet <i>et al.</i> , 2010; Kothari <i>et al.</i> , 2011)
	Minimum spanning tree	(Basavanhally <i>et al.</i> , 2008; Kothari <i>et al.</i> , 2011)



### 3.2.7 Feature Selection

It can be challenging to find out the exact basis on which a human expert makes a particular decision, and so it may be an advantage to generate a relatively large set of potential features despite the fact that many may be redundant. However, this feature set must be reduced to avoid the curse of dimensionality and model overfitting (introduced in section 2.5.4). An exhaustive search of all possible features sets is not viable for large feature initial feature sets and so sequential forward or backwards feature selection or sequential floating feature selection tend to be used (Pudil *et al.*, 1994; Gurcan *et al.*, 2009). More advanced techniques such as genetic algorithms (Sahiner *et al.*, 1996; Li *et al.*, 2011) or boosting (S. Doyle *et al.*, 2012) are increasingly being used; however such techniques are more complex to set up and run than traditional techniques.

### 3.2.8 Classification

In this research, a classification model has been used to predict the correct grading of an image based on a set of feature measurements. There is a significant amount of published research relating the use of classification algorithms in the diagnosis and grading of cancer using histological image analysis. A summary of some of the classification approaches used for cancer grading using histopathology and the reported classification accuracy is given in Table 2.3.

Table 3.2 Classification Approaches used in Histopathology and Published Performance

Method	Tissue	Dataset	Performance	Ref.
Augmented cell graphs	Brain	646 biopsy images	Accuracy 97.1%, sensitivity 97.5%, specificity 93.3% (inflamed), 98.% (healthy)	(Demir <i>et al.</i> , 2005)
Processing (adaptive thresholding, morphological processing) and supervised classification.	Breast	1062 section images	Accuracy of 95.6%	(Sokol Petushi <i>et al.</i> , 2006)
Modified k-Nearest Neighbour	Brain	43 images	Accuracy of 87.8%	(J. Kong <i>et al.</i> , 2009)
Linear Guassian classifier (diagnosis)	Prostate	367 (diagnosis)	Accuracy of 96.7% for diagnosis.	(Tabesh <i>et al.</i> , 2007)
k-Nearest Neighbour (grading)		268 (grading)	Accuracy of 81% for grading	
Support Vector Machine	Prostate	54 biopsy images	Accuracy 92.8% (grade 3 vs stroma), 92.4% (epithelium vs stroma), 76.9% (grade 3 vs grade 4)	(S. Doyle <i>et al.</i> , 2007)
Support Vector Machine (with Guassian kernel), using hyperspectral images	Colon	45,056 features - from 11 image cubes	Accuracy 99.72%	(Rajpoot and Rajpoot, 2004)
AdaBoost (multiclass adaptive boosting classifier)	Breast	34 images	Accuracy 98.3% (non-malignant), 99.3% (invasive) and 90% (non-invasive)	(Oztan <i>et al.</i> , 2013)

The excellent classification accuracies reported by Rajpoot and Rajpoot (2004) reflect the exhaustive optimisation of the SVM kernel functions they have undertaken which improved the accuracy of their classification method from 87% to > 99%. Grade based classification accuracies in the literature vary significantly, and often accuracy for a cancer / non-cancer decision is much higher than accuracy in discrimination of different grades. For instance, Keenan *et al* reported accuracies varying between 62.3%-76.5% for discrimination of different grades in H&E stained cervical tissue (Keenan *et al*, 2000), and while Tabesh *et al* (Tabesh *et al*, 2007) could discriminate between cancer and non-cancer in 96.7% of prostate cancer tissue slides, discrimination between low and high cancer grades was much lower at 81%. The difficulty of differentiating lower cancer grades is also shown in the results of Oztan *et al* (Oztan *et al*, 2013) which show 10% of non-invasive cancers were misclassified as non-malignant, despite reporting accuracies of > 98% for non-malignant and invasive cases.

The literature can also be viewed in terms of whether the favoured methods are generative or discriminative in their approach. Generative methods include Gaussians, naïve Bayes, mixtures of multinomials, mixtures of Gaussians, mixtures of experts, hidden Markov models, Bayesian networks, and Markov random fields. Popular discriminative methods include logistic regression, support vector machines, traditional neural networks, nearest neighbour and conditional random fields. It can be seen that in general discriminative methods have been favoured. This is not unexpected, as there is a tendency in machine learning to favour discriminative models for classification tasks, as they solve the problem directly rather than doing so through an additional intermediate step of modelling the underlying distribution (Ruderman *et al*, 1998). However when Ng and Jordan (1960) compared the two types of models, they showed that while the generative model had a higher asymptomatic error than the discriminative model when the number of training examples became large, the generative model can reach its lowest error with a lower number of training examples.

There are a number of reasons why the generative method of naïve Bayes classification was selected for this research project:

- The number of training examples was limited and so a generative model offered the possibility of achieving a reasonably low error classifier.
- The naïve Bayes classifier is well known to be a robust method, which generally shows good classification accuracy (Zhang, 2004; Demichelis *et al.*, 2006). A number of researchers compared the naïve Bayes classifier with rule based learning algorithms and proved the effectiveness of the naïve Bayes classifier empirically (Clark and Niblett, 1989; Cestnik, 1990).
- The naïve Bayes classifier has been proved robust to noise and irrelevant attributes (Zheng and Webb, 2000), an issue identified as very important in this research problem due to the presence of many image variables unrelated to immune mediated damage.
- It has been reported that domain experts in the field of medicine found the learning theory easy to understand, a point that should be given consideration given that this application was being developed for a company focussed on the biologic and clinical aspects of their technology (Kononenko, 1993).

It is for some of these reasons that the naïve Bayes classifier is often used as a benchmark when new classification methods are being designed. Using a known method provides a useful starting point to assess the success of the image segmentation and feature extraction methods developed in the research and means a usable solution is available for Alcyomics at the earliest opportunity.

Due to the large and dense datasets typically generated in histopathology image classification tasks, the use of ensemble classification methods is becoming prominent in the field. Ensembles of classifiers have been reported to reduce the bias or variance associated with single classifiers and improve classification accuracy (Kuncheva and Whitaker, 2003). While it has not been possible within the scope of this research project, it would be valuable in the future to assess potential improvements in classification accuracy using ensemble methods or extensions to the Naïve Bayes such as the hierarchical approach proposed by Demichelis *et al* (2006) or the non-parametric version used by Soira *et al* (2011).

### 3.2.9 Ground Truth

When designing a computer aided system for image analysis or classification, the final performance is usually validated against an appropriate ground truth or reference standard. The term “ground truth” is used in machine learning and statistics to refer to the *known* data used to train a classification or regression model and to calibrate or measure the accuracy of the new system’s performance against. While manual grading can be used as the ground truth for a computer system attempting to perform the same task, the issues of inter and intra-observer variability have raised some concern (Jannin *et al.*, 2006). One way of improving the ground truth is to use a system based on multiple human experts to generate the known data, for example, Warfield *et al* (2004) use expectation maximisation to estimate a ground truth for segmentation from multiple expert inputs. Realistic simulated images known as phantoms have also been used (Aubert-Broche *et al.*, 2006), however the imagery in histopathology may be too complex to simulate adequately. The ideal situation is the use of clinical data/ patient outcome as ground truth, thus avoiding the uncertainty of the manual grading, however this data may not always be available.

## 3.3 Imaging and automated analysis of skin

### 3.3.1 Alternative Imaging Modalities

Skin imaging is widely used in the cosmetic industry and in dermatology. Digital colour photography is used to create images which are used to analyse skin colour and texture, and investigate facial lesions. While colourimetric staining combined with brightfield microscopy is the standard technique for visualising tissue in histology, there are a number of other imaging modalities which have been used to visualise skin tissue.

There are many imaging modalities that can be used to analyse the skin *in vivo* rather than from surgically excised skin biopsies. Optical coherence topography (OCT) provides cross sectional images of tissue structure *in situ* by measuring back-reflected or back-scattered light. The OCT technique provides images of much lower resolution than the light microscopy images used in this research, but is used widely in dermatopathology applications due to its non-invasive nature and

the opportunity it offers to analyse live tissue *in vivo*. It has been used for the detection of basal cell carcinoma (Avanaki *et al.*, 2009) and the analysis and detection of active inflammation, necrosis, hyperkeratosis and formation of intradermal cavities (Gladkova *et al.*, 2000). The particular changes analysed in the paper by Gladkova would suggest that this technique may be of use in analysing GVHRs *in vivo* to minimise the number of biopsies that need to be taken and provide a more general assessment of damage over a larger area of skin. However since the *in vitro* nature of the Skimune assay is fundamental for its commercial application, the technique does not have a direct application for this research project. In the skin explant assay the GVHRs are only created after very small sections of skin have been incubated in the lab with the test compounds. Applying OCT prior to fixing and staining the skin samples may damage the samples through additional handling and would be unlikely to provide additional information given the limited resolution of the technique.

A large proportion of the published literature relating to image analysis of skin is the detection and analysis of skin lesions and skin cancer. Images are taken from the skin surface using techniques such as epiluminescence microscopy (ELM, or dermoscopy) (Binder *et al.*, 1998; Ganster *et al.*, 2001), digital video microscopy (Seidenari *et al.*, 1999) and confocal scanning laser microscopy (Busam *et al.*, 2002). Although some of these techniques are capable of penetrating the upper layers of the skin, they generally have much poorer resolution than traditional light microscopy on sectioned tissue.

As an extension to standard staining techniques, multi-channel techniques have been used to overlay images generated using different imaging wavelengths and immuno-fluorescent labels have been used to identify, quantify and localise proteins and molecular markers within tissue. Camp *et al* (2002) located subcellular compartments using fluorescently labelled tags in order to identify regions of tumour using co-localisation of tumour-specific antigens. Sequential imaging and registration were used for simultaneous imaging using fluorescent biomarkers and traditional H&E staining using fluorescent and brightfield microscopy respectively (Can *et al.*, 2008).

### 3.3.2 Application of Image analysis techniques for skin histopathology

The only papers identified during this literature review which have applied image analysis techniques in the analysis of GVHR in skin biopsies were by Sahmoud *et al* (Sahmoud *et al.*, 1993) and Fleming *et al* (Fleming *et al.*, 1998). Sahmoud *et al* focussed on the lymphocyte infiltrate in biopsies taken soon after a transplant. The ability of the spatial and texture parameters of the lymphocyte nuclei to discriminate between the high and low risk groups was investigated; a combination of five texture related features resulted in 100% correct classification. Fleming *et al* carried out a similar study investigating whether the size, shape and texture of lymphocyte nuclei could be used as a predictor of GVHD based on a biopsy taken during the early onset non-specific symptoms. In this case, the image analysis was not found to be a useful predictor of GVHD onset. Other uses of image analysis in research relating to bone marrow transplantation and GVHD include: analysis of bone marrow *in situ* following limb/ extremity transplantation (Hewitt *et al.*, 1995; Ramsamooj *et al.*, 1996); measurement of immunohistochemistry (IHC) stained Langerhans cells following BMT (Zambruno *et al.*, 1992); differentiation of *lichen planus* from chronic graft-versus-host disease using quantitative IHC (Hitchins *et al.*, 1997) and morphological analysis of skin thickness and IHC in murine sclerodermatous GVHD.

In summary, image analysis has been used to predict the onset of GVHD from nuclear characteristics of lymphocyte infiltrate, however this is not a parameter that will be investigated in this research. Other research has generally used image analysis as a tool to assist in quantification of IHC or to analyse basic features such as epidermal thickness, this type of use would involve significant user interaction and is quite different from the automated system that is the aim of this research.

Except for the work from Lu and Mandal (2012) and Mokhtari *et al.* (2014) on skin segmentation already presented, there is relatively little reported in the literature. One other source of literature published on image analysis of skin is the University of Graz in Austria. Research in the 1980s and 1990s focussed on the assessment and diagnosis of melanoma through morphological analysis of mononuclear cells (Smolle, 1988), nuclei (Smolle *et al.*, 1989a; Leitinger *et al.*, 1990), vasculature

(Smolle *et al.*, 1989b) and collagen (Smolle *et al.*, 1996), moving from subjective counting and manual identification and grouping of features through to fully automatic measurement procedures. A relatively simple approach consisting of shading correction, a low pass filter and a threshold finding algorithm based on the grey-level histogram was able to segment the epidermis (Smolle and Hofmann-Wellenhof, 1998), however no quantitative data is given relating to how accurate the segmentation is and this type of approach is dependent on good quality staining of the sample. In 2000, a methodology was developed by Josef Smolle known as Tissue Counter Analysis which used hierarchical cluster analysis to classify electronically dissected image sections (called tissue elements) based on texture, colour and grey-level features. Skin structures such epidermis, papillary dermis, and dermal infiltrate were identified and the approach has been applied to the quantification of immunostaining in combination with fractal analysis (Gerger *et al.*, 2004), quantitative analysis of skin biopsies (Smolle and Gerger, 2003) and classification of malignant melanoma in combination with Classification and Regression Trees (CART) (Gerger and Smolle, 2003a; Gerger and Smolle, 2003b; Wiltgen *et al.*, 2007). In combination with CART, the technique resulted in correct classification of the tissue elements at a rate of 91.7% for cellular elements, 90.0% of collagen based elements, 79.9% of fatty elements and 64.3% of other tissue components. The authors identified difficulties in correctly classifying elements at the section margin, in the *stratum corneum* and other histological artefacts.

Image analysis has been used for the quantitative assessment of immunostained eosinophilic granule protein (EGP) in skin tissue (Kiehl *et al.*, 2001). Following additive shading correction, the colourspace transformation and recombination of the greyscale images was carried out according to methods put forward by Smolle (Smolle, 1996) and Ruifrok (Ruifrok, 1997), this was followed by an automated thresholding step based on Otsu's method. This combination of pre-existing methods is common to many of the published studies in histopathology. However it differs from this research in that it is quantifying specific immune based staining rather than looking at structural breakdown.



Given the relative paucity of image analysis literature relating to skin, literature relating to image analysis applications in cancer on H&E stained images is most relevant to this research, due to the structural changes seen in both GVHR and some forms of cancer. This is reflected in the focus of the preceding literature review. Although the results of these applications will be a useful benchmark to assess the success of any developed classification method, it is most important that the classifier provides a useful solution for Alcyomics. This will require that the accuracy is at least as good as that of the existing manual grading method, and that objectivity and reproducibility are improved.

The next chapter will describe the creation of the image dataset used in this research, and this will be followed by chapters describing the development and assessment of an automated image classification method for the images.

## Chapter 4 Data Generation and Image Acquisition

In this chapter, the data set used in the research is described, the selection and optimisation of the image acquisition and initial pre-processing procedures is presented. The manual grading of the dataset is also described and the inter-operator agreement analysed.

### 4.1 Data Source: Skin Explant Assay

While the aim of the research was to develop a classification process to be used as part of the commercial use of the assay, a data set of skin samples generated using chemical and pharmaceutical test compounds was not available when the research project was started. As an alternative, 125 clinical samples generated when using the skin explant assay (described fully in Chapter 2, section 2.3.2) for assessment of donor patient pairs in preparation for bone marrow transplant procedures were provided instead. A further 57 slides were provided later in the research project which had been generated during testing of pharmaceutical test compounds.

For the application of computer based image analysis techniques, information that is traditionally viewed by a human using a microscope must be captured in a digital form. The slides provided by Alcyomics needed to be converted to a digital form and determination of the most appropriate way of doing this necessitated consideration of the specific challenges of the dataset.

### 4.2 Image variation in skin explant data set

There were a number of challenges associated with the images being used in this research project that include: inconsistent sample preparation and histological staining; significant variation in size, shape and orientation of skin sections; and a complex combination of features to differentiate between grades. Despite previous attempts by Alcyomics to reduce variation by optimising sample preparation and staining, this source of variation has not been eradicated. The challenges were an unavoidable part of this particular research problem. Handling this variation was one of the most important aspects of this research.

The length of time between the biopsy being carried out and the start of the assay alongside the storage conditions may result in some degradation of the sample. This potential source of variation is mitigated by dissecting the biopsy into multiple sections and running at least one of the sections as a control in the assay. This is performed by culturing the control sample in culture media in the absence of immune cells or the test substance. Any damage caused by processing techniques during the assay procedure or biopsy storage will be evident in the control sample as a consequence. In the case of a grade II or higher grading for the control sample, the assay will be repeated.

The skin sections are stained using the haematoxylin and eosin staining methodology. Any histological staining process can result in varying colour intensity or saturation when performed with different stain batches, on different samples or on different days. Although great effort has been made to try and make this process more consistent by employing standard operating procedures, a significant amount of variation remains and any automated system developed had to be suitably robust to handle this inherent source of variability. The variation in staining can be observed in Figure 4.1A-D, although it should be noted that the colour variation is a combination of staining differences, lighting variation and biological variation. Dissection of the original biopsy creates variation in terms of sample size, the proportion of epidermis to dermis and sample. The samples vary in size, shape and orientation, but are generally less than 2mm in diameter. In some cases the epidermis forms a fairly linear structure across one edge of the sample (Figure 4.1A); in other samples it curves around the outside edge (Figure 4.1B). In rare cases the epidermis forms an unbroken ring around the edge of the sample (Figure 4.1C). In addition to vacuolisation, cleft formation and the presence of dyskeratotic bodies, a number of the images also included regions of necrotic (dead) tissue, which can be observed in Figure 4.1D. Ordinarily, samples with necrotic tissue are not manually scored and biopsies with such artefacts are excluded in the standard assay readout, however these images were included in this research to enable the software to identify and ignore artefacts or necrotic regions.

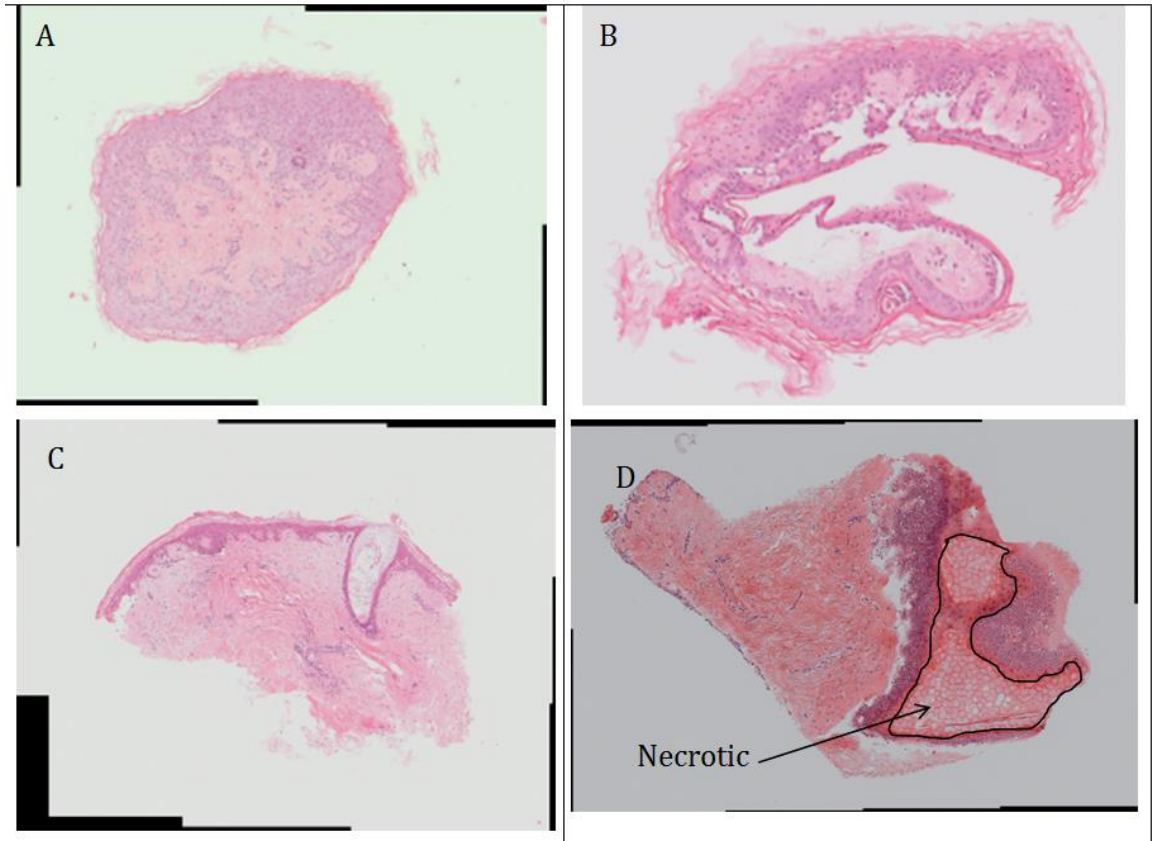


Figure 4.1 Four examples of H&E stained skin images, showing variation in shape, structure and orientation

For some slides, several fragments of tissue make up the sample, this may be due to the sample being particularly small or the person preparing them believing that a single sample was not representative, or it may be the result of a severe GVHR which has broken down the tissue structure and detached the epidermis from the dermis. The sectioning in paraffin at the end of the assay can also introduce variation and artefacts such as tears, particularly in fragile areas such as the DEJ. Tears such as the one indicated by the arrow in Figure 4.2 can be difficult to distinguish from clefts at the DEJ.

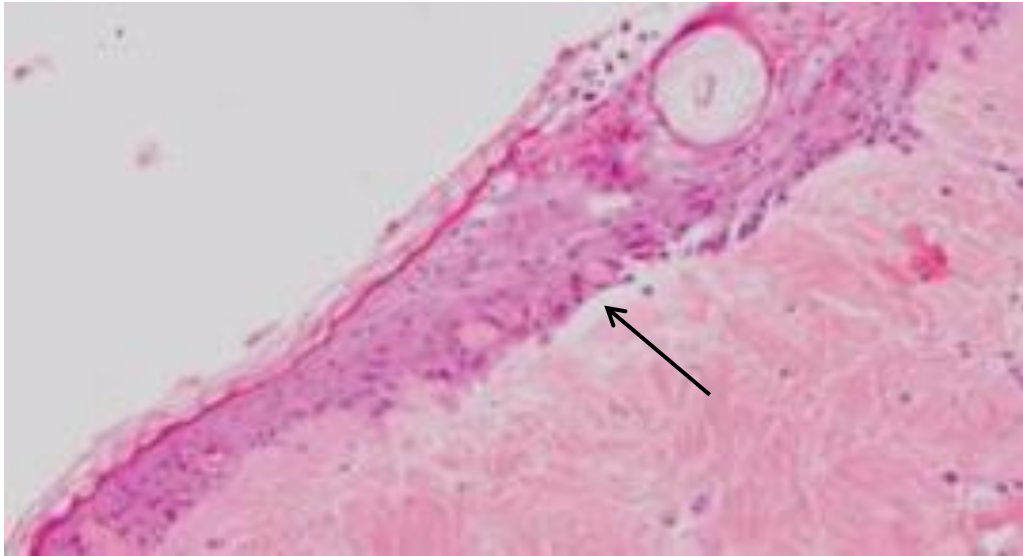


Figure 4.2 Section of H&E stained skin section showing a tear at the dermal epidermal junction, indicated by the arrow.

Regions close to the cut edge often show more damage than internal regions and this can be observed in the sample shown in Figure 4.3, where the affected areas are circled in blue. Any unusual damage located in these areas is discounted during manual grading.

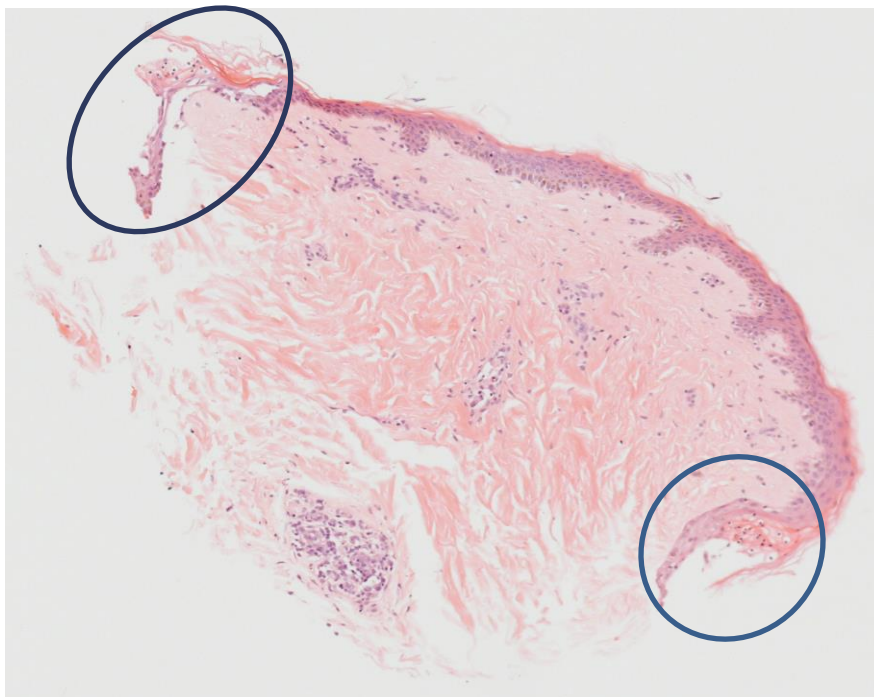


Figure 4.3 H&E stained skin image with grade I damage. Unusual break down in cell and tissue structure at cut sample edges is circled in blue.

### **4.3 Manual Slide Examination**

Considering how information is processed and decisions made during the manual grading process provides a useful starting point when considering how to represent the information in the slide digitally. During the manual grading of histology slides the expert operator quickly ‘scans’ the microscope slide at low magnification and identifies the epidermal tissue and the DEJ. The operator will then typically focus on an area of interest that appears to contain features of skin damage and subsequently switch to a higher magnification lens to confirm the presence and severity of the damage. While this hierarchical examination method is an efficient way of working, the operator is only examining a limited number of image regions in detail and it is possible for some information to be missed at the ‘scanning’ stage. To minimise or eliminate this issue in the automated method, the image acquisition procedure should be designed to eliminate (or minimise) any operator based decision processes.

The type of technology used for slide digitisation is important in terms of costs, ease of use, availability to the industrial partners and image quality. However, before the different systems were tested, the general approach used to capture the information digitally was selected.

### **4.4 Digital Representation Framework**

There are three main frameworks that can be used to capture and represent information in a slide. The first is a sampling based approach, the second uses a multi-resolution acquisition system and the third uses whole slide imaging. An overview of each approach is given in the following sections.

#### **4.4.1 Sampling**

The large size of the image data sets generated in digital histopathology can make it impractical to process, measure and analyse the whole sample, making sampling an attractive proposition. There are a number of ways in which sampling can be carried out. Probability based sampling, which includes random and systematic methods, can be used to avoid operator bias, but critical histological features may be missed as they tend not to be distributed evenly throughout the sample. Figure

4.4 shows an image where most of the sample has the appearance a grade I reaction but small regions of focal grade III changes are also present (circled in green on the figure). A random sampling approach could easily miss these focal grade III regions that the expert evaluators have identified as important indicators of damage.

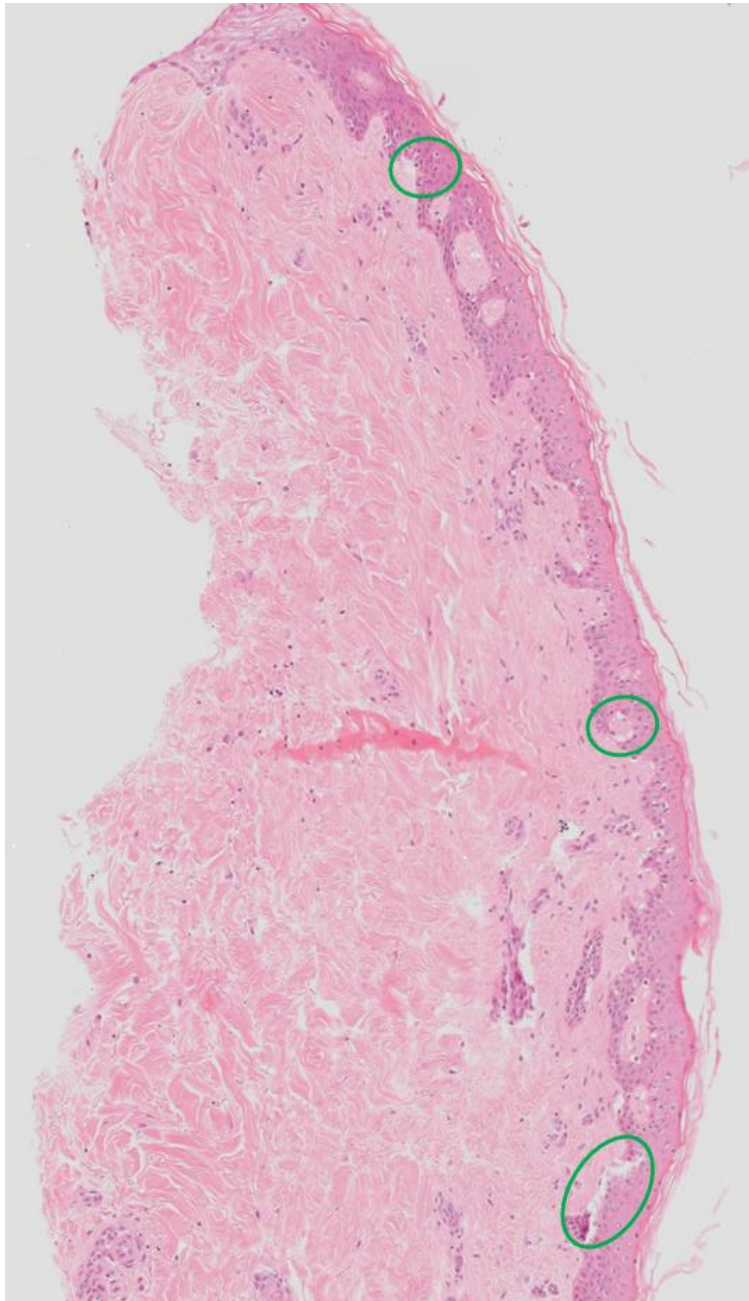


Figure 4.4 H&E stained skin section showing grade I changes with some focal grade III changes, circled in green.

A second sampling approach would be to take a series of images covering the area of interest (AOI), which would be the epidermis in this research. However, this approach would require the location of the epidermis to be specified by an operator which could introduce operator bias into the system from the start of the analysis process. To improve upon the current manual grading, the use of data from the whole sample would be preferable to a sampling based approach as this would remove the issue of sampling bias, and ensure all available information is captured and used in the subsequent classification process.

#### **4.4.2 Multiresolution Image Acquisition**

An alternative to sampling is multi-resolution imaging. In this approach a low resolution image is taken of the whole sample, computer-aided image analysis is used to identify the AOI and finally the identified areas are imaged again at a higher resolution. Alternatively, a series of different resolution images are taken and the low resolution image is used to define which parts of the high resolution image to examine in detail. This method mirrors the manual approach described in section 4.3 and has been described in a number of patents and papers. It aims to limit memory usage by confining high resolution imaging or image analysis to the AOI (Bouman and Liu, 1991; Ong *et al.*, 1996; Ifarraguerri *et al.*, 2003). The multi-resolution approach is an attractive option when using histopathological images, as it offers a framework to handle the significant quantity of data to be analysed. One example described by Madabhushi (2009) of the quantity of input data is the prostate biopsy procedure, where up to 20 biopsy samples may be taken, each of which may contain 225 million pixels once digitised in RGB colour at x40 magnification. High magnification is often required due to some important image features, such as those within the cell nucleus, only being visible at high resolutions. In this research project the features are generally visible at x10 and x20 magnification, so higher magnifications are unnecessary.

#### **4.4.3 Whole Slide Imaging**

In whole sample capture, the whole sample is captured as a single image. This approach is not possible using many traditional microscope imaging systems due to the inability of the lens to capture the entire field of view (FOV) when working



at magnifications high enough to identify the tissue features (at least x10 magnification). More complex imaging systems, such as the Zeiss Axio Imager discussed in section 4.6, are able to take a series of separate images at a high resolution and stitch them together to create a single image. The majority of newer microscope systems include this type of software as a standard feature.

Another solution would be to source and use a dedicated slide scanner to digitise the whole slide at high resolution. Digital slide scanners work by moving an objective lens across the microscope slide and capturing the magnified image scene with technology such as a CCD (charge coupled device). The scanners tend to generate a very high resolution image at high magnification (x40) which can also be viewed at lower resolutions. This size of file creates data handling and storage issues, which is why most scanner manufacturers also offer web hosting and sharing facilities. One potential issue with using a slide scanner is the use of proprietary image formats which make implementation of a novel image analysis approach difficult; scanner manufacturers sell their own image analysis software and frequently make it difficult to develop customised algorithms to use alongside their systems. Despite these issues, digital scanning is a growing technology in histopathology and would be worthy of investigation in the future, however a scanner was not available for this research project and consequently alternative methods were investigated focussing on a system to enable whole slide imaging.

Two different image acquisition systems were investigated. The Leitz Wetzlar/ Canon system is available to Alcyomics at zero cost for the routine analysis of slides, and for reasons of cost, simplicity and ease of access would have been the logical choice for the research project. However a second system is also accessible. The Zeiss Axio Imager is of a higher specification than the Leitz Wetzlar/ Canon system, with additional functionality including autofocus, white balance correction and image tiling. Limitations of this are that it is not available at all times and a cost of £13/hr is incurred each time it is used.

#### 4.5 Assessment of Leitz Wetzlar Microscope/ Canon Digital Camera

The Leitz Wetzlar microscope is currently used by Alcyomics to examine slides in the Skimune assay, the camera is used in cases where a record is required for reporting purposes. The equipment set up is shown schematically in Figure 4.5.

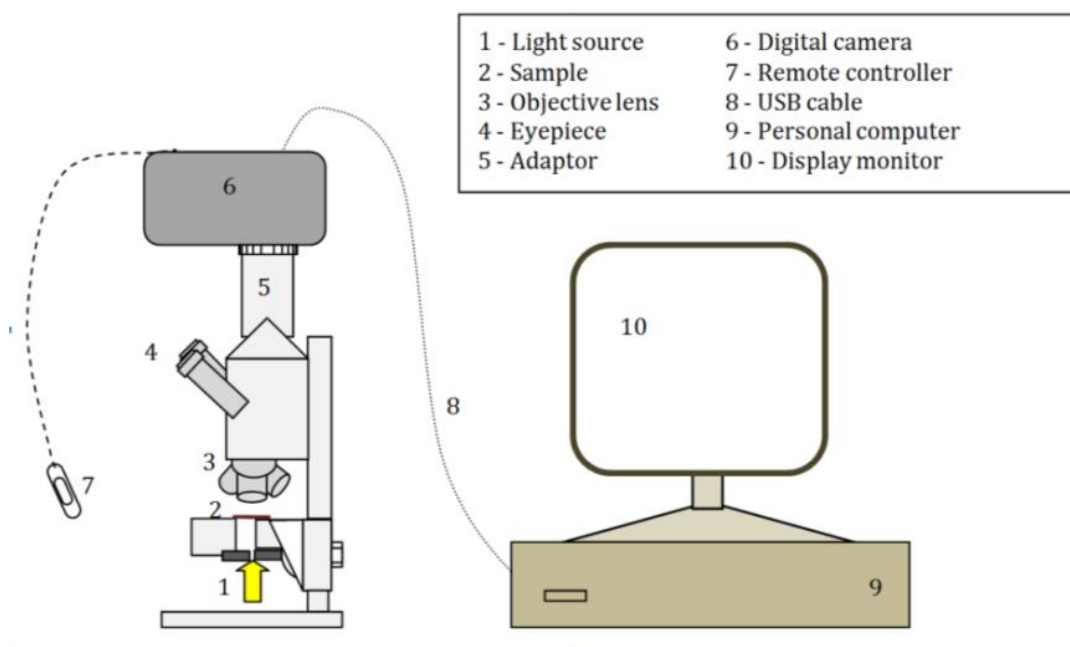


Figure 4.5 Schematic representation of the Leitz Wetzlar microscope and Canon camera image acquisition system.

The microscope uses bright field illumination (1), which was previously introduced in Chapter 2 section 2.1.2, to create image contrast through the absorbance of light by colourimetric dyes in the sample (2). The available objective lenses include x5, x10 and x25 magnification (3). The image can be viewed through the microscope eyepieces (4). An adaptor (5) links the phototube on the trinocular microscope to a Canon EOS 350D digital camera (6). The camera lens has been removed so that the image can be captured directly from the microscope phototube by an 8 megapixel complementary metal–oxide–semiconductor (CMOS) sensor. The image can be taken using a remote switch to avoid camera vibration (7). The digital information is transferred via a USB cable (8) to the computer (9) where a software package displays the captured images on the monitor (10). The camera produces final image size of 3456 x 2306 pixels at 24 bit depth.

### 4.5.1 Field of View

The field of view (FOV) of a microscope is the area of the sample that can be viewed or captured by an optical sensor at one time. Using the lowest magnification lens (x5) the majority of the skin sample could be observed when viewing through the microscope eyepiece, however the camera was only able to capture a portion of the microscope FOV, as can be observed in Figure 4.6.

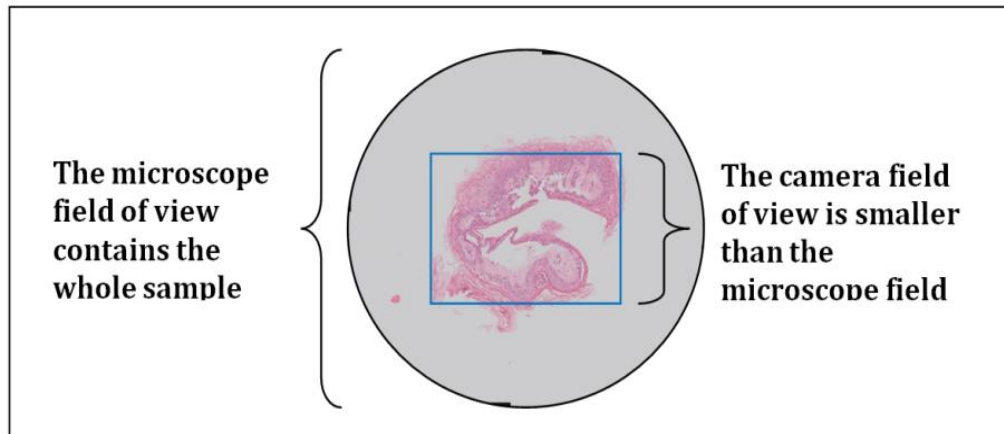


Figure 4.6 Diagram indicating the portion of the microscope field of view that is captured by the camera sensor. Note the image used was not taken with this system.

A whole sample image could only be obtained from this system if several overlapping images of the sample were stitched together using a separate software program. The higher the magnification required for the analysis, the more separate images would be required, and the greater the complexity of the stitching process.

### 4.5.2 Magnification and Image Resolution

When using a digital camera attached to a microscope, the final magnification of the image is the product of the magnification of the microscope objective lens, the camera optics and an enlargement factor dependent on the size of the pixels in the final viewed image. For the determination of the appropriate set up of the microscope and camera a range of images were taken of the same slide using different objective lenses. The lower the objective lens magnification (e.g. x5), the lower the resolving power of the microscope. The microscope magnification selected determines the optimal camera resolution. There is no advantage in using a higher camera resolution than that of the microscope. The requirement for

resolution must be balanced with the need for a large field of view (FOV) on which to perform the analysis.

Four images were taken using the x5, x10, x25 and x50 objective lenses. Figure 4.7 shows the same field of view cropped from each image and enlarged to a set image size to allow comparison of the detail and resolution at each magnification.

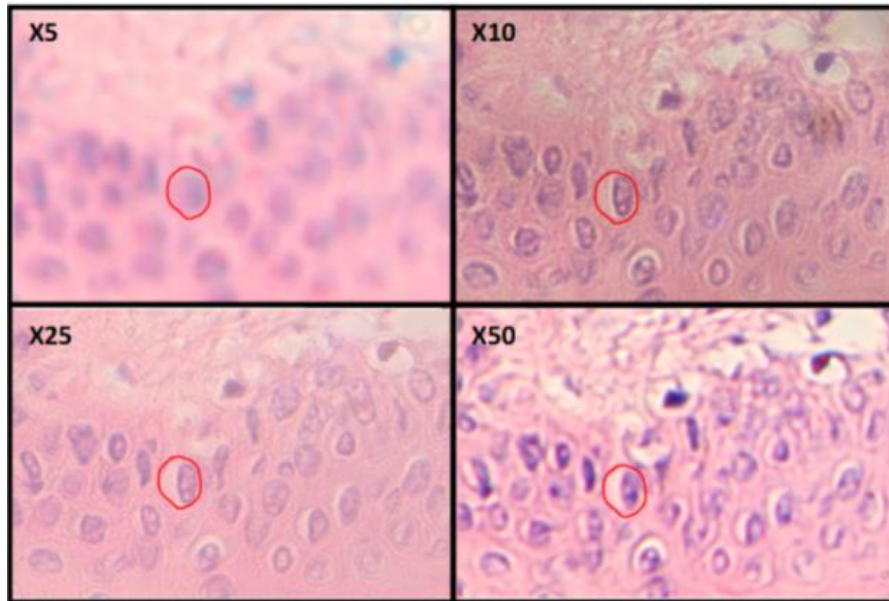


Figure 4.7 Comparison of image resolution with four different microscope objective lenses.

The main observation from Figure 4.7 is that the image taken with the x5 objective lens has lost a significant amount of detailed information about the individual cells compared to the other images. The resolution of the other images is more difficult to explain as there appears to be a drop in resolution when the x25 lens is used when compared to the x10 lens. This is unexpected, as the x25 lens should provide more detail. While it is possible that the microscope wasn't focussed correctly, this is unlikely as the images appeared well focussed when viewed through the eyepiece. A mismatch in the focus of the microscope and camera caused by the camera being mounted at the wrong height would explain a general lack of focus, but this would be consistent across all the images. It is clear that there are other factors limiting the resolution when using the x25 objective lens. The colour contrast is also poor in this image and so one possible explanation is that the lens itself was flawed in some way. The colour difference must result from the lenses, as no other changes were made to the system.

The inconsistency observed raises questions about the ability of this system to produce high quality images. The x10 objective lens is producing focussed images which show some detail of the cells and tissue structure, however if this image acquisition system were to be selected, further investigations into the focussing inconsistencies would need to be made.

#### **4.6 Assessment of Zeiss Axio Imager A2 System**

The Zeiss Axio Imager A2 system is a high quality imaging system which can be rented by the hour, and is based in the Bio-Imaging Unit, Newcastle University Medical School. The system offers transmitted light, fluorescence and confocal-like modes of operation. For the H&E stained slides used in this project, the transmitted light mode is the most suitable as H&E is not a fluorescent stain and the 3-dimensional images produced during confocal microscopy are not required for this application. The transmitted light mode works in the same way as the previously described Leitz Wetzlar microscope; however the camera is built into the system and can be controlled using a touchscreen on the system or through a linked desktop computer. Additionally, the Zeiss microscope has automatic autofocussing, background correction and white balance adjustment, and a motorised stage to enable automated tiling and stitching of multiple images post-capture using in-built MosaiX software. The available objective lenses range from x2.5 to x100. The colour camera on the system has a resolution of 1388 x 1040 pixels (a lower resolution than the Canon camera which had a resolution of 3456 x 2306 pixels).

##### **4.6.1 Magnification and Image Resolution**

Initially single fields of view were taken to compare the level of detail that could be resolved at different magnifications. Single 1388 x 1040 pixel images were taken using x2.5, x10, x20 and x40 objective lenses, resulting in a 4.12MB TIFF file for each image. The most suitable magnifications were x10 and x20. The x2.5 magnification did not show the individual cell components in sufficient detail to clearly see vacuoles, while the x40 magnification did not provide significantly more relevant cellular and structural detail than the x10 or x20 magnification. The increased spatial resolution at x20 magnification results in more sharp and

detailed images than those created using the x10 magnification. The x10 and x20 images are shown in Figure 4.8 and Figure 4.9 respectively, both were able to capture the main detail of individual cells including the darker purple central nuclei, white vacuoles and pink cytoplasm. The x20 images show the internal detail of the cells more clearly.

