Automatic Severity Assessment of Hand Eczema

Tim Havinga

August 13, 2010

Abstract

The classification of hand dermatitis (HD) images is an area in which multiple attempts have been made to provide a good manual classification scheme. These methods vary in their approach by the degree of complexity and the range of output classifications. This paper tries to provide a computer severity assessment, based on photographs of the hand, which should become an aid to the dermatologists to make reliable severity assessments. Different classification methods and regression methods are tested, as well as different colour bands, to come to an optimal severity assessment. The error made by this severity assessment is compared to the error made by the dermatologists, to see if the proposed severity assessment can compete with that of the dermatologists. Using the L*a*b* colour space as best tested colour space, bagged regression trees and the recent Limited Rank GMLVQ method prove to be valuable regression methods. They are favourable over classification methods and approach the error made by the dermatologists themselves within 13 %.

Supervisors:

Dr. M.H.F. Wilkinson
University of Groningen

Prof. P.J. Coenraads
University Medical Centre Groningen

Contents

Chapter 1

Introduction

Hand eczema (HE), sometimes called hand dermatitis (HD) is a disease in which the hand is affected with dermatitis. This can be a chronic disease, in which the patient suffers from pain and a combination of visual characteristics on the hand. The disease can have social implications, interfere with the functioning in a job or in domestic tasks or even imply permanent disability. In this thesis, we suggest an automatic severity assessment which should ultimately lower the work burden for the patients and the dermatologists by setting a baseline for severity assessment and reducing the number of necessary hospital visits.

1.1 About hand eczema

Hand eczema has several characteristics by which it is recognised by the dermatologists. They are listed below to give the reader an impression of the disease. The definitions are partially taken from [6]. Hand eczema is diagnosed when a patient has several of the below characteristics.

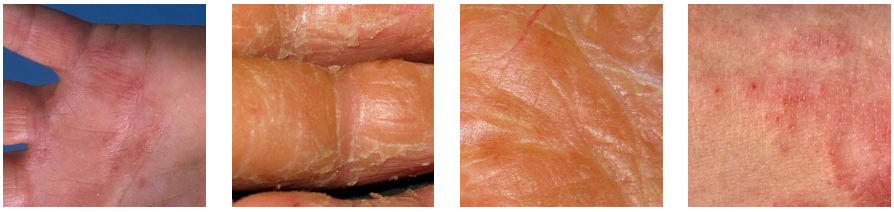
- **Erythema** is redness of the skin, which can vary from slight redness to a deep intense red colour.
- **Papules** are small bumps in the skin. They are easily recognised by feeling the skin, but hard to see.
- **Scaling** is flaking of the skin, varying from fine scales over a limited area to desquamation (shedding of the outer skin layers) with coarse thick scales covering up to 30% of the hand area.
- **Hyperkeratosis and linchenification** is thickening of the skin (hyperkeratosis), exaggerating skin lines (linchenification), varying from mild thickening in limited areas to prominent thickening over widespread areas with exaggeration of normal skin markings.
- **Vesiculation** implies small blisters (vesicles) on the hand, scattered in mild cases, more clustered with erosion (remains of a vesicle that has lost

its fluids) or excoriation (dents in the skin where vesicles used to be) in more severe cases. They are most prominent between the fingers and on the hand palm.

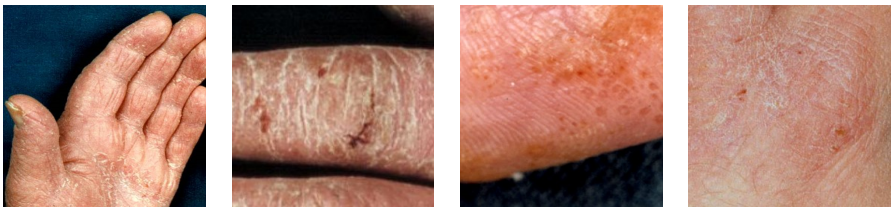
- **Oedema** is an accumulation of fluids beneath the skin, with noticeable thicker and firmer skin in more severe cases.
- **Fissures** are cracks in the skin, which are usually narrow but deep. The condition varies from superficial cracked skin to fissures that cause bleeding and pain.
- **Dryness** of the skin.
- **Pruritus** (itching) and **pain** varying from slight discomfort a few times a day to persistent pain that can interfere with sleep.