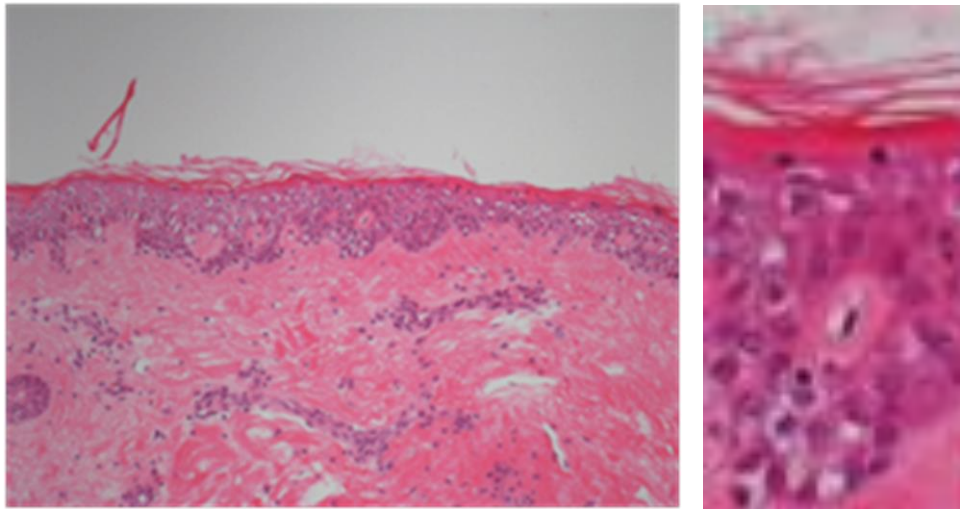


Figure 4.8 A 1388 x 1040 pixel image captured using a x10 microscope objective lens. The image on the right is an enlarged section to show cellular detail

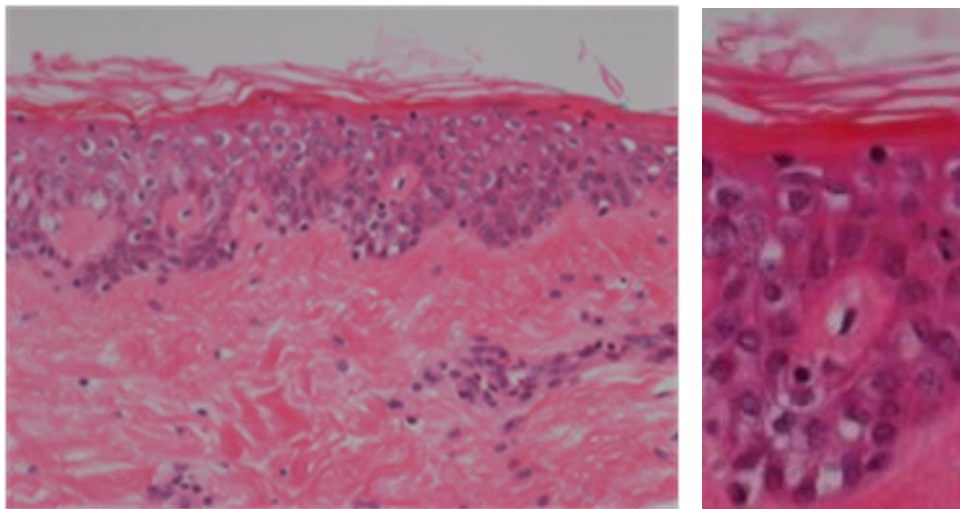


Figure 4.9 A 1388 x 1040 pixel image captured using a x20 microscope objective lens. The image on the right is an enlarged section to show cellular detail

The MosaiX software supplied with the Axio Imager system was used to define the area that would contain the whole sample, and control the motorised stage and camera to take digital images of multiple overlapping FOV covering the defined

area. The software includes an image stitching algorithm which matches the overlapping parts of the images to create a single image of the whole sample. Image tiling was attempted at both x10 and x20 magnifications. Although the x20 magnification provided very good resolution a major disadvantage to using it was that the number of image tiles increased significantly. The camera resolution is fixed in this system, and so when the system is set to capture the whole sample, the total image size is dependent on the physical size of the skin sample. In theory an image at x20 magnification will require four times more tiles than an image taken at x10 magnification. In practice the increase can be less than this due to differing amounts of background included in the image, for instance an image captured at x10 magnification in 16 tiles required 56 tiles at x20 magnification. Using a x20 magnification would increase the time of the acquisition by a factor of four, and use four times more memory. While the impact of the acquisition time would only become an issue if Alcyomics wanted to increase their sample throughput significantly, the increased image size would have increased the time and processing power required for all the subsequent image processing steps.

#### **4.6.2 Image Tiling, White Balance Correction and Background Correction**

A number of images were taken, altering the settings for white balance, background correction and magnification. In Figure 4.10 the colour balance has a blue tint, however this was addressed by using an interactive “colour picker” tool to select an area of white background against which to normalise the other colours. The results of this procedure can be seen in Figure 4.11 where the blue cast has been removed. The separate image tiles can be seen clearly in Figure 4.10 and Figure 4.11; this is due to variation in lighting across the microscope FOV. This was resolved by taking an image of a blank area of the slide once the microscope had been set up, and using an automatic built in software tool to subtract this “background image” from each subsequent image taken with the same settings. The process can be compared to a baseline correction procedure used in signal processing. The result of the background correction can be seen in Figure 4.12. The final step of image acquisition was to utilise the image stitching feature in the Zeiss system software. When the stitching was not carried out, the overlapping image tiles did not always align, as can be observed in Figure 4.10 at the point indicated



by the arrow. When the image tiling was used in combination with white balance correction and background correction, the process resulted in a high quality image that represented the sample accurately, as shown in Figure 4.12.

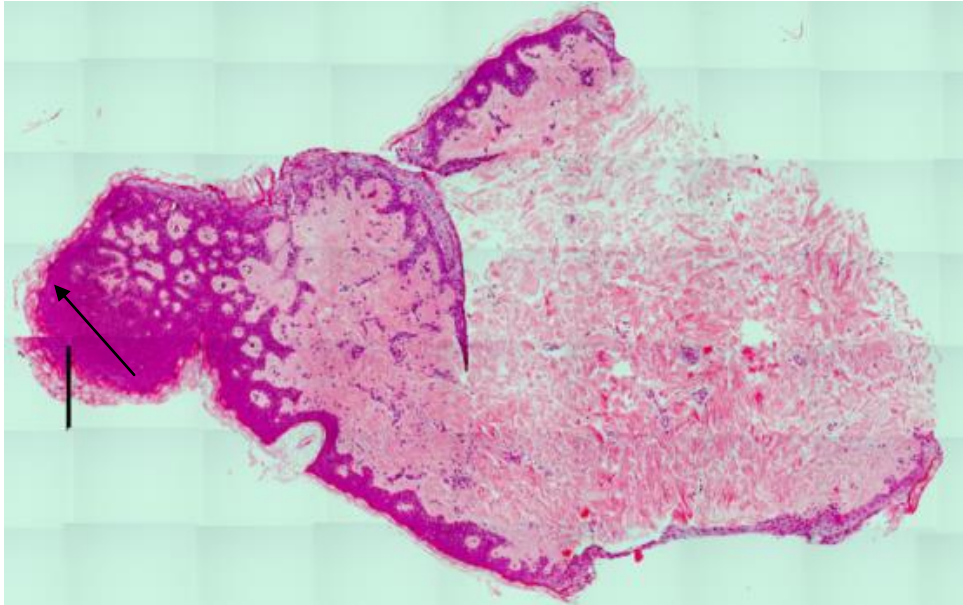


Figure 4.10 Skin sample image, requiring 56 tiles at x20 magnification (objective). White balance, background correction and image stitching NOT applied.



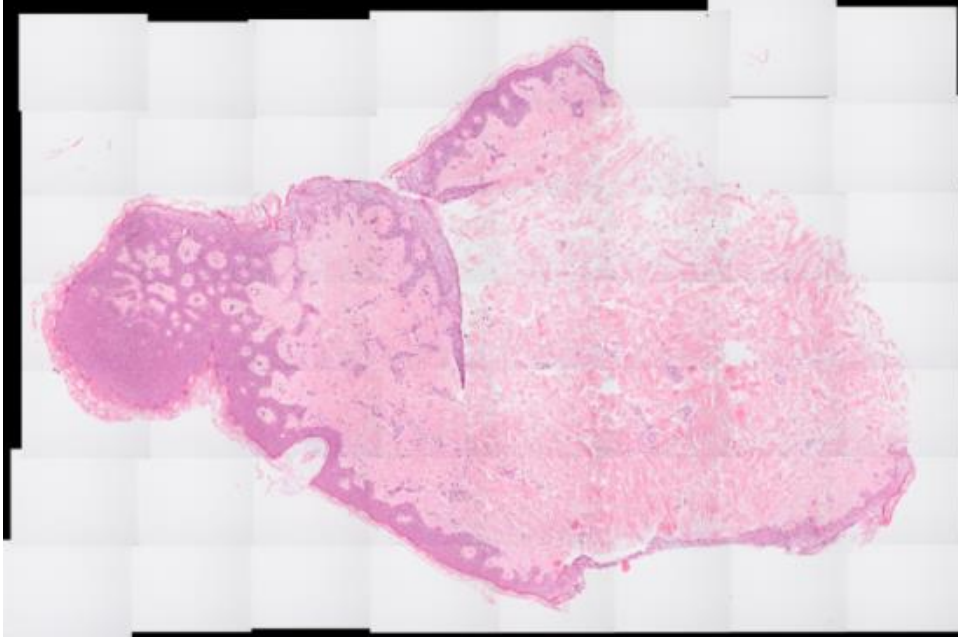


Figure 4.11 Skin sample image, requiring 56 tiles at x20 magnification (objective). White balance applied, background correction and image stitching NOT applied.

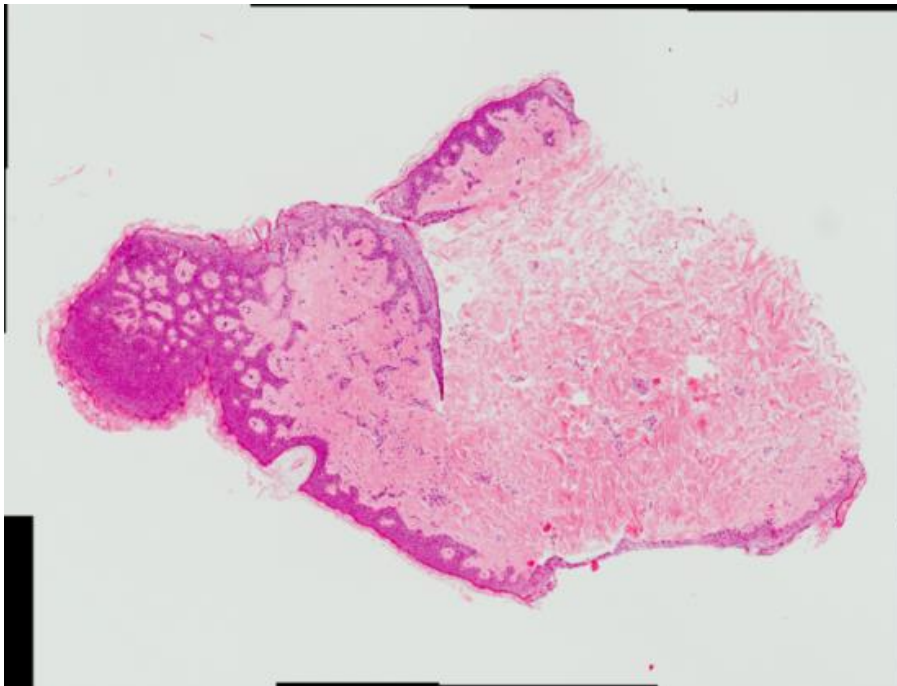


Figure 4.12 Skin sample image, requiring 16 tiles at x10 magnification (objective). White balance, background correction and image stitching applied

A standard operating procedure (SOP) was prepared for Alcyomics to enable them to capture images using the appropriate methodology. This SOP is attached in Appendix A.

#### 4.7 Image Grading using Manual Approach

The manual grading process was performed on all images provided by Alcyomics, but a full grading assessment was performed on the initial 125 image. The images were initially assessed at the microscope by the author and checked by an experienced histopathologist. After further discussions it became obvious that certain images were difficult to grade and were a source of disagreement in grading between different experts. Since this grading was being used as the basis of the training information for the classification model, it was vital that this information was as accurate as possible.

To get the most accurate grading (and to check the precision of manual grading), two experts with significant experience in grading GVHRs were asked to grade 125 whole slide images independently to avoid an agreement measurement bias. Of the 125 samples, eight were deemed unsuitable, either because they contained significant artefacts (e.g. necrotic tissue, tears), had missing tissue, faint staining or were generally atypical tissue sections. Table 4.1 shows the reasons why each of the eight images were deemed unsuitable and excluded from further analysis.

Table 4.1 Images excluded from further analysis and reason for exclusion

Image ID	Reason for Exclusion
16	Large area of necrotic tissue in epidermis
28	Large area of necrotic tissue, and little normal epidermis.
31	Dermis completely detached and not present on slide
35	Tear in epidermis has artificially split sample
36	Staining of dermis very pale and subsequently difficult to see.
59	Tear in epidermis has artificially split sample
91	Very small sample of epidermis
95	Unusual morphological structure to epidermis

The grading of the remaining 117 samples was analysed using kappa statistics (Cohen, 1960). Kappa statistics can be used to analyse agreement of multiple operators evaluating the same samples. Kappa measures agreement between operators and is a ratio:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} \quad \text{Equation 4.1}$$

where  $P(A)$  is the proportion of times the two operators agree,  $P(E)$  is the proportion of times the operators would be expected to agree by chance. Kappa values range from -1 to +1, with a kappa value of 1 indicating perfect agreement and a kappa value of 0 indicating the same agreement as would be expected by chance. Kappa values of less than zero are rare, and indicate the agreement is weaker than would be expected by chance.

Table 4.2 shows the grading agreement between the two expert operators, XW and AD when grading 117 skin explant samples. The table shows the correlation in grading scores between the two experts. For instance, the table shows there were 16 cases where both experts agreed a particular sample was grade II and 2 cases when AD gave a grading of IV and XW gave a grading of III.

Table 4.2 Grading agreement of skin explant samples by two expert operators

		AD				
		Grade I	Grade II	Grade III	Grade IV	All
XW	Grade I	29	13	1	0	43
	Grade II	1	16	1	0	18
	Grade III	0	1	40	2	43
	Grade IV	0	0	0	13	13
	All	30	30	42	15	117

The observed agreement  $P(A)$  is the percentage of all images for which the two operators' evaluations agree and was calculated in the following way:

$$P(A) = \frac{29 + 16 + 40 + 13}{117} = \frac{98}{117} = 0.838 \quad \text{Equation 4.2}$$

The agreement that would be expected to be present by chance alone,  $P(E)$ , was determined using the following calculation:

$$P(E) = \left[ \frac{43}{117} \times \frac{30}{117} \right] + \left[ \frac{18}{117} \times \frac{30}{117} \right] + \left[ \frac{43}{117} \times \frac{42}{117} \right] + \left[ \frac{13}{117} \times \frac{15}{117} \right] \quad \text{Equation 4.3}$$
$$= 0.094 + 0.039 + 0.132 + 0.014 = 0.280$$

This results in a kappa value of 0.833, indicating good agreement between the operators. While the general agreement is good and all but one of the disagreements are by a single grade, there is one interesting observation around the grade I and II border. In 13 cases, XW gave a grading of I and AD gave a grading of II; the opposite situation with XW grading II and AD grading I only occurred once. This indicates that at this critical border between a negative and positive result, AD is much more likely to give a positive grading than XW. The decision between grade I and II was described anecdotally to be the most difficult to make due to the absence of a clear differentiating feature such as cleft formation or complete separation of the tissue layers. This analysis supports this conclusion showing that there was disagreement at the grade I/II borderline in 15 of the 117 images (a rate of 12.8%). In the cases of disagreement, a discussion with both experts present was used to decide on the final grading of each training sample.

In the course of this discussion the experts also suggested a simplified binary classification grading scales during classification development. A binary classification of negative (grade I), vs positive (grade II, III or IV) was determined to be the most important and useful classification. While a secondary multiclass classification consisting of negative (grade I), mild reaction (grade II) or severe reaction (grade III or IV) would be a useful additional classifier, it was decided that this would form the basis of future work beyond the scope of this project.

#### **4.8 Discussion of Data Generation and Image Acquisition**

In the images obtained using the Leitz Wetzlar system the colour balance was skewed towards yellow tones, which means a white balance correction would be needed prior to any other image processing. Unusual colour casts typically result from the type of illumination used or faults with the camera sensor. There were

also focusing issues with this system, probably resulting from a mismatch in the focus of the microscope and camera. This mismatch is likely to have been caused by the height at which the camera is mounted and solving the problem would have required a new adaptor tube. While this was not a significant barrier to the use of this system, there were other issues which made the Zeiss system a better choice. First, the reliance on manual focussing when using the Leitz system would introduce more variation into the images than when using the Zeiss autofocus. The FOV limitations of the Leitz would also require images to be taken at the x10 magnification and then stitched together using a separate program either sourced or developed in MATLAB. Background correction would also be required to correct uneven illumination across the FOV, which would require a separate image of a blank area of the slide to be taken at each imaging session, and later subtracted from each sample image. While it would have been possible to develop new methods for image tiling and image correction, this would not have been an effective use of research time. Overall the images from Leitz system would require a significant amount of processing before they could be used in image analysis and there is no guarantee that the image focussing problems could be addressed.

The Zeiss Axio Imager II system offered a more consistent approach to obtaining images than the Leitz Wetzlar system. In-built features were available for autofocus, automatic white balance, background correction and image tiling and stitching. These low level image processing techniques all helped to improve the performance of the system and reduce operator associated variation. The Zeiss Axio Imager II was chosen to create the image data set, and this was done in 5 sessions over a period of several months. The image dimensions and aspect ratio varied because of the variation in sample sizes, with heights of 1047 to 4819, and widths from 2676 to 5254 pixels. The slight differences in microscope set up, and lighting in particular led to some variation in colour balance between sets of images created on the same day. This issue would be solved by using a digital slide scanner, and while this was not an available option during this research project, it would be a useful technique to use in the future.

The potential differences between skin samples generated using chemical and pharmaceutical test compounds compared to those generated when the assay used for clinical assessment of donor patient pairs are worthy of consideration. The clinical samples appear to show the same GVHRs as are produced in the commercial assay, however there is a possibility that there may be some differences between the traditional GVHRs and those produced when a test compound or drug is present. Without the commercial dataset, it was not possible to rule out differences at the start of the research project. An additional set of 56 test images provided by Alcyomics towards the end of the study were all generated in the commercial assay. While there are no obvious visual differences in the clinical and commercial sample sets, visual comparison of the two image groups is a very subjective assessment.

While a multi-class classification system that classifies new images as either grade I, II, III or IV would be the preferred solution in this research, it was decided that the initial focus should be to develop an accurate binary classifier due to the challenges of the grading process. A binary classification provides the information required by Alcyomics to determine whether a test compound causes an immunogenic reaction. More specifically, when a manual grading is carried out, grade I is quoted as a negative result whilst a grade II, III or IV is determined to be a positive result. A positive result indicates that the test compound has caused a significant immune based reaction in the skin. This binary classification is a challenging problem as it is this classification that is the most difficult for a human grading manually to determine, mainly due to the lack of a clear differentiating feature between grade I and II damage.

This chapter has provided information on the image set used in the research and the optimisation of the process by which the images were digitised. The challenges of the dataset and the rationale behind the decision to focus on binary classification have also been presented. Finally an assessment of the current manual grading process was made. The next chapter will describe the development of the procedure used to pre-process the images and identify the regions of interest showing immune mediated damage.

## Chapter 5 Image Processing and Segmentation

In this chapter, the development of the methodology used to process the images and identify (or segment) the regions of interest is described. The segmentation of particular regions in an image is the fundamental basis on which all subsequent steps in the automation process depend.

The objective of the research project was to successfully classify input images into particular grades based on the severity of immune mediated histological reactions. The grade is based on the presence and extent of particular features including epidermal vacuolisation, clefts at the DEJ and dyskeratotic bodies. Successful classification relies on being able to identify these features within an image. The histological features associated with GVHRs are found exclusively within particular tissue types and hence the first step must be to locate the different tissue types in the skin sample, prior to identifying specific features. The image processing, segmentation and feature extraction process that has been developed is split into a number of separate operations and when carried out in series, the process segments the image in a hierarchical manner. This approach works on the basis of feature scale; the first segmentation splits the image into background and sample, the next splits the sample into epidermis and dermis, and the final segmentation splits the epidermis down into vacuoles, clefts and normal tissue, and the dermis down into clefts and normal tissue. The hierarchical process is summarised in Figure 5.1.

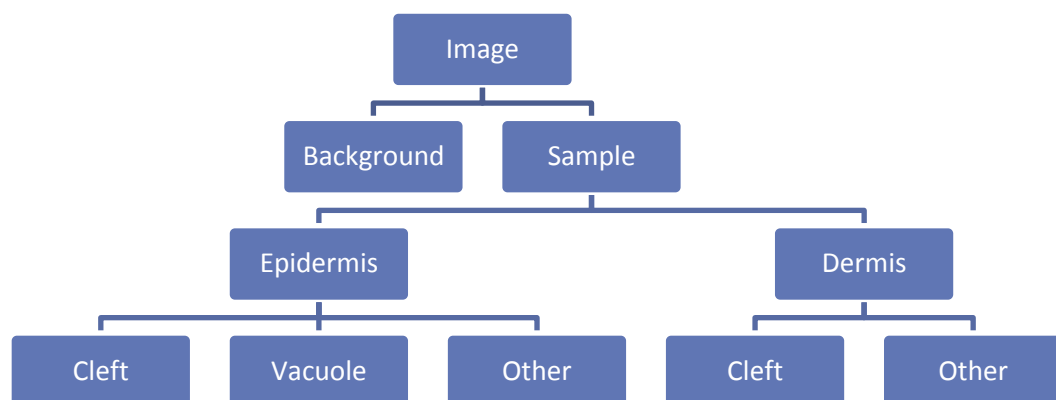


Figure 5.1 Hierarchical structure of segmentation process, starting with the whole image and resulting in the segmentation of the critical histological features, vacuoles and clefts.

A three-stage platform process for the segmentation of histological images is presented in this chapter. The initial method development and optimisation focussed on sample and epidermal segmentation and this is presented in section 5.1. The performance of this segmentation is evaluated in section 5.2. The outputs from the sample and epidermal segmentation step are used as the starting point for dermal segmentation, which is described in section 5.3. The modifications of the platform segmentation process required for the identification of clefts and vacuoles are described in sections 5.4 and 5.5 respectively. An alternative set of criteria to identify clefts and vacuoles is presented in section 5.6 and finally the research outputs are discussed in section 5.7.

There are many parameters which were optimised during the development of this process, some of which are dependent on the exact image spatial resolution and the specific staining and lighting properties of the image data set used to develop the process. In order to aid future development and application of the method, a list of all parameters which would needed to be re-optimised if images of a different spatial resolution or colour profile were being analysed are presented in Appendix C.

### **5.1 Sample and Epidermal Segmentation: Method Development**

The three-stage platform process for image segmentation consists of: (1) colour image pre-processing primarily for the purpose of contrast enhancement, (2) Otsu thresholding and (3) morphological processing and object classification of the binary segmentation mask. The proposed method is a novel approach to enabling highly variable sets of complex histopathological images to be segmented using the well-known Otsu thresholding method. Unlike the multi-resolution approach of Wang *et al* (2007a) which requires images at x2 and x40 magnification, this procedure can be performed on a single image at x10 magnification. The thresholding approach was improved by pre-processing the colour image prior to thresholding and post-processing the binary image produced by the thresholding operation. This is a similar approach to that described by Eramian *et al* (2011), who included a pre-segmentation colour standardisation and post-segmentation processing step based on domain specific rules.



The previously published methods for tissue segmentation all use classification or clustering of single pixels or pixel sets based on a feature vector of properties, including support vector machines (Y. Wang *et al.*, 2007a), Hierarchical Self-organizing Maps (Datar *et al.*, 2008) and a graph cut method (Eramian *et al.*, 2011).

The process was optimised and tested on whole slide images of H&E stained human skin sections that exhibited varying levels of histopathological damage including vacuolisation, sub-epidermal cleft formation, dyskeratosis and necrosis. The presence of varying levels of structural damage and image variation created by inconsistencies in tissue preparation and staining means accurate segmentation of specific tissue types is particularly challenging for this dataset. The development of a segmentation algorithm able to handle the challenges of this dataset is essential to the creation of a computer assisted process for histological grading.

Figure 5.2 summarises the main stages of the segmentation algorithm. The algorithm is based on the differences in the colour and intensity of staining in the different tissue types, the texture within each tissue and the overall shape and size of tissue regions. A colour normalisation step has also been included to handle colour variations in the images resulting from differences in sample thickness, staining procedure, and lighting during sample preparation and image acquisition. The text on the right of the figure describes the main functions of the processing steps shown in the flow diagram. The algorithm was implemented using the Image Processing Toolbox™ in MATLAB®, Version 7.11, R2010b (The MathWorks, Inc.).

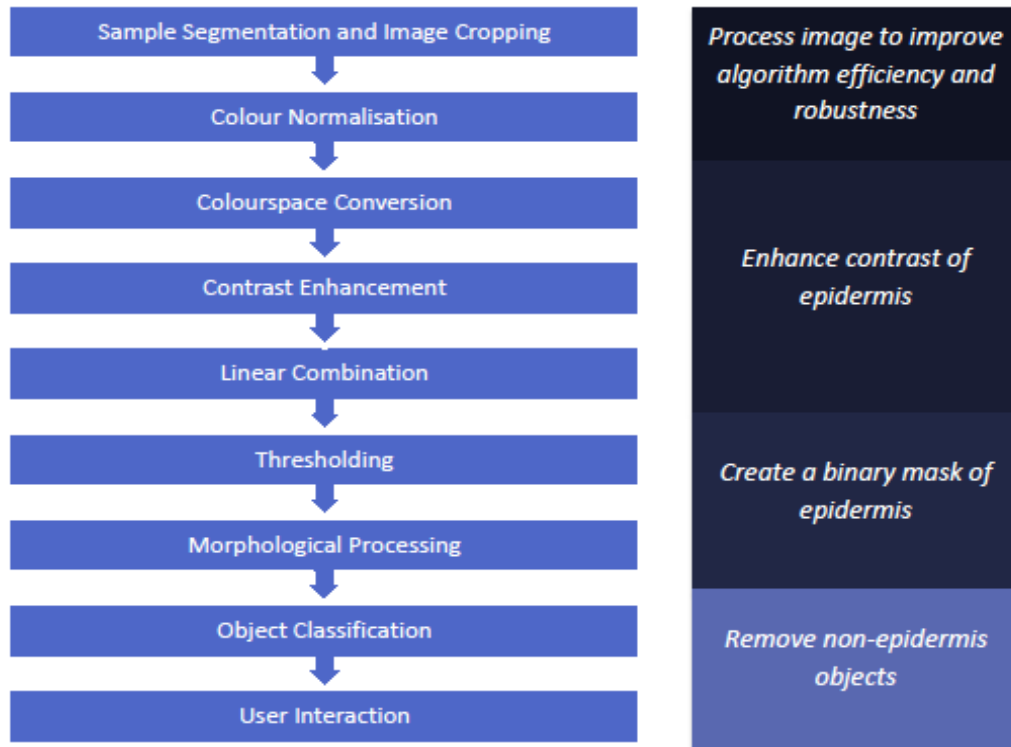


Figure 5.2 Main processing steps in the algorithm to segment the epidermis from a digital image of an H&E stained skin section. The text boxes on the right describe the function of the processing steps throughout the algorithm.

### 5.1.1 Sample segmentation and Image Cropping

The first stage in the algorithm is the segmentation of pixels in the image representing the skin sample. This first segmentation increases the efficiency of the algorithm by limiting the number of pixels being processed during subsequent steps. While segmentation of the skin sample could be achieved by locating either the background or the sample pixels, the background pixels are used as they have lower intra-image variance. Background pixels within a single image show very little variation in colour or intensity as there is no tissue present and illumination correction to remove variation resulting from the microscope lighting is performed during acquisition. The main variation present in the background is due to very small tissue fragments or dust. The background pixels are located by creating a composite image,  $K$ , which is the summation of the red, green and blue ( $R$ ,  $G$  and  $B$ ) intensities for each pixel in the  $RGB$  image (Eq. 5.1).

$$\mathbf{K} = \mathbf{R} + \mathbf{G} + \mathbf{B} \quad \text{Equation 5.2}$$

Black pixels are present at the image edges due to the image tiling procedure, as can be seen in Figure 4.12. The location of these black pixels is not fixed and so they are located by finding all zero elements in  $\mathbf{K}$ . The most frequently occurring value in  $\mathbf{K}$ , excluding the black pixels, is used to approximate the background colour and this value as the background threshold,  $bg_{thresh}$ . The black pixels at the image edge are replaced with the background threshold intensity,  $bg_{thresh}$ , to create a consistent background.

The calculation of  $bg_{thresh}$  can be written as:

$$bg_{thresh} = mode(\{k_{ij} | k_{ij} \text{ is an element of } \mathbf{K} \text{ and } k_{ij} > 0\}) \quad \text{Equation 5.3}$$

When the  $bg_{thresh}$  values of the first 50 images obtained were examined it was observed that while the majority of images had a  $bg_{thresh}$  of 640-710, there was a second group with values in the range of 531- 561. On closer examination it was found that all of the images with the lower  $bg_{thresh}$  values were taken on the same day; it is thus likely that the lighting on the microscope was set at a slightly different level. Figure 5.3 shows the  $bg_{thresh}$  values of the 50 images with the images taken on 31/05/2012 highlighted in red. This type of variation is typical in this application, and must be accounted for in solution developed. The fact that the difference in lighting can be seen in the  $bg_{thresh}$  value confirms the validity of the  $bg_{thresh}$  measurement.



Figure 5.3 Bar chart showing the mode composite intensity,  $bg_{threshv}$  values for 50 images.

To reduce memory requirements in the implementation of this algorithm, excess outer rows and columns of background pixels which do not intersect the sample are cropped. For an  $m \times n$  size image, this is done by cropping any rows where the sum of composite pixels in the row is less than  $bg_{thresh} * n$  (an approximation of the sum of values in one row of the composite image,  $K$ , assuming only background pixels are present), and cropping any columns where the sum of composite pixels in the column is less than  $bg_{thresh} * m$ . The result of cropping is shown in Figure 5.4 for a particularly challenging image showing grade IV damage. Images with many small tissue fragments are the most prone to errors using this approach as small fragments can be mistaken for normal background variation. The original image is shown in Figure 5.4a and in Figure 5.4b all excess background has been cropped without cropping more than a few pixels of tissue at the sample edges. The cropped sample pixels are either part of the *stratum corneum*, which is the only part of the epidermis that is not deemed important in GVHR, or those located at the cut edges of the sample, which are prone to artefacts and generally discounted from the analysis.

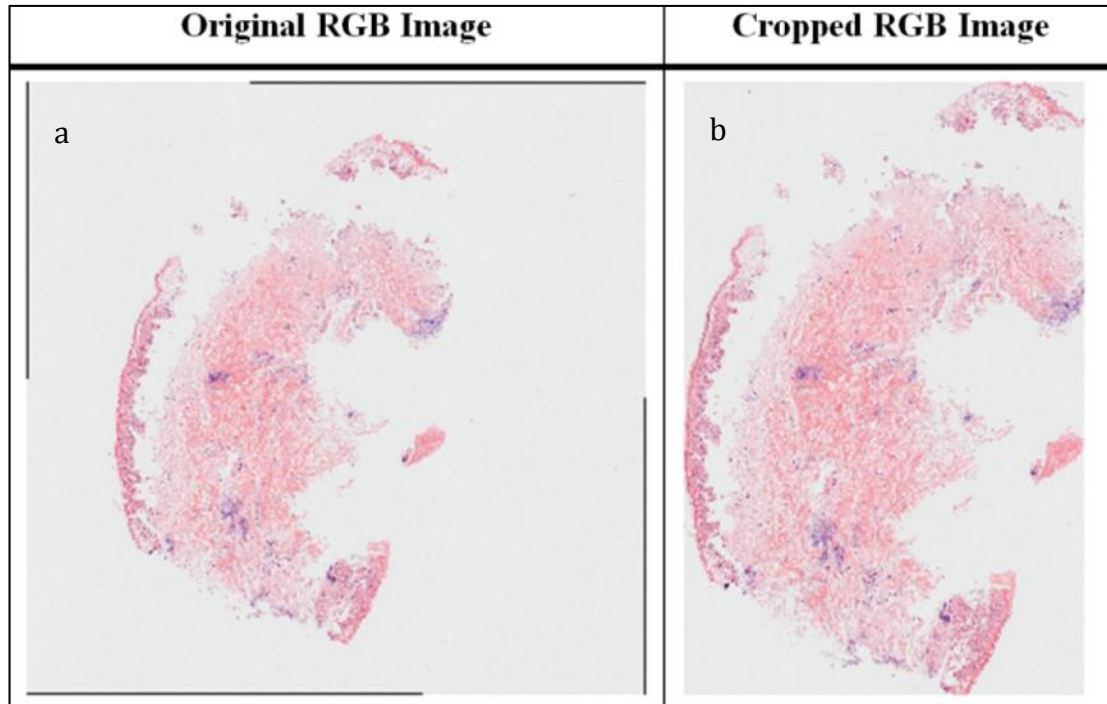


Figure 5.4 Effect of the automated image cropping procedure on an image which includes a number of small tissue fragments.

Prior to thresholding, the colour image is smoothed using a mean filter that replaces each image pixel with the mean value of its pixel neighbours in each colour channel. This approach was introduced in Chapter 2, section 2.4.8. The operation is performed using convolution with a kernel filter to represent the pixel neighbourhood ( $K_{smoothed} = K * \text{kernel}$ ). Mean filtering reduces variation within the background and sample pixel sets and facilitates the choice of threshold when creating the binary sample mask. A new binary image **sMask** is created by thresholding the smoothed image,  $K_{smoothed}$ . The value of the threshold was based on  $bg_{thresh}$ , however since  $bg_{thresh}$  is a measure of central tendency, the actual threshold used must be lower to ensure the majority of background pixels fall below it. Calculation of the appropriate threshold was based on the standard deviation of the background pixels in 40 smoothed composite images. The value lay between 1.2 and 3.2, and hence subtracting three standard deviations of the image with the highest standard deviation ( $3*3.2$ ) from  $bg_{thresh}$  will mean that 99.8% of the background pixels should be thresholded correctly:

$$sMask_{i,j} = \begin{cases} 1 & \text{if } K_{i,j} > bg_{thresh} - 9.6 \\ 0 & \text{if } K_{i,j} \leq bg_{thresh} - 9.6 \end{cases} \quad \text{Equation 5.4}$$

The effect of the changing the parameters in the mean filtering operation on the subsequent sample thresholding was investigated. The effect of the mean filter and its size on the thresholding operation is shown in Figure 5.5. A variety of kernel filter sizes were tested to determine their effect on the thresholding step. The figure shows the post-thresholding binary masks created after thresholding was performed with no smoothing, and when thresholding was performed after smoothing with 9x9, 29x29 and 49x49 mean filters. These sizes were empirically selected based on the typical size of vacuoles and clefts in the images. Although more filter sizes were tested, only three examples are shown.

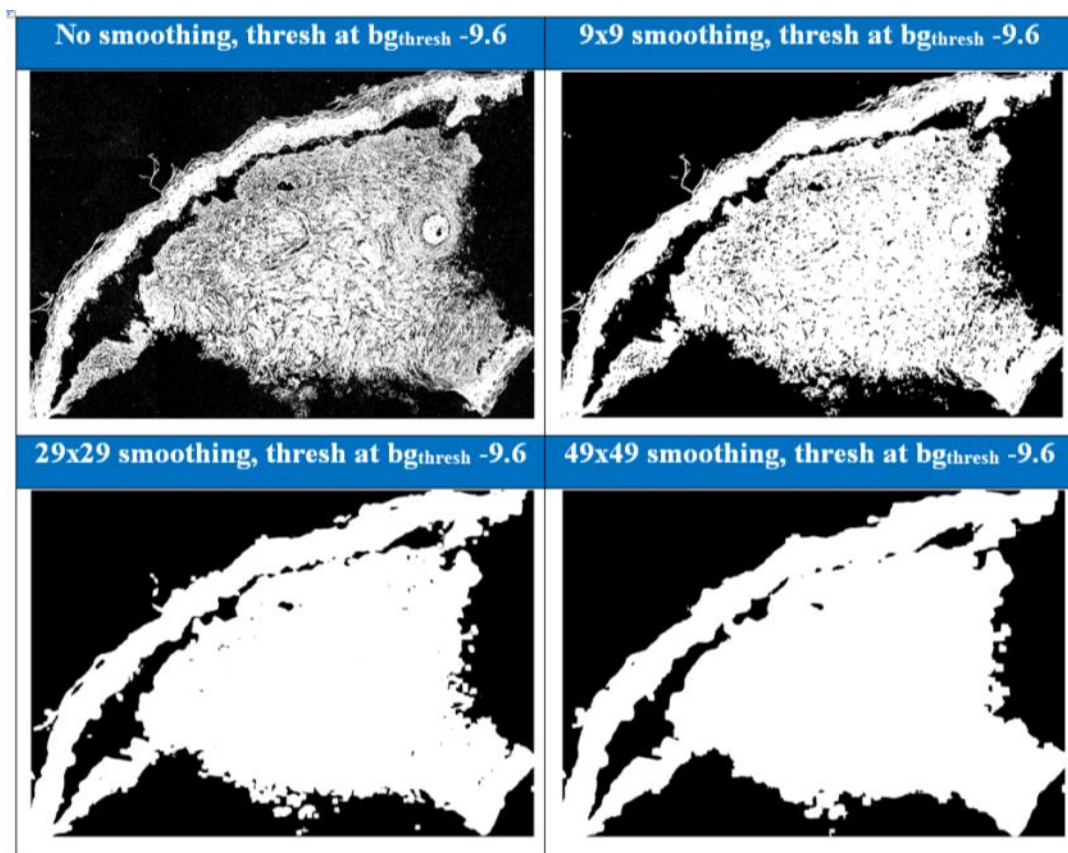


Figure 5.5 Effect of a pre-thresholding smoothing step on the subsequent thresholding operation. The figure shows the binary mask created by the thresholding operation without smoothing, and when the smoothing step is performed using a 9x9, 29x29 and 49x49 sized filter.

The dimensions of the mean filter must be large enough to smooth the coarse texture in the lower parts of the dermis and clefts at the DEJ so that these features are included in the sample mask, while still able to prevent the loss of accuracy at the perimeter of the sample. Figure 5.5 shows that without any smoothing, the thresholding results in a mask which is very detailed but does not include any

vacuoles or clefts in the white mask foreground. By smoothing, a simpler mask is created that includes increasingly more clefts, vacuoles and white regions within the dermis as the filter size is increased. While mean filters of sizes between 9 and 49 could be used successfully, an intermediate value of 29 was selected. The rationale for this was that the thresholded masks produced using the 29 x29 filter had fewer separate objects than when lower order filters were used and improved segmentation accuracy at the sample perimeter compared to the higher order filters.

The original RGB image and the image after smoothing with the 29x29 mean filter are shown in Figure 5.6. The filtering causes a reduction in variation within the internal parts of the tissue as a means of facilitating the thresholding of sample and background pixels. This can be seen as a strong blurring effect in the figure.

In some images the clefts at the DEJ are very large and the smoothing operation is not sufficient to include them as foreground objects in the *sMask*. This can be seen Figure 5.5, where the large clefts appear as black regions within the mask after thresholding. Mathematical Morphology (Chapter 2, section 2.4.13) is used to in-fill these “holes” in the binary sample mask and also to remove small objects such as dust or tissue debris on the slide which have been captured during thresholding, but which are not informative for subsequent analysis. The sequence of operations used to refine the segmentation is as follows.

- **Fill holes** - Fills internal regions of background pixels within foreground objects in the binary image using the MATLAB, *imfill* function, an implementation of morphological reconstruction described by Soille (1999).
- **Remove small objects** - Removes foreground objects that consist of less than 25,000 connected pixels. This value was chosen so that the smallest fragments of tissue typically found in the image set used in this research were not excluded, but the value was large enough to exclude smaller objects such as dust or other tissue debris.



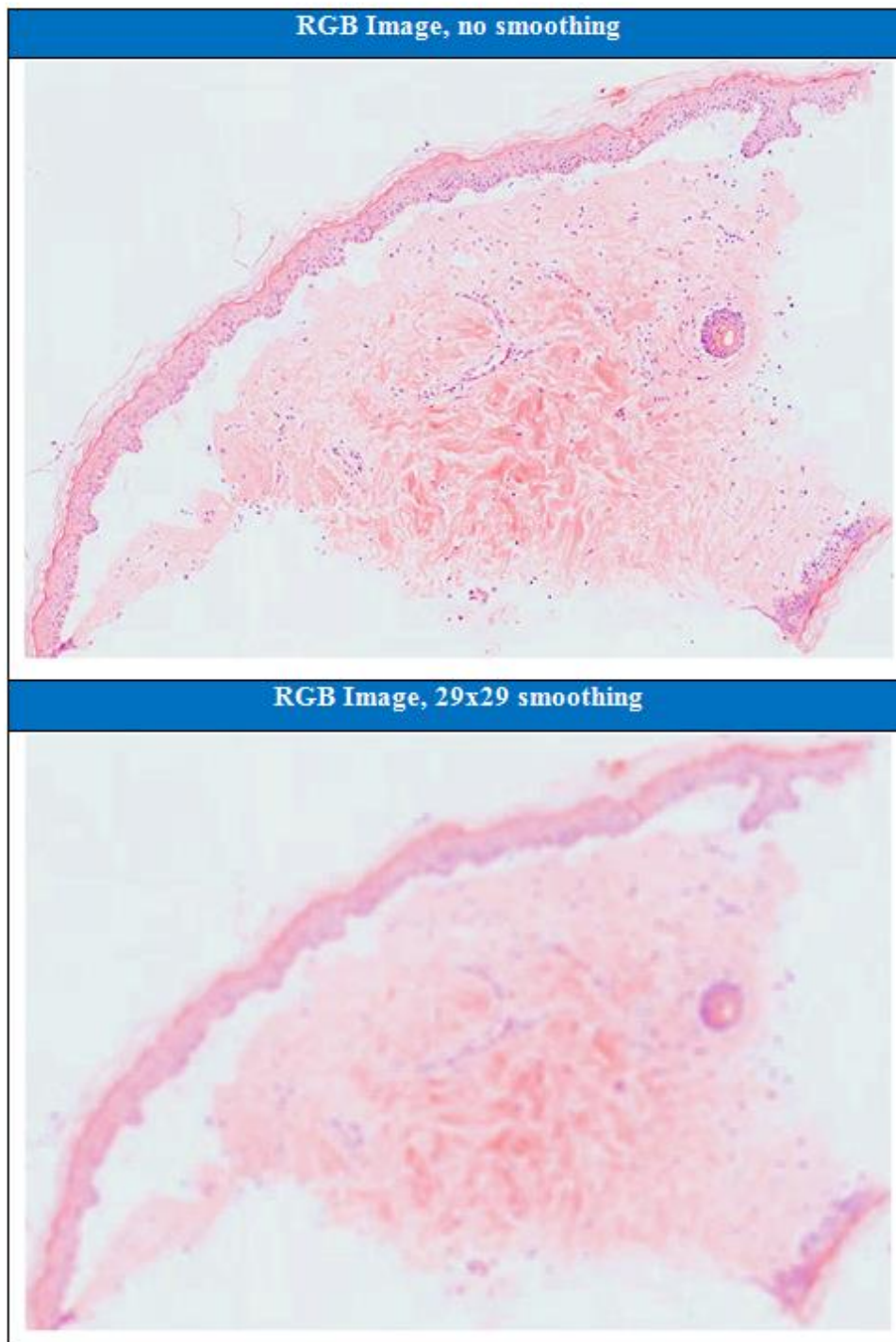


Figure 5.6 The effect of a mean filtering step using a 29x29 filter on an RGB image.

### 5.1.2 Colour Normalisation

The initial optimisation of the following epidermal segmentation method and parameters was carried out without colour normalisation, however the addition of this step was found to improve the performance of the epidermal segmentation in terms of sensitivity, specificity and overall accuracy. The relative improvement is



discussed in the results section 5.2.5. The mean filtering is used only to facilitate the segmentation of the sample pixels to create the **sMask**. The subsequent epidermal segmentation uses the original cropped RGB image retaining the fine resolution of the internal tissue texture. Staining inconsistencies in the input images are addressed by mapping the histogram for each individual colour channel of the cropped RGB image to those of a target image,  $I_{ref}$ , identified as well stained by an expert histopathologist. Only the sample pixels identified in the appropriate **sMask** are included in this colour normalisation step.

The colour normalisation is performed by application of a greyscale transformation,  $T$ , to all the sample pixel intensities,  $k$ , in the image. The pixels are located using the **sMask**. A transform is calculated for each colour plane in the RGB image so as to minimise the difference between the cumulative histogram,  $c_{input}$ , of the transformed input image intensities and the cumulative histogram  $c_{ref}$  of the well stained target image  $I_{ref}$ . The function to be minimised is:

$$|c_{input}(T(k)) - c_{ref}(k)| \quad \text{Equation 5.5}$$

This can be implemented in MATLAB using the function *histeq* (section 2.4.9).

The effect of the colour normalisation on two images with significant differences in staining and lighting is shown in Figure 5.7. The two original images are shown in column (a) in Figure 5.7 and the images with the normalised sample pixels are shown in column (b). The non-sample pixels have been changed to white in Figure 5.7b. The two normalised images have a similar contrast between the epidermis and dermis, and a similar range of colour hues and saturation.

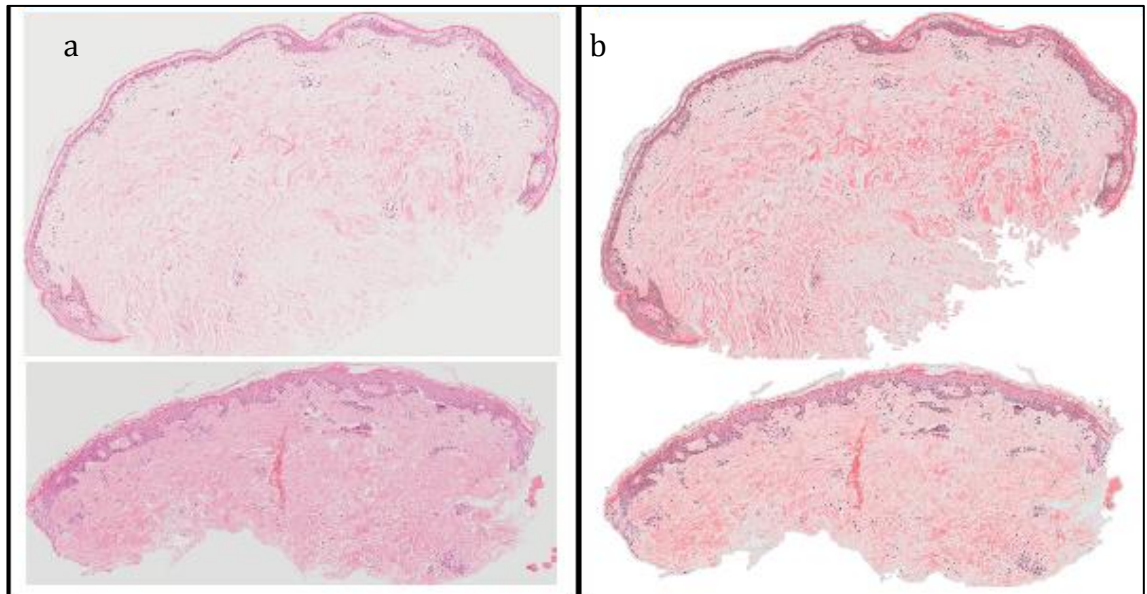


Figure 5.7 Effect of colour normalisation on RGB skin images showing two RGB skin images before and after colour normalisation with different staining contrast, lighting during acquisition, overall colour hues, and proportions of epidermis and dermis tissue. The non-sample pixels have been changed to white in the normalised images.

### 5.1.3 Colourspace Conversion

Following colour normalisation in the RGB colourspace, the next part of the segmentation procedure is a coarse segmentation of the epidermis based on the thresholding of a high contrast image. A number of colourspace were investigated to identify a representation that would maximise the contrast between the epidermis and the rest of the skin tissue. Those tested included **RGB**, **CMYK** (cyan, magenta, yellow and black) which is based on subtractive colour mixing, **HSV** (hue, saturation and value), **YCbCr** (luminance, blue chrominance and red chrominance) and the **L\*a\*b\*** (lightness, red/green, yellow/blue) colourspace. The contrast between the epidermis and dermis in each of the colourspace planes was assessed visually by two independent observers. The two observers, experienced in identifying the two tissue types, scored each of 20 images a 1 if the contrast was good and a 0 if the contrast was poor. The results are summarised in Table 5.1.

Table 5.1 Visual analysis of dermal epidermal contrast in 5 colourspace.

	RGB			HSV			YCbCr			CMYK				L*a*b*		
	R	G	B	H	S	V	Y	Cb	Cr	C	M	Y	K	L*	a*	b*
Scorer 1	8	6	8	0	4	12	8	18	6	8	2	0	0	10	6	15
Scorer 2	8	8	8	0	4	11	8	16	6	6	3	0	0	9	6	18
Total	16	14	16	0	8	23	16	34	12	14	5	0	0	19	12	33

The **Cb** channel in the **YCbCr** colour space and the **b\*** channel in the **L\*a\*b\*** colour space were selected as those providing the best contrast between the epidermis and the rest of the tissue. While this is a fairly subjective method, the two colour channels selected had significantly better contrast than the others tested. The **Cb** (blue chrominance) and the **b\*** (yellow/blue) colour channels both highlight the blue staining of haematoxylin which stains the nuclei in the cells of the epidermis. Although there are nuclei-containing cells present in the dermis, they are few in number. When the contrast enhancement was performed on a second set of 20 images selected to include images with different damage levels, staining and lighting levels, the **L\*a\*b\*** colour space enhanced the contrast of the tissue types more successfully than the **YCbCr** colour space. More specifically, when the contrast enhancement was performed on some of the images with lower overall illumination, the **Cb** channel of the **YCbCr** colour space did not enhance the contrast between the epidermis and dermis tissue as much as when the same procedure was performed on the images in the **b\*** channel of the **YCbCr** colour space. The **b\*** plane of the **L\*a\*b\*** colour space was therefore chosen for use in the algorithm due to its ability to show contrast between epidermal and dermal tissue despite differing levels of illumination.

It was noted that during optimisation of the subsequent contrast enhancement (section 5.1.4) and thresholding (section 5.1.6) stages that a contrast enhanced greyscale image (**G**) could provide useful additional information to the **b\*** colour channel. More specifically, the contrast enhanced greyscale images displayed good contrast in the few images where the **b\*** colour channel was displaying poor contrast. The images showing poor contrast of the epidermis in the **b\*** colour channel tended to have weak nuclear staining by haematoxylin, which appears as a strong blue/purple colour and therefore stands out in this yellow/blue colour channel. The complement image of the greyscale representation highlights more

intensely stained areas, but is not specific to a particular colour. It therefore tends to highlight both the pink cytoplasm and blue nuclei in the epidermis which are usually stained more intensely than the dermis tissue. When testing a set of 20 images, the image variation meant that in different images the epidermis was highlighted best in either the  $\mathbf{b}^*$  or the  $\mathbf{G}$  image, it was therefore decided that a combination of the data in the greyscale and  $\mathbf{b}^*$  images could be used to enhance the robustness of the following steps.

#### 5.1.4 Contrast Enhancement

Contrast enhancement was applied to the  $\mathbf{b}^*$  and the  $\mathbf{G}$  image to increase the intra-class variance of pixels in the epidermis and dermis. A linear transformation preserving the intensity histogram shape difference was selected (see section 2.4.9, equation 2.3). Only sample pixels were included in the contrast enhancement process, as the aim was to maximise the contrast between the dermis and epidermis, and background pixels are not relevant to the rest of the process. Remapping a narrow, more specific band of intensities was investigated to try to improve the contrast. When tested manually using a variety of absolute intensity levels as penetration points, the optimal intensity band for remapping to enhance contrast of the epidermis varied significantly for different images. This issue was addressed by determining penetration points based on the cumulative percentage histogram so that a set percentage of low and high intensity pixels were saturated in the final image. Removing a percentage of low and high pixel intensities is a better way of handling any remaining staining variation and pixel intensity outliers than choosing absolute intensity levels, and the approach was used on both the  $\mathbf{b}^*$  and greyscale image planes to create new contrast enhanced images,  $\mathbf{b}'$  and  $\mathbf{G}'$ . The optimal values for the upper and lower penetration points for the  $\mathbf{G}$  and  $\mathbf{b}^*$  images were determined using a Design of Experiments approach. This optimisation is described in detail in section 5.2.3. The usual 0 – 255 intensity scale is changed to a mapping from 0 to 1 for these steps as this is a requirement to perform these operations in Matlab. The optimal values determined in the DoE study are used in the remapping functions:

$$\mathbf{G}'_{i,j} = INT \left\{ \frac{1}{1 - 0.2743} [\mathbf{G}_{i,j} - 0.2743] \right\} \quad \text{Equation 5.6}$$

$$\mathbf{b}'_{i,j} = INT \left\{ \frac{1}{1 - 0.4034} [\mathbf{b}^*_{i,j} - 0.4034] \right\} \quad \text{Equation 5.7}$$

Following the contrast enhancement, the two images  $G'$  and  $b'$  were smoothed using an averaging mean filter, as described for sample segmentation in section 5.1.1. The operation is performed using convolution with a kernel filter to represent the pixel neighbourhood ( $\mathbf{K}_{smoothed} = \mathbf{K} * \text{kernel}$ ). This has the effect of reducing variation within the sample pixels and smoothing minor variations within the epidermis and dermis regions. This reduction in intra-class variation in the epidermis and dermis pixel sets was sought in order to emphasise the inter-class variation and facilitate the choice of threshold (section 5.1.6).

The size of this smoothing mean filter was optimised using the Design of Experiments study described in section 5.2.3. Based on the optimisation a mean filter size of 40x40 (where each element is  $1/(40*40)$ ) was chosen during the optimisation.

### 5.1.5 Linear Combination

Once the colourspace conversion and contrast enhancement of both images has been performed, the information from both must be combined in a single image to be thresholded to create a binary image. The binary image (sometimes referred to as a mask) contains information on regions of interest; in this case it will identify the location of epidermis pixels and non-epidermis pixels. A set of 30 images was used to assess the effect of different weightings in the linear combination and to find a threshold value. Three different linear combinations of the two enhanced and smoothed images  $G'$  and  $b'$  were tested ( $G'/2 + b'$ ,  $G'/1.5 + b'/1.5$ , and  $G' + b'/2$ ) before each was thresholded at 100. Using a qualitative visual assessment, an equal addition of the two images was found to result in good and specific segmentation of the epidermis for the highest number of images. The optimal combination is very much dependent on the staining properties of the individual image and it was decided that detailed optimisation using an image subset was

unlikely to prove useful. These parameters were not therefore included in the DoE optimisation.

The equal weighted linear combination applied to the two enhanced and smoothed images  $\mathbf{G}'$  and  $\mathbf{b}'$  to create a new image,  $\mathbf{Gb}$  is given as:

$$\mathbf{Gb}_{i,j} = 0.5 \times \mathbf{G}'_{ij} + 0.5 \times \mathbf{b}'_{ij} \quad \text{Equation 5.8}$$

Combining the two images captures both staining intensity and colour information in a single greyscale image. The effect can be observed in Figure 5.8.

In image 1 of Figure 5.8 the greyscale image shows good contrast between the epidermis on the right edge of the sample and the dermis. The  $b^*$  colour channel contains regions of high intensity in the epidermis and internal regions of the dermis. Combining the two colour channels retains high intensity in the epidermis and results in lower intensity and reduced intra-class variance in dermis. In image 2, the  $b^*$  colour channel shows the whole of the epidermis as high intensity, whereas the greyscale image does not have high intensity in epidermal regions which are not stained as intensely (indicated by green arrows). A similar difference can be seen in image 3. It is important that these regions of less intense staining are included in the epidermal mask as they are often areas with significant vacuolisation or cleft formation. Although the difference between the greyscale and  $b^*$  colour channels varies between images, a general effect is that the combination of the two data sources has the effect of cancelling out some of the intra-class variance, an effect that helps to maximize inter-class variance and facilitate the subsequent thresholding step.

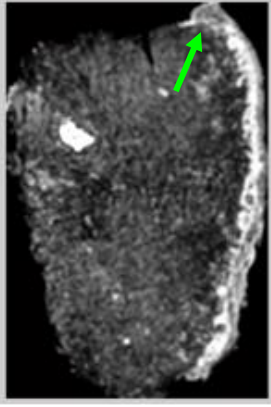
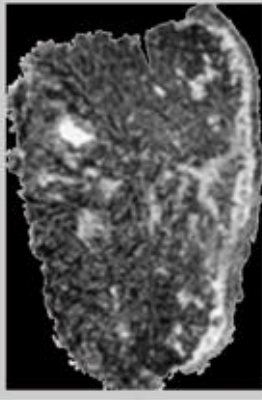
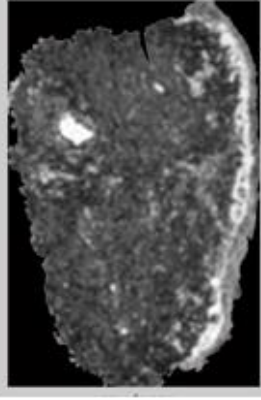
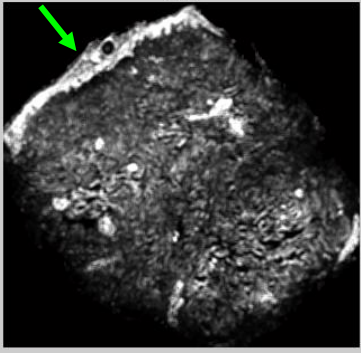
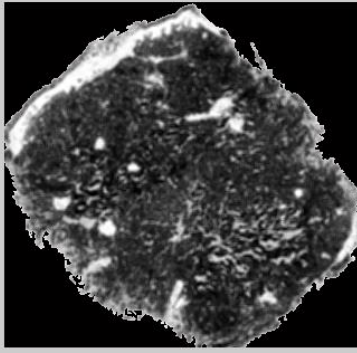
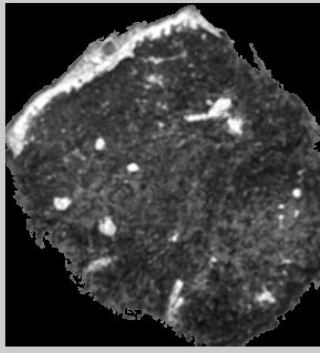
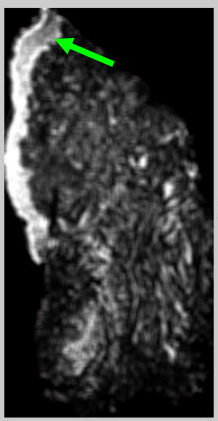
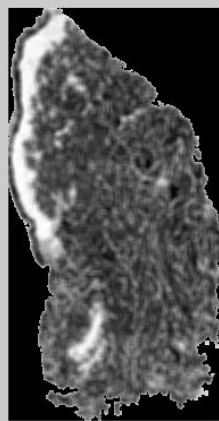
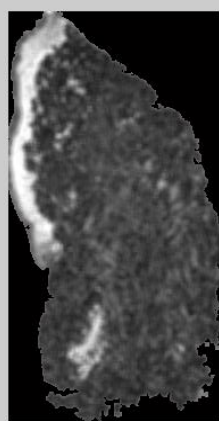
	Greyscale image	b* image	Equal weighted combination
Image 1			
Image 2			
Image 3			

Figure 5.8 Effect of linear combination of three sets of greyscale and b\* images.

### 5.1.6 Thresholding

Following colour normalisation and contrast enhancement, Otsu's automated thresholding method (section 2.4.12) was applied to determine the optimal threshold based on the intensity distribution of sample pixels in the combined, enhanced image,  $GB'$ . The Otsu method uses discriminant analysis to determine a threshold,  $t$ , which maximises the separability of two pixel classes by minimising the intra-class variance. In this process, the aim is to maximise the separability of

the dermis and epidermis pixel sets.  $GB'$  is converted into a binary image,  $BW$ , using the threshold,  $t$ . Any non-sample pixels are changed to black (as background):

$$BW_{i,j} = \begin{cases} 1 & \text{if } GB'_{i,j} > t \\ 0 & \text{if } GB'_{i,j} \leq t \end{cases} \quad \text{Equation 5.9}$$

Figure 5.9 shows a histogram of a typical  $GB'$  image. The method assumes an approximately bimodal distribution. The Otsu threshold,  $t$ , attained using the method described in section (section 2.4.12) is labelled in the figure at the intersection of the upper and lower intensity components. The threshold does not intersect at the valley of the two intensity peaks, but at a mid grey level of 137. The higher intensity pixels (between 215-255) represent the cell nuclei, which are mainly found in the epidermis, however the cytoplasm and weakly stained nuclei are also above the threshold in this case.

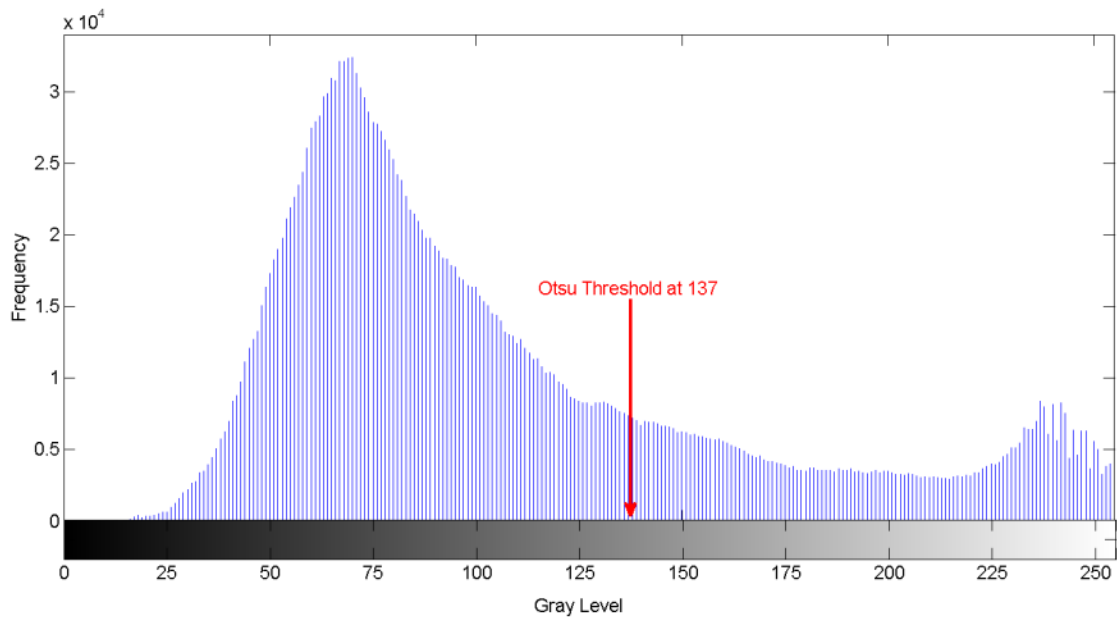


Figure 5.9 Histogram of enhanced additive image showing Otsu threshold.

### 5.1.7 Morphological Processing

Morphological processing is used to further process the binary image,  $BW$ , by removing small misclassified objects such as groups of cells within the dermis, merging multiple objects, in-filling holes and closing gaps. These operations are applied to the whole image, however once the operations are completed, any non-sample pixels which may have been affected are reverted to black. The sequence of



operations summarised below is used to refine the segmentation. The choice of structuring element size (radius = 6) for the morphological closing and opening steps was optimised based on the final sensitivity and specificity of the algorithm as described in section 5.2.3. A disk shaped structuring element was used as this shape reflects biological structures more accurately than sharp angles or linear shapes.

- **Morphological closing** - Morphological closing (dilation then erosion) enlarges the boundaries of foreground (bright) objects in the image and closes gaps between them, and shrinks background-coloured holes in the foreground objects. A disk shaped structuring element with *radius = 6 pixels* is utilised.
- **Morphological opening** - Morphological opening (erosion then dilation) removes some of the foreground (bright) pixels from the edges of foreground objects, breaking fine bridges between objects while preserving the object size. A disk shaped structuring element with *radius = 6 pixels* is utilised.

Steps 1 and 2 combine to smooth the objects edges without changing the size of objects. Smooth object perimeters are more reflective of the tissue edges seen in the real images.

- **Remove small objects** – Removal of foreground objects that comprise fewer than 4000 connected pixels. The threshold of 4000 pixels was selected based on the number of pixels contained within the regions of dermis identified incorrectly as epidermis objects prior to this step. The majority of correctly identified epidermis regions at this stage included more than 4000 pixels.
- **Fill holes** – In-fills internal regions of background pixels within foreground objects in the binary image that comprise fewer than 7000 connected pixels. A threshold is required as in some images there are regions of dermis tissue surrounded by epidermis tissue (due to tissue slicing technique) and if these regions are filled the specificity of the final algorithm is compromised. Again the value was selected based on the typical size of enclosed dermis regions within the epidermis which were misclassified.

### 5.1.8 Object Classification

Following morphological processing, the binary mask,  $BW$ , includes objects that are not part of the epidermis. These include collections of cells within the dermis that have a similar appearance to epidermis tissue and parts of the dead surface layer, the *stratum corneum*, which can be segmented with the epidermal tissue in cases where it is highly stained. For each object,  $Z$ , the object area,  $Z_{Area}$ , and the area of the object's bounding box,  $Z_{BoundingBox}$ , are determined (for definition, see section 2.5.1). The ratio of  $Z_{Area}$  to  $Z_{BoundingBox}$  gives the extent,  $Z_{Extent}$ , of the object:

$$Z_{Extent} = \frac{Z_{Area}}{Z_{BoundingBox}} \quad \text{Equation 5.10}$$

The  $Z_{Area}$  and  $Z_{Extent}$  can both be used to classify the remaining objects as either epidermis or non-epidermis. While the area provides information on region size, the extent is a shape based measure that can differentiate between the long thin objects of the epidermis and the more compact, circular clusters of cells within the dermis. Including the area measurement prevents very small regions being classified as epidermis.

The thresholded objects are either retained or removed based on their area and extent and hence the impact of adjusting the area and extent thresholds on the sensitivity and specificity of the algorithm was investigated. The two thresholds are critical values, and the exact values were determined in an optimisation study described fully in Section 5.2.4. The values used in this classification were determined once all the other critical parameters had been set. The parameters determined first were the upper and lower histogram penetration points used for the greyscale and  $b^*$  contrast enhancement, the size of this smoothing mean filter used after the contrast enhancement, and the size of the SE used for morphological processing after thresholding. Based on the optimisation, the following classification rule was used to classify each object pixel,  $z$ , in the binary mask:

$$z = \begin{cases} 1 & \{ \forall z \in Z \mid Z_{Extent} < 0.44, Z_{Area} > 20000 \} \\ 0 & \text{else} \end{cases} \quad \text{Equation 5.11}$$

where  $z$  are the pixel elements in the object  $Z$ .

The final binary mask showing the location of the epidermis pixels, ***eMask***, is the image, ***BW***, which has been morphologically processed using the steps in section 5.1.7, and then subjected to further processing by the conversion of any pixels, *z*, to either 1 or 0 based on Equation 5.11.

### 5.1.9 User Interaction

The object classification step can be used to fine tune the specificity and sensitivity of the final algorithm, however the algorithm also includes the option for the user to interact with the programme and select or remove objects in the final epidermis mask. Epidermis segmentation is critical to the performance of the subsequent steps in the skin damage classification process and hence this optional interaction step is included to improve the performance of the algorithm if required. It is relatively straightforward for a user to determine whether a given object is part of the epidermis when shown next to an image of the RGB image. This is confirmed by the expert histopathologists at Alcyomics, who agree that identifying epidermis and dermis tissue can be mastered by a non-expert after a short period of training looking at a selection of skin images. The user has the option to (1) approve the object selection, (2) remove objects that are incorrectly classified, or (3) select additional objects, which were removed during the object classification step described in section 5.1.8. For a fully automated process, this step can be excluded. The effect of the user interaction step on algorithm performance is given in section 5.2.5.

## 5.2 Epidermal Segmentation Optimisation and Evaluation

This section first describes the optimisation of six key parameters in the epidermal segmentation algorithm. They are the upper and lower histogram penetration points used for the ***G'*** and ***b'*** contrast enhancement, the size of this smoothing mean filter used after the contrast enhancement, and the size of the SE used for morphological processing after thresholding. Following this optimisation, the performance of the final epidermis segmentation approach was evaluated with and without the optional user interaction step.

The optimisation and final evaluation were performed using “ground truth” images created through the manual mark-up of the epidermis in a set of images made up of equal numbers of grade I, II, III and IV images with varying staining and lighting. The 40 image set included 25 images used for the initial optimisation and an additional 15 “validation” images for evaluation.

Before the optimisation and evaluation is described, the procedure for generating the ground truth images and performance metrics is given.

### **5.2.1 Generation of a Ground Truth Data set**

The manual mark-up was achieved by drawing the boundary of the epidermis onto the original **RGB** images in green with the aid of a graphics tablet (Wacom Bamboo Fun S Pen and Touch Digitiser). The high colour contrast boundary was easily identified using a thresholding procedure on the red channel of the RGB image. The outlined regions were flood-filled with a morphological reconstruction algorithm implemented using the MATLAB function, *imfill*. The *stratum corneum*, the epidermal surface layer which appears as a looser collection of flaky layers was excluded from the manual epidermis mark-up as it consists of dead cells that do not provide useful information about the state of damage in the tissue. Evaluation of the segmentation procedure was undertaken by comparing the area of the algorithm-segmented epidermis with the “true” epidermis area generated during manual segmentation.

### **5.2.2 Performance Metrics**

The total number of pixels identified as part of the epidermis in the manual mark-up and generated segmentation mask, and the total number of pixels in each image were used to determine:

- True positive – Epidermis in mark-up, epidermis in generated mask.
- True negative – Not epidermis in mark-up, not epidermis in generated mask.
- False positive – Not epidermis in mark-up, epidermis in generated mask.
- False negative – Epidermis in mark-up, not epidermis in generated mask.

The total pixel number in the image was based on the cropped image, to avoid an excessive number of background pixels skewing the results. This is the case because although both background pixels and dermis pixels are non-epidermis pixels, the background pixels are more likely to be classified correctly as non-epidermis pixels than dermis pixels due to the simple segmentation of background and sample pixels. If  $A_s$  and  $A_t$  represent the pixel sets identified as epidermis by the algorithm and manual methods respectively, the various fractions can be calculated as follows:

$$\text{False Negative (FN)} = (\text{ImageArea} - A_s) \cap A_t \quad \text{Equation 5.12}$$

$$\text{False Positive (FP)} = A_s \cap (\text{ImageArea} - A_t) \quad \text{Equation 5.13}$$

$$\text{True Positive (TP)} = A_s \cap A_t \quad \text{Equation 5.14}$$

$$\text{True Negative (TN)} = (\text{ImageArea} - A_s) \cap (\text{ImageArea} - A_t) \quad \text{Equation 5.15}$$

These fractions were used to calculate the percentage sensitivity, specificity and accuracy of the automated segmentation for each image by comparing the algorithmic method to manual segmentation. The three metrics were calculated as follows:

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \times 100 \quad \text{Equation 5.16}$$

$$\text{Specificity} = \frac{TN}{(TN + FP)} \times 100 \quad \text{Equation 5.17}$$

$$\text{Overall Accuracy} = \frac{(TP + TN)}{(TP + FP + FN + TN)} \times 100 \quad \text{Equation 5.18}$$

In combination, the three metrics provide an indication of the performance of the segmentation algorithm. Sensitivity is a measure of the algorithms ability to identify epidermis pixels, while specificity measures the ability to identify non-epidermis pixels. In the skin explant assay, a balance between sensitivity and specificity is required. Typically an increase in one will lead to a decrease in the other. The accuracy measurement combines the two metrics within one measurement, and quantifies the percentage of pixels correctly classified as epidermis and non-epidermis when compared to manual segmentation.

### 5.2.3 Optimisation of Algorithm Parameters

The six key parameters in the algorithm were optimised by executing the algorithm without user interaction and assessing the effect that changing their values had on the mean sensitivity and specificity. The sensitivity and specificity were calculated as described in section 5.2.2. The algorithm was optimised using 25 of the H&E stained skin sections. The optimisation was carried out using a Design of Experiments (DOE) approach using the software program MINITAB v16.2.4. This approach was adopted so that the interaction between the various parameters could be assessed. Initially a 2-level Fractional Factorial design was used to screen the six parameters (called factors in DOE). Factorial designs change two or more factors in a single experiment and they are used to determine the effect of multiple variables on a single response, or output. A full factorial study investigating six factors at two levels would require  $2^6$ , or 64 experiments (or in this case algorithm executions). Fractional factorial experiments use a carefully prescribed and representative subset of a full factorial design to reduce the number of experiments required. A Resolution IV design was utilised, in which  $\frac{1}{4}$  of all possible factor combinations were tested. The design included 16 experiments where the six factors were set at high or low levels, and one with the factors set at the mid-point between the low and high levels (called a centre point).

The high and low levels were selected by iteratively changing each factor on a set of 20 images and assessing the outcome visually. For example, for the ***b'*** image penetration points were selected to accentuate the blue/ purple pixels of the keratinocyte cell nuclei within the epidermis, while the ***G'*** image values were selected to highlight the whole of the epidermis including the cytoplasm and cell membranes. The upper and lower size limits of the smoothing mean filter were selected to reduce in variation within the epidermis and dermis pixel sets.

Morphological processing using the SE aims to smooth object edges to create a more biologically meaningful representation. To do this, the upper and lower size limits of the structuring element (SE) were set by measuring the pixel dimensions of cells within the tissue images. Figure 5.10 shows the dimensions of normal and vacuolised cells within the epidermis and maximum and minimum values chosen to cope with biological variation.

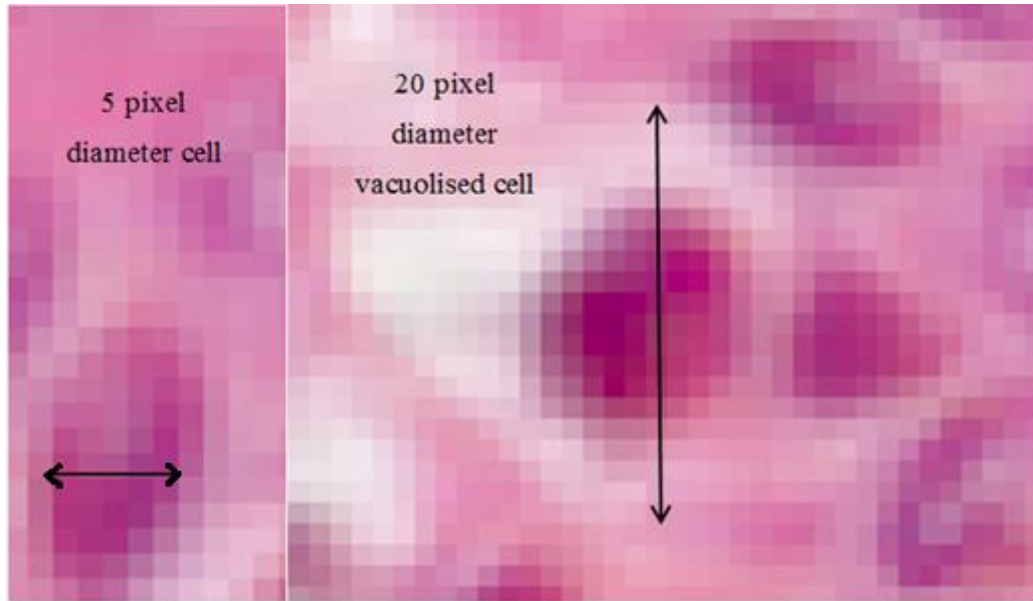


Figure 5.10 Enlarged sections of images showing H&E stained epidermal cells. The arrows highlight the diameter of normal and vacuolised cells

Each set of experimental conditions was used to test epidermal segmentation on 25 images. The values of the factors used for an initial screening run and the mean sensitivity and specificity of epidermal segmentation for the 20 images are shown in Table A in Appendix B. The best performing sets of factors were identified in runs 2, 11, 15 and 16, which all had sensitivity values of > 73% and specificity values of > 97%. These runs are starred in the first column of Table A.

The effects and coefficients of the main effects and interactions were analysed to determine the relative strength of the factors (the effect is two times the coefficient). The effect data is shown in Table 5.2. The factors with effects of the greatest magnitude (positive or negative) have the greatest effect on the responses. Varying the factors within the range tested had a small effect on specificity with all of the runs resulting in specificities of between 93% and 99%. The effect of the factors on sensitivity was much greater. The B term, which is the upper threshold for contrast enhancement of the G' image, had the greatest effect on both specificity (1.99) and sensitivity (21.85), increasing both outputs when set at the higher level. The next most important effects on sensitivity were the positive effect of increasing A (the lower G' threshold) and the negative effect of increasing F (the size of the SE). The most important interaction effects on sensitivity were A\*B and A\*D. The sensitivity model had an R<sup>2</sup> value of 99.45% and the specificity

model had an  $R^2$  value of 96.08%. This measure describes the amount of variation in the observed response values explained by the factors in the model; the high values indicate the models explain the data well.

*Table 5.2 Effects and Coefficients for Factors and Interactions in the Sensitivity and Specificity Screening*

<b>Term</b>	<b>Factor Description</b>	<b>Specificity Effect</b>	<b>Sensitivity Effect</b>
<b>Constant</b>			
<b>A</b>	Lower penetration point, $G'$	-0.69	13.30
<b>B</b>	Upper penetration point, $G'$	1.99	21.85
<b>C</b>	Lower penetration point, $b'$	-0.62	1.52
<b>D</b>	Upper penetration point, $b'$	0.12	0.95
<b>E</b>	Mean Filter Kernel Size	-0.08	6.50
<b>F</b>	Radius of SE	0.86	-13.71
<b>A*B</b>	Interaction of Factors A and B	1.31	-4.89
<b>A*C</b>	Interaction of Factors A and C	0.32	-1.05
<b>A*D</b>	Interaction of Factors A and D	0.14	4.91
<b>A*E</b>	Interaction of Factors A and E	0.05	-0.03
<b>A*F</b>	Interaction of Factors A and F	-0.30	2.82
<b>B*D</b>	Interaction of Factors B and D	-0.28	0.27
<b>B*F</b>	Interaction of Factors B and F	-0.86	2.12

The p-values in the analysis of variance table were used to find the statistically significant effects. Considering the interaction effects first, none of the interactions had p-values lower than the threshold of 0.05 usually used to indicate significance. The interaction effect p-values were between 0.086 and 0.919 for specificity and between 0.094 and 0.989 for sensitivity. For the main effects, only factor B had a p-value  $< 0.05$  for specificity and factors A, B and F had p-values  $< 0.05$  for sensitivity.

The results of the screening study indicated that only factors A, B and F were significant, however a feature of fractional factorial designs is the presence of confounding, which means that one or more of the effects cannot be estimated separately from each other and are said to be aliased. In the Resolution IV design



used the main effects are confounded with three-way interactions and two-way interactions are confounded with other two-way interactions. For example, it cannot be determined whether an effect is due to factor A or the combined effect of factors B, C and D. It was decided that a more detailed experiment was required to investigate interactions given the potential confounding masking complex interactions. The screening study indicated that factor B should be set at the higher level and as the value of 1 used for the upper threshold was the highest possible, this factor was fixed in the optimisation study.

A more detailed optimisation was undertaken on factors A, C, D, E and F, fixing the upper grey threshold (factor B) at the higher value of 1, but varying the other five factors using a Response Surface Design. This type of design allows the optimisation of one or more outputs which are influenced by several independent variables (factors). The ability of this design to detect curvature (nonlinearity) in the responses means it can be used to find the factor settings which optimise the responses. This second optimisation considered additional levels and combinations, including extreme points to investigate the relationship between the factors, including possible interactions and curvature in the data. A Central Composite Design was selected that uses the two level factorial design as a base with additional axial points to investigate extreme conditions and centre points to enable curvature and second order responses to be investigated. A representation of the initial cube points, the centre point and the axial points is shown in Figure 5.11 for a two factor, two level study.

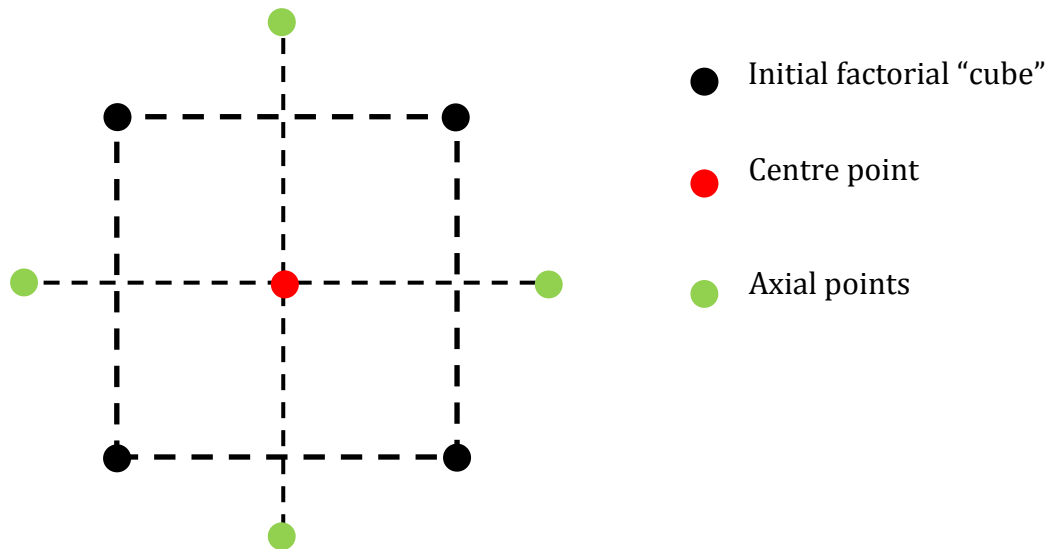


Figure 5.11 Representation of box, centre and axial points in a Central Composite Design

A full quadratic response surface model was fitted to the data using MINTAB. Considering specificity first, the significant square and interaction terms (with p-values < 0.05 in the Analysis of Variance table) were E\*F, A\*A, E\*E and F\*F and the significant linear terms were A, C, E and F. In the coefficients table the interaction term E\*F had a p-value of 0.000, the squared terms A\*A, E\*E and F\*F had p-values of 0.001, 0.000 and 0.009 respectively, and the linear terms A, C, E and F had p-values of 0.000, 0.003, 0.000 and 0.001 respectively. Variance inflation factors (VIF) can be used to evaluate correlation between factors. VIFs inflate the variance of the coefficients and although theoretically a VIF > 1 could indicate correlation between factors, in practice values < 5 tend to indicate that the estimation of the regression coefficient is acceptable. All VIF values were between 1 and 2, giving reasonable confidence in the results.

For sensitivity the significant square and interaction terms (with p-values > 0.05 in the Analysis of Variance table) were A\*F, C\*E, E\*F, A\*A, E\*E and F\*F and the significant linear terms were A, C, E and F. In the coefficients table the interaction terms A\*F, C\*E, E\*F had p-values of 0.000, 0.037 and 0.000 respectively, the squared terms A\*A, E\*E and F\*F had p-values of 0.000, 0.000 and 0.000 respectively, and the linear terms A, C, E and F had p-values of 0.000, 0.007, 0.000 and 0.000 respectively. All VIF values were between 1 and 2.

Squared and interaction terms which were insignificant for both specificity and sensitivity were removed sequentially, starting with the term with the highest p-value. This process was performed sequentially as the significance of terms can change as insignificant terms are removed and the model becomes more accurate. After all the squared and interaction terms with p-values > 0.05 had been removed the following terms were still included in the model:

- Linear – A, C, D, E and F
- Squared – A\*A, E\*E and F\*F
- Interaction – A\*E, A\*F, C\*E and E\*F

The only insignificant term remaining was the linear term, D, which had a p-value of 0.822, and so this term was removed. The final regression equations for sensitivity and specificity were:

*Mean sensitivity*

$$\begin{aligned}
 &= 64.28 + 51.9A - 5.17C + 0.512 - 0.927F \\
 &- 90.0(A^2) - 0.00947(E^2) - 0.03438(F^2) \\
 &- 0.384(AE) + 1.972(AF) + 0.02001(EF)
 \end{aligned}
 \tag{Equation 5.19}$$

*Mean specificity*

$$\begin{aligned}
 &= 94.885 + 15.28A - 1.60C + 0.1001 - 0.0588F \\
 &- 25.13(A^2) - 0.001802(E^2) - 0.002715(F^2) \\
 &- 0.0309(AE) + 0.0120(AF) + 0.003500(EF)
 \end{aligned}
 \tag{Equation 5.20}$$

The models for specificity and sensitivity were analysed next. For sensitivity, the R<sup>2</sup> value showed that 95.67% of the variation in specificity was explained by the model. The predicted R<sup>2</sup> highlights potential overfitting when it is much lower than the R<sup>2</sup> and it was 88.66% for sensitivity, which indicates a good model. For specificity, the R<sup>2</sup> value showed that 91.02% of the variation in specificity was explained by the model, however the predicted R<sup>2</sup> was 20.07%. This may indicate that the model requires further simplification, however it is not of great concern as specificity was varying within the small range of 95.29% and 98.24%. The more important consideration was to find conditions which increased sensitivity, which varied between 48.56% and 81.44% during the optimisation runs. Table 5.3 shows

the effect of each term in the final model on sensitivity and specificity. The magnitude of the effect is greater for sensitivity, and the most influential effects are A, E and F which also show curvature and interaction. The magnitude of the effects on specificity is much smaller.