Visual representations of these characteristics are presented in figure 1.1.

For our research, the dermatologists specifically asked for a recognition method that does not try to identify these characteristics, as the dermatologists themselves can do this perfectly. They asked to create other – image processing – features of the hand images on which the diagnosis would be based, to see what are the most prominent features for a computer to classify an image on. That way these classifications could potentially be useful to the dermatologists. Besides this, the computer classification is intended to give a more robust classification, as dermatologists could potentially be biased in their scoring because of a previous case they have seen, but the computer has no such memory.



(a) Erythema (b) Scaling (c) Hyperkeratosis (d) Papules



(e) Oedema (f) Fissures (g) Vesicles (h) Dryness

Figure 1.1: The characteristics on which the severity of hand eczema can be assessed. Images courtesy UMCG.

1.2 Previous work

This thesis is a continuation of the master thesis by B. van de Wal [26] who made a start in the field of automatic classification of images of hand affected with hand eczema. Several issues were thoroughly researched and examined in this work, which are listed below.

1.2.1 Colour models

A focus in the previous work are the different available colour models. There was finally decided to perform the experiments only with the red colour band, because "the red colour band gave the best results in preliminary tests" [26, 5.5.3]. In the mean time, we have acquired extra photographs with their corresponding classification, such that we want to test the performance of other colour bands, including the ones from other colour models, again. We will also try to run the classification with the three colour bands of each colour model concatenated into a feature vector that is three times as long. Previous experiments in this area [10] proved that the curse of dimensionality [1] was not present during testing. This might be because of the high correlation between features in that specific test, we will investigate the occurrence in our research.

1.2.2 Preprocessing

Some research went into the preprocessing of images. This was done by first removing the blue neutral background that all provided images have. For this purpose, k-means clustering [18] was used, clustering the image into a foreground or hand section and a background section, which would later be removed. This binary mask was filled up to remove any holes, and eroded with a disc to remove the border pixels, which are frequently blurred and do not contain useful information. See figure 1.2 for a preprocessing example.

1.3 Automatic classification

1.3.1 Automatic classification versus dermatologist's classification

The major disadvantage in the classification of hand dermatitis by dermatologists, is that their classification can vary because the dermatologist is influenced by previous classifications or by difference in experience between dermatologists. For this reason, all previous researches that create a classification scheme are dependent on the classifying dermatologist. Therefore, many of them feature statistical tests, in one form or another, that measure the interobserver and intraobserver ratings. The interobserver rating is the difference in classification of the same hand between two dermatologists. The intraobserver rating is the difference between two classifications by the same dermatologist of the same