*Table 5.3 Effect of each term in the final model on sensitivity and specificity*

<b>Term</b>	<b>Factor Description</b>	<b>Sensitivity Effect</b>	<b>Specificity Effect</b>
<b>A</b>	Lower penetration point for G'	5.81	0.89
<b>C</b>	Lower penetration point for b'	1.84	-0.30
<b>E</b>	Order of Mean Filter kernel	4.95	0.53
<b>F</b>	Radius of SE	-11.87	-0.29
<b>A*A</b>	Curvature of A	-1.80	-0.50
<b>E*E</b>	Curvature of E	-1.90	-0.36
<b>F*F</b>	Curvature of F	-3.87	-0.31
<b>A*E</b>	Interaction between A and E	-0.77	-0.06
<b>A*F</b>	Interaction between A and F	2.96	0.02
<b>C*E</b>	Interaction between C and E	0.87	-0.02
<b>E*F</b>	Interaction between E and F	3.00	0.53

Residuals show the difference between observed and fitted response values and trends in residuals can indicate if underlying assumptions of the model have been satisfied and highlight problems with the model. Figure 5.12 and Figure 5.13 show four residuals plots for the sensitivity and specificity models respectively. The normal probability graph plots the actual residuals versus their expected values when the distribution is normal; it can highlight non-normality, skewness, outliers and unidentified variables. A normal distribution is an underlying assumption of this analysis, however some deviations are typical. The histogram of the residuals shows their distribution and provides information on the spread, variation and distribution of the data and can be used to identify unusual values or outliers. The plot of residuals versus fitted values is used to look for constant variance, another assumption of this analysis, which should result in residuals scattered randomly around zero. This plot can also be used to highlight missing higher order terms, outliers or influential points. Finally, the residuals are plotted versus run order.

This plot is useful for identifying systematic effects in the data over time or data collection, and is particularly useful if runs are not randomised. In this analysis the runs were not randomised because they were being run automatically on a computer, and the same conditions run at a different time would always give the same result.

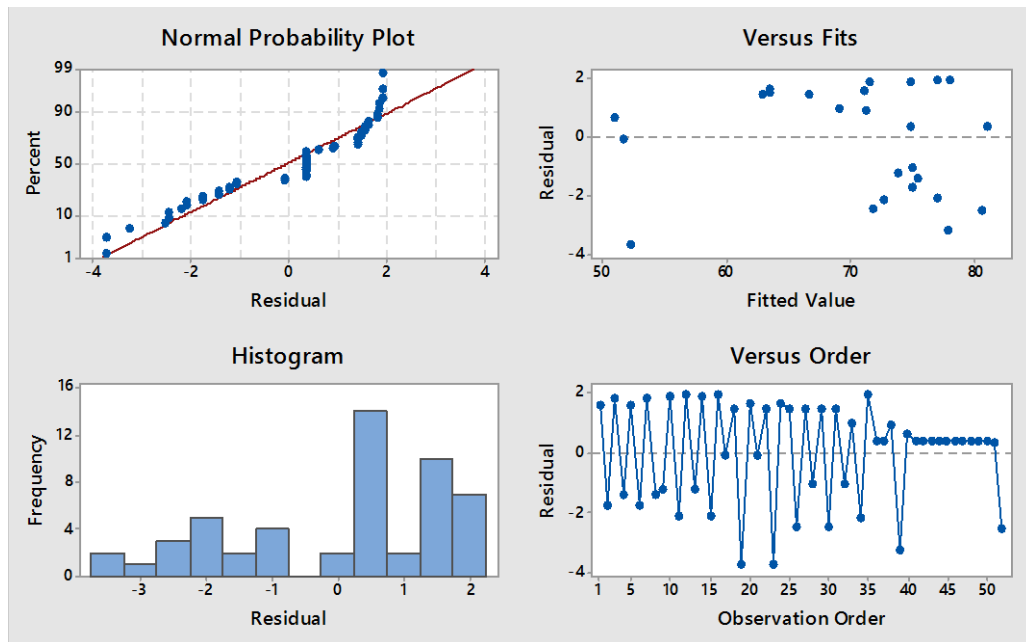


Figure 5.12 Residuals plots for model of key factor effect on mean sensitivity

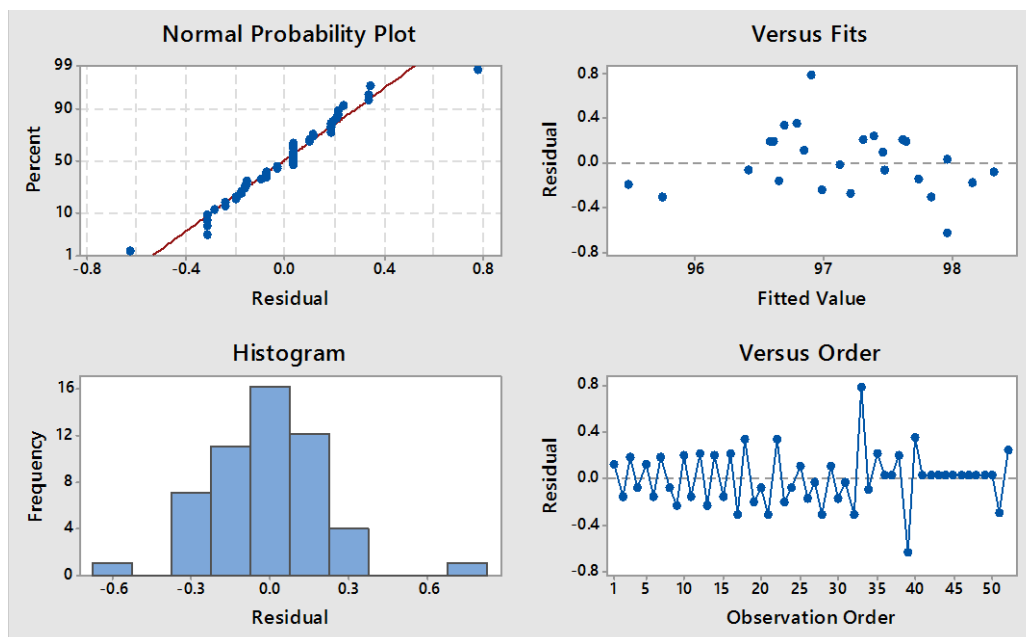


Figure 5.13 Residuals plots for model of key factor effect on mean specificity.

For sensitivity, the normal plot shows curvature at the tails, which could either be normal variation or indicate skewness. The histogram shows that skewness is the most likely explanation, as the distribution of residuals is skewed towards positive residuals. The residuals versus fits shows that while the positive residuals are spread evenly across all fitted values, the negative residuals occur mainly at higher fitted values. For specificity, the residuals plots indicate a normal distribution, constant variance, and an even spread of residuals versus fits. The plots also indicate that there are two outliers with higher residuals. These two runs used extreme conditions for factor A and factor F, with each factor at a particularly low level in one of the runs.

The models are not a perfect representation of the epidermis segmentation algorithm, however they were used to estimate optimal settings for the factors which were then tested in the algorithm. The Response Optimiser in MINITAB was used to find the optimal set of parameters. This algorithm is based on a reduced gradient approach with multiple starting points to identify the combination of input values that maximise the desired response.

The optimiser identified the combination of factor settings that would maximise both specificity and sensitivity, using the two regression models. The sensitivity was given a higher importance than the specificity, which ensured it had greater influence on the final measure of composite desirability. This measurement combines the desirability of both sensitivity and specificity, and weights the combination according to the importance set. The suggested optimal value for each of the four parameters and the range that was tested in the screening and optimisation study is shown in Table 5.4.

*Table 5.4 Range tested and optimised values for the four factors included in the model*

<b>Factor</b>	<b>Low level</b>	<b>High level</b>	<b>Optimised level</b>
<b>Lower penetration point for G'</b>	0.0723	0.350	0.2743
<b>Lower penetration point for b'</b>	0.0466	0.4034	0.4034
<b>Order of Mean Filter kernel</b>	15	65	40
<b>Radius of SE</b>	5	30	6

The models indicate that if the algorithm was run with these four optimised parameters, a sensitivity of 81.4% could be achieved with 95% confidence limits of 79.8% and 83.0%, and a specificity of 97.5% could be achieved with 95% confidence limits of 97.3% and 97.8%. Figure 5.14 contains three plots for each variable. The plots in the top row indicate how the composite desirability changes with each factor in the model, the plots in the second and third row indicate how the sensitivity and specificity change as each factor is varied. The differing impact of the factors on sensitivity and specificity can be seen. An interesting observation is that if the lower b' threshold increases, it causes an increase in sensitivity, and a decrease in specificity. It may be possible to tune the algorithm sensitivity and specificity using this parameter. The effect of the importance value set for the different responses can be seen in Figure 5.14, in cases where the optimal setting of a factor to maximise sensitivity and specificity is different (e.g. with the low b' threshold, labelled B Low), the value is set to maximise sensitivity.

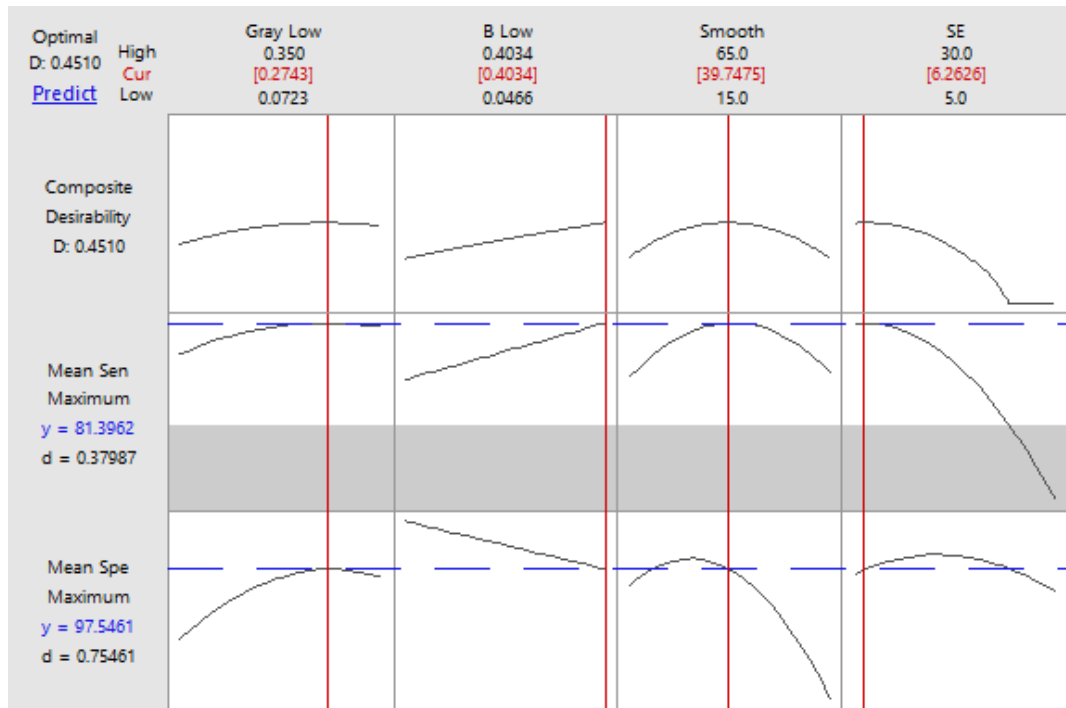


Figure 5.14 Optimisation plot for key parameters to maximise sensitivity

For the final implementation and evaluation of the epidermal segmentation algorithm the upper threshold used in the G' contrast enhancement (factor B) was set at the upper level of 1 as indicated by the screening study results. The upper

threshold of the  $b^*$  contrast enhancement was deemed insignificant in the subsequent optimisation study, but it was set at the higher level of 1 as sensitivities tended to be slightly higher.

#### **5.2.4 Optimisation of Object Classification Rules**

The algorithm was further enhanced by optimising the object classification rules described in section 5.1.8. These rules are completely dependent on the output from the previous parts of the image processing and segmentation process. They were not included in that optimisation as they would require impractically large factor ranges to be tested. Using the 25 image test set and optimised parameters from section 5.2.3, the algorithm was run without user interaction varying the area threshold between 15000 and 100000 pixels and the extent threshold between 0.36 and 0.46 in a full factorial design. The values included in the optimisation were determined based on the typical extent and area of epidermis objects (measured manually) after all prior processing steps had been completed on the 25 image test set. The resulting sensitivity and specificity measurements are given in Table 5.5 and the data is displayed as a contour plot in Figure 5.15.

The response optimiser and models indicated that the optimised factors should result in sensitivities of  $\sim 81.4\%$  ( $\pm 1.5$ ) and specificities of  $\sim 97.5\%$  ( $\pm 0.3$ ). In this study the object classification thresholds were set at 0.36 (extent) and 80000 (area), based on typical sizes and extents of objects in previous experiments. When tested at these conditions the sensitivity was 79.4% and the specificity was 97.3%, as shown in Table 5.5. This result shows the model gave an accurate prediction for specificity and slightly overestimated sensitivity. When the object classification thresholds were varied in this study the sensitivities varied between 75.0% and 87.0% and the specificities varied between 95.3% and 97.4%.

The contour plot (Figure 5.15.) is darker in colour when the sensitivity or specificity is higher. The plots shows that decreasing the area threshold and increasing the extent threshold improve sensitivity, however some combined settings which result in improved sensitivity do this at the cost of specificity. As the specificity was greater than 95% for all runs the aim with this optimisation was to try and increase sensitivity. An extent threshold of 0.44 and area threshold of



20000 were chosen to improve the sensitivity (86.9%) and maintain high specificity (95.3%). These settings do this by reducing the number of objects that are removed during the object classification stage, but ensuring that those that are retained have a high probability of being epidermis objects based on their shape and size.

Table 5.5 Effect of the tested area and extent thresholds on algorithm sensitivity and specificity

<b>Extent Thresh</b>	<b>Area Thresh</b>	<b>Mean Specificity</b>	<b>Mean Sensitivity</b>
0.36	60000	97.32	76.39
0.38	60000	96.29	80.59
0.4	60000	95.94	82.76
0.36	80000	97.34	79.39
0.38	80000	96.35	80.59
0.4	80000	96.00	82.76
0.36	100000	97.40	74.98
0.38	100000	96.41	79.18
0.4	100000	96.07	81.36
0.42	60000	95.75	83.63
0.42	80000	95.86	82.76
0.42	100000	95.92	81.36
0.44	40000	95.66	86.10
0.44	60000	95.75	84.78
0.44	80000	95.85	83.91
0.42	40000	95.72	84.95
0.44	20000	95.34	86.94
0.4	40000	96.34	84.34
0.38	20000	96.26	82.67
0.4	20000	96.18	84.85
0.42	20000	95.99	85.79
0.43	30000	95.59	85.71
0.46	20000	95.31	86.94
0.44	15000	95.32	87.00
0.43	18000	95.44	85.79
0.43	23000	95.48	85.71

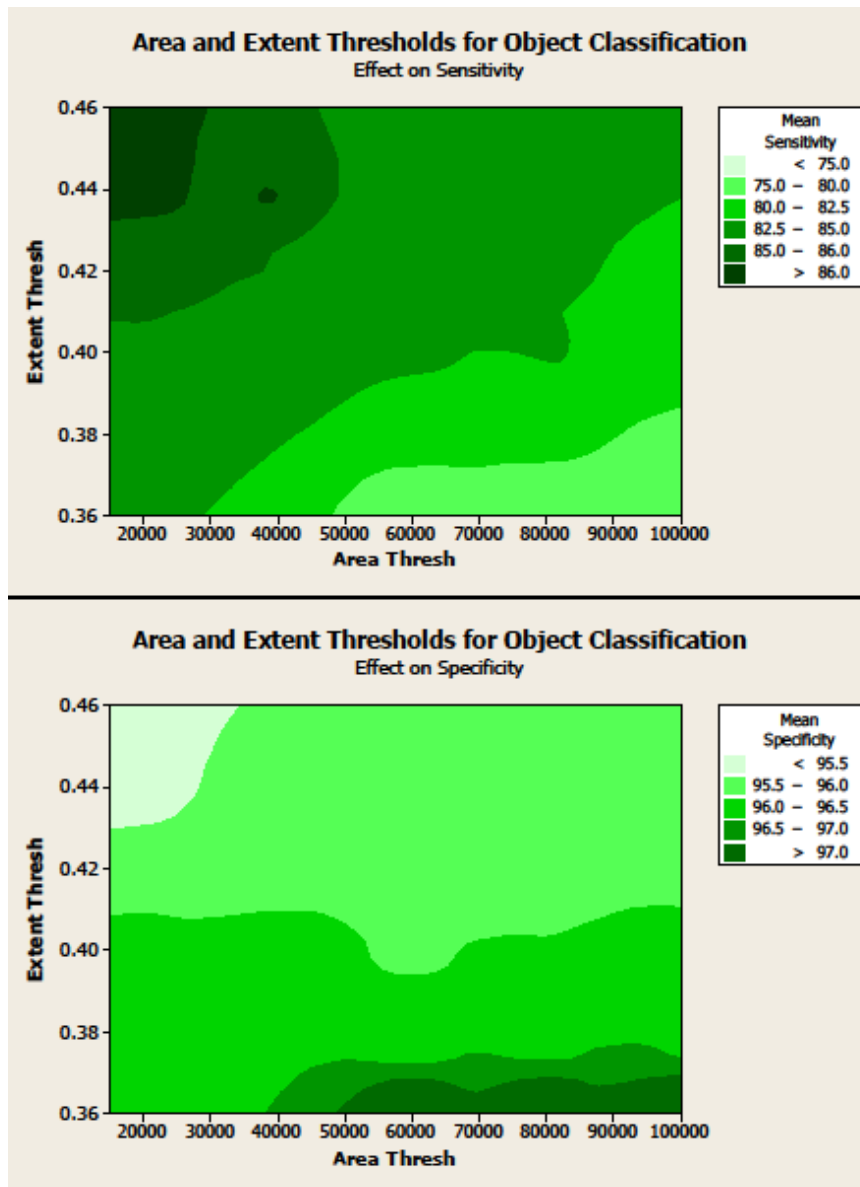


Figure 5.15 Contour plot of the effect of area and extent object classification thresholds on mean sensitivity and specificity for segmentation of epidermis.

### 5.2.5 Final Performance Evaluation for Epidermal Segmentation

The optimised parameter values and final method were tested on a new set of 40 images with associated manual mark-ups, both with and without the user interaction step. This image set included the 25 images used for the optimisations and an additional 15 images which were used a validation set.

When this final evaluation was performed it was noted that some of the unseen images had significantly lower sensitivities and that these images had unusual lighting or staining colour profiles. Figure 5.16 shows three images for which sensitivities of less than 60% were achieved (A, B and C) and one image (D) with good staining and contrast for which a sensitivity of 81% was achieved.

In Figure 5.16, images A, B and C have poor contrast between the epidermis and dermis, image B is also weakly stained and image C has a slightly altered colour cast, potentially due to the illumination level during image acquisition.

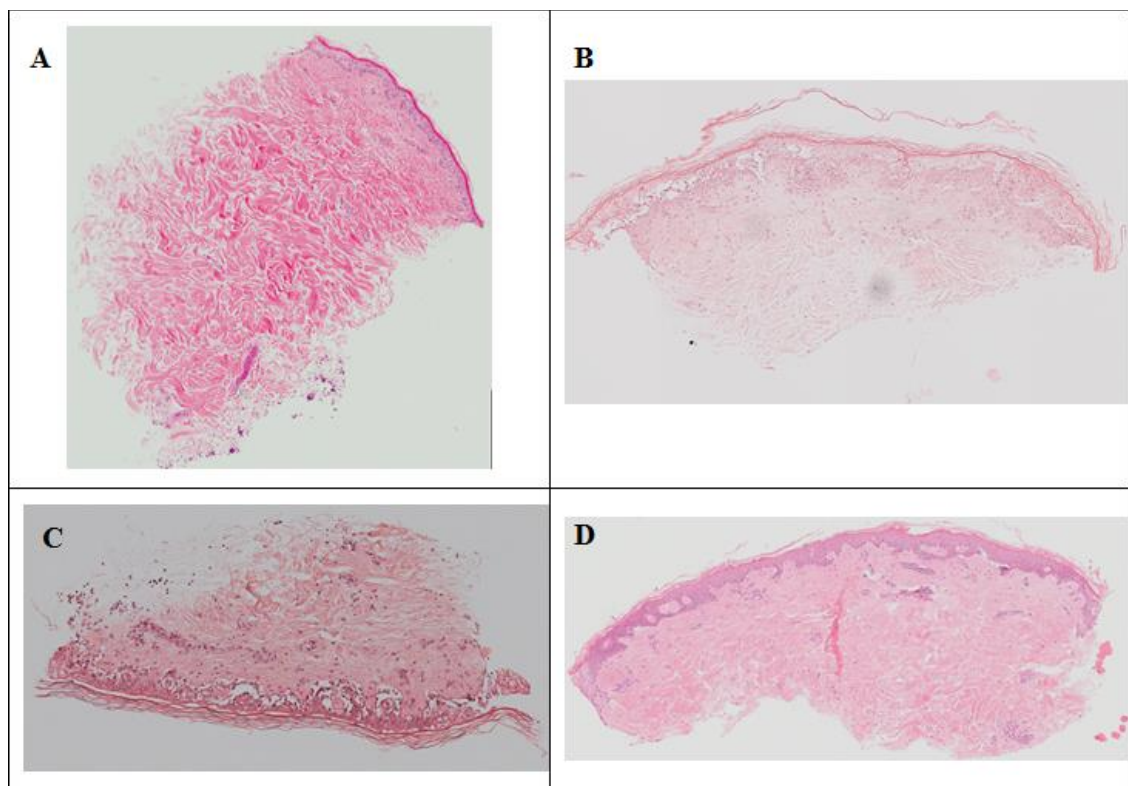


Figure 5.16 Four H&E stained images showing varying staining and lighting. A, B and C had sensitivities of < 60% and D had a sensitivity of 81%.

To address this issue, a colour normalisation step was added (section 5.1.2) and the 40 images were retested with the new colour normalisation step both with and without the user interaction step. Table 5.6 shows average accuracy, sensitivity and specificity for the training set of 25 images, the test set of 15 images and the average for the whole 40 image dataset. The standard error of the mean and minimum for each metric is also given. The standard error of the mean takes into

account both the standard deviation of the dataset and the sample size and is calculated thus:

$$\text{Standard Error of the Mean (SEM)} = \frac{\sigma}{\sqrt{n}} \quad \text{Equation 5.21}$$

where  $\sigma$  is the population standard deviation and  $n$  is the sample size. The SEM tends to decrease as the number of samples in the dataset increases. It is an informative measure in this case because the samples size of the training and test set are different.

Table 5.6 Summary of statistics for accuracy, sensitivity and specificity performance metrics

		No User Interaction			With User Interaction			
		Spec	Sens	Acc	Spec	Sens	Acc	
Without colour normalisation	Training set (n=25)	Mean	95.34	86.94	93.89	96.99	88.26	95.64
		SEM	1	2.04	1.08	0.73	1.55	0.84
		Min	70.91	48.32	71.56	76.79	61.35	77.18
	Test set (n=15)	Mean	97.11	80.28	94.26	97.47	89.23	96.43
		SEM	0.39	4.07	1.01	0.22	0.87	0.29
		Min	90.68	0	72.69	94.27	80.29	92.74
	All (n=40)	Mean	96	84.44	94.03	97.17	88.63	95.93
		SEM	0.83	2.96	1.04	0.59	1.33	0.69
		Min	70.91	0	71.56	76.79	61.35	77.18
		No User Interaction			With User Interaction			
		Spec	Sens	Acc	Spec	Sens	Acc	
With colour normalisation	Training set (n=25)	Mean	97.69	91.79	96.75	98.14	92.87	97.25
		SEM	0.29	1.32	0.37	0.25	1.07	0.36
		Min	91.43	68.6	88.11	94.15	75.63	88.41
	Test set (n=15)	Mean	97.7	85.32	95.96	97.63	87.76	96.13
		SEM	0.22	2.25	0.34	0.2	1.32	0.33
		Min	94.7	43.17	91.66	94.7	74.75	91.66
	All (n=40)	Mean	97.69	89.37	96.45	97.95	90.95	96.83
		SEM	0.26	1.77	0.36	0.23	1.22	0.36
		Min	91.43	43.17	88.11	94.15	74.75	88.41

Considering all 40 images, with no user interaction, the addition of the colour normalisation step increases the specificity by 1.69% and the sensitivity by 4.93%, resulting in a mean accuracy increase from 94.03% to 96.45%. The impact of the colour normalisation step is particularly apparent when the minimum sensitivities and specificities in the image set are examined. Whereas without colour

normalisation one image had a sensitivity of zero, the worst performing image with colour normalisation had a sensitivity of 43.17%. Normalising staining intensity ensures even very weakly stained epidermal tissue can be correctly identified and segmented.

A comparison of the training and test set data with and without user interaction is presented as a boxplot in Figure 5.17. The boxplot defines the median, interquartile range and highlights outlying results, i.e., those that are more than 2.7 standard deviations beyond the mean.

There is a slightly reduced sensitivity in the test set (85.32%) compared to the training set (91.79%), as well as an increase in interquartile range and range of the test set compared to the training set. The difference in test and training set is slightly less when the user interaction step is included (87.76% compared to 92.87%). Overall the training and test sets are similar which suggests that the model developed using the training set was not overfitted, and the optimised values parameters are likely to be valid for new images. Across the whole dataset of 40 images, including user interaction, the mean specificity is 97.95%, mean sensitivity is 90.95% and mean accuracy 96.83%. Without user interaction, the mean specificity is 97.69%, mean sensitivity is 89.37% and mean accuracy 96.45%. The user interaction step improves the algorithm by reducing the level of variability. More specifically, the standard error of the mean sensitivity for the test set reduces from 14.2 to 8.4 when the user interaction step is added.

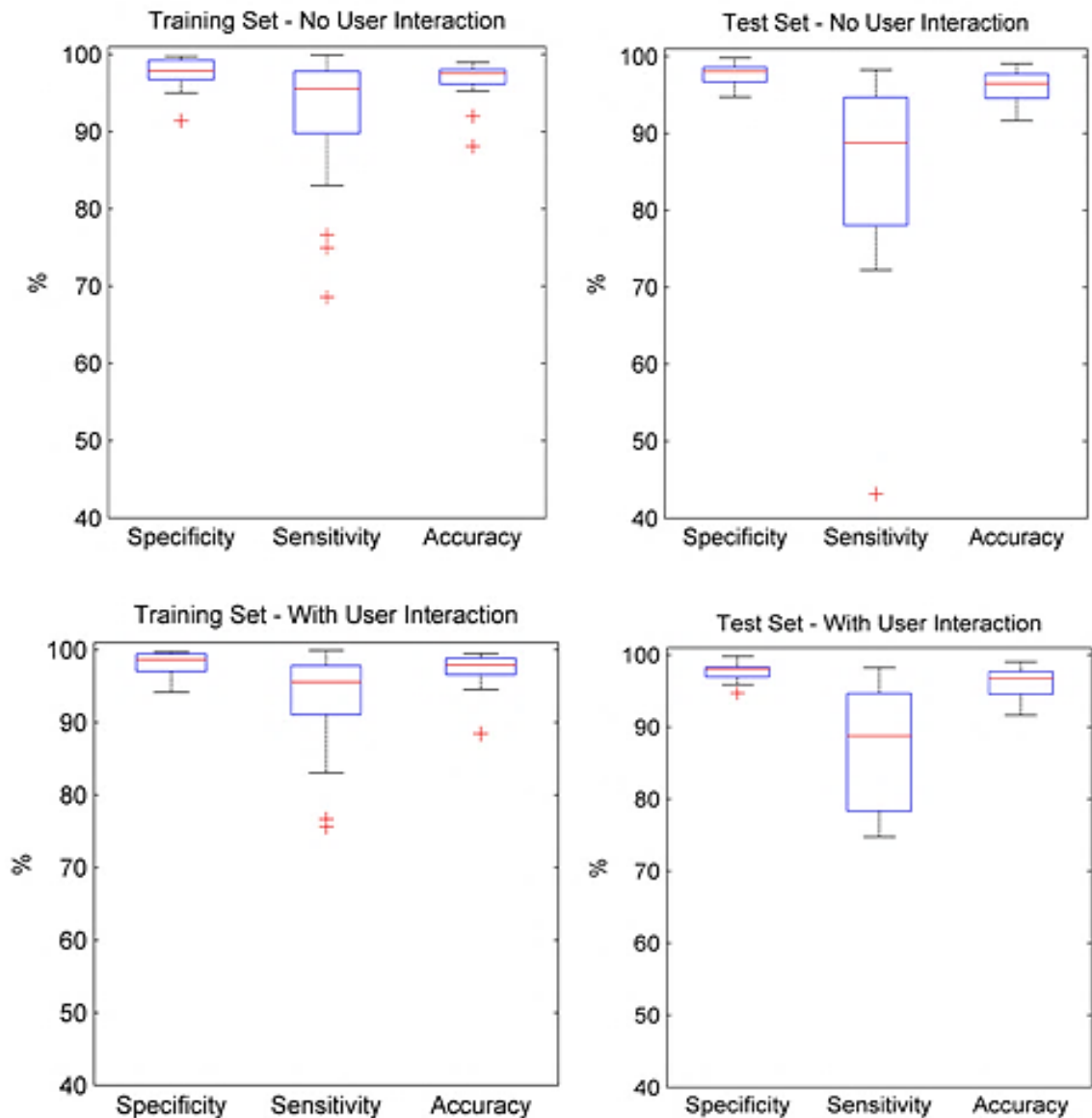


Figure 5.17 Boxplot of specificity, sensitivity and accuracy for epidermal segmentation in training and test sets– with and without user interaction

The six worst segmentations in the dataset of 40 had sensitivities of 75-78% and these were not specific to a particular class of damage with two grade I, two grade II, one grade III and one grade IV. Figure 5.18 shows boxplots for specificity and sensitivity grouped by grade of damage.

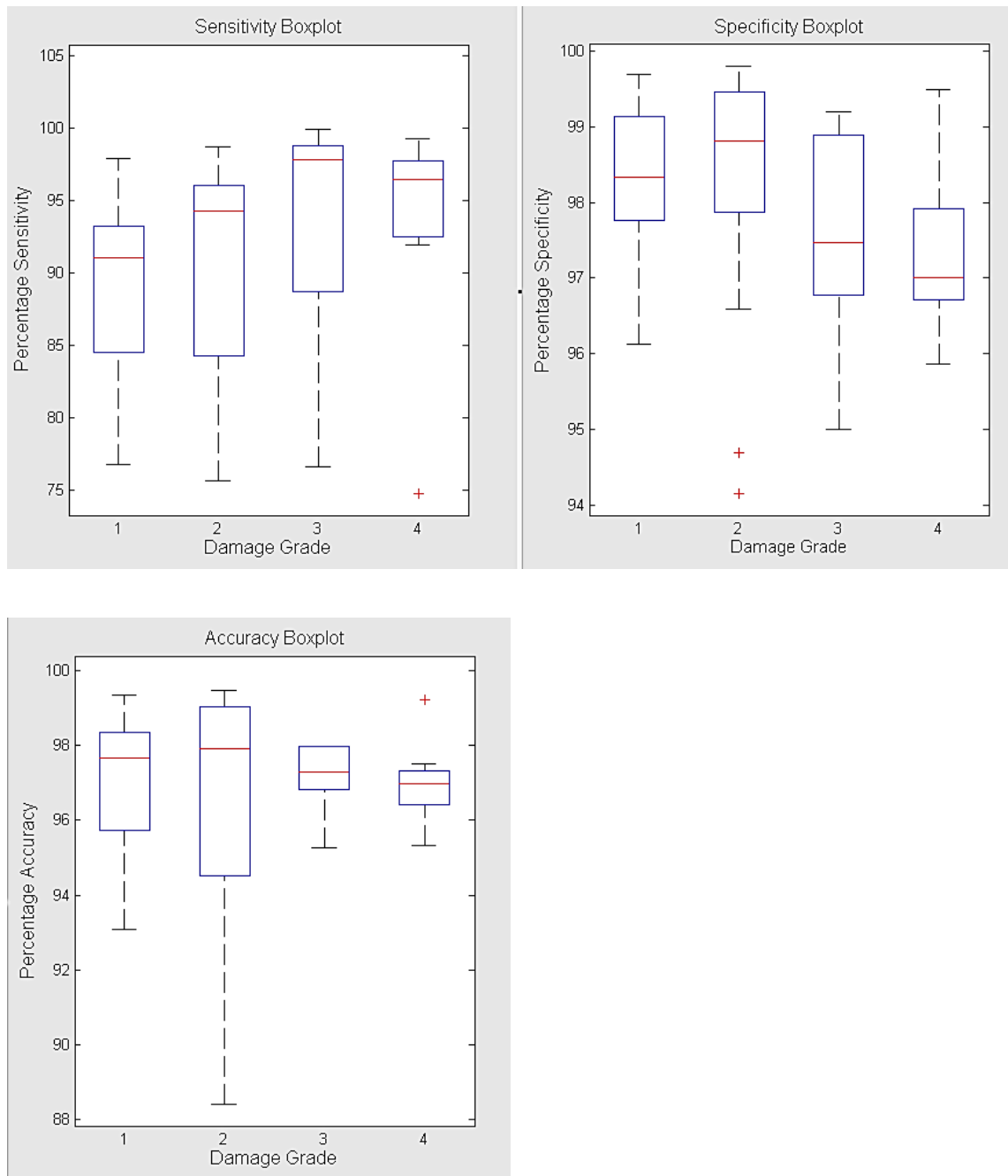


Figure 5.18 Boxplots showing effect of damage grade on specificity, sensitivity and accuracy of epidermal segmentation.

From these results, it can be observed that there are small differences between the data, but no major differences. The 40 image dataset contained 12 grade I, 14 grade II, 6 grade III and 8 grade IV images. Considering the relatively small sample size once the data set is split into the four damage types, it is difficult to draw definite conclusions about the distributions. However, one observation is that as the damage grade increases the specificity appears to decrease and the sensitivity

increases. While this may be due to chance, it is possible that the algorithm may be misclassifying more dermis tissue as epidermis when there is damage present. While this non-specificity is not ideal, identifying dermis tissue as epidermis would be more of an issue if it was happening in the grade I images, as this may lead to false positive identification of spaces in the dermis as clefts or vacuoles.

Once the epidermis had been segmented, the next task was to accurately segment the dermis tissue so that the DEJ could be identified. Locating the DEJ correctly is essential for objects to be identified accurately as clefts because location at the DEJ is a specific property of clefts. The epidermal segmentation was optimised using performance metrics that were based on an accurate manual mark-up of the “correct” epidermal pixels by an expert. While this was a labour intensive process, it provided the best method for optimising the multiple parameters included in the algorithm utilising multiple images. The drawback of such a method is the potential inaccuracy and human bias inherent with such a process due to the manual annotation. In practice, annotation of the outer edge of the epidermis was relatively straightforward. The boundary of the epidermis was relatively smooth and small errors of < 10 pixels in the location of the marked-up boundary had minimal impact due to the relatively large dimensions of the epidermis (20,000 to 80,000 pixels).

Creating a mark-up for the dermal, cleft and vacuole segmentation would be a significantly more challenging task. The dermis perimeter is often highly convoluted, making an accurate and repeatable mark-up difficult. For vacuoles and clefts, the small size and large number make the potential for error during mark-up much greater due to the high ratio of perimeter to total area. Individual vacuoles comprising between 100 and 500 pixels may have perimeters of between 50 and 300 depending on their shape. Small (1-2 pixel) errors in the mark-up of the boundary could have a significant effect on the final segmentation and feature measurements. Although the segmentation of these features is very important, it would be extremely time consuming and error prone to mark-up vacuoles and clefts in sufficient images to get a representative sample.



In light of the issues, it was decided that as opposed to utilising a quantitative approach, a qualitative visual assessment would be used to optimise the dermal, cleft and vacuole segmentation based on 16 images and a set of 30 images would be used to assess the final accuracy of the segmentation. The images were selected to include varying cleft and vacuole densities, and variable staining.

### 5.3 Dermal Segmentation

The dermis segmentation starts with a simple logical operation that locates the dermis pixels using the assumption that any sample pixels that are not epidermis pixels, must therefore be dermis pixels. Initially this was followed by a classification of objects based on major and minor axis length and object area, the aim of which was to exclude any very long and thin objects such as the long sections of *stratum corneum*. A binary image containing information on the location of all dermis pixels was then used during cleft segmentation to identify the DEJ. However this simple approach to dermal segmentation was not successful in avoiding misclassification of the *stratum corneum* as part of the dermis, and led to false identification of clefts at the boundary of the main epidermis and the *stratum corneum*.

The *stratum corneum* is the top layer of the epidermis and although it can be linear in nature or curve around the sample, it is always located at the sample perimeter. Firstly, as has already been stated, the *stratum corneum* should be excluded from both the dermis and epidermis masks even though it is part of the epidermis. However excluding this part of the tissue is difficult because the thickness, structure and staining intensity of the layer varies significantly between samples making identification challenging. One reason for the variation is the areas of necrotic (dead) tissue that sometimes build up at the top of the epidermis around the *stratum corneum*. This layer of tissue can become very thick and irregular in structure, and because it is dead tissue it should not be included in either the dermis or epidermis masks. Figure 5.19 shows three images with the *stratum corneum* outlined in green. The images give an indication of the typical variation observed within the data set. An area of necrotic tissue is outlined in blue and filled with blue hatching. The necrotic tissue is pale pink in colour and has sparsely

distributed purple/blue nuclei; these are also properties of dermal tissue and so it can easily be misclassified.

A perimeter masking step was designed to address the issues of misclassification of *stratum corneum* as dermis tissue, and the presence of necrotic tissue. This step is described in the following section.

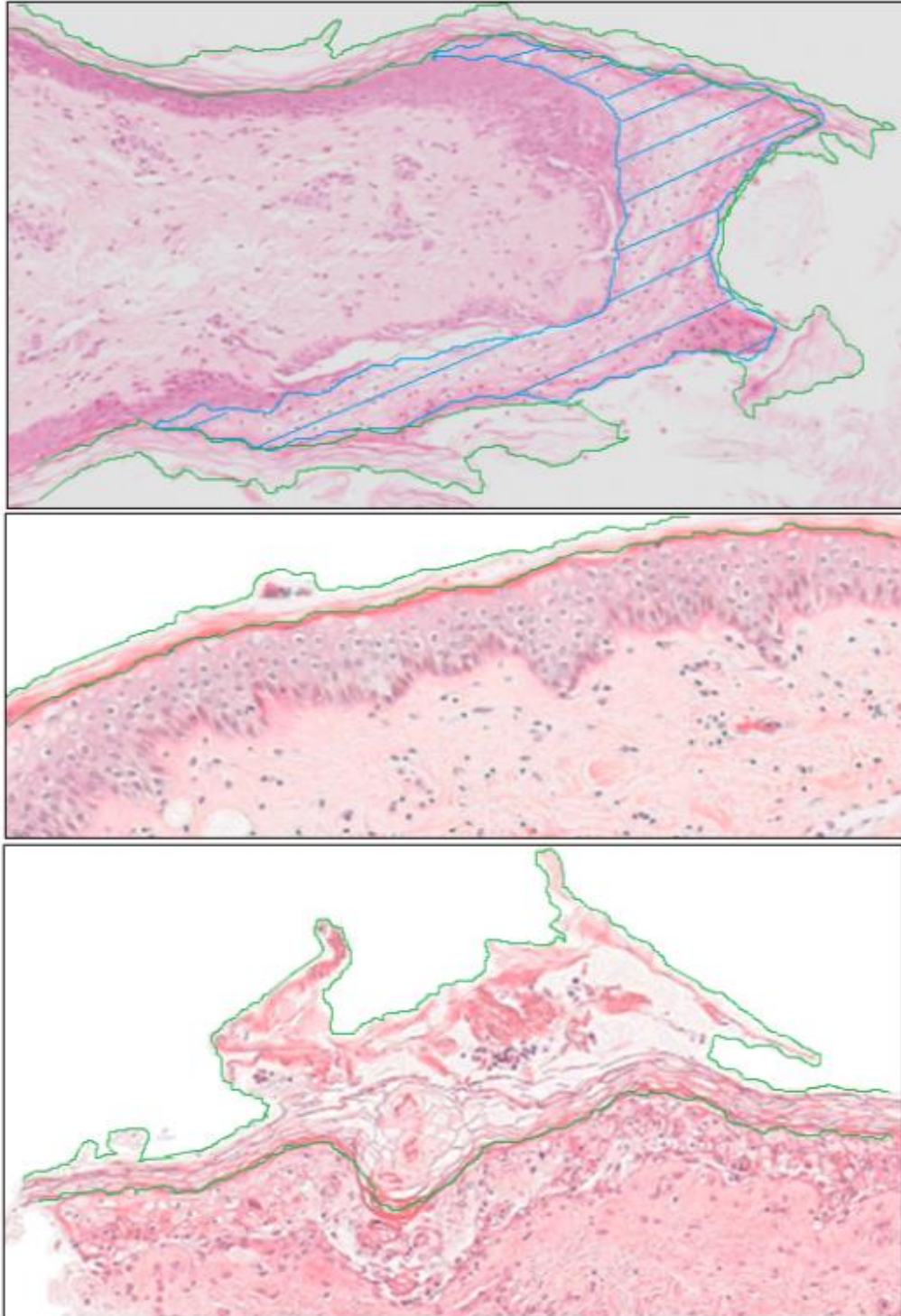


Figure 5.19 A selection of images highlighting differences in the stratum corneum and areas of necrotic tissue.

### 5.3.1 Sample Perimeter Masking

A mask based on the perimeter of the sample mask,  $sMask$  (Equation 5.4), was tested to determine whether it could be used to remove the *stratum corneum* without removing important areas near the DEJ. The perimeter mask of the tissue

sample was created by subtracting a version of *sMask* eroded by one pixel from the original version of *sMask*. The perimeter mask was then morphologically thickened by adding pixels to the exterior of the perimeter. Some sample masks had very convoluted perimeters due to the loose and fibrous structure of the dermis tissue. When these convoluted edges were thickened it had the effect of masking out too much tissue, in particular the dermis pixels at the DEJ. To avoid this issue a morphological closing step was performed on the *sMask* to smooth the edges prior to the perimeter extraction. Figure 5.20 illustrates the benefit of adding the smoothing step when creating the perimeter mask.

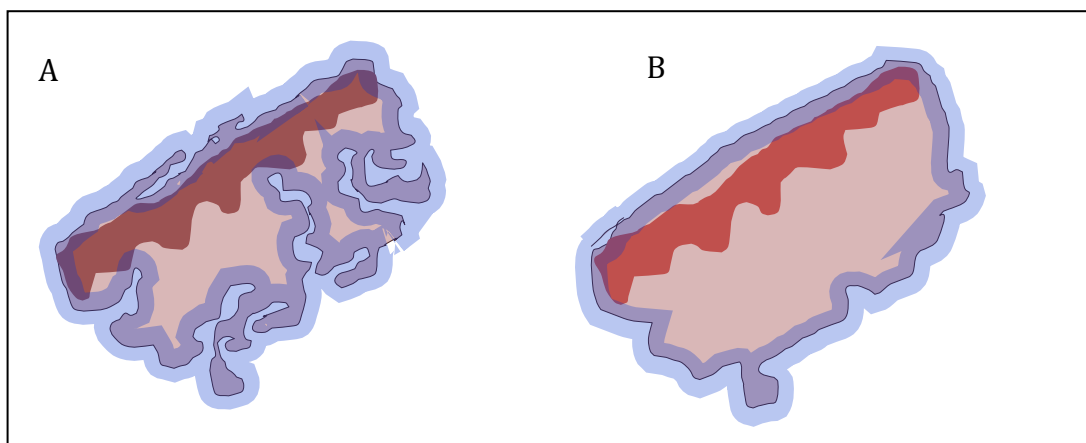


Figure 5.20 Diagram illustrating the effect of smoothing on the perimeter masking step.

In diagram A, the dermis perimeter is highly convoluted and at certain points the perimeter reaches deep within the tissue sample and masks out some pixels at the DEJ. In diagram B the perimeter has been smoothed, reducing the convolution and ensuring that the *stratum corneum* is masked but the DEJ pixels are not.

Figure 5.21 shows an overlay of a thickened, smoothed perimeter over the original image. In this case, the perimeter mask overlays the *stratum corneum*, some of the epidermis, and also parts of the dermis. The masking of the dermis pixels deeper in the tissue (indicated by the black arrows in Figure 5.21) is not an issue as these pixels are not adjacent to the epidermis and therefore not required to locate the DEJ. The masking of the epidermis pixels also does not have an impact, since the location of these pixels is already known from the epidermal segmentation procedure. If they are masked out, they can be recovered using the epidermis mask, *eMask*, described in section 5.1.8. The side edges of the tissue (identified

within rectangles in Figure 5.21), have adjacent epidermis and dermis pixels, however the masking of these regions is a positive outcome as histopathology experts tend to disregard these areas from the analysis due to the high probability of artefacts from the sample slicing procedure being present.

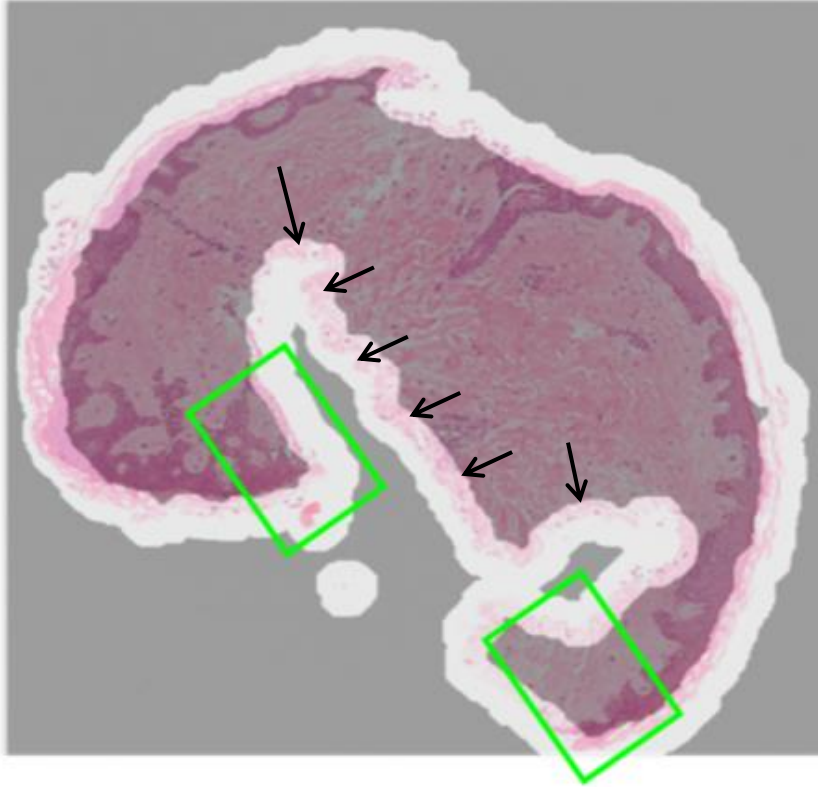


Figure 5.21 RGB image of H&E stained skin sample overlaid with mask of thickened perimeter

### 5.3.2 Dermal Segmentation: Method

The correct segmentation of the dermis tissue and the disregarding of the *stratum corneum*, requires a variety of processing steps to be used in series. First, a new binary mask  $dMask$  is created containing only pixels within the sample mask  $sMask$  that are not present in the epidermis mask  $eMask$ .

$$dMask_{i,j} = \begin{cases} 1 & \text{if } sMask_{i,j} = 1, eMask_{i,j} = 0 \\ 0 & \text{otherwise} \end{cases} \quad \text{Equation 5.22}$$

The outer contour of the sample mask is smoothed using morphological closing:

$$sMask_{closed} = sMask \cdot SE = (sMask \oplus SE) \ominus SE \quad \text{Equation 5.23}$$

where SE is a disk shaped structuring element with a radius of 20 pixels.

Structuring elements of different sizes were tested, but the selection of SE with radius 20 was made based on the fact that this was the smallest size able to close the small openings in the perimeters of the dermis tissue. A larger size was avoided to minimise the loss of information in the image.

The perimeter of **sMask<sub>closed</sub>** is found by subtracting a new version of **sMask<sub>closed</sub>** which has been morphologically eroded,  $\oplus$ , using a structuring element with a radius of 1 pixel, SE1, from the original **sMask<sub>closed</sub>**. The resulting image is then dilated using a disk shaped structuring element of radius 60, SE2, to thicken it and create the perimeter mask, **pMask**. The value of 60 was selected based on the average depth of the *stratum corneum* layer in all but the most extreme examples. If the examples with the very thick layers of *stratum corneum* were included in this calculation then the mask would get rid of unnecessarily large parts of the tissue samples in the more typical examples :

$$pMask_{i,j} = \begin{cases} 1 & \text{if } (((sMask_{closed} \oplus SE1) - sMask_{closed}) \ominus SE2)_{i,j} = 1 \\ 0 & \text{otherwise} \end{cases} \quad \begin{array}{l} \text{Equation} \\ 5.24 \end{array}$$

The perimeter mask is subtracted from the dermis mask to remove regions of *stratum corneum* and pixels near the cut edge of the sample:

$$dMask_{i,j} = \begin{cases} 1 & \text{if } pMask_{i,j} = 0 \\ 0 & \text{otherwise} \end{cases} \quad \begin{array}{l} \text{Equation} \\ 5.25 \end{array}$$

For each object  $Z$  in **dMask**, the object area, the major axis length and minor axis length are determined. The ratio of  $Z_{MajorAxisLength}$  to  $Z_{MinorAxisLength}$  gives the dimension measurement,  $Z_{Dim}$ , of the object, which is a measure that can be used to identify the thin objects which make up the *stratum corneum*. The following classification rule based on the object area and dimension was used to classify each object pixel,  $z$ , in the **dMask**:

$$z = \begin{cases} 1 & \{ \forall z \in Z | Z_{Dim} < 7.7, Z_{Area} < 42500 \} \\ 0 & \text{else} \end{cases} \quad \begin{array}{l} \text{Equation} \\ 5.26 \end{array}$$

The values used in this classification were chosen based on the area and dimension measurements of dermis and non-dermis objects from 16 images. Figure 5.22 is the histogram of dimension measurements for all dermis and non-dermis objects in the 16 versions of *dMask*. Figure 5.23 is the histogram of area measurements for all dermis and non-dermis objects in the 16 versions of *dMask*.

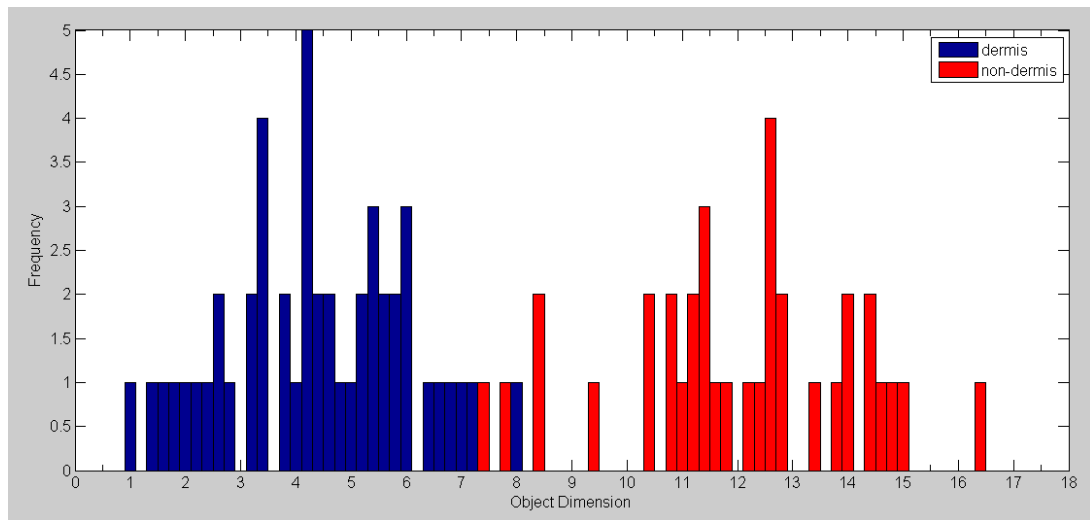


Figure 5.22 Histogram of dimension measurements for dermis and non-dermis objects

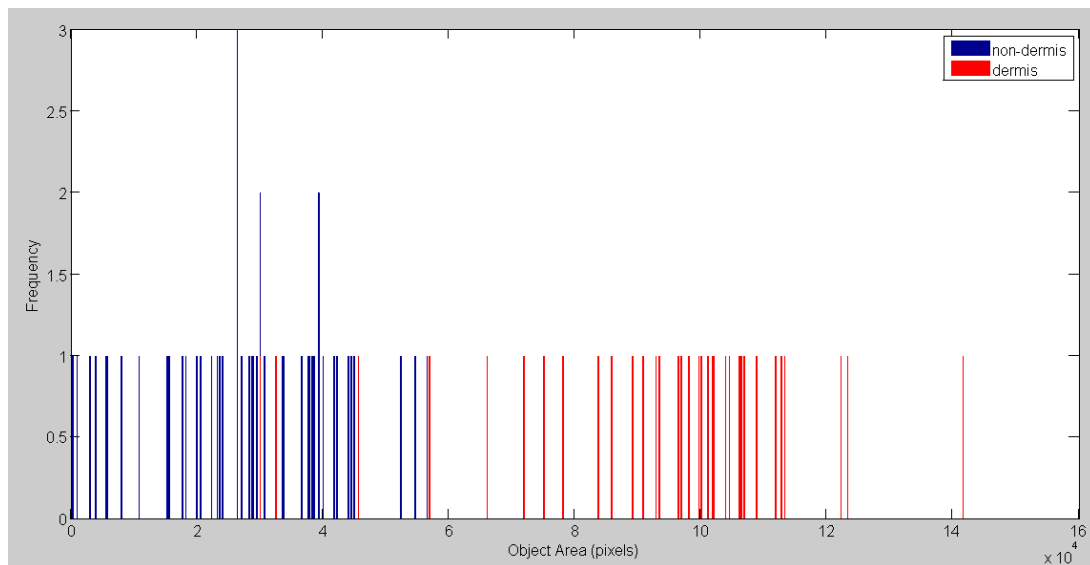


Figure 5.23 Histogram of area measurements for dermis and non-dermis objects

Visual examination of the 16 image data set and assessments of the histograms shown in Figure 5.22 and Figure 5.23 were used to determine appropriate threshold values for dimension and area. The area threshold of 42,500 was set at a level which excluded the majority of non-dermis objects and included all but the very smallest objects. The smaller dermis objects, sized between 20,000 and

42,500 pixels, were small fragments of tissue which were non-informative with regards to the DEJ location. The extent threshold was easier to select manually as the separation of the dermis and non-dermis objects was greater. The highest dimension of a dermis object (8.1) and the lowest dimension of a non-dermis object (7.3) were identified, and the mean calculated to arrive at the final threshold of 7.7.

### 5.3.3 Dermal Segmentation: Results

The following series of figures show the impact of each step in the dermis segmentation. A particularly challenging example has been chosen to illustrate the reason for each step's inclusion. Figure 5.24 shows the subtraction of the epidermis mask from the sample mask (Equation 5.22), and illustrates how it breaks the large sample objects in the mask into smaller objects, which are comprised of actual dermis tissue and other regions such as the *stratum corneum* and necrotic tissue.



Figure 5.24 Subtraction of epidermis mask from sample mask

Subtracting the thickened perimeter mask (Equation 5.25), either thins or completely removes objects around the edge of the sample including the *stratum corneum* and necrotic tissue. The radius of the SE used to thicken the perimeter mask is 60 pixels and so objects that are smaller than this in size are removed completely, larger objects are thinned. In Figure 5.25, the necrotic tissue combined



with the *stratum corneum* makes it extremely thick and so the mask thins rather than removes these objects.

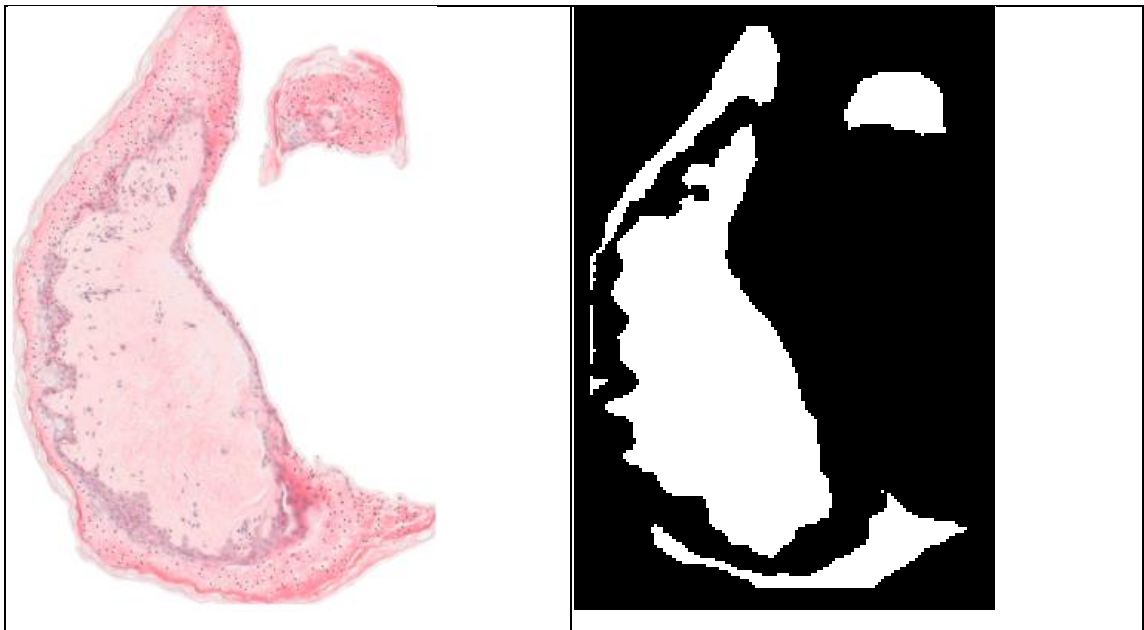


Figure 5.25 Subtraction of thickened sample perimeter mask

Figure 5.26 illustrates the selective removal of the thinned objects using an object classification step based on object shape and size (Equation 5.26). The arrows indicate objects which are being misclassified as dermis objects at the end of this process.

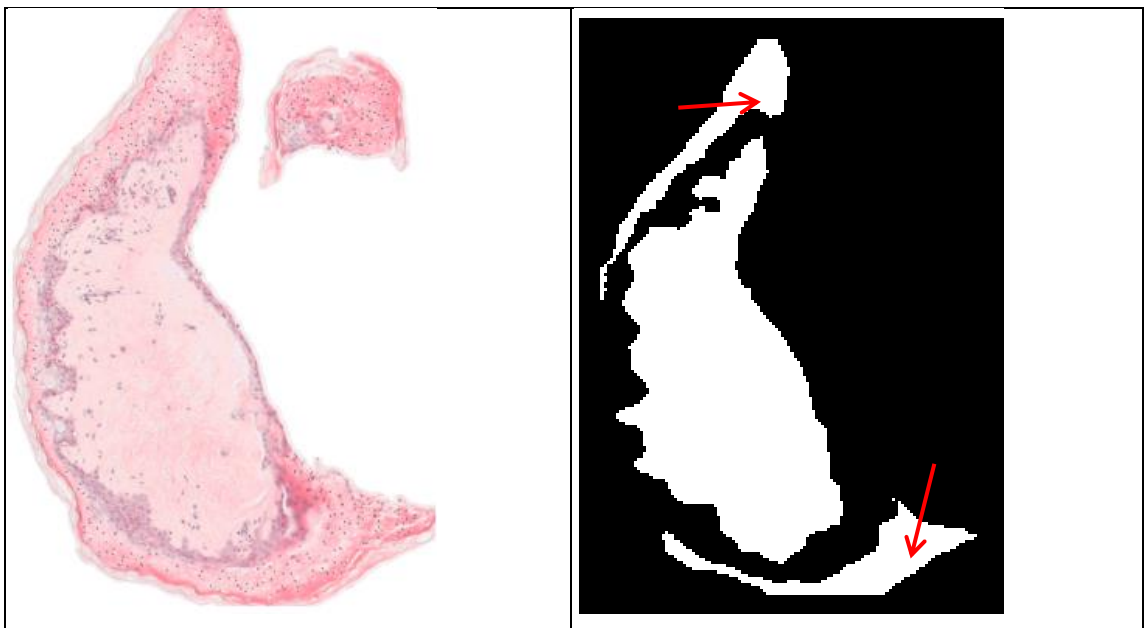


Figure 5.26 Removal of non-dermis objects with a classification rule

The dermal segmentation method was tested on a set of 16 images and a qualitative analysis performed in which the dermal segmentation was deemed to be inaccurate if it would lead to an over or underestimation of the true DEJ by  $\sim 25\%$ . The biological and staining variation of the samples resulted in inaccurate dermal segmentation in  $\sim 10\%$  of cases.

Although additional optimisation may be able to improve dermal segmentation, a user interaction step was added to create a workable solution. The overall process still works without the user interaction stage, but by incorporating the step it is possible to remove any regions that are still misclassified as dermis (such as the large areas of necrotic tissue indicated by arrows in Figure 5.26). Figure 5.27 shows final dermis mask in white and perimeter of the epidermis mask in green, following the user interaction step.

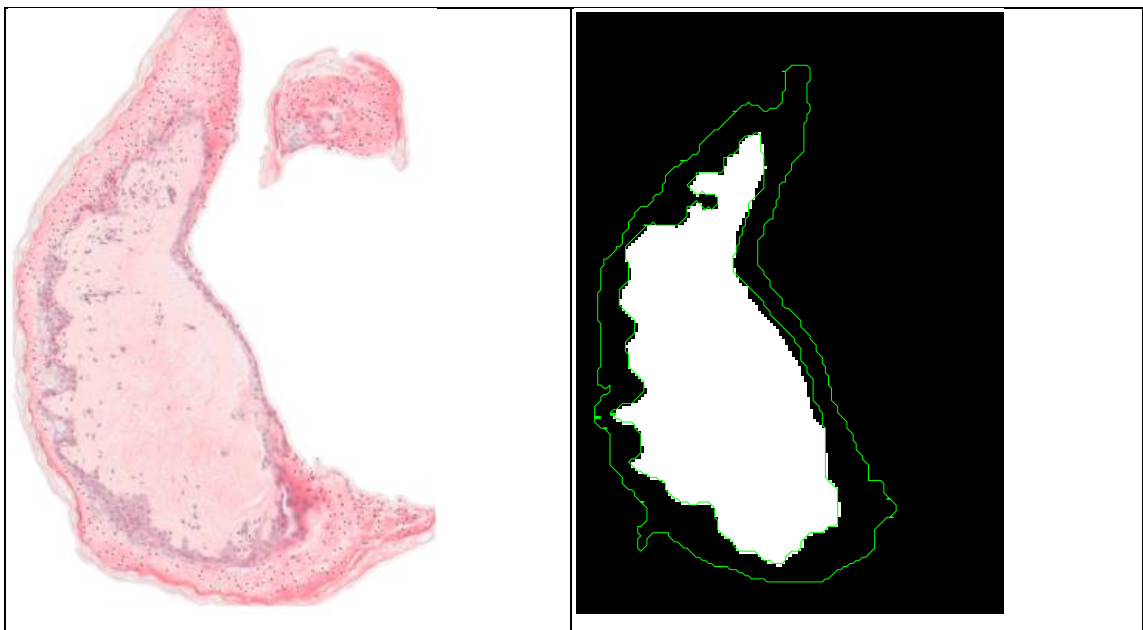


Figure 5.27 User-interactive removal of any remaining misclassified objects

#### 5.4 Cleft Segmentation

The next stage of the process was to identify clefts at the DEJ. The clefts appear as white spaces between the dermis and epidermis when the damage reaction reaches Grade III severity. When examining the colour normalised images it was noted that the internal clefts sometimes appeared to have uneven colouring, rather than the consistent background colour that would be expected. This effect is an artefact of the histogram matching procedure. Despite the input images having

differing proportions of colours, they are matched to a target image with a set proportion of colours. As a result the consistent background colour of clefts is sometimes replaced with a greater range of light colours. This can be seen in the cleft outlined in Figure 5.28. In the original RGB image there is very little variation in colour throughout the cleft whereas in the normalised image there are more pale pink and grey pixels present. For this reason, the original, non-normalised images were used to identify the clefts.

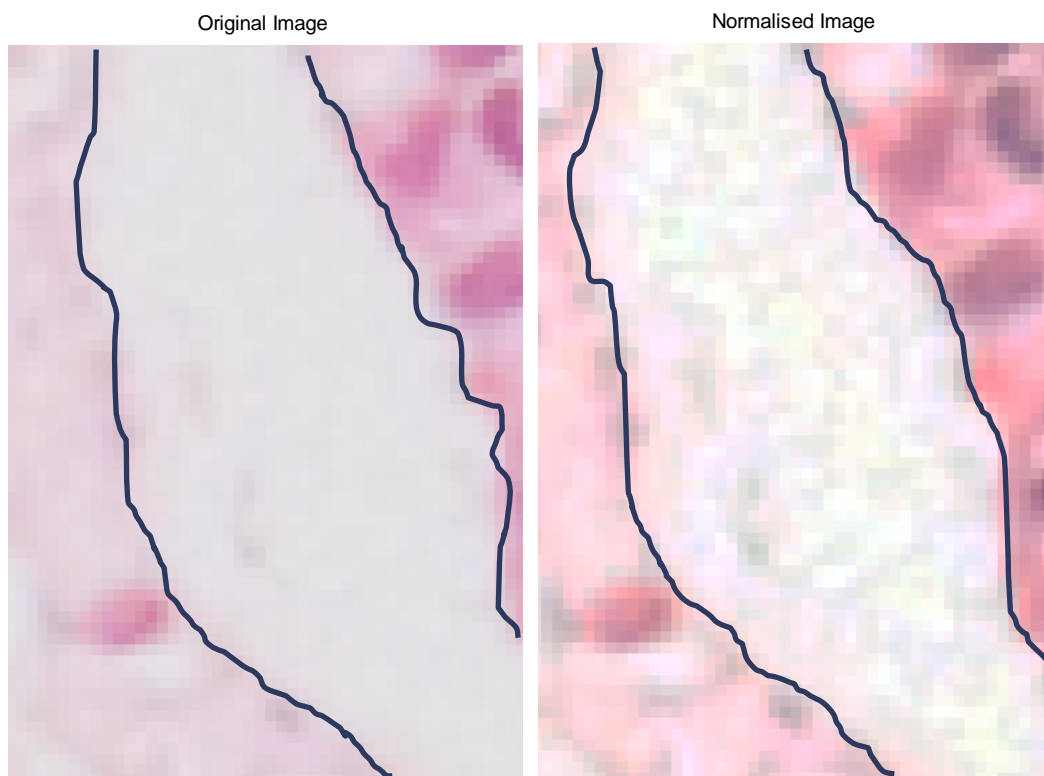


Figure 5.28 A sub-epidermal cleft in an original RGB image and an image which has been normalised using histogram matching.

#### 5.4.1 Cleft Segmentation: Method

The basic platform process for cleft segmentation was the same as that developed for epidermal segmentation, i.e., colour image pre-processing for contrast enhancement, thresholding, morphological processing and size and shape based object classification.

For the colour image pre-processing step, it would be expected that a measure such as luminance would highlight the clefts since they are visualised by light

passing directly through spaces in the tissue. To do this, the RGB image was converted to the  $L^*a^*b^*$  colour space and the luminance ( $L^*$ ) image contrast stretched using a linear mapping function to create a new image,  $L$ , utilising the full dynamic range (see section 2.4.9, equation 2.4). This operation is a type of normalisation and is of value since the images being used have not been colour normalised. The function remapped the intensities in  $L^*$  to exclude 1% of the lowest intensities and 1% of the highest intensities. The penetration points  $P_{min}$  and  $P_{max}$  required to identify the lower and upper bounds of the intensities to be included in the remapped image were determined using the cumulative percentage histogram (section 2.4.9). The remaining pixels were remapped to utilise the full dynamic range of 0 to 1:

	$L_{i,j} = INT \left\{ \frac{1 - 0}{P_{max} - P_{min}} [L^*_{i,j} - c] \right\}$	Equation 5.27
--	----------------------------------------------------------------------------------	---------------

The  $INT$  function converts the output into an integer which determines the intensity of a pixel in the new image.

The next step developed was a thresholding operation on the  $L$  image to identify potential clefts. The threshold  $C_{thresh}$  is based on the value of the mode intensity of all pixels in  $L$ . In a similar manner to sample thresholding (Equation 5.4), a lower threshold than the mode value was chosen to account for variation in the intensity of the background pixels:

$$C_{thresh} = mode(\{l_{ij} | l_{ij} \in L\}) - 20 \quad \text{Equation 5.28}$$

The decision to subtract 20 from the mode value was made by iteratively changing the value subtracted and then visually assessing the effect of changing the threshold on the proportion of cleft spaces identified on multiple images. This process is described fully in the results section 5.4.2 and the impact of changing this threshold on the segmentation of clefts is shown in Figure 5.31.

The operation that is described next is applied twice to identify all potential clefts. The reason that cleft regions must be identified in two stages is that sometimes

clefts at the DEJ are included within the dermal mask and at other times they are included as part of the epidermal mask. First,  $C_{thresh}$  is used to threshold the dermis mask, **dMask**, and create a new binary mask **cObjD** containing all background coloured pixels within the dermis:

$$cObjD_{i,j} = \begin{cases} 1 & \text{if } L_{i,j} > C_{thresh} \text{ AND } dMask_{i,j} = 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{Equation 5.29}$$

The contiguous objects in **cObjD** are then classified based on whether they are located at the boundary of the epidermis and dermis, the DEJ. More specifically, any pixels part of a contiguous object in **cObjD** and within a 5 pixel distance of the epidermis are identified by utilising the **eMask** and four translated versions of the **cObjD** mask. The translated versions are **cObjD** masks which have been shifted 5 pixels up, down, right or left. Pixel locations which have an intensity of 1 in both the **eMask** and a translated **cObj** mask must therefore be dermis pixels that have (1) been identified as being part of a potential cleft, and (2) be within a 5 pixel distance of the epidermis. The identified pixel locations in each **cObjD** mask are then subjected to a reverse translation operation and added into a single matrix, **eAdjacent**. Any object in **cObjD** that contains at least one of the adjacent pixels in **eAdjacent** is retained in a new mask, **cMaskDerm**.

The thresholding and location classification procedure is then repeated to identify clefts on the epidermal side of the DEJ. More specifically, epidermis pixels are thresholded using  $C_{thresh}$  to create a mask **cObjE** that identifies background coloured pixels within the epidermis. Pixels within a 5 pixel distance of the dermis are located using the translation operation described above. Any objects in **cObjE** that are within a 5 pixel distance of the epidermis are saved in a new mask, **cMaskEpi**.

The two masks, **cMaskEpi** and **cMaskDerm**, are added together into a single binary mask, **cMask**. Figure 5.29 shows the dermal clefts objects identified adjacent to the epidermis (Figure 5.29a), the epidermal clefts adjacent to the dermis (Figure 5.29b), and the combination of both sets of clefts (Figure 5.29c).

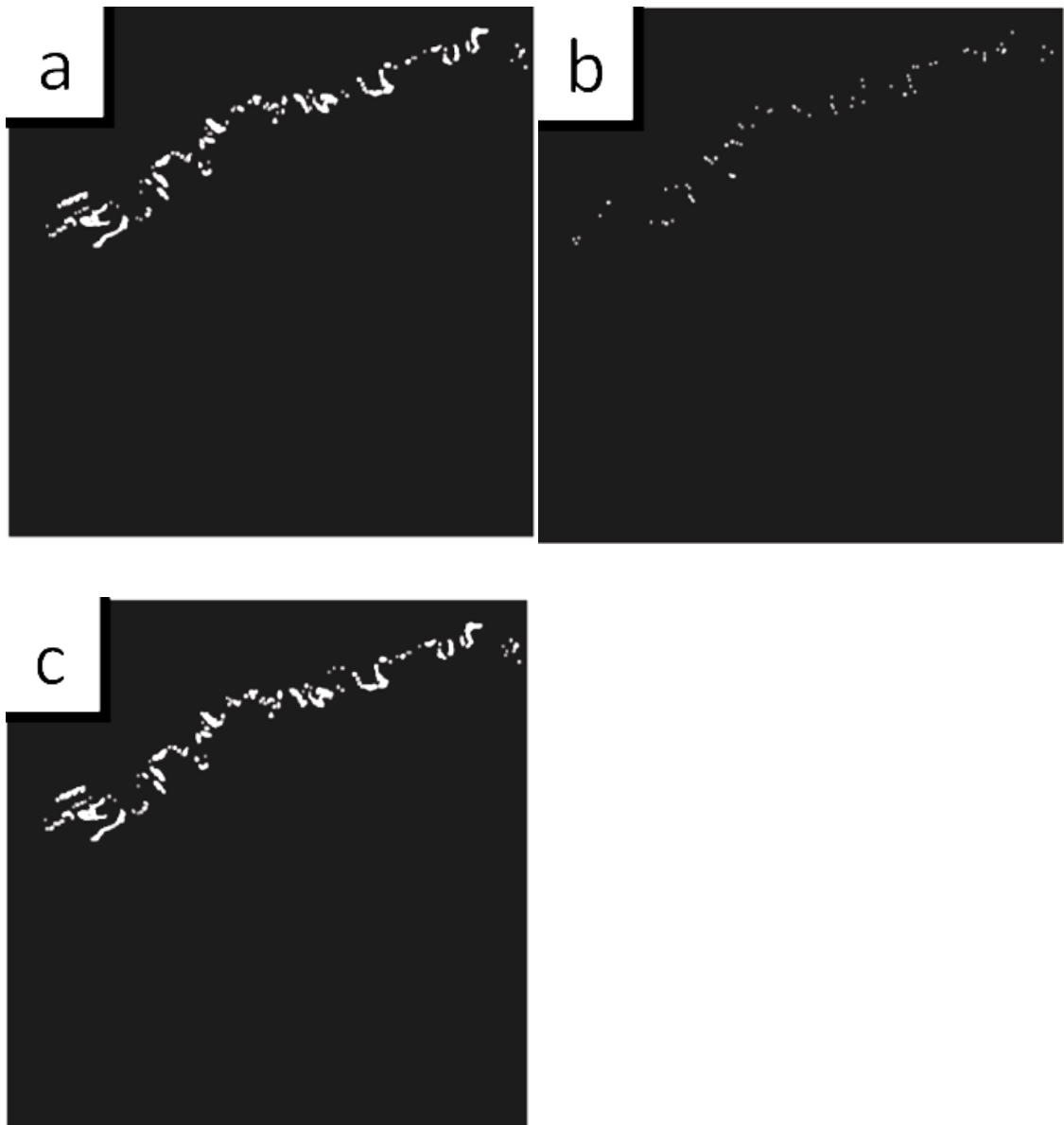


Figure 5.29 Dermal cleft objects adjacent to the epidermis (Figure 5.29a ), the epidermal cleft objects adjacent to the dermis (Figure 5.29b), and the combination of both sets of cleft objects (Figure 5.29c).

Clefts positioned near the edges of the tissue sample may be artefacts due to the slicing procedure, so the sample perimeter mask  $pMask$  (Equation 5.24) was applied next to remove them. Figure 5.30 shows a skin sample with a tear at one of the cut edges, indicated on the image with an arrow. The perimeter mask is shown masking this artefact.

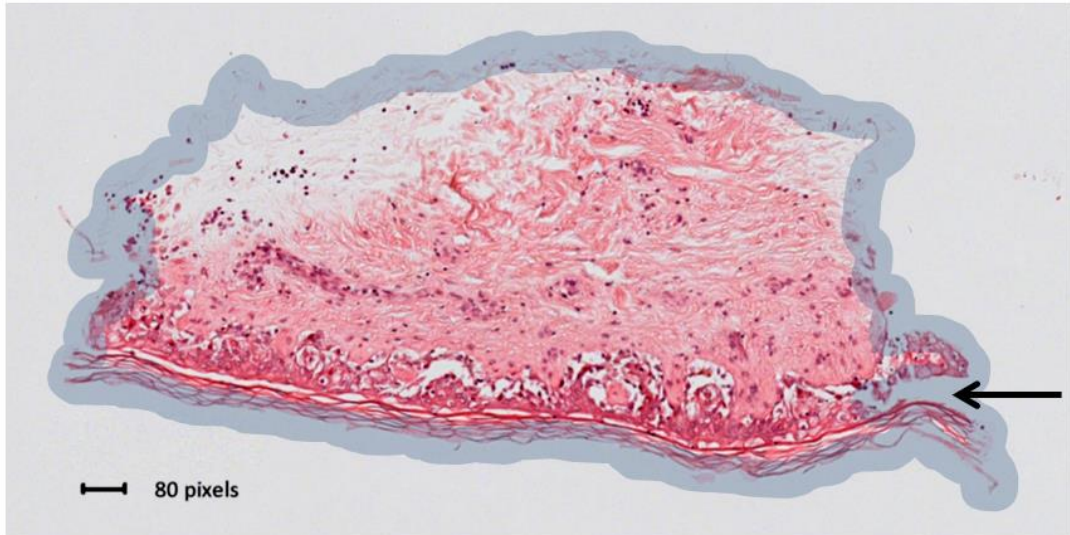


Figure 5.30 A skin sample showing grade III damage, with clefts at the DEJ. The thickened sample perimeter mask is shown masking a tear at one cut edge of the tissue.

Any objects in *cMask* that are fully or partially masked by the perimeter mask are removed from further analysis.

#### 5.4.2 Cleft Segmentation: Results

Figure 5.31 shows the effect of changing  $C_{thresh}$  (Equation 5.28) on the cleft thresholding operation on two image sections, one with clefts and one without. A total of 8 thresholds were tested, but the results of three are shown to demonstrate the impact of this step. The first image in Figure 5.31 does not have any clefts and none of the thresholds considered result in false clefts being identified at the DEJ. More of the dermis tissue is included in the cleft mask as the threshold is decreased, as observed for both images in Figure 5.31. In the second image, clefts are included in the mask when a subtraction of 20 or 40 is applied to the mode luminance value. A subtraction of 20 was selected to create the binary mask from contrast enhanced luminance image. This threshold resulted in clefts being identified and the majority of the areas of tissue with high luminance being excluded.



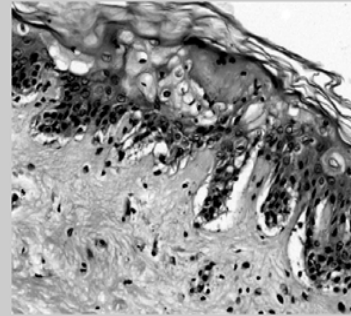
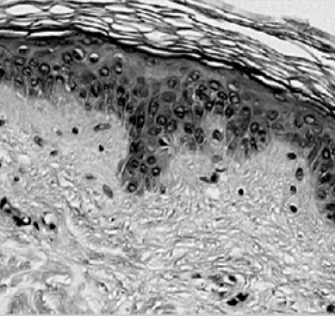



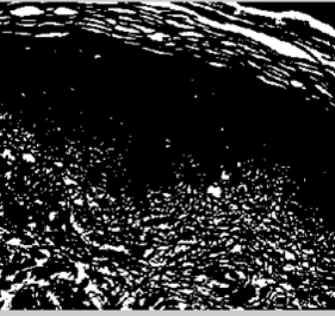

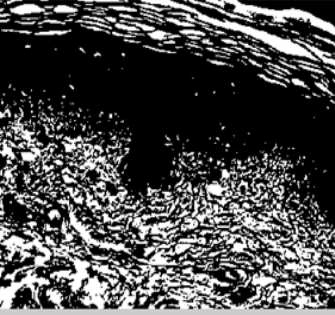
Image 2	Image 1	
		Luminance channel ( $L^*$ ) image
		Threshold at $L^*$ mode
		Threshold at $L^*$ mode minus 20
		Threshold at $L^*$ mode minus 40

Figure 5.31 The effect of different thresholds on the binary cleft mask created during thresholding of two luminance images, one containing clefts and one with no clefts.



Figure 5.32 shows a section of the original RGB image which has been processed using the full cleft segmentation, with the cleft boundaries plotted in green over the colour normalised RGB image. The procedure has identified the clefts in this image accurately. There are two regions within the epidermis that have a similar appearance to the clefts, but as they are not at the DEJ they have not been classified as clefts.

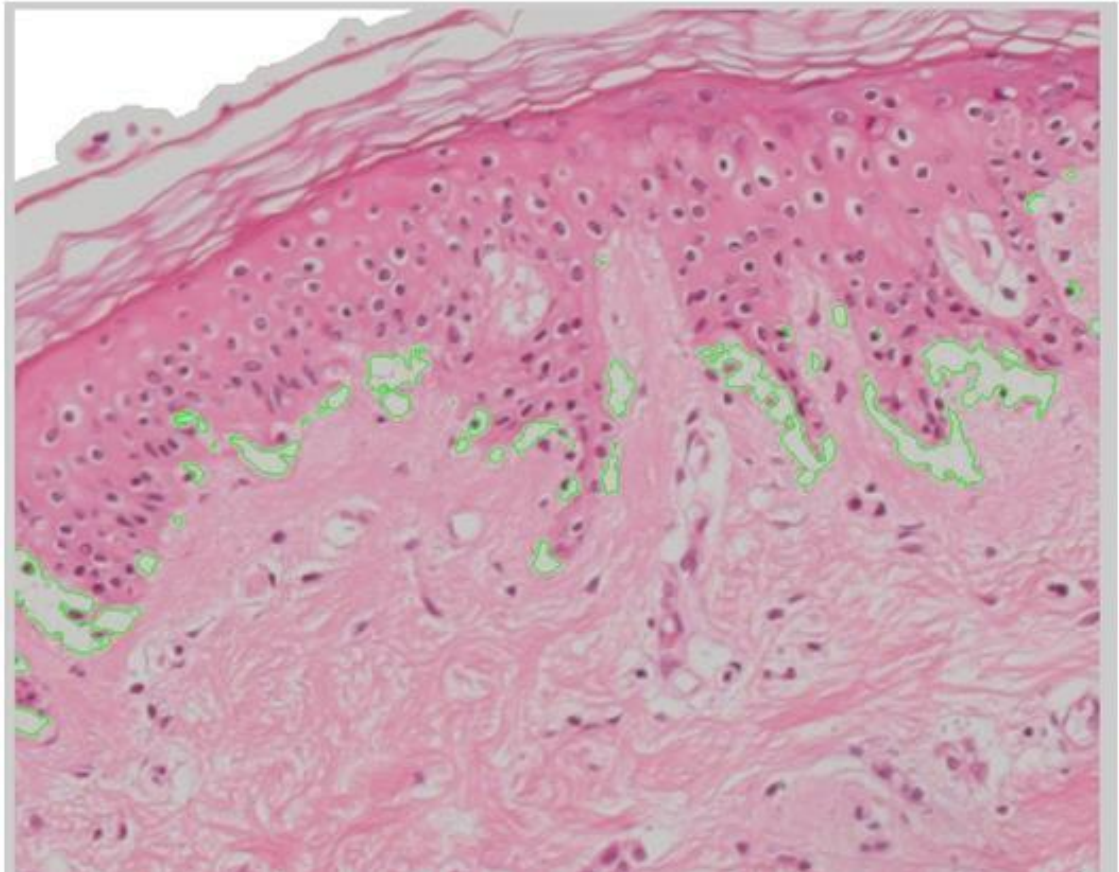


Figure 5.32 A section of the original RGB image, with the cleft boundaries identified using the cleft segmentation procedure plotted in green over the RGB image.

### 5.5 Vacuole Segmentation

Vacuoles are membrane bound cavities within the cells of the epidermis and they are not stained by H&E, meaning they are approximately the same colour as the background and lighter in colour than the surrounding cytoplasm. As with clefts, the mode luminance is used as the basis of a thresholding operation. The vacuoles do not have as strong a contrast against the surrounding tissue as the clefts, and so the threshold needed to be lower to ensure all the vacuoles were appropriately segmented. The slight difference in colour of the vacuoles compared to the

background and clefts may be because the vacuoles contain fluid and the refractive index of this fluid may be affecting the passage of light through the sample.

### 5.5.1 Vacuole Segmentation: Method

The identification of the vacuoles within the epidermis is important as they are the second major feature of damage after the DEJ clefts. As this vacuolisation occurs exclusively in the cells of the epidermis, this is the only region that needs to be analysed. As with the cleft segmentation, the contrast enhanced luminance image  $\mathbf{L}$  (Equation 5.27) is used as a starting point. The vacuoles do not have as strong contrast against the surrounding tissue as the clefts and so a larger value (100) was subtracted from the mode luminance to obtain the optimal threshold,  $V_{thresh}$ . The impact of varying the value subtracted from the threshold is described in section 5.5.2 and summarised in Figure 5.33:

$$V_{thresh} = mode(\{l_{ij} | l_{ij} \in \mathbf{L}\}) - 100 \quad \text{Equation 5.30}$$

Potential vacuoles in the epidermis were located by thresholding the epidermis pixels in  $\mathbf{L}$  at  $V_{thresh}$  to create a new binary mask  $\mathbf{vObj}$ :

$$\mathbf{vObj}_{i,j} = \begin{cases} 1 & \text{if } \mathbf{L}_{i,j} > V_{thresh} \text{ AND } \mathbf{eMask}_{i,j} = 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{Equation 5.31}$$

There was no need to perform this operation on the dermis pixels as vacuolisation occurs exclusively in the epidermis. Within the *stratum corneum* there are numerous gaps between the layers that are small and background-coloured and they can be incorrectly classified as vacuoles as a result of their similarity in appearance to them. The misclassified vacuoles occur in images where some of the *stratum corneum* has been previously misclassified as part of the epidermis. A slightly altered version of the perimeter mask,  $\mathbf{pMask}$  (Equation 5.24) was applied to mask out the *stratum corneum* in those images where it has not previously been removed. This version  $\mathbf{pMask}$  was created by locating the perimeter of the  $\mathbf{sMask}$  and performing a morphological dilation using a radius 45 'disk' shaped SE to thicken the perimeter mask. The size of SE was reduced compared to the radius 60

version used previously to exclude as much of the *stratum corneum* as possible without excluding real vacuoles located in the top layers of the epidermis. Any *vObj* objects overlapping with pixels in the *pMask* were discarded. A final mask, *vMask*, showing the location of the vacuoles was created by discarding any objects in *vObj* that were identified as clefts in *cMask*. This avoided double counting of objects on the epidermal side of the DEJ as both clefts and vacuoles.

The object area of object *Z* in *vObj*, was then determined and any object greater than 1000 pixels in area were excluded. These very large objects are likely to be artefacts or misidentified clefts. The following classification rule based on the object area was used to classify each object pixel, *z*, in *vObj*:

$$z = \begin{cases} 1 & \{ \forall z \in Z \mid Z_{Area} < 1000 \} \\ 0 & \text{else} \end{cases} \quad \begin{array}{l} \text{Equation} \\ 5.32 \end{array}$$

### 5.5.2 Vacuole Segmentation: Results

The effect of changing the threshold,  $V_{thresh}$ , on the final vacuole segmentation procedure was tested by running the procedure on a set of 16 images showing differing levels of vacuolisation, staining and contrast. The effect of changing the threshold on two typical images with different staining hues and intensities is shown in Figure 5.33. Not all tested thresholds are shown, with the actual thresholds tested being the mode luminance, and mode luminance -15, -20, -40, -60, -80, -100, and -120.

When the mode luminance was used as the threshold to segment the vacuoles, many of the vacuoles were not adequately segmented. This can be seen particularly in the second image (image 2) in Figure 5.33. When a value of 80 was subtracted most of vacuoles were segmented accurately. When a value of 100 was subtracted from mode luminance, areas of lightly stained cytoplasm were misclassified as vacuoles in some cases. These misclassified regions are indicated by the yellow arrows in Figure 5.33. The impact of the final processing step using the cleft mask, *cMask*, can also be observed in Figure 5.33. The white region in the bottom left corner of image 2 is located at the DEJ and was classified as a cleft in

the previous step. It has therefore been excluded from the vacuole mask to avoid counting this region as a cleft and a vacuole.

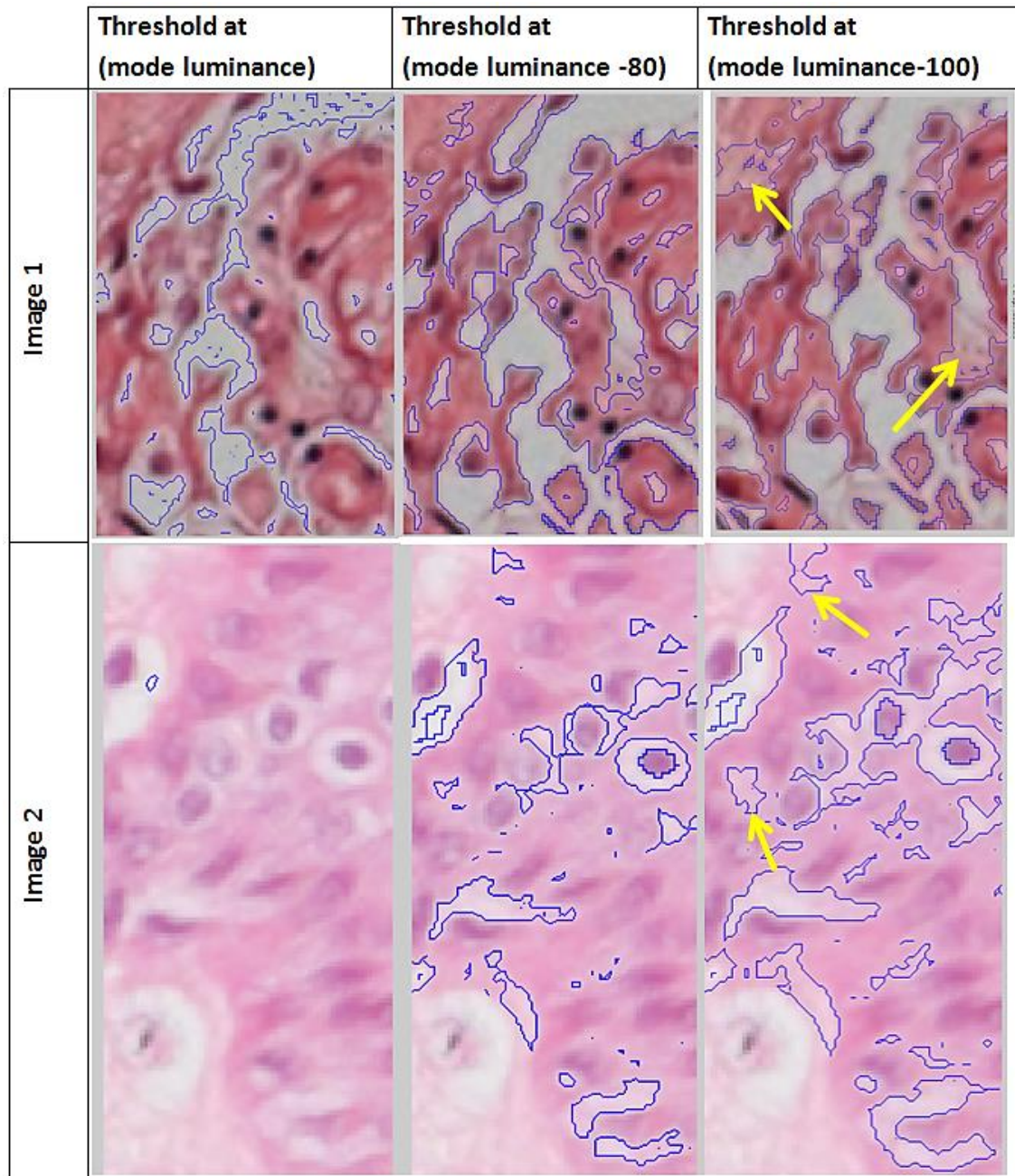


Figure 5.33 Sections of two RGB images, which have been through the whole vacuole segmentation procedure using thresholds of mode luminance, mode luminance – 80 and mode luminance -100. The final vacuole boundaries are plotted in blue over the RGB image. Yellow arrows indicate regions misclassified as vacuoles when using the lower threshold.

## 5.6 Size-based Classification of Vacuoles and Clefts

Once the initial segmentation of vacuoles and clefts had been tested it was noted that some of the vacuoles appeared to be larger than some of the clefts. This was

unexpected as clefts are formed when a number of vacuoles fuse together, meaning that clefts should be larger in size than vacuoles. The two regions within the epidermis highlighted in the discussion of Figure 5.32 were classified as vacuoles due to their location, but appear more like intra epidermal clefts or vacuole fusions. The cleft and vacuole segmentation procedures described in sections 5.4.1 and 5.5.1 differentiate between vacuoles and clefts based on their location; an object is classified as a cleft if it is at the DEJ, other objects within the epidermis are classified as vacuoles.

An alternative means of classifying the objects was developed based entirely on size. This method was tested by adding all the clefts and vacuoles identified in **cMask** and **vMask** using the procedure in sections 5.4.1 and 5.5.1, to create a new mask of all the potential vacuoles and clefts, **faultMask**:

$$faultMask_{i,j} = \begin{cases} 1 & \text{if } vMask_{i,j} = 1 \text{ AND } cMask_{i,j} = 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{Equation 5.33}$$

The objects in **faultMask** were then re-classified as vacuoles or clefts based on their area in pixels. The individual areas of clefts and vacuoles in a number of images were measured to determine an appropriate threshold. There was not a clear size threshold where a vacuole at the DEJ clearly transitions to a cleft. Instead the threshold was set at the maximum area of a “fault object” made up of 3 fused vacuoles measured within the epidermis tissue. This was done by manually removing objects at the DEJ from **vMask** using a user interaction step like that described in section 5.1.9. This was performed on the whole 16 image data set and the maximum vacuole area identified in the image set was 150 pixels in area.

Two copies of **faultMask** were created and renamed **vac\_b** and **cleft\_b**. For each object  $Z$  in **vac\_b**, each pixel  $z$  within  $Z$  was changed to a zero if the area of the object was greater than 150 pixels:

$$z = \begin{cases} 1 & \{ \forall z \in Z | Z_{Area} \leq 150 \} \\ 0 & \text{otherwise} \end{cases} \quad \text{Equation 5.34}$$

For each object  $Z$  in *cleft\_b*, each pixel  $z$  within  $Z$  was changed to a zero if the area of the object was less or equal to 150 pixels:

$$z = \begin{cases} 1 & \{ \forall z \in Z | Z_{Area} > 150 \} \\ 0 & otherwise \end{cases} \quad \begin{array}{l} \text{Equation} \\ 5.35 \end{array}$$

The segmentation and image processing procedures described in this chapter created a number of output images used to extract features. The images include the epidermis pixel mask, *eMask*, the cleft and vacuoles masks based on location and size, *cMask*, and *vMask* and finally the alternative cleft and vacuole masks based on object area, *cleft\_b* and *vac\_b*.

## 5.7 Discussion of Image Processing and Segmentation

The aim of this part of the research was to develop an algorithm capable of segmenting the epidermis, dermis and features of importance including clefts and vacuoles in images of H&E stained skin that exhibited varying degrees of histological damage. Epidermal segmentation is the first step in the automated procedure for the detection and classification of histological damage caused by immune responses within the skin. This first step is critical, because the epidermis is the part of the sample where the damage is manifested. Vacuolisation occurs within the keratinocytes (the main cells of the epidermis) and clefts form at the DEJ at the base of the epidermis. The proposed method is robust in terms of its ability to segment the epidermis even in cases where the morphology and structure has broken down, as evidenced by a mean epidermal segmentation accuracy of between 96% and 97% for sets of images showing grade I, II, III and IV damage. The three-stage process presented enables the use of a traditional and well understood thresholding technique in a challenging domain in which it would not ordinarily give good results.

After image cropping and an initial segmentation of sample pixels to improve algorithm efficiency, the main processes can be implemented. After a colour normalisation step based on histogram matching to a well-stained target image in the RGB colour space, pixel colour and staining intensity information is captured through a linear combination of two image representations. Colour information



relating to the staining is captured using a contrast enhanced  $b^*$  plane from the  $L^*a^*b^*$  colour space, and general staining intensity information is captured using a contrast enhanced greyscale image. The two image representations are then mean filtered to remove some of the cellular detail within the different tissue types and emphasise the variation between the two tissue types. Information from both image representations is combined before Otsu thresholding is performed. The segmentation is fine-tuned using morphological processing, and a final object classification step based on size and shape is applied.

The proposed method segments the epidermis from whole slide skin images with a mean specificity of 98.0%, a mean sensitivity of 91.0% and a mean accuracy of 96.8% when the performance is tested on 40 images. It offers improved performance over Lu and Mandal's (2012) multi-resolution global thresholding and shape analysis (GTSA) approach which had a 92% sensitivity rate, 93% precision and 97% specificity rate when tested on 16 images. It is also an improvement on previously published segmentation approaches for epithelial tissues such as those reported by Wang *et al* (2007a) who achieved accuracies of 94.9 – 96.3%, and Eramian *et al* (2011) who achieved an accuracy of 85%, and mean sensitivity and specificity of 91.4% and 84.6% respectively. The balance of sensitivity and specificity required is dictated by the particular application. While attempts were made to increase sensitivity as much as possible by optimising parameters on the training set, doing this at the expense of specificity could easily result in false vacuole and cleft identification and an unacceptable number of false positive results in the subsequent classification process.

The time efficiency is difficult to compare accurately with other published methods as it is dependent on the computer system. However an approximate number of pixels processed per second can be used to compare methods. The algorithm (without user interaction) processes approximately 866,432 pixels per second with a standard deviation of 124,764. On average, it takes 11.4 seconds (with a standard deviation of 5.1) to process a typical image in this dataset using an Intel quad-core 3.4GHz processor with 8GB RAM. The processing efficiency compares favourably with Eramian *et al* (2011) who quoted an average runtime of 7.2s per

image, which can be scaled to  $\sim 189,583$  pixels per second. Wang *et al* (2007a) were processing much larger images, so despite a reported runtime of 21 minutes, the pixel processing per second is of the order of  $\sim 7,619,048$ . Both Wang *et al* and Lu and Mandal's (2012) fast processing time 3,764,705 pixels per second are explained by their use of a multi-resolution approach.

In the segmentation procedure developed in this research additional time efficiencies could be gained if the some of the more time consuming functions such as colourspace conversions were translated to MEX-files. As computer processing power and speed continues to improve there is also the option to run the algorithm using parallel or cloud computing.

A colour normalisation step was included prior to the implementation of the main segmentation algorithm to enable the method to handle staining and lighting variation in the input images. The inclusion of a colour normalisation step is a trade-off between retaining as much colourimetric information as possible within the images and managing the variation resulting from sample preparation, staining and imaging. Mapping to an ideal target image can be problematic since each image has differing proportions of background to sample, and also of epidermis to dermis tissue. The effect of carrying out a mapping between differing images is that some differences are smoothed out, while others are enhanced. These effects were mitigated in this study by confining the colour mapping to sample pixels in the target and test images. The ability of the algorithm to achieve high accuracy, sensitivity and specificities despite significant variation in the input images shows the approach was effective.

Applying colour normalisation prior to the colourspace conversion and contrast enhancement steps that follow ensures that the effects of staining and lighting variation in the input image are addressed early in the process and prior to the subsequent contrast enhancement and thresholding steps. The initial colour normalisation also enhances the contrast between epidermis and dermis in images where there is poor contrast between the two tissue types. This normalisation step enables the following contrast enhancement to be more finely tuned. Without the normalisation step, variations in overall colour hue, saturation and intensity



(caused by staining and lighting differences) have a significant impact on subsequent processing steps leading to reduction in final segmentation accuracy and sensitivity. When colour normalisation was added to the process, the specificity increased by 1.69% and sensitivity by 4.93%, resulting in a mean accuracy increase from 94.03% to 96.45%.

Key parameters for the segmentation were simultaneously optimised because of the known interactions between them. For example, the mean filtering of the greyscale and  $b^*$  images affects the scale and resolution of variation within the image, and therefore impacts on the size of the structuring element required to fine tune the thresholding. Once these parameters were optimised, the method was sufficiently robust enough to work effectively on the images in the dataset.

Since the author's paper on epidermal segmentation was published in 2014, a number of other papers in this area have been published. The GTSA approach first proposed by Lu and Mandal's (2012) was improved to achieve a sensitivity of 98.0%, specificity of 99.6% and precision of 96.0% on 61 images (Lu and Mandal, 2014). In a further paper published by the same authors in 2015 on a larger set of 105 skin sample set, there appears to have been a drop in performance with a sensitivity of 95.7%, specificity of 99.4% and precision of 93.1%. Another technique for epidermal segmentation using a morphological closing and global thresholding-based technique (MCGT) was proposed by Mokhtari et al as part of research paper developing a method to measure melanoma depth of invasion.(Mokhtari *et al.*, 2014). As the epidermal segmentation was not the main subject of the paper, the accuracy of this individual step was not quoted.

In 2015 a new technique was proposed which refined the global thresholding and shape analysis (G TSA) used by Lu *et al* by adding an epidermal thickness check and a k-means classification (Xu and Mandal, 2015). The paper compared the performance of their new technique against the original GTSA approach, Mokhtari's MCGT approach and the approach developed and described in this thesis, referred to as the CET (contrast enhancement and thresholding) technique. The CET technique when reproduced and tested on a set of skin biopsies produced using a slide scanner showed precision, sensitivity and specificity of 56.5%, 91.4%

and 95.1% on the training set and 49.9%, 91.3% and 93.8% on the test set. The performance was negatively affected by the very high proportion of cell nuclei in the dermis area, which led to some of the dermis being misclassified as epidermis and resulting in a low specificity. The images used in the paper had more intense staining and a different colour profile in comparison to the training images used for this research. While it is a positive sign that the technique has been applied elsewhere and showed improved performance over other techniques in the literature, the results reinforce the point that most techniques in the area of histopathology image analysis are very application specific and would require tuning to be appropriate to different image sets.

The algorithm proposed and described in this chapter has application beyond the grading of adverse immune reactions, and is a useful framework on which to build any skin segmentation, such as epidermal segmentation prior to epidermal thickness measurements, detection of melanoma, or diagnosis of dermatological conditions such as psoriasis. Furthermore the approach could be applied to other types of tissue, in particular other epithelial tissues with H&E staining. The four critical parameters identified and optimised in the Design of Experiments study would probably need to be re-optimised for different tissues and images generated in different ways (e.g. slide scanners). Additionally, the sensitivity and specificity could be tuned using the object classification step depending on whether it was more important to minimise false positives or negatives in the new application.

The dermal segmentation is dependent on the epidermal segmentation and by focussing on the accuracy of epidermal segmentation, the accuracy of the dermal segmentation was improved. When the dermal segmentation method was tested on a set of 16 images the biological and staining variation of the samples resulted in at least a 25% over or underestimation of the DEJ in ~10% of cases. It is important that the DEJ is located accurately as this location is used to classify the clefts correctly and clefts are a fundamental differentiator between grade II and grade III skin damage. A variety of other methods including edge detection and morphological processing were tested to improve the method further, however none of the methods worked effectively for all the images and they have not been

included in this thesis due to space constraints. The performance of this step was affected by biological variability in the samples with large differences in the appearance and morphology of the *stratum corneum* and the presence of necrotic tissue necessitating the inclusion of a user interaction step to improve the accuracy of subsequent steps. It is possible that further work in this area could result in a method for automatically classifying and excluding the necrotic tissue from further analysis, including the use of a texture filter utilising the regular spacing of the dead cells' condensed nuclei and the high contrast they have in relation to the rest of the pale pink necrotic tissue.

The basic structure of the image processing and segmentation procedure developed during the epidermal segmentation was applied successfully for the segmentation of clefts and vacuoles. Using an appropriate colour channel (luminance), the subsequent contrast enhancement, thresholding, morphological processing and object classification was effective and relatively quick to develop and optimise. This would support the argument that the epidermal segmentation procedure provides a useful framework on which to build other tissue segmentations.

Two different classification methods were used to identify clefts and vacuoles; one based on location and the other using size and location, resulting in two sets of potential clefts and vacuoles. While biological knowledge could have been used to decide which classification was a better representation of the biology, formation of vacuoles and clefts is a continuous biological process and identification of these incurs a significant degree of subjectivity. Instead an objective, automated method was used to select the best type of classification. During the feature selection process described in Chapter 6 (section 6.3), measurements based on vacuoles and clefts classified by both methods were tested, and those which differentiated best between different grades of damage were automatically included in the final classification model.

The research described in this chapter is one of the main academic contributions in the thesis. There are very few published methods for the segmentation of epidermal tissue, the most similar were highlighted in the literature review and

used as benchmarks to evaluate the success of this method. The development of a new methodology is a useful contribution in the areas of dermatology, tissue segmentation, and *in vitro* assay technology. The robustness is shown by the method's high accuracy in segmentation of a challenging dataset of epidermis tissue from H&E images of human skin showing varying degrees of histological damage.

The next chapter shows how the information extracted from the skin explant images was transformed further into quantitative and histologically meaningful measures representative of the image. The information in the multiple tissue and feature masks is extracted and reduced so that it can be represented as a single column of numbers and used in an automated classification system.

## Chapter 6 Feature Extraction, Selection and Classification

The previous chapter described the segmentation of epidermal tissue, DEJ clefts and vacuoles within images of skin at various stages of immune-based damage and structural breakdown. In this chapter, the design and extraction of a set of representative feature measurements from the skin images is given in section 6.1. Section 6.2 describes the training of a classification model using these features and section 6.3 describes the validation of a classification model using an unseen feature subset. There are many parameters which were optimised during feature extraction, most of which are not dependent on the exact image spatial resolution and the specific staining and lighting properties of the image data set used to develop the process. However, in order to aid future development and application of the method, a list of parameters which would needed to be re-optimised if images of a different spatial resolution were being analysed are presented in Appendix C.

### 6.1 Feature Extraction and Selection

Following the completion of the epidermis, cleft and vacuole segmentation, a range of measurements derived from these objects were calculated to populate a feature vector for each image. Two types of features were extracted from the images, both of which were introduced in chapter 2, section 2.5.1:

- **Morphological Features:** These measurements were designed to closely reflect the expert knowledge and histological guidelines and descriptions used by histopathologists to grade images. The features included measures of the number, size, shape and variability of the vacuoles and clefts in the image.
- **Texture Features:** This set of feature measurements are mathematical descriptors of texture calculated using the grey level co-occurrence matrix (GLCM) of the epidermis regions. They offer a more abstract way of capturing the breakdown of tissue based on image texture.

#### 6.1.1 Morphological Features

Standard qualitative histological descriptions of GVHRs include diffuse or severe vacuolisation, the presence of clefts at the DEJ and the complete separation of the

dermis from the epidermis a result of cleft formation. The appearance of vacuoles at the base of the epidermis is the first indication of damage, and as the damage becomes more severe these vacuoles fuse together and clefts are formed between the epidermis and dermis. This means that vacuole and cleft based features are linked.

Vacuolisation increases in severity with increasing damage and this is shown by an increase in vacuole number and size. Features were created to attempt to capture and measure this quantitatively, based on counting the number of vacuoles normalised for epidermis area, determining the average area of a vacuole in an image and calculating the percentage area of the epidermis made up of vacuoles. Vacuole shape also appears to become more irregular with increasing damage as multiple vacuoles fuse together and so features such as extent and eccentricity were used to measure this change in shape. In addition to number, size and shape based measures the severity of clefts was also assessed based on the proportion of the DEJ affected. This was estimated using the major axis length and an approximation based on the halved perimeter of the epidermis. Finally, statistical descriptors such as inter-quartile range, skewness or kurtosis were used to describe the distribution and variability of these shape and size based characteristics across the vacuole or cleft population in an image.

The point at which a vacuole becomes a cleft is not specified by the traditional histological criteria and is an issue which must be addressed in the developed algorithm. The only definite difference in the traditional manual grading criteria between clefts and vacuoles is that a cleft must be located at the DEJ, while a vacuole can occur anywhere in the epidermis. In the previous chapter, this issue was addressed by creating four masks based on two different classifications, the first used size and specific location, whilst the second primarily used size as a classifier and did not discriminate based on feature location. By creating these new classifications it is possible to include additional information in the classification process compared to the traditional approach.

Size and location based discrimination of vacuoles and clefts

- *Vac\_a*: These are transparent regions found within the epidermis with an area of less than 1000 pixels. A fairly high size threshold was chosen to ensure regions of fused vacuoles within the epidermis were included. As clefts were identified first, they were subsequently excluded from the set of vacuole objects to avoid double counting.
- *Cleft\_a*: These are transparent regions, of any size, located at the DEJ. This set will include even very small faults which could potentially be described as vacuoles, but only if they are located at the DEJ.

#### Size based discrimination of vacuoles and clefts

- *Vac\_b*: These are transparent regions at the DEJ or in the epidermis with an area of less than 150 pixels. The lower threshold ensures that only single vacuoles and clusters of two or three fused vacuoles are included.
- *Cleft\_b*: These are transparent regions at the DEJ or in the epidermis with an area of greater than 150 pixels. This classification method ensures that very small faults are counted as vacuoles, even if they are located at the DEJ.

For each of the four sets of objects, a number of properties were measured:

- *Number* – The total number of a specific type of object in the image was determined and normalised by the area of epidermis in the image.
- *Area* – For each set of objects in each image, the sum of all object areas was divided by the epidermis area and used to calculate the percentage area of the epidermis covered by this type of object. In addition the mean, maximum, median, standard deviation, interquartile range, skewness and kurtosis of the object areas were calculated.
- *Eccentricity* – For each set of objects, in each image, the mean, maximum, median, standard deviation, interquartile range, skewness and kurtosis of the object eccentricities were calculated.
- *Extent* – For each set of objects, in each image, the mean, maximum, median, standard deviation, interquartile range, skewness and kurtosis of the object extents were calculated.

In addition, for the two sets of clefts the following additional properties were measured:

- *Cleft coverage of DEJ* – For each set of clefts, in each image, the total pixel distance covered by clefts was approximated in two ways. First by dividing the sum of all cleft perimeters by 2, and also by calculating the sum of all cleft major axis lengths. These two approximations were each normalised using two approximations for DEJ length. The first was the epidermis perimeter divided by 2 and the second was the sum of the epidermis major axis lengths.
- *Cleft Major Axis Length* - For each set of clefts, in each image, the mean, maximum, median, standard deviation, interquartile range, skewness and kurtosis of cleft major axis lengths were calculated.
- *Cleft Dimension* – For each set of clefts, in each image, the sum of all cleft major axis lengths was divided by the sum of all cleft minor axis lengths.

Approximations, including major axis length and perimeter divided by 2, were used to estimate the proportion of the DEJ covered by clefts. A direct measurement of the DEJ based on adjacent dermis and epidermis pixels would be subject to the errors caused by misclassification of the *stratum corneum* discussed in chapter 5, section 5.3.3. To avoid these errors having an impact on the extraction of features indicating cleft coverage of DEJ, alternative approximations were used. A variety of approximations were used purposely, with the aim of identifying the best objectively in the subsequent feature selection step.

### **6.1.2 Texture Features**

Texture is a property of image areas rather than individual pixels and involves the spatial distribution of grey level or colour. First order statistics describe properties of individual pixel colours, without considering interactions with neighbouring pixel values. These statistics therefore measure the likelihood of observing a given grey level or colour at a given location. Second order statistical features describe pixel grey level or colour relationships, and are properties of pairs of pixels, providing a quantitative measurement that can be used to describe the texture of an image. These features quantify the likelihood of observing two specific grey level or colour values at a given distance and orientation from one another using



the GLCM of an image. The mathematical theory behind the GLCM was described in chapter 2, section 2.5.1.

The texture features were calculated based only on the epidermis regions, as this is the main area of damage. This means that any clefts or vacuoles near the DEJ which are included as part of the dermis segmentation rather than the epidermis segmentation did not contribute to the texture measurements. However, these features at the DEJ should have been captured in the morphological measures. The primary purpose of including the texture measurement was to provide additional features to differentiate between grade I and II damage by quantifying subtle changes in texture and colour patterns caused by vacuolisation. The grade I/ II classification is the most difficult and error prone for human operators (see chapter 4, section 4.3 for evidence of this). It depends on an increase in vacuolisation and the occasional presence of dyskeratotic bodies rather than the definite presence or absence of a particular feature, making the decision prone to subjectivity. Dyskeratotic bodies are only sometimes present in grade II samples and are difficult to identify. As a result they were not a major focus within this research. It is however possible that dyskeratotic bodies, which generally have more intense pink staining and a smaller, darker nucleus than surrounding cells in the epidermis, could be identified by one of the texture features being extracted in addition to changes due to vacuolisation.

Each texture measure was determined for the intensity distributions in six colour channels: red, green and blue channels of the mapped RGB image, and the luminance ( $L^*$ ), red-green chromacity ( $a^*$ ) and yellow-blue chromacity ( $b^*$ ) of the transformed  $L^*a^*b^*$  image. The GLCM was calculated based on a version of the image scaled to 10 intensity levels. The rationale for this was to reduce the high spatial frequency components relating to noise in the image, and focus on major textural changes. The features were based on pixel pairs spaced at a distance of five pixels in four directions, as shown in Figure 6.1.

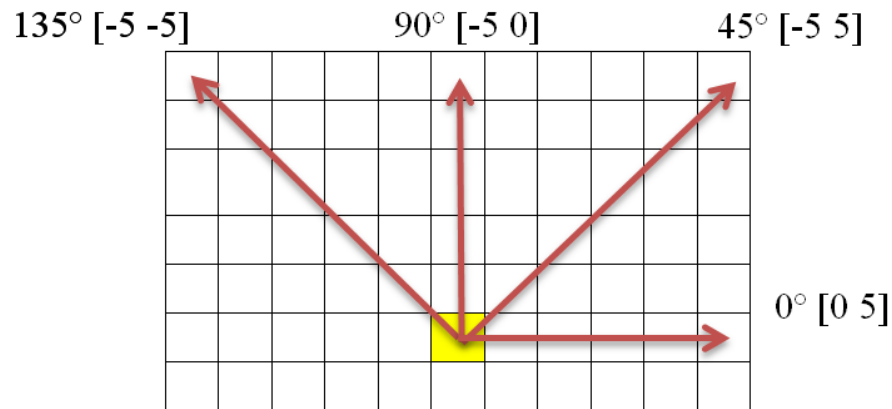


Figure 6.1 Four directions and sets of pixels pairs used to calculate texture features

Including four directions ensured intensity distribution patterns occurring in different orientations were captured. This is important because the cells in the epidermis are not in a fixed orientation either between images or within a single image. The spacing of 5 pixels was chosen to capture variation within a single cell.

Four texture features were calculated based on the intensity distributions within the epidermis, from 6 colour channels, yielding a total of 24 texture features for each image:

- *Contrast*: Measure of intensity variance or contrast between the pixel pairs over the epidermal pixels (Chapter 2, Equation 2.17). This provides a measure of image smoothness.
- *Correlation*: Measure of joint probability occurrence or correlation of the pixel pairs over the epidermal pixels (Chapter 2, Equation 2.18). Linear structures in a given direction will tend to result in a large correlation value in this direction.
- *Energy*: The sum of squared elements in the GLCM, which is the angular second moment and is a measure of uniformity (Chapter 2, Equation 2.19). The fewer grey level transitions within the epidermis, the larger the energy.
- *Homogeneity*: Measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal, and is also measure of uniformity in the epidermis (Chapter 2, Equation 2.20). The measure reflects the degree of repetition amongst the grey level pairs.

The full feature set included 23 features based on *Vac\_a*, 35 features based on *Cleft\_a*, 23 features based on *Vac\_b*, 35 features based on *Cleft\_b* and 24 GLCM texture features. A full list of all the different features extracted and the objects sets they were extracted from is given in Appendix B. The set of 140 features generated included features likely to be highly correlated. For example, the measurement sets extracted from the two vacuole sets were likely to be correlated due to the sets containing many of the same objects. The following section describes how this feature set was reduced using an objective, mathematical approach to remove uninformative, correlated and redundant features and identify those features which, in combination, provide the greatest level of information to differentiate between grades of damage.

### **6.1.3 Feature Selection**

The main categories of feature selection methods, introduced in chapter 2, section 2.5.5, are filter, wrapper and hybrid methods. The filter approach was not selected as it ranks features based on their individual relevance to the classification task using a univariate approach, and hence does not account for complex inter-dependence and correlation of features or account for feature redundancy. The reliance on individual features ignores interaction of features which is likely to be prominent in the skin image dataset due to the link between vacuolisation and cleft formation described in section 6.1.1. In addition, the failure of filter methods to consider redundancy could result in many very similar features being selected, for example, the five most relevant features identified using the filter approach may all be related to cleft area or shape, however using these features would be unlikely to help differentiate between grade I and II damage, where no clefts are present.

Wrapper methods account for feature inter-dependence and feature redundancy and are therefore better suited than filter methods for selecting features when building a classifier. Although they require more computational power than filter methods the feature set in this research is small enough that this approach is feasible. In the wrapper approach adopted, the feature selection algorithm searches for the subset of features which will maximise the predictive performance of a Naïve Bayes classifier using sequential forward and backward feature

selection (introduced in chapter 2, section 2.5.5). The classifier is used to estimate the predictive accuracy of the classifier with each potential subset of features. Cross-validation was employed to prevent make the best use of the limited dataset and avoid overfitting of the model.

#### **6.1.4 Data Preparation and Use in the Classification Task**

##### **X Data: The matrix of feature vectors**

A 181 x 140 data matrix,  $\mathbf{X}$ , was prepared using all the feature vectors. Each row represented one of the images, including 125 from the original dataset and 56 from the additional validation set provided by Alcyomics. Each column represented one of the 140 features.

A review of these images with expert histopathologists from Alcyomics resulted in 8 of these images being excluded from further analysis as they were deemed to be either too poorly stained, unrepresentative or “ungradeable” (see chapter 4, section 4.4 for full manual grading results). On review of the validation set, four images were removed for similar reasons as above. Once these samples were removed the data matrix,  $\mathbf{X}$ , was 169 x 140 containing 61 negative (grade I) images and 108 positive (grade II, III, IV) images.

Some entries contained “Not a Number” (NaN) values, which occurred where there was missing data for the calculation of a measurement, for example when calculating the cleft area normalised by epidermal area in images with no clefts, the calculation of  $0/\text{epidermal area}$  would produce a NaN value. One approach considered was to replace these values with the mean or median for the particular measurement relevant to the specific grade of damage. However it would not be possible to do this for new observations presented to the algorithm, as the grade of damage for these new samples would be unknown. Introducing a different pre-processing procedure for training and test images would most probably lead to poor generalisation performance. Instead, it was decided that the NaN values would be left as they were for the feature selection process as the methods used for feature selection and model training are able to cope with the presence of this type of data. The next step was to scale the data so all the features were mapped to a common scale between 0 and 1:

$$b = \frac{(a - \min(\mathbf{a}))}{\max(\mathbf{a}) - \min(\mathbf{a})} \quad \text{Equation 6.1}$$

where  $a$  is the original value and  $b$  the scaled value.

### **Y Data: The vector of grade labels**

A 169 x 1 vector,  $\mathbf{y}$ , was created containing the correct grade for each image in  $\mathbf{X}$ , agreed by two expert histopathologists. The manual grading process, including an assessment of the intra-observer agreement is presented in chapter 4, section 4.7. A second 169 x 1 vector,  $\mathbf{yb}$ , was created containing a binary label of 0 for the Grade I images (a negative result in the assay), and 1 for the Grade II, III and IV images (a positive result).

For the final, independent validation of the classifier, 20 images with the same proportion of positive and negative grades as in the original data set were removed. The remaining 149 images were used for feature selection and classifier optimisation. The validation set will be referred to as  $\mathbf{Xv}$  and the training set as  $\mathbf{X}$ .

#### **6.1.5 Methodology for Selecting Features**

To determine the feature subset, an initial state, termination condition and search strategy were defined. In the first round of feature selection, the initial state was an empty feature set, sequential forwards feature selection was used as a search strategy, and feature selection was terminated once 25 features had been selected. The criterion for adding or removing features was an improvement in the classification accuracy on the test set, determined using 10-fold cross-validation. The following section describes the method for this feature selection.

For selection of first feature

- a) Create a random split of  $\mathbf{X}$  into training data and test data, putting 90% of the images into  $\mathbf{X\_Train}$  (149 x 140), and the remaining 10% in  $\mathbf{X\_Test}$  (20 x 140).
- b) Identify the equivalent grading data and store as  $\mathbf{yb\_Train}$  and  $\mathbf{yb\_Test}$ .
- c) Estimate the parameters of the Naïve Bayes model (see section 6.1.6 for further explanation of this process) using column 1 of  $\mathbf{X\_Train}$  to predict the outputs in  $\mathbf{yb\_Train}$ .

- d) Test the predictive accuracy of the model (see section 6.1.7 for further explanation of this process) by comparing predictions from **X\_Test** with the actual grades in **yb\_Test** and calculating the rate of misclassification.
- e) Repeat steps c and d for each of the remaining features, calculating the misclassification rate in **X\_Train** each using each feature.
- f) Repeat steps (a – e) 10 times using a different split of the data each time, so all images have been used once in the test set.
- g) Calculate the average misclassification rate for each feature over the ten cross validation runs.
- h) Identify the feature with the lowest average misclassification rate, add to a new feature subset, **Z** and remove this feature from the dataset **X**.

To select each subsequent feature (up to a maximum of 25)

- a) Create a random split of **X** into training data and test data, putting 90% of the images into **X\_Train**, and the remaining 10% in **X\_Test**.
- b) Identify the equivalent grading data and store as **yb\_Train** and **yb\_Test**.
- c) Estimate the parameters of the Naïve Bayes model using the feature/ features in **Z** plus the first column of **X\_Train** to predict the outputs in **yb\_Train**.
- d) Test the predictive accuracy of the model by using **X\_Test**, comparing the predictions with the actual grades in **yb\_Test** and calculating the rate of misclassification.
- e) Repeat steps c and d for each of the 139 remaining features in **X\_Train**, calculating the misclassification rate when each feature is added.
- f) Repeat steps (a – e) 10 times using a different split of the data each time in step a, so all images have been used once as the test set.
- g) Calculate the average of the misclassification rates for each feature in **X\_Train** over the ten cross validation runs.
- h) Identify the feature that resulted in lowest average misclassification rate when it was tested in the feature subset **Z**. Add the feature to **Z** and remove from **X**.
- i) Repeat steps six to ten until **Z** contains 25 features.

The whole subset selection as described above can be repeated to overcome the variance in prediction accuracy which results from choosing random subsets for training and testing during cross validation.

#### **6.1.6 Model Estimation of the Naïve Bayes Classifier**

For model estimation of the Naïve Bayes model, the training data was used to estimate the parameters of a probability distribution, assuming conditional independence of features given the class. The prior probabilities were estimated using the relative frequency of each class in the training set. The data was modelled using a kernel smoothing density estimate, which is suitable for features with a continuous distribution, but does not assume there is a normal distribution. It is therefore suitable when the distribution is skewed or has multiple peaks or modes. This is ideal, as the distribution of the features used in this research have complex and non-normal distributions. For each feature, a separate kernel density estimate was made for each class. Initial test runs suggested that a normal (Gaussian) kernel type showed similar or better results than other distribution types such as Box, Triangle or Epanechnikov and so this distribution was used during feature selection. It was decided that other distribution types would be tested more thoroughly when the model was being fine-tuned after feature selection. The classifier automatically selected a kernel width for each feature and class.

#### **6.1.7 Testing of Predictive Accuracy**

Once the model had been created, it was used to compute the posterior probability of a new observation belonging to each class. Each new sample was then classified into the class for which it had the highest posterior probability. Once all observations in the test set were classified, the number of observations which were classified incorrectly (based on the expert manual classification) were summed and the misclassification rate calculated by dividing the number of misclassified images by the total number of images in **X\_Test**.

#### **6.1.8 Results for Forwards Feature Selection using Cross Validation**

Figure 6.2 shows how the misclassification rate changes as features are added. Only three of the ten runs are shown in the interests of clarity. In this forward

feature selection the misclassification rate drops as features which help the model to differentiate between observations are added. However as more features are added the problem of nesting, where a local minima rather than the real minima is reached, begins to have an impact on the results. The effect of nesting can be seen in run 3; a local minima was reached at 5 features, followed by an increase in misclassification at the addition of the 13th feature, and a subsequent further drop in misclassification rate to its lowest level once 25 features had been added.

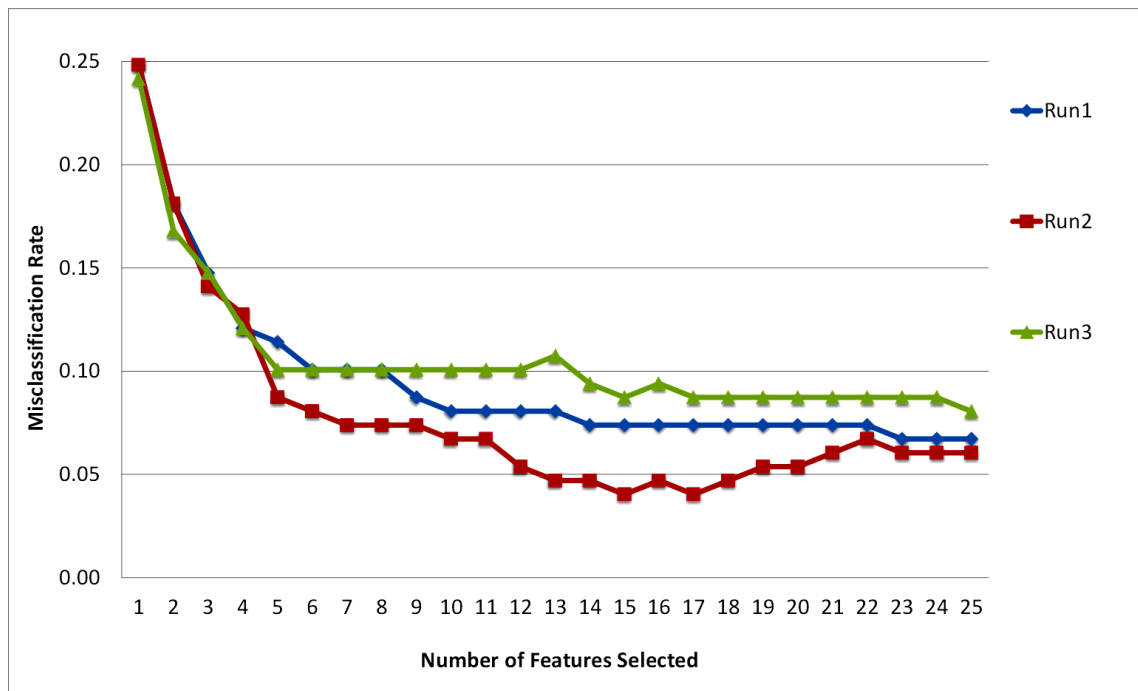


Figure 6.2 Change in misclassification rate as the 25 best features are added sequentially.

The nesting issue makes it difficult to know if a true minima has been reached, but a clear sign is the peaking phenomenon where a sharp increase in misclassification rate is seen with the addition of features after the true minima has been reached. It was expected that the effect of peaking would be seen clearly in this experiment, with a sharp increase in misclassification rate caused by overfitting of the model as more features were added. However, since this was not the case, the experiment was repeated adding a total of 50 features in each run.

As can be seen in Figure 6.3, there was no sharp increase in misclassification rate once the initial minima had been reached even when up to 50 features were added. One explanation for the absence of peaking is that the initial features chosen were so dominant that the influence of the other features was masked.



A reduced subset of features was created, including the following features

- Those present in Run 2 of the 25 feature test when the misclassification rate was at its minimum, 0.0403.
- Those present in Runs 1 and 5 of the 50 feature test when the misclassification rate was at its minimum, 0.0604.

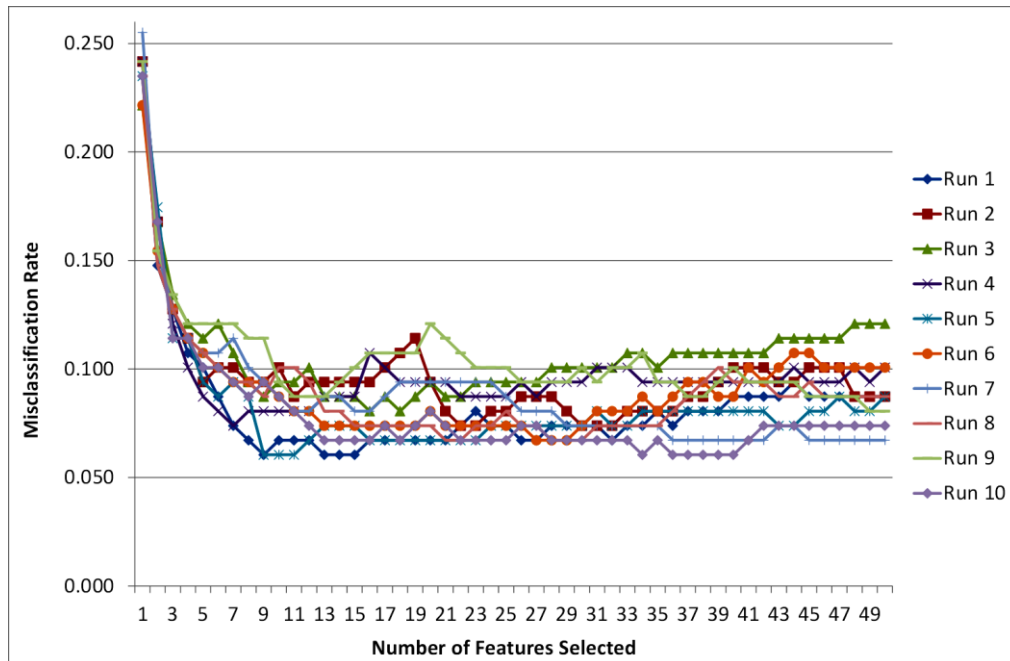


Figure 6.3 Change in misclassification rate as the 50 best features are added sequentially.

The feature selection process had created a reduced feature subset of 31 features.

To test whether the subset could be reduced further without affecting the classification accuracy, backward feature selection (or feature removal) was carried out fitting with a normal kernel distribution.

### 6.1.9 Wrapper-based Backwards Feature Selection using Cross Validation

Rather than allowing the algorithm to stop the feature selection early, the removal of features was continued until only one feature remained to obtain more information and increase the likelihood of finding the real minima. Although it is usually accepted in the literature that 10 fold cross validation balances bias and variance when estimating prediction error in classification tasks (Kohavi, 1995; Rodríguez *et al.*, 2010), an experiment was carried out to check this assertion and test whether the models had high bias or variance when the feature sets were changed significantly. The backwards feature selection of the 31 feature subset was

repeated using  $k$  values of 2, 5, 10, 20 and 40 for the  $k$ -fold cross-validation. Each cross-validation train and test run was repeated 3 times, making a total of 15 runs. Figure 6.4 shows how the mean misclassification rate drops as redundant features are removed; the data plotted is the average over the three runs. The figure shows that the misclassification rate decreased initially as the first 10-15 features were removed (the first measurement made is on the right side of the graph, with all 31 features). The misclassification rate then increased as the number of features still included in the model approached zero. This increase occurred because there was not enough information in the training data to teach the model to discriminate between classes.

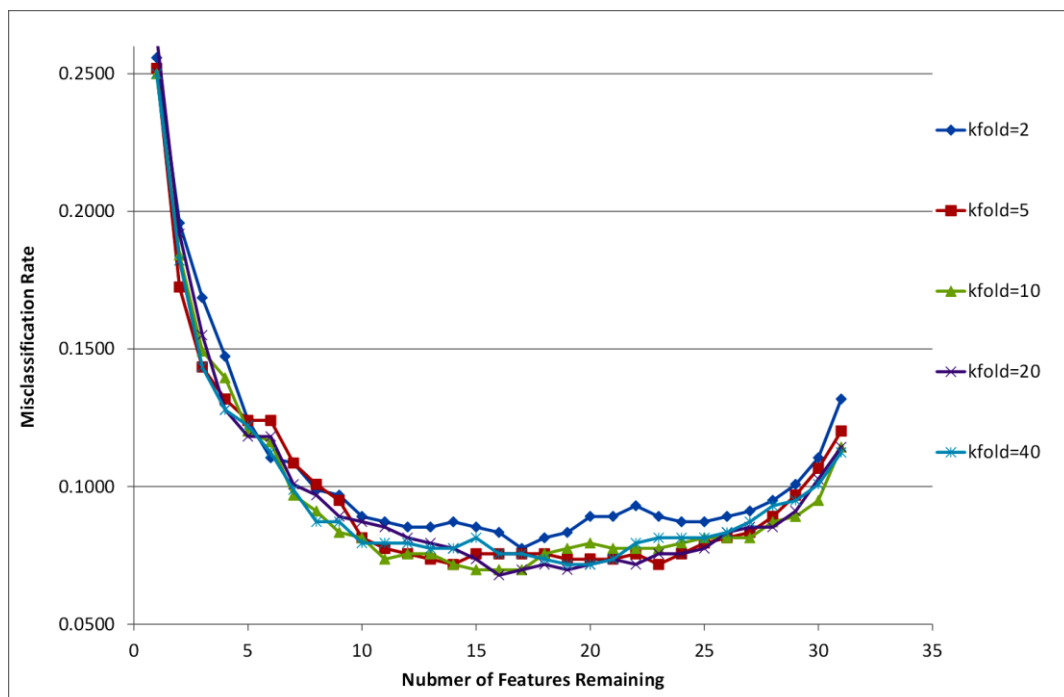


Figure 6.4 Change in misclassification rate as features are removed from the 31feature subset. There was little difference in the average misclassification rates at a given number of features when using the different  $k$  values, however the lowest average misclassification rates resulted from the 10 and 20 fold cross validation experiments.

The variance in the 3 runs when using the different  $k$  values was also investigated. The standard error of the mean (SEM) of the misclassification rate across the three runs is plotted in Figure 6.5. The SEM is the sample estimate of standard deviation divided by the square root of the number of samples.

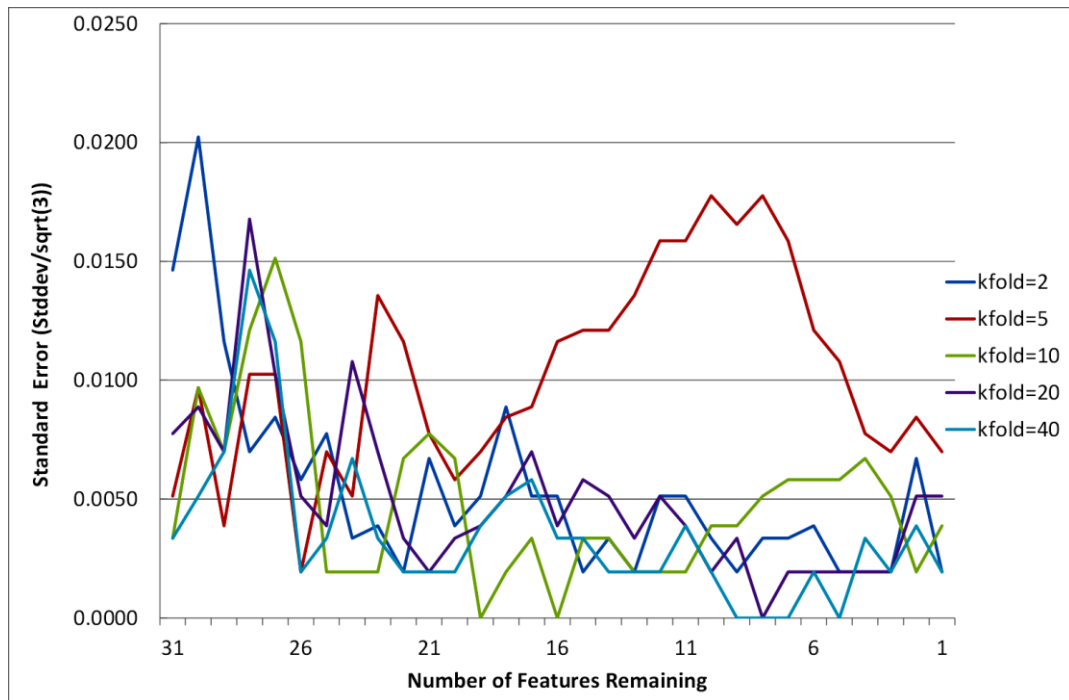


Figure 6.5 Change in standard error of the mean as features are removed from the 31 feature subset

It would be expected that there would be more variance at low values of  $k$ , when the influence of training-test set split is greater and less variance at high values of  $k$ . This can be seen in the cross-validation when  $k=5$ , where one run had a very different misclassification profile to the others, creating the large increase in variation between the runs as the final 20 features were removed. However in general there was not a clear trend of increasing variance with increased  $k$  value in this limited study. The other thing worthy of note in Figure 6.5 is that there was more variation between the runs at the start when the most redundant features were being removed; there was then a trend of decreasing variance as more features were removed. This is likely to be because the majority of the features were not particularly relevant to the task and so the choice of which one to remove was variable.

While the difference in variance between the different cross-validation runs was not immediately obvious, the data does show that the lowest misclassification rates were achieved using the lower  $k$  values to split the data. While this may be the result of an unusual model being trained with a more novel training set, it is possible that it is simply the result of a favourable data split and so for this reason

the best runs from the 20 and 40-fold cross validation experiments were also included.

The feature subset was reduced, by including only the following features

- The 19 features present in Run 2 of the 5-fold cross-validation test when the misclassification rate was at its minimum, 0.0523. The smallest feature set to achieve this misclassification rate was chosen.
- The 11 features present in Run 1 of the 10-fold cross-validation test when the misclassification rate was at its minimum, 0.0581.
- The 15 features present in Run 2 of the 20-fold cross-validation test when the misclassification rate was at its minimum, 0.0640. The smallest feature set to achieve this misclassification rate was chosen.
- The 13 features present in Run 2 of the 40-fold cross-validation test when the misclassification rate was at its minimum, 0.0698. The smallest feature set to achieve this misclassification rate was chosen.

This resulted in a reduced subset of 22 features (some of the features were present in the more than one of the best performing runs). Although the feature set was reduced further, it is of interest to consider which type of features were still included at this stage as they had been selected as the most relevant to the classification task.

#### **6.1.10 Analysis of Feature Subset**

Table 6.1 shows the features included in the 22 feature subset. The features relating to the vacuoles in *Vac\_a* describe the distribution of their areas and eccentricities. Those associated with *Cleft\_b* relate to the area of clefts compared to the area of epidermis and also the shape measurements of extent and eccentricity. The features relating to the vacuoles in *Vac\_b* describe the distribution of their areas and eccentricities, as with vacuoles set a, but there are also feature relating to the extent measurements. Those associated with *Cleft\_b* relate to the area of clefts compared to the area of epidermis, the number of clefts and the proportion of the DEJ they cover. Finally there are six texture features, five of which were extracted from the  $L^*a^*b^*$  image, a colourspace also found to be useful during the epidermal, cleft and vacuole segmentation. In summary, the remaining features are

a set of measurements that describe vacuolisation, cleft formation and changes in the texture of the epidermis. There remain features likely to be correlated, particularly those measuring the same property in the two vacuoles sets or the two clefts sets.

Table 6.1 Summary of features retained in the 22 feature subset

Feature Group	Feature Description
<b>Vac_a (size and location based discrimination of vacuoles)</b>	Inter-quartile range of vacuole (a) areas in image
	Skewness of vacuole (a) areas in image
	Mean eccentricity of vacuole (a) areas in image
	Standard deviation of vacuole (a) eccentricities in image
	Kurtosis of vacuole (a) eccentricities in image
<b>Cleft_a (size and location based discrimination of clefts)</b>	Percentage area of epidermis covered by clefts (a)
	Median extent of clefts (a) in the image
	Max eccentricity of cleft (a) areas in image
<b>Vac_b (size based discrimination of vacuoles)</b>	Standard deviation of vacuole (b) areas in image
	Inter-quartile range of vacuole (b) areas in image
	Mean eccentricity of vacuole (b) areas in image
	Skewness of vacuole (b) extents in image
	Kurtosis of vacuole (b) extents in image
<b>Cleft_b (size based discrimination of clefts)</b>	Percentage area of epidermis covered by clefts (b)
	Number of clefts (b) in image, normalised for epidermis area
	Sum of all cleft (b) major axis lengths, divided by sum of all epidermis object major axis lengths
<b>Texture based features</b>	Contrast of L*channel (epidermis pixels only)
	Contrast of a* channel (epidermis pixels only)
	Correlation of a* channel (epidermis pixels only)
	Energy of a* channel (epidermis pixels only)
	Contrast of b* channel (epidermis pixels only)
	Contrast of blue channel (epidermis pixels only)

### 6.1.11 Wrapper-based Backwards Feature Selection using 10 fold Cross Validation

The 22 feature subset contained features which were likely to be correlated. A second backwards feature selection was carried out to see if the 22 feature subset could be further reduced without affecting the classification accuracy. 10-fold cross validation was used and run a total of 10 times, as previous experiments had not provided any clear evidence that an alternative approach was superior. The data is plotted in Figure 6.6.

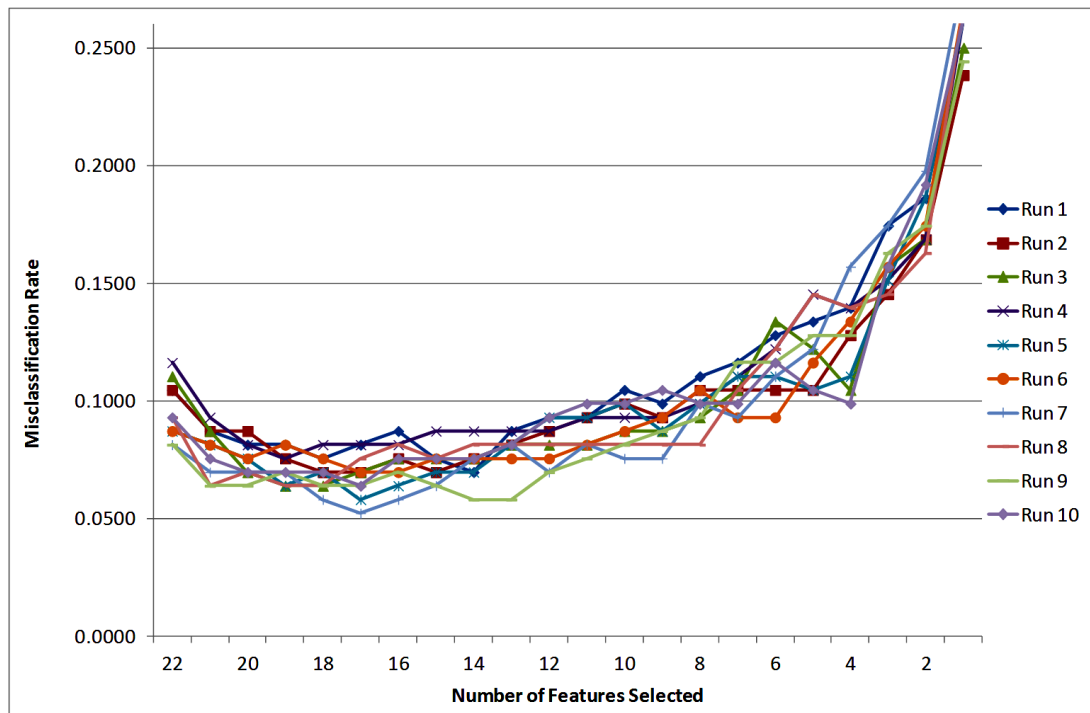


Figure 6.6 Change in misclassification rate as features are removed from the 22 feature subset

As a whole, the data suggest that the optimal number of features is between 18 and 16. The minimum misclassification rate seen was 0.0523, when the number of features had been reduced to 17 in Run 7, which is the same as the minimum seen during the previous round of feature selection, suggesting that the features removed were redundant. The 17 features present in Run 7 when the misclassification rate was 0.0523 were selected as the final feature set. In reducing the feature set from 22 to 17, two of the features removed related to the distribution of vacuole areas, of which two measurements remained. Another was a measure of cleft extent, and the final two features removed were extracted from the b\* and blue images planes, meaning all remaining texture measurements were extracted from the L\* (luminance) and a\* (red-green) colour channels.

### 6.1.12 Final Feature List

The final feature subset selected for the classification model is shown in Table 6.2.

Table 6.2 Final feature subset used in classification model

Feature Group	Feature Description
<b>Vac_a (size and location classification)</b>	Skewness of vacuole (a) areas in image
	Mean eccentricity of vacuole (a) areas in image
	Standard deviation of vacuole (a) eccentricities in image
	Kurtosis of vacuole (a) eccentricities in image
<b>Cleft_a (size and location classification)</b>	Percentage area of epidermis covered by clefts (a)
	Max eccentricity of cleft (a) areas in image
<b>Vac_b (size classification)</b>	Inter-quartile range of vacuole (b) areas in image
	Mean eccentricity of vacuole (b) areas in image
	Skewness of vacuole (b) extents in image
	Kurtosis of vacuole (b) extents in image
<b>Cleft_b (size classification)</b>	Percentage area of epidermis covered by clefts (b)
	Number of clefts (b) in image, normalised for epidermis area
	Sum of all cleft (b) major axis lengths, divided by sum of all epidermis object major axis lengths
<b>Texture based features</b>	Contrast of L*channel (epidermis pixels only)
	Contrast of a* channel (epidermis pixels only)
	Correlation of a* channel (epidermis pixels only)
	Energy of a* channel (epidermis pixels only)

## 6.2 Final Model Training

Having selected a set of representative feature measurements from the skin images, a classification model was trained using these features. The training was carried out multiple times and the best classification model was selected by estimating the error of each model that was produced. The resubstitution error is the proportion of misclassified images in the training set. This measure tends to be an optimistic indicator of future performance because it is based on the same training data used for learning by the classifier. For this reason an alternative measurement of error was used to select the optimal classifier. The cross validation error measures the proportion of misclassified images in a test set not

used to train the classifier. Cross-validation produces an effectively unbiased error estimate, but the estimate can be highly variable.

The final model was selected using 10 fold cross validation on 149 training images. The cross validation error was estimated based on 20 runs and results are shown in Table 6.3.

Table 6.3 Misclassification error estimated using 10 fold cross validation

Run number	Misclassification CV error on the 15 test samples
1	0.0667
2	0
3	0
4	0.0667
5	0.0667
6	0.0667
7	0.1333
8	0.0667
9	0.0667
10	0.0667
11	0
12	0
13	0
14	0.2000
15	0.0667
16	0.0667
17	0.0667
18	0.1333
19	0
20	0.0667

The variation in cross validation error reflects the fact that there was between 0 and 3 images misclassified for all runs and this varied as the training set was changed. The average of the misclassification errors was taken to estimate the cross validation error resulting in a final cross validation error of 0.060 which means a classifier trained using these features will be 94% accurate. The



resubstitution error on the 149 training set was calculated to be 0.046, suggesting a 95% accuracy.

### 6.3 Final Model Validation

The final performance of the classification model was then validated by using the model to classify 20 observations removed from the data set prior to feature selection and thus not involved in any of the model training. In this final test, 3 of the 20 images were misclassified which equates to misclassification error of 0.15. The final model therefore had 85% accuracy on the validation set.

To summarise the performance of the classifier:

- Accuracy of classifier on 149 training set = 94.1%
- Accuracy of classifier on 20 unseen validation images = 85.0%
- Accuracy of classifier on 169 image set = 94.1%

#### 6.3.1 Investigation into Misclassified Images

To investigate the performance in greater detail, all 10 images that were misclassified in the 169 image data set were examined to determine whether there were any common factors. Table 6.4 summarises the results of this analysis. It shows that there were two false positive predictions and eight false negative predictions. Eight of the misclassifications occurred at the boundary of grade I and II damage, the most difficult but critical boundary. In four of these cases the experts had expressed uncertainty about the correct classification as positive or negative, and on discussion deciding on a positive grading.

Table 6.4 Data on manual grading of the 10 misclassified images in the 169 image dataset

Image ID	Predicted by Classifier	Manual Grading (Binary)	Manual Grading (Multiclass)	Notes on manual grading
5	Positive	Negative	1	
6	Positive	Negative	1	
40	Negative	Positive	2	Experts initially disagreed on whether it was Grade I or II
90	Negative	Positive	4	
114	Negative	Positive	2	
123	Negative	Positive	2	Experts initially disagreed on whether it was Grade I or II
141	Negative	Positive	2	Experts initially disagreed on whether it was Grade I or II
167	Negative	Positive	2	Experts initially disagreed on whether it was Grade I or II
172	Negative	Positive	2	Presence of necrotic tissue, would usually re-test
175	Negative	Positive	3	Experts initially disagreed on whether it was Grade II or III

As this research concerns an assay to be used to predict potential immunogenicity reactions, it was important to minimise false negatives. A false negative could allow an unsafe compound to progress to clinical trials. In an attempt to reduce the number of false negative results while maintaining the best sensitivity (true positive rate) and specificity (true negative rate) an experiment was performed to investigate the effect of changing the prior probabilities in the classifier. Previously the priors had been automatically set based on the distribution of positive and negatives samples in the data set. When estimated based on the 169 images dataset, these priors would be 0.36 for the negative prior and 0.64 for the positive prior.

### 6.3.2 Investigation into Effect of Prior Probabilities on Classifier Performance

The final training procedure described in section 6.2 was repeated, this time selecting the model with the lowest 10 fold cross validation error over 20 runs and calculating the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) rates. The 20 runs were repeated with a range of prior probabilities for positive and negative classes. As can be seen in Figure 6.7, altering the prior probabilities changes the performance of the final classifier. The red vertical line in the figure shows the performance when the priors are based on the actual distribution of negative and positive images.

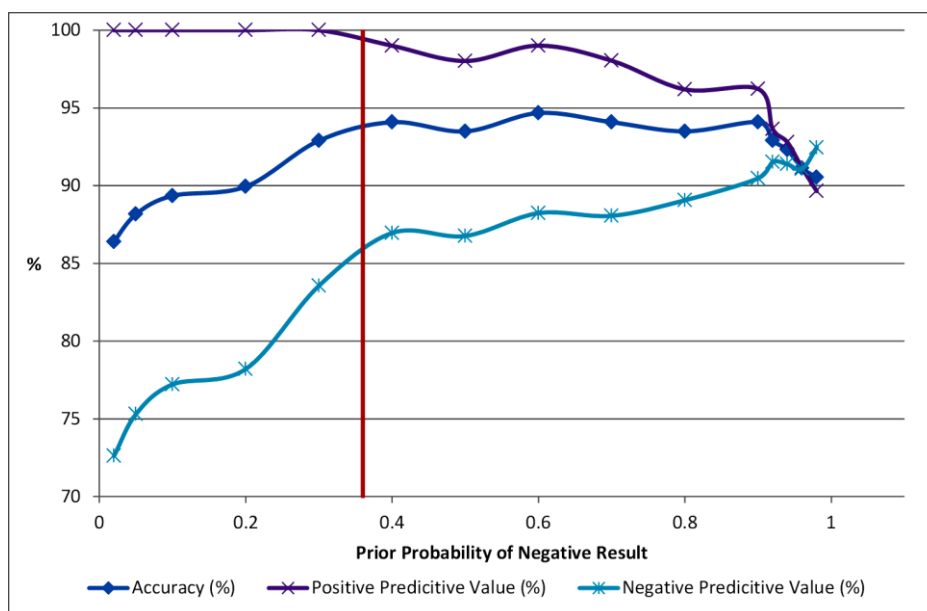


Figure 6.7 The effect of changing the prior probabilities the performance of the final classifier.

The negative predictive value is the percentage of all negative results which are classified correctly and the positive predictive values is the percentage of all positive results which are classified correctly. The prior probabilities offer an opportunity to tune the performance of the classifier and alter the probability of false negatives or positives. To minimise the number of false negatives, the negative predictive value must be maximised. The figure shows that if the prior for a negative result is increased from 0.36 to 0.90 the negative predictive value can be increased from ~85% to 90.5% while the overall accuracy is maintained at 94%. Using the new priors of 0.1 (positive) and 0.9 (negative), the performance of the classifier for the whole 169 image dataset is shown below.

To summarise the final performance of the classifier once the prior probabilities had been optimised for this specific application:

<b>False Positive</b>	4
<b>False Negative</b>	6
<b>True Positive</b>	102
<b>True Negative</b>	57
<b>Accuracy (%)</b>	94.1
<b>Sensitivity (%)</b>	94.4
<b>Specificity (%)</b>	93.4
<b>Positive Predictive Value (%)</b>	96.2
<b>Negative Predictive Value (%)</b>	90.5

This performance is on the whole 169 image dataset, the accuracy on the 149 image training set and 20 image validation set are 94.1% and 85.0% respectively, as quoted previously.

Changing the priors did not alter the overall accuracy; ten of the 169 images were still misclassified. However instead of there being two false positives and eight false negatives, there were four false positive and six false negatives. This is an improvement for the application this classifier is being developed for.

### **6.3.3 Final Industrial Application of Image Processing, Feature Selection and Classification Algorithm**

A Matlab script was written to upload a new image, perform image processing and segmentation of epidermis, dermis, clefts and vacuoles, extract 17 numerical image features, train a model using a 149 image training set and classify the new image. A Graphical User Interface was also developed to facilitate the user interaction step during the segmentation stage. The speed of the final algorithm is dependent on the computer on which it is run, but when run on an Intel quad-core 3.4GHz processor with 8GB RAM an image could be classified in ~40 seconds. The training

of the model was included as it would be preferable to continue updating the model as the training set of images of known grade is built up.

#### **6.4 Discussion of Feature Extraction, Selection and Classification**

The traditional skin explant classification relies on differences in the severity of vacuolisation, the presence or absence of dyskeratotic bodies, and the extent of cleft formation at the DEJ. Human operators have an ability to interpret qualitative descriptions like this, but they also tend to find additional criteria that support their decisions through years of experience and learning. Obtaining a full explanation of all the information that an experienced operator uses to make decisions is challenging, and consequently as many potential measures as possible were generated in the hope that a subset of these features would incorporate the information that a human expert uses.

The set of 140 features generated included features likely to be highly correlated. For example, measurement sets extracted from the two vacuole sets were likely to be correlated due to the sets containing many of the same objects. No attempt was made to select between these different measures during design of the feature set as this would have introduced subjectivity and one of the aims of the research was to reduce subjectivity in the grading process. Instead, the feature set was reduced using an objective, mathematical approach to remove uninformative, correlated and redundant features and identify those features which, in combination, provide the greatest level of information to differentiate between grades of damage. While it would have been possible to analyse each feature individually, this approach ignores the interaction and interdependence of the features. Additionally, the number and similarity of the features in the full feature set would make this approach highly subjective and time consuming. Consequently, automated feature selection was chosen to utilise the strengths of the computer (objectivity, quantitation and processing power) over the human operator and allow multiple features to be examined simultaneously.

The final set of features included 8 features relating to the amount of vacuolisation in the image, 5 features relating to the clefts, and 4 related to texture of the

epidermis tissue. The objective feature selection method resulted in a spread of different feature types being retained in the final feature type.

The vacuole descriptive features focus on the distribution of vacuole area, extent and eccentricity. The inclusion of those relating to area was unsurprising given that the main criteria in the manual grading between a positive and negative result is the amount of vacuolisation. The extent is an interesting addition to this. Extent is a measurement of the region area divided by the area of the bounding box, it therefore changes as the vacuoles become more circular and compact rather than the eclipse or crescent shapes that they tend to be when the vacuolisation is very mild. This shape property is not present in the original grading criteria; however it is something that is visible when examining the images. The eccentricity measurements included in the final feature set are probably capturing the same change as the extent, as they measure circularity.

The percentage area and number of clefts and one of the proposed surrogate measurements for cleft coverage of the DEJ were also present in the final feature set. The combined length and length distribution of the clefts, their total area and their shape is captured in these features.

The final 4 features are texture features which are calculated only on image regions showing the epidermis tissue. The luminance contrast measurement may capture information about clefts and vacuoles in the epidermis, while the three different measurements of texture in the red-green colour channel may reflect changes in the tissue uniformity as the tissue begins to break down. The selection of the texture features confirms the validity of including both morphometric and texture based features. While the texture of the epidermis very obviously changes as it breaks down, it is difficult to capture the visual appearance in a written set of criteria. The inclusion of texture based features was one way of capturing this knowledge. The use of graph based features would be an interesting extension to this work, as they offer an alternative way of capturing this structural information. For example, mapping the spatial distribution or connectivity of the cleft or vacuole centre points could capture information in a useful form.

The final evaluation of the image processing, feature extraction and selection was performed by testing the set of features in a well-known classifier structure, namely the Naïve Bayes classifier. With minimal model optimisation, the error rate of a Naïve Bayes classifier using the defined feature set selected was 0.06 when tested on the whole data set. A higher error rate of 0.15 was seen when classifying a 20 image validation set not used in any of the training or feature selection procedures. This could be a result of the small size of the validation set or it could indicate that the design of the image analysis and feature selection was biased by the feature set, resulting in poor generalisation ability. It is the author's opinion that a much larger training set of images is required to develop an accurate and robust classifier for such a complex classification task, however this minimal dataset has enabled a classifier of high accuracy to be created. The performance of the final classifier is good when the challenges of the classification task are considered. The manual grading analysis (section 4.7) showed that there was disagreement in the grading of samples as either positive or negative in 12.8% of cases. A large study inter-observer variability for the original skin explant assay (Sviland *et al.*, 2001), with 503 slides graded also showed disagreement in 8% of cases mainly attributed to difficulties grading at the grade I/ II borderline.

There is no published data of which the author is aware of automated classification of graft versus host reactions to which the developed classification algorithm can be compared. However the performance of other histopathological classification algorithms, reviewed in chapter 3, section 3.2.8, can be used as a benchmark. Classification accuracies for binary classifications of brain, prostate, breast and colon tissue varied from 87.8% to 99.7%. It is difficult to compare performance as an independent test has not always been used to validate final performance. Often the cross-validation error on the whole set is quoted and by this comparison, the performance of 94% is fairly typical to that seen in the literature. The excellent classification accuracies (99.7% on the test set) reported by Rajpoot and Rajpoot (2004) reflect the exhaustive optimisation of the SVM kernel functions they have undertaken which improved the accuracy of their classification method from 87% to 99.7%. They were working with a training set of 11,000 samples and a test set of 34,056 samples. This suggests that with a larger dataset and further optimisation

of the classification method, the tissue segmentation and feature extraction developed in this research could deliver significantly higher classification accuracies, particularly on an independent test set.



## Chapter 7 Final Discussion and Conclusions

The overall aim of this industrial research project was to develop an automated system to enable non-expert users to grade histological skin damage using Alcyomics' Skimune assay with a comparable level of accuracy, repeatability and reproducibility to that achieved through expert manual grading. This final discussion explains the research approach taken and assesses how well the developed solution has fulfilled the original industrial research objectives and summarises the contributions of the work. Finally, future work to continue the research or improve the industrial solution is presented.

### 7.1 Discussion of Research Approach

The ultimate aim of the research was to develop an automated classifier of skin images; however the majority of effort and research was focussed on identifying the specific features that could be used in the classifier and minimising the impact of non-relevant variation in the images. During the initial assessment of the research problem it became clear that the system and method used to capture the images was important. The system and method selected ensured the whole sample could be used in the subsequent analysis and included various approaches to minimise image variation due to microscope focussing and illumination. While access to a commercial slide scanner would improve the robustness of this step further, this was not available during this research project.

An early decision was made to focus the research on the segmentation of the epidermis. All the significant histological changes being assessed during the analysis were either in the epidermis tissue or directly adjacent to it and segmentation ensured that all other variation in the image background and the dermis tissue was excluded from the analysis. The particular image set used in the research was representative of a "real world" data set, with significant variation in the proportion of epidermis and dermis tissue and in the overall sample shape, structure and staining. These challenges needed to be overcome and this shaped the research approach taken. Multiple procedures to "normalise" the images were used, including background cropping, colour normalisation and contrast enhancement. An alternative colourspace ( $L^*a^*b^*$ ) proved to offer a more

consistent representation of colour than the standard RGB colour space for the particular images and features being analysed in this research.

The applicability of Mathematical Morphology (MM) and shape analysis to image analysis in histology was also demonstrated. Alongside colour and location, shape is one way in which humans recognise the central features being analysed in histology and MM and quantitative shape analysis provided an objective way of using this information. MM also proved a useful technique to remove non-relevant areas or features of the image.

The quantitative information extracted during shape analysis formed the basis of the morphological features extracted from and used to represent the images. Morphological features were originally chosen because of their similarity to standard histological feature descriptions, however the strengths of computer based systems were utilised to extract many more complex quantitative features and feature population descriptors than a human could analyse. The choice to include more abstract texture features, such as entropy or correlation of the grey level co-occurrence matrix, was an attempt to capture global changes seen across the tissue rather than at an individual feature level. Recognising the pattern of structural breakdown is something that a human can do, but it is a challenge for them to analyse this consistently from image to image.

In the final parts of the research, there was an even stronger emphasis on computational methods. The feature selection and classification approaches were designed to select features and create models objectively based on their ability to classify the test images correctly. Standard methods employed in the literature were used at this stage. While it may be possible to improve the performance of the classifier by further optimising the feature selection and classification, it is the author's opinion that extracting the subset of features identified as most relevant to the classification task in section 6.1 from a much larger and more representative feature set would be a priority for future work.

## 7.2 Assessment of Performance against Industrial Research Objectives

The next section assesses how successfully the developed solution answers the original industrial research objectives.

### 7.2.1 Automation

In the current method used to grade samples on slides, the whole process is manual and there is no automation. The developed solution can be run in a fully automated manner once the images have been digitised and saved on a computer. A standard operating procedure has also been developed and provided to Alcyomics for image acquisition using a Zeiss AxioImager and is attached in Appendix A. This standard procedure uses automated procedures for background correction, focussing, white balance correction and image capture and stitching. It should therefore minimise variation from the image acquisition process. The main image processing, feature extraction and classification process can be fully automated, however introducing user interaction during the segmentation of epidermal and dermal tissue was found to improve the algorithm by reducing the level of variability. More specifically, the standard error of the mean sensitivity for the test set reduced from 14.2 to 8.4 when the user interaction step is added. While the introduction of this step means the process is not fully automated, the end user has a choice to include it. The risk of introducing subjectivity is fairly low as long as the user has a basic knowledge of skin tissue structure and appearance.

### 7.2.2 Non-expert user

The algorithm has been developed in Matlab and a basic Graphical User Interface created to show how a non-expert user could upload images and then run the algorithm with a single button click. This is in contrast to the current manual grading method which requires the operator to have had significant training in histopathology. If user interaction is involved, the user will need basic training to be able to differentiate between dermis, epidermis and *stratum corneum* tissue but this is a fairly simple task which does not require detailed knowledge of histopathology. This solution is suitable for internal use by the company, but further software development would be required to create a robust package which could be distributed or used by customers.

### 7.2.3 Accuracy

The accuracy of the current manual grading process is difficult to estimate, however reported disagreements in grading vary between 8% and 13% in the skin explant assay. A large study inter-observer variability for the original skin explant assay (Sviland *et al.*, 2001), with 503 slides graded also showed disagreement in 8% of cases mainly attributed to difficulties grading at the grade I/ II borderline. The manual grading analysis (reported in section 4.3) showed that there was disagreement in the grading of samples as either positive or negative in 13% of cases.

Converting the disagreement into a measurement of accuracy of manual grading is not straightforward. One could assume that when two operators disagree there is an equal chance of each of them being wrong, if this is the case then the manual classification (or grading) error rate of a single operator can be estimated as being 6.5%. However, this assumption ignores the variation between operators. It is equally possible that one experienced operator makes no errors and another is making errors 13% of the time. It is a better assessment to state that manual grading of GVHRs has been demonstrated to have variable accuracy of between 87% and 100%. The accuracy of the manual process is dependent on a number of factors including the operator experience and the number of samples at the grade I/II borderline. As all operators used in the manual grading studies described were experienced, we can assume that an inexperienced operator could have a grading accuracy of less than 87%.

The automated classification algorithm developed in this research has an accuracy of 94% on an image set of 169 images. Using the more stringent criteria of classification accuracy for an unseen 20 image validation set, the classification accuracy was 85%. This performance compares favourably with the manual process. Access to a larger dataset in the future would be likely to improve the accuracy and generalisation ability of the classifier when using the defined feature set developed in this research project.

Taking into account the known issues of inter-observer variability and the specific challenges of the qualitative Lerner grading scale, it would be preferable to use

alternative data as ground truth to train the classification model. The ideal situation for this research project would be to use a set of samples which had been exposed in the assay to a variety of compounds of known toxicities at a variety of doses. By knowing what the assay should be predicting, feature selection and training of the classification model would not be biased or affected by human grading error and variability. As Alcyomics is a relatively young company, this data was not available and so samples generated when the assay was used to predict GVHD in patient-donor pairs were used instead. Data on the clinical outcome was not available and so the expert manual grading was used to label the data. As the company generates more data on typical skin reactions exposed to a variety of compounds, this could be used to improve the developed solution and even develop a new set of grading criteria based on quantitative measures extracted from the images.

#### **7.2.4 Repeatability and Reproducibility**

The developed grading system, when run without user interaction, will produce the same grade when run repeatedly on the same image. Using a different computer will not alter the result, although the time taken to classify an image will be dependent on the processing power of the computer being used. As such the repeatability and reproducibility are improved when compared to the current manual grading system where an individual operator may grade a borderline case differently on different occasions, and different operators are known to have biases, evidenced by the manual grading study and in the multi-centre study by Sviland et al (2001). When the user interaction step is included a small element of bias is introduced, however the task for the user is simple and so the risk of introducing significant issues of repeatability and reproducibility is low.

#### **7.2.5 Robustness**

The system is able to grade images with differing morphology, staining intensity and background lighting. The image training set was purposely created to include “difficult” images rather than be an idealised image set. The classification system makes the most errors at the boundary between grade I and II. This is also known to be a difficult judgement to make for a human operator, which is unfortunate

considering it is the main discriminator used to identify a positive or negative reaction.

### 7.3 Discussion of Academic Research Contributions

The main research contributions described in this thesis are:

- The development of a new methodology for epidermal segmentation able to identify epidermis tissue from H&E stained skin sections showing varying degrees of histopathological damage. Although many methods have been described for segmentation of histology images, most are for cell, gland or nuclear segmentation rather than tissue segmentation. The epidermis segmentation algorithm is a useful addition to this small but growing area of research and has already been reproduced by another research paper as a benchmark technique (Xu and Mandal, 2015). It provides a useful framework for segmentation of other epithelial tissues and (noting the requirement for appropriate parameter tuning and optimisation) it is a useful contribution in the areas of dermatology, tissue segmentation, and *in vitro* assay technology. The robustness is shown by the method's high accuracy in segmentation of a challenging dataset of epidermis tissue from H&E images of human skin showing varying degrees of histological damage. The author is unaware of any segmentation methods that have been applied to images showing severe histological damage such as graft versus host type reactions. This part of the work has been published in the peer reviewed open access academic journal, BMC Medical Imaging, where it has been classified as highly accessed. The paper is available at <http://www.biomedcentral.com/content/pdf/1471-2342-14-7.pdf>.
- A novel set of object and spatial level quantitative features have been defined and a method for their extraction created. The extracted feature measurements are relevant to the expert grading criteria for histological damage but add a quantitative dimension. While this has direct application to the grading of the Skimune assay, this set of feature measurements could also be applied in an automated version of the Lerner grading used in the diagnosis and prediction of graft versus host disease.

- An approach to histopathological tissue classification, which combines expert domain knowledge in the design of potential features, with a fully objective feature selection and classification approach. In this research, the influence of domain knowledge and the bias that this may bring is not disputed, one of the main hypotheses of this research was that incorporating such knowledge into the early stages of the image processing and feature extraction would enable variation relevant to skin damage to be distinguished from non-relevant image variation.
- A new image analysis and classification method for the automated classification of H&E images of human skin showing positive or negative graft versus host reaction. The author is not aware of any other automated image analysis and classification method for this application.

#### 7.4 Future Work

The most important future work would be to use the developed image processing and feature extraction approaches on a more representative image set consisting of samples that have been exposed in the assay to a variety of compounds of known toxicities at a variety of doses. By knowing what the assay should be predicting, feature selection and training of the classification model is not biased or affected by human grading error and variability.

Dyskeratotic bodies are a differentiating factor between the difficult grading boundary between a grade I and grade II result. They are difficult to identify and not always present and were disregarded from this research on the advice of experienced histopathologists. It would be an interesting extension of the research to investigate potential methods to extract these key features, potentially using a colour and texture based feature extraction methodology to identify these structures with bright pink cytoplasm and condensed nuclei.

At the start of the research project, the focus was on choosing biologically relevant features, using the hypothesis that this would be the best way to capture damage-related variation rather than that resulting from the staining, sample preparation, lighting and biological processes. As the project has progressed, it became clear

that the written criteria capture only a fraction of the knowledge used to grade the samples. While it was vital to ensure that the initial epidermal and dermal segmentation was correct to limit assessment and feature extraction to the relevant areas of tissue, it would seem sensible in any future work to include a wide variety of other mathematical features including other texture features, graph-based features and wavelet based features to attempt to capture more of the tacit knowledge used during manual grading. In a time limited project this was not possible; however this would offer a potential route to improvement.

Due to the large and dense datasets typically generated in histopathology image classification tasks, the use of ensemble classification methods is becoming prominent in the field. Ensembles of classifiers have been reported to reduce the bias or variance associated with single classifiers and improve classification accuracy (Kuncheva and Whitaker, 2003). It has not been possible to include the development of such an ensemble method within the scope of this research project; however, it would be valuable to assess potential improvements in classification accuracy using such methods in the future. There is potential to develop and optimise the classification further using ensemble methods or extensions to the Naïve Bayes such as the hierarchical approach proposed by Demichelis *et al* (2006) or the non-parametric version used by Soira *et al* (2011).



## Appendix A

### Zeiss AxioImager Standard Operating Procedure


#### *Set Up*

1. Turn on if not already set up: In the order, PC, Monitor, Power Supply, Microscope. Note only brightfield is needed, not fluorescence.
2. Log on: user name, no password required – click OK.
3. In Windows Explorer make a new folder for your images:  
E drive\UserData\month\YourName
4. Open Axiovision Rel. 4.8 from desktop.
5. Click Brightfield on top toolbar
6. On right menu, click camera and select colour.
7. Make sure objective lens is set to x10 in left menu.
8. Make sure cap is off light source at bottom of microscope.

#### *Microscope Set Up (if not already done, or changing from a different objective lens)*

9. Illumination Iris: Viewing through eyepiece and using button next to F (on right). Make iris smaller, then ensure it is sharp and centred. Then enlarge by opening iris until it clears field of view.
10. Stage Iris: set until you can just see edges using button on front of microscope.
11. Swing out condenser should be in (up position) for x10 objective lens.

#### *Load Slide*

12. On microscope display click **load** (top right corner of microscope display) to bring stage down to loading position, add slide then adjust stage to approximately correct position.
13. Click  to bring stage back up once slide is on.
14. Find sample either using eyepiece or viewing on monitor.
  - a. To view image through eyepiece: click brightfield, eyes. (NB if you can't see anything, try clicking 'make it visible' on microscope display).
  - b. To get image on monitor: In standard workflow, select camera (colour) and click live.
  - c. Click 'make it visible' on microscope display, which brings settings back to default.

#### *Optimise Image*

15. Get approximate focus manually, then click autofocus.
16. To set colour, go into colour set up in left menu, and under white balance click interactive, then click on a region of white background.

17. If image looks too light/ dark try clicking exposure (bottom toolbar).
18. To correct shading in background (usually only need to do this on 1<sup>st</sup> image of a session). Move field of view to a clear section of background. Go to properties (bottom toolbar), general, shading correction, make sure shading correction is ticked. Untick and then click shading correction to re-tick, this ensures it updates.

### *Image Tiling – Mosaic*

19. Go to Acquisition menu, then click on Mosaic acquisition, set to autofocus every 3 tiles.
20. Click set up: Reset imaging field by clicking box with red cross on. Go to edges of sample and click 4 arrow icon to set bounding points using centre of crosshair, once the whole sample is enclosed click OK (bottom right corner), then start button in in Mosaic acquisition menu.  
NB: Include all of main sample except very small tissue debris. If 4-arrow is greyed out, that area is already included in field of view.
21. Once image acquisition is finished, close live view to look at tiled image. Check if the image tiles look OK (cover whole sample, no shading issues)
  - a. If shading isn't right, you may need to re-do shading correction step.
  - b. If colour balance looks wrong (very bright or very dark), try clicking histogram icon in 2D view toolbar at bottom of screen (gamma correction and max range) – the aim is good contrast between epidermis and dermis.
22. To and do this go to tileview and click stitch (green square icon).

### *To Export file*

23. Go to File, export, navigate to your new folder in E Drive. Change filename if required (e.g. Date\_01, Date\_02)
  - a. Make sure it is set to only save merged files.
  - b. Set to save as a TIFF file.
  - c. Make sure convert to 8bit is clicked and compression is set to 0%.
24. Click start to save.

Repeat from step 10 (no need to re-do shading correction and export setting should now stay as you have set them).

Once all images are complete, save all files to USB drive. At end of session, close Axiovision software and log off computer.

## Appendix B

*Table A. Factor values and responses (mean sensitivity and specificity) for fractional factorial screening test*

<b>Run</b>	<b>Lower P point, G' (A)</b>	<b>Upper P point, G' (B)</b>	<b>Lower P point, b' (C)</b>	<b>Upper P point, b' (D)</b>	<b>Mean Filter Kernel Size (E)</b>	<b>Radius of SE (F)</b>	<b>Mean Sensitivity (%)</b>	<b>Mean Specificity (%)</b>
<b>1</b>	0.2	0.875	0.225	0.95	30	12.5	61.51	96.43
<b>* 2</b>	0.3	1	0.3	0.9	40	5	76.88	97.09
<b>3</b>	0.1	1	0.15	0.9	40	20	58.49	97.56
<b>4</b>	0.3	0.75	0.15	0.9	40	5	61.45	93.77
<b>5</b>	0.1	0.75	0.15	1	20	20	20.56	98.29
<b>6</b>	0.3	1	0.15	0.9	20	20	62.59	97.57
<b>7</b>	0.3	0.75	0.15	1	40	20	54.42	94.91
<b>8</b>	0.1	0.75	0.15	0.9	20	5	43.82	95.20
<b>9</b>	0.3	0.75	0.3	1	20	5	62.84	93.68
<b>10</b>	0.1	0.75	0.3	0.9	40	20	35.70	96.59
<b>* 11</b>	0.1	1	0.15	1	40	5	72.43	97.03
<b>12</b>	0.3	0.75	0.3	0.9	20	20	41.01	94.48
<b>13</b>	0.1	0.75	0.3	1	40	5	46.89	94.76
<b>14</b>	0.1	1	0.3	0.9	20	5	70.36	96.90
<b>* 15</b>	0.3	1	0.15	1	20	5	74.25	97.77
<b>* 16</b>	0.3	1	0.3	1	40	20	73.85	97.59
<b>17</b>	0.1	1	0.3	1	20	20	52.65	96.07

*P point = Penetration point*

Table B Description of features extracted and the sets of objects used to calculate the features.

<b>Feature Description</b>	<b>Object sets or colour channels from which the measurement is taken</b>
Percentage area of epidermis covered by object	Vacuole sets a and b, cleft sets a and b
Mean area of objects in the image	Vacuole sets a and b, cleft sets a and b
Max area of objects in the image	Vacuole sets a and b, cleft sets a and b
Median area of objects in the image	Vacuole sets a and b, cleft sets a and b
Standard deviation of object areas in image	Vacuole sets a and b, cleft sets a and b
Inter-quartile range of object areas in image	Vacuole sets a and b, cleft sets a and b
Skewness of object areas in image	Vacuole sets a and b, cleft sets a and b
Kurtosis of object areas in image	Vacuole sets a and b, cleft sets a and b
Number of objects in image	Vacuole sets a and b, cleft sets a and b
Mean eccentricity of objects areas in image	Vacuole sets a and b, cleft sets a and b
Max eccentricity of objects areas in image	Vacuole sets a and b, cleft sets a and b
Median eccentricity of objects areas in image	Vacuole sets a and b, cleft sets a and b
Standard deviation of object eccentricities in image	Vacuole sets a and b, cleft sets a and b
Inter-quartile range of object eccentricities in image	Vacuole sets a and b, cleft sets a and b
Skewness of object eccentricities in image	Vacuole sets a and b, cleft sets a and b
Kurtosis of object eccentricities in image	Vacuole sets a and b, cleft sets a and b
Mean extent of objects in the image	Vacuole sets a and b, cleft sets a and b
Max extent of objects in the image	Vacuole sets a and b, cleft sets a and b
Median extent of objects in the image	Vacuole sets a and b, cleft sets a and b
Standard deviation of object extent in image	Vacuole sets a and b, cleft sets a and b
Inter-quartile range of object extent in image	Vacuole sets a and b, cleft sets a and b
Skewness of object extent in image	Vacuole sets a and b, cleft sets a and b
Kurtosis of object extent in image	Vacuole sets a and b, cleft sets a and b
Sum of all cleft perimeters, divided by epidermis perimeter (all epidermis objects)	Cleft sets a and b
Sum of all cleft perimeters, divided by sum of all epidermis object major axis lengths	Cleft sets a and b
Mean major axis length of clefts in image	Cleft sets a and b
Max major axis length of clefts in image	Cleft sets a and b

Median major axis length of clefts in image	Cleft sets a and b
Standard deviation of cleft major axis lengths in image	Cleft sets a and b
Inter-quartile range of cleft major axis lengths in image	Cleft sets a and b
Skewness of cleft major axis lengths in image	Cleft sets a and b
Kurtosis of cleft major axis lengths in image	Cleft sets a and b
Sum of all cleft major axis lengths, divided by epidermis perimeter ( all epidermis objects)	Cleft sets a and b
Sum of all cleft major axis lengths, divided by sum of all epidermis object major axis lengths	Cleft sets a and b
Median cleft dimension in image (dimension = major/minor axis length)	Cleft sets a and b
Contrast (epidermis pixels only)	Each colour channel of L*a*b* and RGB images
Correlation (epidermis pixels only)	Each colour channel of L*a*b* and RGB images
Energy (epidermis pixels only)	Each colour channel of L*a*b* and RGB images
Homogeneity (epidermis pixels only)	Each colour channel of L*a*b* and RGB images

## Appendix C

### Hard Coded Parameters Dependent on Image Spatial Resolution

These parameters are dependent on image spatial resolution. These parameters should be scaled appropriately if using an image with a different resolution.

- Mean filter used in preparation of the sampleMask: 29 x 29
- Threshold for removal of small objects in sampleMask: 25,000
- Mean filter used in preparation of the epiMask: 40 x 40
- Structuring element used for morphological operations of the epiMask: 6
- Threshold for removal of small objects in epiMask: 4000
- Threshold for filling of small holes in epiMask: 7000
- Object area,  $Z_{Area}$ , used during final object classification in epiMask: 20,000
- The size of the structuring element used to smooth the dermMask prior to creation of the perimeter mask, pMask: 20
- The size of the structuring element used to thicken the perimeter mask used to exclude the *stratum corneum* in the preparation of the dermMask: 60
- The size of the structuring element used to thicken the perimeter mask used to exclude the *stratum corneum* during the vacuole identification: 45
- Object area,  $Z_{Area}$ , used during final object classification in dermMask: 42,500
- The pixel distance used to identify potential cleft objects as being at the dermal epidermal junction: 5 pixels
- Object area,  $Z_{Area}$ , used during final object classification of faults into vac\_a: 1000
- Object area,  $Z_{Area}$ , used during final object classification of faults into sets vac\_b and cleft\_b : 150
- Pixel pair spacing used to calculate texture features in the GLCM: 5 pixels

### Hard Coded Parameters Dependent on Image Staining, Lighting and Imaging

The choice of optimal colourspace (grayscale and b\*), the linear combination parameters (0.5 + 0.5) and the appropriate upper and lower thresholds used for contrast enhancement during epidermal segmentation would need to be re-

optimised if using images with very different staining properties or colour profiles to the training images used in this research.

The luminance threshold used during identification of clefts and vacuoles (mode – 20 and mode – 100 respectively) would need to be re-optimised if the image lighting conditions or the tissue thickness was very different from the training set used in this research.

## References

- Adiga, U., Malladi, R., Fernandez-Gonzalez, R. and de Solorzano, C. O. (2006) 'High-throughput analysis of multispectral images of breast cancer tissue', *IEEE Transactions on Image Processing*, 15(8), pp. 2259-2268.
- Al-Kadi, O. S. (2010) 'Texture measures combination for improved meningioma classification of histopathological images', *Pattern Recognition*, 43(6), pp. 2043-2053.
- Angenent, S., Pichon, E. and Tannenbaum, A. (2006) 'Mathematical methods in medical image processing', *Bulletin of the American Mathematical Society*, 43(3), pp. 365-396.
- Antani, S., Kasturi, R. and Jain, R. (2002) 'A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video', *Pattern Recognition*, 35(4), pp. 945-965.
- Attarwala, H. (2010) 'TGN1412: From Discovery to Disaster', *Journal of Young Pharmacists*, 2(3), pp. 332-336.
- Aubert-Broche, B., Griffin, M., Pike, G. B., Evans, A. C. and Collins, D. L. (2006) 'Twenty new digital brain phantoms for creation of validation image data bases', *IEEE Transactions on Medical Imaging*, 25(11), pp. 1410-1416.
- Avanaki, M. R. N., Hojjat, A. and Podoleanu, A. G. (2009) 'Investigation of computer-based skin cancer detection using optical coherence tomography', *Journal of Modern Optics*, 56(13), pp. 1536-1544.
- Basavanhally, A., Agner, S., Alexe, G., Bhanot, G., Ganesan, S. and Madabhushi, A. (2008) 'Manifold learning with graph-based features for identifying extent of lymphocytic infiltration from high grade, her2+ breast cancer histology', *MMBIA workshop in conjunction with MICCAI 2008*. Available at: <http://engineering.case.edu/centers/ccipd/> (Accessed: 20 October 2015).
- Bellman, R. E. (1961) *Adaptive Control Processes*. Princeton, NJ: Princeton University Press.
- Binder, M., Kittler, H., Seeber, A., Steiner, A., Pehamberger, H. and Wolff, K. (1998) 'Epiluminescence microscopy-based classification of pigmented skin lesions using computerized image analysis and an artificial neural network', *Melanoma Research*, 8(3), pp. 261-266.
- Bins, J. and Draper, B. A. (2001) 'Feature selection from huge feature sets', *8th IEEE International Conference on Computer Vision*. Vancouver, BC. IEEE, pp. 159-165.
- Bishop, C. (2010) 'Embracing Uncertainty: The New Machine Intelligence ' *IET/BCS Turing Lecture*. 18 March 2010. [Online] Available at: <http://tv.theiet.org/technology/infopro/turing-2010.cfm> (Accessed: 10 October 2015).



- Bouman, C. and Liu, B. (1991) 'Multiple resolution segmentation of textured images', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(2), pp. 99-113.
- Bruner, J. M., Inouye, L., Fuller, G. N. and Langford, L. A. (1997) 'Diagnostic discrepancies and their clinical impact in a neuropathology referral practice', *Cancer*, 79(4), pp. 796-803.
- Busam, K. J., Charles, C., Lohmann, C. M., Marghoob, A., Goldgeier, M. and Halpern, A. C. (2002) 'Detection of intraepidermal malignant melanoma *in vivo* by confocal scanning laser microscopy', *Melanoma Research*, 12(4), pp. 349-355.
- Camp, R. L., Chung, G. G. and Rimm, D. L. (2002) 'Automated subcellular localization and quantification of protein expression in tissue microarrays', *Nature Medicine*, 8(11), pp. 1323-1328.
- Can, A., Bello, M., Cline, H. E., Xiaodong, T., Ginty, F., Sood, A., Gerdes, M. and Montalto, M. (2008) 'Multi-modal imaging of histological tissue sections', *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. Paris, 14-17 May 2008. IEEE, pp. 288-291.
- Cestnik, B. (1990) *European Conference on Artificial Intelligence. ECAI 90*. Stockholm. Available at: [http://www.temida.si/~bojan/resources/Cestnik\\_Estimating\\_probabilities.pdf](http://www.temida.si/~bojan/resources/Cestnik_Estimating_probabilities.pdf) (Accessed: 10 October 2015).
- Chen, C., Ozolek, J. A., Wang, W. and Rohde, G. K. (2011) 'A general system for automatic biomedical image segmentation using intensity neighborhoods', *Journal of Biomedical Imaging*, 2011(ID 606857) [Online]. Available at: <http://www.hindawi.com/journals/ijbi/2011/606857/> DOI: 10.1155/2011/606857
- Clark, P. and Niblett, T. (1989) 'The CN2 induction algorithm', *Machine Learning*, 3(4), pp. 261-283.
- Cohen, J. (1960) 'A Coefficient of Agreement for Nominal Scales', *Educational and Psychological Measurement*, 20(1), pp. 37-46.
- Cox, N. H. and Coulson, I. H. (2010) 'Diagnosis of Skin Disease', in Burns, T., Breathnach, S., Cox, N. and Griffiths, C. (eds.) *Rook's Textbook of Dermatopathology*. Wiley-Blackwell.
- Dalton, L. W., Pinder, S. E., Elston, C. E., Ellis, I. O., Page, D. L., Dupont, W. D. and Blamey, R. W. (2000) 'Histologic Grading of Breast Cancer: Linkage of Patient Outcome with Level of Pathologist Agreement', *Modern Pathology*, 13(7), pp. 730-735.
- Datar, M., Padfield, D. and Cline, H. (2008) 'Color and texture based segmentation of molecular pathology images using HSOMS', *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. Paris, 14-17 May 2008. IEEE, pp. 292-295.

de Vet, H. C. W., Knipschild, P. G., Schouten, H. J. A., Koudstaal, J., Kwee, W.-S., Willebrand, D., Sturmans, F. and Arends, J. W. (1990) 'Interobserver variation in histopathological grading of cervical dysplasia', *Journal of Clinical Epidemiology*, 43(12), pp. 1395-1398.

Demichelis, F., Magni, P., Piergiorgi, P., Rubin, M. A. and Bellazzi, R. (2006) 'A hierarchical Naive Bayes Model for handling sample heterogeneity in classification problems: an application to tissue microarrays', *BioMed Central Bioinformatics*, 7(1), p. 514.

Demir, C., Gultekin, S. H. and Yener, B. (2005) 'Augmented cell-graphs for automated cancer diagnosis', *Bioinformatics*, 21(suppl 2), pp. ii7-ii12.

Di Cataldo, S., Ficarra, E., Acquaviva, A. and Macii, E. (2009) 'Achieving the way for automated segmentation of nuclei in cancer tissue images through morphology-based approach: A quantitative evaluation', *Computerized Medical Imaging and Graphics*, 34(6), pp. 453-461.

Diamond, J., Anderson, N. H., Bartels, P. H., Montironi, R. and Hamilton, P. W. (2004) 'The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia', *Human Pathology*, 35(9), pp. 1121-1131.

Dickinson, A. M., Sviland, L., Jackson, G., Dunn, J., Stephens, S. and Proctor, S. J. (1994) 'Monoclonal anti-TNF-alpha suppresses graft vs host disease reactions in an in vitro human skin model', *Cytokine*, 6(2), pp. 141-146.

DiMasi, J. A., Feldman, L., Seckler, A. and Wilson, A. (2010) 'Trends in Risks Associated With New Drug Development: Success Rates for Investigational Drugs', *Clinical Pharmacology & Therapeutics*, 87(3), pp. 272-277.

Dobson, L., Conway, C., Hanley, A., Johnson, A., Costello, S., O'Grady, A., Connolly, Y., Magee, H., O'Shea, D., Jeffers, M. and Kay, E. (2010) 'Image analysis as an adjunct to manual HER-2 immunohistochemical review: a diagnostic tool to standardize interpretation', *Histopathology*, 57(1), pp. 27-38.

Doyle, S., Feldman, M., Tomaszewski, J. and Madabhushi, A. (2012) 'A Boosted Bayesian Multiresolution Classifier for Prostate Cancer Detection From Digitized Needle Biopsies', *IEEE Transactions on Biomedical Engineering*, 59(5), pp. 1205-1218.

Doyle, S., Hwang, M., Shah, K., Madabhushi, A., Feldman, M. and Tomaszewski, J. (2007) 'Automated grading of prostate cancer using architectural and textural image features', *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. Arlington, VA, 12-15 April. pp. 1284-1287.

Doyle, S., Madabhushi, A., Feldman, M. and Tomaszewski, J. (2006) 'A Boosting Cascade for Automated Detection of Prostate Cancer from Digitized Histology', in Larsen, R., Nielsen, M. and Sporring, J. (eds.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*. Springer Berlin Heidelberg, pp. 504-511.

- Eaden, J., Abrams, K., McKay, H., Denley, H. and Mayberry, J. (2001) 'Inter-observer variation between general and specialist gastrointestinal pathologists when grading dysplasia in ulcerative colitis', *The Journal of Pathology*, 194(2), pp. 152-157.
- Elston, C. W. and Ellis, I. O. (1991) 'Pathological prognostic factors in breast cancer. The value of histological grade in breast cancer: experience from a large study with long-term follow-up', *Histopathology*, 19, pp. 403 - 410.
- Elston, C. W. and Ellis, I. O. (2002) 'Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Comment on: C. W. Elston & I. O. Ellis. *Histopathology* 1991; 19; 403-410', *Histopathology*, 41(3A), pp. 151-2, discussion 152-3.
- Eramian, M., Daley, M., Neilson, D. and Daley, T. (2011) 'Segmentation of epithelium in H&E stained odontogenic cysts', *Journal of Microscopy*, 244(3), pp. 273-292.
- Farmer, E. R., Gonin, R. and Hanna, M. P. (1996) 'Discordance in the histopathologic diagnosis of melanoma and melanocytic nevi between expert pathologists', *Human Pathology*, 27(6), pp. 528-531.
- Filipczyk, P., Kowal, M. and Obuchowicz, A. (2011) 'Automatic breast cancer diagnosis based on K-means clustering and adaptive thresholding hybrid segmentation', in Choraś, R. (ed.) *Image Processing and Communications Challenges 3*. Berlin: Springer Berlin Heidelberg, pp. 295-302.
- Fleming, M. G., Khona, D. and Kohler, S. (1998) 'Image cytometry in early graft-versus-host disease', *American Journal of Dermatopathology*, 20(5), pp. 459-462.
- Fraga, G. (2012) 'The Skin', in Haider, S. A. (ed.) *Atlas of Histopathology*. India: Jaypee Brothers Medical Publishers, pp. 297-331.
- Frazier, J. M. (1992) 'General perspectives on in vitro toxicity testing', in Frazier, J. M. (ed.) *In Vitro Toxicity Testing: Applications to Safety Evaluation*. New York: Marcel Dekker, pp. 1-11.
- Freinkel, R. K. and Woodley, D. T. (eds.) (2001) *The Biology of the Skin*. New York: The Parthenon Publishing Group.
- Frierson, H. F., Jr., Wolber, R. A., Berean, K. W., Franquemont, D. W., Gaffey, M. J., Boyd, J. C. and Wilbur, D. C. (1995) 'Interobserver reproducibility of the Nottingham modification of the Bloom and Richardson histologic grading scheme for infiltrating ductal carcinoma', *American Journal of Clinical Pathology*, 103(2), pp. 195-8.
- Fuchs, T. J. and Buhmann, J. M. (2011) 'Computational pathology: Challenges and promises for tissue analysis', *Computerized Medical Imaging and Graphics*, 35(7-8), pp. 515-530.

- Ganster, H., Pinz, A., Rohrer, R., Wildling, E., Binder, M. and Kittler, H. (2001) 'Automated melanoma recognition', *IEEE Transactions on Medical Imaging*, 20(3), pp. 233-239.
- Garbay, C., Brugal, G. and Choquet, C. (1981) 'Application of colored image analysis to bone marrow cell recognition', *Analytical and Quantitative Cytology*, 3(4), pp. 272-280.
- Gartner, L. P., Hiatt, J. L. and Strum, J. M. (2007) *Cell Biology and Histology*. 5 edn. Baltimore: Lippincott Williams & Wilkins.
- Geladi, P. and Grahn, H. (1996) *Multivariate Image Analysis*. Chichester, UK: John Wiley and Sons Ltd.
- Gerger, A., Bergthaler, P. and Smolle, J. (2004) 'An automated method for the quantification and fractal analysis of immunostaining', *Cellular Oncology*, 26(3), pp. 125-134.
- Gerger, A. and Smolle, J. (2003a) 'Diagnostic imaging of melanocytic skin tumors', *Journal of Cutaneous Pathology*, 30(4), pp. 247-252.
- Gerger, A. and Smolle, J. (2003b) 'Diagnostic tissue elements in melanocytic skin tumors in automated image analysis', *American Journal of Dermatopathology*, 25(2), pp. 100-106.
- Ghaznavi, F., Evans, A., Madabhushi, A. and Feldman, M. (2013) 'Digital Imaging in Pathology: Whole-Slide Imaging and Beyond', *Annual Review of Pathology: Mechanisms of Disease*, 8(1), pp. 331-359.
- Gladkova, N. D., Petrova, G. A., Nikulin, N. K., Radenska-Lopovok, S. G., Snopova, L. B., Chumakov, Y. P., Nasonova, V. A., Gelikonov, V. M., Gelikonov, G. V., Kuranov, R. V., Sergeev, A. M. and Feldchtein, F. I. (2000) 'In vivo optical coherence tomography imaging of human skin: norm and pathology', *Skin Research and Technology*, 6(1), pp. 6-16.
- Gonzalez, R. C. and Woods, R. E. (2008) *Digital Image Processing*. 3rd edn. New Jersey: Pearson Prentice Hall.
- Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M. and Yener, B. (2009) 'Histopathological Image Analysis: A Review', *IEEE Reviews in Biomedical Engineering*, 2, pp. 147-171.
- Hackam, D. G. and Redelmeier, D. A. (2006) 'Translation of research evidence from animals to humans', *The Journal of the American Medical Association*, 296(14), pp. 1727-1732.
- Hafiane, A., Bunyak, F. and Palaniappan, K. (2008) 'Fuzzy Clustering and Active Contours for Histopathology Image Segmentation and Nuclei Detection', *Proceedings of the 10th International Conference on Advanced Concepts for Intelligent Vision Systems*. Juan-les-Pins, France. Springer-Verlag Berlin Heidelberg, pp. 903-914.

- Haralick, R. M., Shanmugam, K. and Dinstein, I. H. (1973) 'Textural Features for Image Classification', *IEEE Transactions on Systems, Man and Cybernetics*, 3(6), pp. 610-621.
- Haralick, R. M., Sternberg, S. R. and Zhuang, X. (1987) 'Image Analysis Using Mathematical Morphology', *IEEE Transactions on Pattern Analysis and Machine Intelligence* (4), pp. 532-550.
- Hastie, T., Tibshirani, R. and Friedman, J. (2011) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd Edition edn. New York: Springer.
- Hastrup, N., Clemmensen, O. J., Spaun, E. and Sondergaard, K. (1994) 'Dysplastic naevus: histological criteria and their inter-observer reproducibility', *Histopathology*, 24(6), pp. 503-509.
- Hewitt, C. W., Englesbe, M. J., Tatem, L. D., Strande, L. F., Doolin, E. J., Dalsey, R. M. and DeLong, W. G. (1995) 'Graft-versus-host disease in extremity transplantation: digital image analysis of bone marrow in situ', *Annals of Plastic Surgery*, 35(1), pp. 108-112.
- Hitchins, L., Fucich, L. F., Freeman, S. M., Millikan, L. E. and Marrogi, A. J. (1997) 'Immunophenotyping as a diagnostic tool to differentiate lichen planus from chronic graft-versus-host disease: diagnostic observations on two patients', *Journal of Investigative Medicine*, 45(8), pp. 463-468.
- Horn, T. D., Bauer, D. J., Vogelsang, G. B. and Hess, A. D. (1994) 'Reappraisal of Histologic Features of the Acute Cutaneous Graft-Versus-Host Reaction Based on an Allogenic Rodent Model', *Journal of Investigative Dermatology*, 103(2), pp. 206-210.
- Ifarraguerri, A. I., O'Brien, G., Shen, W., Thompson, B. D., Harris, W., Freund, P. and Cascisa, R. (2003) *Computerised image capture of structures of interest within a tissue sample*. WO03105675A2
- Jannin, P., Krupinski, E. and Warfield, S. K. (2006) 'Validation in medical image processing', *IEEE Transactions on Medical Imaging*, 25(11), pp. 1405-1409.
- Jarvis, M., Schulz, U., Dickinson, A. M., Sviland, L., Jackson, G., Konur, A., Wang, X. N., Hromadnikova, I., Kolb, H. J., Eissner, G. and Holler, E. (2002) 'The detection of apoptosis in a human *in vitro* skin explant assay for graft versus host reactions', *Journal of Clinical Pathology*, 55(2), pp. 127-132.
- John, G. H. and Langley, P. (1995) *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*. Montreal, Quebec. Morgan Kaufmann Publishers Inc.
- Jondet, M., Agoli-Agbo, R. and Dehennin, L. (2010) 'Automatic measurement of epithelium differentiation and classification of cervical intraneoplasia by computerized image analysis', *Diagnostic Pathology*, 5(1), pp. 1-10.

- Kass, M., Witkin, A. and Terzopoulos, D. (1988) 'Snakes: active contour models', *International journal of computer vision*, 1(4), pp. 321-331.
- Keenan, S. J., Diamond, J., Glenn McCluggage, W., Bharucha, H., Thompson, D., Bartels, P. H. and Hamilton, P. W. (2000) 'An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (CIN)', *Journal of Pathology*, 192(3), pp. 351-362.
- Kempf, W., Hantschke, M., Kutzner, H. and Burgdorf, W. H. C. (2008) *Dermatopathology*. Steinkopff-Verlag Heidelberg (Accessed: 10 July 2015).
- Kiehl, P., Falkenberg, K., Vogelbruch, M. and Kapp, A. (2001) 'Tissue eosinophilia in acute and chronic atopic dermatitis: A morphometric approach using quantitative image analysis of immunostaining', *British Journal of Dermatology*, 145(5), pp. 720-729.
- Kohavi, R. (1995) 'A study of cross-validation and bootstrap for accuracy estimation and model selection', *International Joint Conference on Artificial Intelligence*. Montreal, Canada. pp. 1137-1145.
- Kola, I. (2008) 'The State of Innovation in Drug Development', *Clinical Pharmacology & Therapeutics*, 83(2), pp. 227-230.
- Kong, J., Sertel, O., Shimada, H., Boyer, K. L., Saltz, J. H. and Gurcan, M. N. (2009) 'Computer-aided evaluation of neuroblastoma on whole-slide histology images: Classifying grade of neuroblastic differentiation', *Pattern Recognition*, 42(6), pp. 1080-1092.
- Kong, J., Shimada, H., Boyer, K., Saltz, J. and Gurcan, M. (2007) 'Image analysis for automated assessment of grade of neuroblastic differentiation' *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. 12-15 April 2007 Arlington, VA, pp. 61-64.
- Kononenko, I. (1993) 'Inductive and Bayesian learning in medical diagnosis', *Applied Artificial Intelligence an International Journal*, 7(4), pp. 317-337.
- Kothari, S., Phan, J. H., Young, A. N. and Wang, M. D. (2011) 'Histological Image Feature Mining Reveals Emergent Diagnostic Properties for Renal Cancer', *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Atlanta, GA, 12-15 Nov. 2011. IEEE, pp. 422-425.
- Kothari, S., Phan, J. H., Young, A. N. and Wang, M. D. (2013) 'Histological image classification using biologically interpretable shape-based features', *BMC Medical Imaging*, 13(9) [Online] DOI: 10.1186/1471-2342-13-9.
- Kuncheva, L. and Whitaker, C. (2003) 'Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy', *Machine Learning*, 51(2), pp. 181-207.
- Lehr, H.-A., van der Loos, C. M., Teeling, P. and Gown, A. M. (1999) 'Complete Chromogen Separation and Analysis in Double Immunohistochemical Stains Using

- Photoshop-based Image Analysis', *Journal of Histochemistry & Cytochemistry*, 47(1), pp. 119-126.
- Leitinger, G., Cerroni, L., Soyer, H. P., Smolle, J. and Kerl, H. (1990) 'Morphometric diagnosis of melanocytic skin tumors', *American Journal of Dermatopathology*, 12(5), pp. 441-445.
- Lerner, K. G., Kao, G. F. and Storb, R. (1974) 'Histopathology of graft vs. host reaction (GvHR) in human recipients of marrow from HL A matched sibling donors', *Transplantation Proceedings*, 6(4), pp. 367-371.
- Levenson, R. M. (2004) 'Spectral Imaging and Pathology: Seeing More', *Laboratory Medicine*, 35(4), pp. 244-251.
- Li, S., Wu, H., Wan, D. and Zhu, J. (2011) 'An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine', *Knowledge-Based Systems*, 24(1), pp. 40-48.
- Lu, C. and Mandal, M. (2012) 'Automated segmentation and analysis of the epidermis area in skin histopathological images', *Annual International Conference of the IEEE in Engineering in Medicine and Biology Society (EMBC 2012)*. San Diego, CA. IEEE, pp. 5355-5359.
- Lu, C. and Mandal, M. (2014) 'Efficient epidermis segmentation for whole slide skin histopathological images', *Annual International Conference of the IEEE in Engineering in Medicine and Biology Society (EMBC 2014)*. Chicago, IL. IEEE, pp. 5546-5549.
- Mack, A. and Rock, I. (1998) *Inattentive blindness*. Cambridge, MA: The MIT Press.
- Madabhushi, A. (2009) 'Digital pathology image analysis: opportunities and challenges', *Imaging*, 1(1), pp. 7-10.
- Magee, D., Treanor, D., Crellin, D., Shires, M., Smith, K., Mohee, K. and Quirke, P. (2009) 'Colour Normalisation in Digital Histopathology Images', *Proc. Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*. London. Daniel Elson, pp. 100-111.
- Massi, D., Franchi, A., Pimpinelli, N., Laszlo, D., Bosi, A. and Santucci, M. (1999) 'A reappraisal of the histopathologic criteria for the diagnosis of cutaneous allogeneic acute graft-vs-host disease', *American Journal of Clinical Pathology*, 112(6), pp. 791-800.
- Matheron, G. (1975) *Random Sets and Integral Geometry*. New York: Wiley.
- Mokhtari, M., Rezaeian, M., Gharibzadeh, S. and Malekian, V. (2014) 'Computer aided measurement of melanoma depth of invasion in microscopic images', *Micron*, 61, pp. 40-48.

- Morris, J. A. (1994) 'Information and observer disagreement in histopathology', *Histopathology*, 25(2), pp. 123-128.
- Naik, S., Doyle, S., Agner, S., Madabhushi, A., Feldman, M. and Tomaszewski, J. (2008) 'Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology', *IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI)*. 14-17 May 2008. pp. 284-287.
- Ong, S. H., Jin, X. C., Jayasooriah and Sinniah, R. (1996) 'Image analysis of tissue sections', *Computers in Biology and Medicine*, 26(3), pp. 269-279.
- Otsu, O. (1979) 'A threshold selection method from gray-level histogram', *IEEE Transactions on Systems, Man and Cybernetics*, 9(1), pp. 62 - 66.
- Oztan, B., Shubert, K. R., Bjornsson, C. S., Plopper, G. E. and Yener, B. (2013) 'Biologically-driven cell-graphs for breast tissue grading', *10th International Symposium on Biomedical Imaging (ISBI)*. San Francisco, CA, 7-11 April 2013. IEEE, pp. 137-140.
- Paizs, M., Engelhardt, J. I. and Siklos, L. (2009) 'Quantitative assessment of relative changes of immunohistochemical staining by light microscopy in specified anatomical regions', *Journal of Microscopy*, 234(1), pp. 103-112.
- Perel, P., Roberts, I., Sena, E., Wheble, P., Briscoe, C., Sandercock, P., Macleod, M., Mignini, L. E., Jayaram, P. and Khan, K. S. (2007) 'Comparison of treatment effects between animal experiments and clinical trials: systematic review', *British Medical Journal*, 334(7586), p. 197.
- Petushi, S., Garcia, F., Haber, M., Katsinis, C. and Tozeren, A. (2006) 'Large-scale computations on histology images reveal grade-differentiating parameters for breast cancer', *BMC Medical Imaging*, 6(1), p. 14.
- Petushi, S., Katsinis, C., Coward, C., Garcia, F. and Tozeren, A. (2004) 'Automated identification of microstructures on histology slides', *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. Arlington, VA, 15-18 April 2004 pp. 424 - 427.
- Pilette, C., Rousselet, M. C., Bedossa, P., Chappard, D., Oberti, F., Rifflet, H., Maiga, M. Y., Gallois, Y. and Cales, P. (1998) 'Histopathological evaluation of liver fibrosis: Quantitative image analysis vs semi-quantitative scores: Comparison with serum markers', *Journal of Hepatology*, 28(3), pp. 439-446.
- Prewitt, J. M. S. and Mendelsohn, M. L. (1966) 'The analysis of cell images', *Annals of the New York Academy of Sciences*, 128(3), pp. 1035-1053.
- Price, G. J., McCluggage, W. G., Morrison, M. L., McClean, G., Venkatraman, L., Diamond, J., Bharucha, H., Montironi, R., Bartels, P. H., Thompson, D. and Hamilton, P. W. (2003) 'Computerized Diagnostic Decision Support System for the Classification of Preinvasive Cervical Squamous Lesions', *Human Pathology*, 34(11), pp. 1193-1203.



- Pudil, P., Novovicova, J. and Kittler, J. (1994) 'Floating search methods in feature selection', *Pattern Recognition Letters*, 15(11), pp. 1119-1125.
- Rajpoot, K. and Rajpoot, N. (2004) 'SVM Optimization for Hyperspectral Colon Tissue Cell Classification', in Barillot, C., Haynor, D. and Hellier, P. (eds.) *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004*. Springer Berlin Heidelberg, pp. 829-837.
- Ramsamooj, R., Llull, R., Tatem, L. D., Black, K. S., Lotano, V., Dalsey, R. M., Born, C. T., DeLong, W. G. and Hewitt, C. W. (1996) 'Graft-versus-host disease in limb transplantation: digital image analysis of bone marrow and TGF-beta expression in situ using a novel 3-D microscope', *Transplantation Proceedings*, 28(4), pp. 2029-31.
- Reinhard, E., Adhikhmin, M., Gooch, B. and Shirley, P. (2001) 'Color transfer between images', *Computer Graphics and Applications, IEEE*, 21(5), pp. 34-41.
- Ridler, T. W. and Calvard, S. (1978) 'Picture thresholding using an iterative selection method', *IEEE Transactions on Systems, Man and Cybernetics*, 8(8), pp. 630-632.
- Robertson, A. J., Anderson, J. M., Beck, J. S., Burnett, R. A., Howatson, S. R., Lee, F. D., Lessells, A. M., McLaren, K. M., Moss, S. M. and Simpson, J. G. (1989) 'Observer variability in histopathological reporting of cervical biopsy specimens', *Journal of Clinical Pathology*, 42(3), pp. 231-238.
- Rodríguez, J. D., Perez, A. and Lozano, J. A. (2010) 'Sensitivity analysis of k-fold cross validation in prediction error estimation', *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(3), pp. 569-575.
- Rousselet, M. C., Michalak, S., Dupre, F., Croue, A., Bedossa, P., Saint-Andre, J. P. and Cales, P. (2005) 'Sources of variability in histological scoring of chronic viral hepatitis', *Hepatology*, 41(2), pp. 257-264.
- Rubin, R., Strayer, D., Rubin, E. and McDonald, J. (2007) *Rubin's Pathology: Clinicopathologic Foundations of Medicine*. 5th edn. Baltimore: Lippincott Williams & Wilkins.
- Ruderman, D. L., Cronin, T. W. and Chiao, C. C. (1998) 'Statistics of cone responses to natural images: Implications for visual coding', *Journal of the Optical Society of America A: Optics and Image Science, and Vision*, 15(8), pp. 2036-2045.
- Ruifrok, A. C. (1997) 'Quantification of immunohistochemical staining by color translation and automated thresholding', *Analytical and Quantitative Cytology and Histology*, 19(2), pp. 107-113.
- Ruifrok, A. C. and Johnston, D. A. (2001) 'Quantification of histochemical staining by color deconvolution', *Analytical and Quantitative Cytology and Histology*, 23(4), pp. 291-299.

- Russell, W. M. S. and Burch, R. L. (1959) *The Principles of Humane Experimental Technique*. London: Methuen.
- Sahiner, B., Chan, H. P., Wei, D., Petrick, N., Helvie, M. A., Adler, D. D. and Goodsitt, M. M. (1996) 'Image feature selection by a genetic algorithm: application to classification of mass and normal breast tissue', *Medical Physics*, 23(10), pp. 1671-1684.
- Sahmoud, T. M., Heudes, D., Irinopoulou, T., Gluckman, E., Nguyen, Q., Brocheriou, C., Devergie, A., Rigaut, J. P. and Mary, J. Y. (1993) 'Towards an objective prognostic index of acute graft-versus-host disease', *Analytical Cellular Pathology*, 5(5), pp. 289-297.
- SCCP (Scientific Committee on Consumer Products) (2007) *Memorandum on the in vitro test EPISKIN™ for skin irritation testing, 18 December 2007*. European Commission.
- Seidenari, S., Pellacani, G. and Giannetti, A. (1999) 'Digital videomicroscopy and image analysis with automatic classification for detection of thin melanomas', *Melanoma Research*, 9(2), pp. 163-171.
- Sena, E. S., van der Worp, H. B., Bath, P. M. W., Howells, D. W. and Macleod, M. R. (2010) 'Publication Bias in Reports of Animal Stroke Studies Leads to Major Overstatement of Efficacy', *PLoS Biology*, 8(3), p. e1000344.
- Sertel, O., Kong, J., Catalyurek, U., Lozanski, G., Saltz, J. and Gurcan, M. (2009) 'Histopathological Image Analysis Using Model-Based Intermediate Representations and Color Texture: Follicular Lymphoma Grading', *Journal of Signal Processing Systems*, 55(1), pp. 169-183.
- Sezgin, M. and Sankur, B. (2004) 'Survey over image thresholding techniques and quantitative performance evaluation', *Journal of Electronic Imaging*, 13(1), pp. 146-168.
- Shinde, V., Burke, K. E., Chakravarty, A., Fleming, M., McDonald, A. A., Berger, A., Ecsedy, J., Blakemore, S. J., Tirrell, S. M. and Bowman, D. (2014) 'Applications of Pathology-Assisted Image Analysis of Immunohistochemistry-Based Biomarkers in Oncology', *Veterinary Pathology Online*, 51(1), pp. 292-303.
- Smolle, J. (1988) 'Mononuclear cell patterns in the skin. An immunohistological and morphometrical analysis', *American Journal of Dermatopathology*, 10(1), pp. 36-46.
- Smolle, J. (1996) 'Optimization of linear image combination for segmentation in red-green-blue images', *Analytical and Quantitative Cytology and Histology*, 18(4), pp. 323-329.
- Smolle, J., Fiebiger, M., Hofmann-Wellenhof, R. and Kerl, H. (1996) 'Quantitative morphology of collagen fibers in cutaneous malignant melanoma and melanocytic nevus', *American Journal of Dermatopathology*, 18(4), pp. 358-363.

- Smolle, J. and Gerger, A. (2003) 'Tissue counter analysis of tissue components in skin biopsies: Evaluation using CART (Classification and Regression Trees)', *American Journal of Dermatopathology*, 25(3), pp. 215-222.
- Smolle, J. and Hofmann-Wellenhof, R. (1998) 'Quantitative assessment of epidermal Langerhans cells using automated image analysis', *Skin Research and Technology*, 4(1), pp. 37-40.
- Smolle, J., Hofmann-Wellenhof, R., Soyer, H. P., Stettner, H. and Kerl, H. (1989a) 'Nuclear size and shape parameters correlate with proliferative activity in cutaneous melanocytic tumors', *Journal of Investigative Dermatology*, 93(1), pp. 178-182.
- Smolle, J., Soyer, H. P., Hofmann-Wellenhof, R., Smolle-Juettner, F. M. and Kerl, H. (1989b) 'Vascular architecture of melanocytic skin tumors. A quantitative immunohistochemical study using automated image analysis', *Pathology Research and Practice*, 185(5), pp. 740-745.
- Soille, P. (1999) *Morphological Image Analysis: Principles and Applications*. New York: Springer-Verlag.
- Soria, D., Garibaldi, J. M., Ambrogi, F., Biganzoli, E. M. and Ellis, I. O. (2011) 'A 'non-parametric' version of the naive Bayes classifier', *Knowledge-Based Systems*, 24(6), pp. 775-784.
- Standish, R. A., Cholongitas, E., Dhillon, A., Burroughs, A. K. and Dhillon, A. P. (2006) 'An appraisal of the histopathological assessment of liver fibrosis', *Gut*, 55(4), pp. 569-578.
- Sviland, L. and Dickinson, A. M. (1999) 'A human skin explant model for predicting graft-versus-host disease following bone marrow transplantation', *Journal of Clinical Pathology*, 52, pp. 910-913.
- Sviland, L., Dickinson, A. M., Carey, P. J., Pearson, A. D. J. and Proctor, S. J. (1990) 'An *in vitro* predictive test for clinical graft-versus-host disease in allogeneic bone marrow transplant recipients', *Bone Marrow Transplantation*, 5(2), pp. 105-109.
- Sviland, L., Hromadnikova, I., Sedlacek, P., Cermakova, M., Stechova, K., Holler, E., Eissner, G., Schulz, U., Kolb, H. J., Jackson, G., Wang, X. N. and Dickinson, A. M. (2001) 'Histological correlation between different centers using the skin explant model to predict graft-versus-host disease following bone marrow transplantation', *Human Immunology*, 62(11), pp. 1277-1281.
- Tabesh, A., Teverovskiy, M., Ho-Yuen, P., Kumar, V. P., Verbel, D., Kotsianti, A. and Saidi, O. (2007) 'Multifeature Prostate Cancer Diagnosis and Gleason Grading of Histological Images', *IEEE Transactions on Medical Imaging*, 26(10), pp. 1366-1378.
- Tadrous, P. J. (2010) 'On the concept of objectivity in digital image analysis in pathology', *Pathology*, 42(3), pp. 207-211.

Taylor, C. R. and Levenson, R. M. (2006) 'Quantification of immunohistochemistry - Issues concerning methods, utility and semiquantitative assessment II', *Histopathology*, 49(4), pp. 411-424.

The Mathworks (2010) 'Image Processing Toolbox Users Guide', *MATLAB R2010b Documentation*, [Online]. Available at: <http://www.mathworks.co.uk/help/images/>.

Toussaint, G. T. (1980) 'The relative neighbourhood graph of a finite planar set', *Pattern Recognition*, 12(4), pp. 261-268.

Trussell, H. J. (1979) 'Comments on "Picture Thresholding Using an Iterative Selection Method"', *IEEE Transactions on Systems, Man and Cybernetics*, 9(5), pp. 311-311.

van den Bent, M. J. (2010) 'Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective', *Acta neuropathologica*, 120(3), pp. 297-304.

Van Putten, P. G., Hol, L., Van Dekken, H., Han van Krieken, J., Van Ballegooijen, M., Kuipers, E. J. and Van Leerdam, M. E. (2011) 'Inter-observer variation in the histological diagnosis of polyps in colorectal cancer screening', *Histopathology*, 58(6), pp. 974-981.

Vogelsang, G. B., Hess, A. D. and Berkman, A. W. (1985) 'An *in vitro* predictive test for graft versus host disease in patients with genotypic HLA-identical bone marrow transplants', *New England Journal of Medicine*, 313(11), pp. 645-650.

Wang, Y., Crookes, D., Diamond, J., Hamilton, P. and Turner, R. (2007a) 'Segmentation of squamous epithelium from ultra-large cervical histological virtual slides', *9th Annual International Conference of the IEEE in Engineering in Medicine and Biology Society (EMBS 2007)* 2007, pp. 775-8.

Wang, Y. Y., Chang, S. C., Wu, L. W., Tsai, S. T. and Sun, Y. N. (2007b) 'A color-based approach for automated segmentation in tumor tissue classification', *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS 2007)*, 2007, pp. 6577-80.

Warfield, S. K., Zou, K. H. and Wells, W. M. (2004) 'Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation', *IEEE Transactions on Medical Imaging*, 23(7), pp. 903-921.

Whitney, A. W. (1971) 'A direct method of nonparametric measurement selection', *IEEE Transactions on Computers*, 20(9), pp. 1100-1103.

Wilkins, B. S., Erber, W. N., Bareford, D., Buck, G., Wheatley, K., East, C. L., Paul, B., Harrison, C. N., Green, A. R. and Campbell, P. J. (2008) 'Bone marrow pathology in essential thrombocythemia: interobserver reliability and utility for identifying disease subtypes', *Blood*, 111(1), pp. 60-70.

Wiltgen, M., Gerger, A., Wagner, C., Berghaler, P. and Smolle, J. (2007) 'Evaluation of texture features in spatial and frequency domain for automatic

- discrimination of histologic tissue', *Analytical and Quantitative Cytology and Histology*, 29(4), pp. 251-263.
- Xu, H. and Mandal, M. (2015) 'Epidermis segmentation in skin histopathological images based on thickness measurement and k-means algorithm', *EURASIP Journal on Image and Video Processing*, 2015(1), pp. 1-14.
- Zambruno, G., Girolomoni, G., Manca, V., Andreani, M., Galimberti, M., Lucarelli, G. and Giannetti, A. (1992) 'Epidermal Langerhans cells after allogeneic bone marrow transplantation: depletion by chemotherapy conditioning regimen alone', *Journal of Cutaneous Pathology*, 19(3), pp. 187-92.
- Zhang, H. (2004) 'The optimality of naive Bayes', *American Association for Artificial Intelligence*, 1(2), p. 3.
- Zheng, Z. and Webb, G. I. (2000) 'Lazy learning of Bayesian rules', *Machine Learning*, 41(1), pp. 53-84.
- Zhou, R., Hammond, E. H. and Parker, D. L. (1996) 'A multiple wavelength algorithm in color image analysis and its applications in stain decomposition in microscopy images', *Medical Physics*, 23(12), pp. 1977-1986.
- Zitová, B. and Flusser, J. (2003) 'Image registration methods: A survey', *Image and Vision Computing*, 21(11), pp. 977-1000.