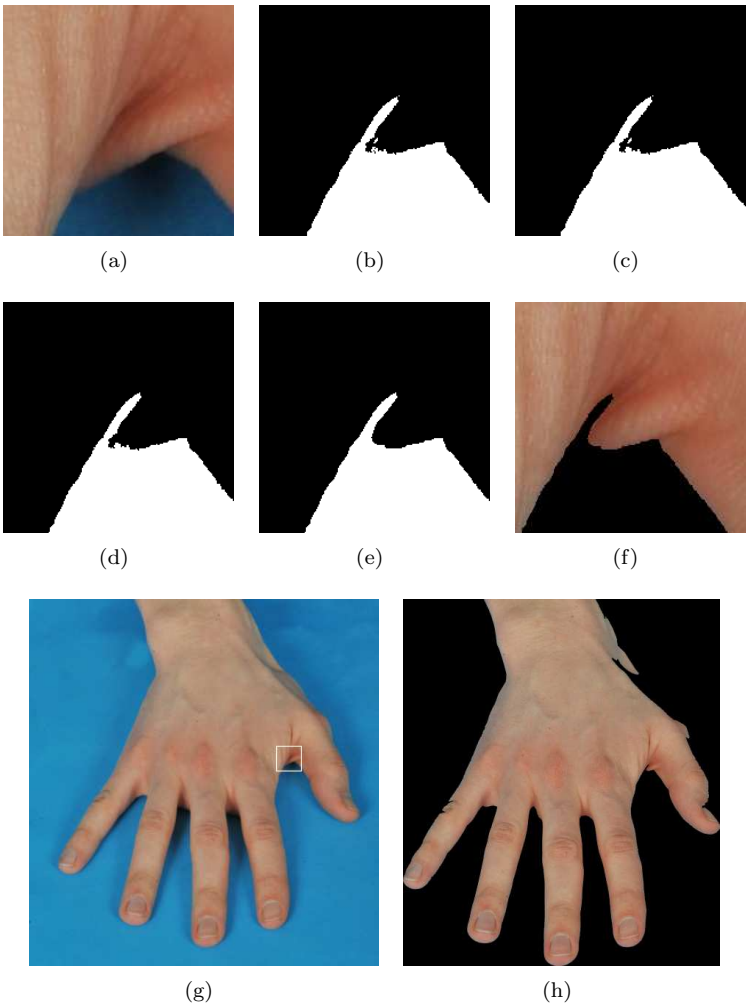


Figure 1.2: The segmentation of a hand image, in phases. Segmentation is shown for a small region for clarity. Figure (a) shows the original image, (b) is the result of the 2-means segmentation, in (c), the holes in the white (outer) region are filled, in (d), the holes in the black (inner) region are filled, (e) shows the smoothed mask by an opening with a disc of radius 5, and (f) shows the result. Figures (g) and (h) show the original image and the segmentation result (trimmed to the hand component), respectively. The white rectangle indicates the zoomed area shown in (a) t/m (f). Original image is courtesy UMCG.

hand, but on different times. These tests need to be performed because of the human element.

For our purposes, there is no need to measure inter- or intraobserver reliability, because the computer will always give the same classification when presented with the same image. (Image processing difficulties like scaling are discussed later.) It is unthinkable that the computer would be influenced by previous decisions. However, the computer is dependent on the data that is given to him. Therefore, we must provide a reliable basis for classification by providing a large amount of example images.

Because we only have a single classification for our data set, some effort will be made to obtain other dermatologist's classifications, such that we can measure the error made by the dermatologists and incorporate this error margin into our results.

Furthermore, because the classification of our data set by this dermatologist sets the standard for our computer severity assessment, there is no need to cross-evaluate this assessment with classifications given by other dermatologists, once we know this error measure.

1.3.2 The benefits of automatic classification

The research into automatic classification of hand dermatitis has several grounds which support the need for research.

First of all, the development of an automatic classification scheme would provide a better efficiency in hospital work, by reducing the work load of dermatologists to assess the hand eczema severity of patients, and only requiring the patients to visit the hospital for a less frequent check-up or, for example, when the classification system spots an interesting shift in the severity of the patient's disease.

Secondly, it is a challenge to test if the classification system that is developed is robust enough to work with the images that are taken by less qualitative cameras, for example a mobile phone camera or a computer web camera.
Issues in this matter are the measurement of the size of the hand, the lighting conditions and the background. However, this is material for further research, when the current research proves its value.

Thirdly, if the severity assessment by the computer becomes more reliable than an assessment made by a dermatologist, this means that we can use it to objectively quantify the results of treatment, and reliably monitor the patient's progress over time. Without having to take into account the difference in classification that can occur between dermatologists, or between classifications of the same dermatologist over time.

Chapter 2

Related work

According to [4], the severity scoring of skin diseases has been neglected. Research is done, but results are not comparable because of the different severity scoring systems created by the researchers, because no standard scoring system exists. Researchers can choose the best fitting scoring system for their needs, which makes meaningful interpretation and comparison of results very difficult. In the following we describe the most important existing systems.

2.1 Manual classification schemes

The developed manual classification schemes can be grouped into photographic scales and written or verbal scales.

2.1.1 Written or verbal scales

Written or verbal scales are a clinical assessment of hand dermatitis severity, grading several of the characteristics listed in Section 1.1. This is done overall, per region (palm, dorsal, fingers), or for both hands separately.

The HECSI score [14] is a scoring system based on disease symptoms and extent in different hand areas. The HECSI score gives five hand regions a score from $0-3$ for several disease symptoms and a score from $0-4$ for the overall extent in that region. Multiplying the sum of the feature scores with the extent, and summing this for all regions gives a total HECSI score in the region $0-360$.

The subdiagnosis scale [8] measured the medical history, HECSI score, and an outcome of the ESS patch test. Using this information, they try to connect the symptoms to the possible options for HE sub diagnoses (ACD, ICD, AHE, discoid, vesicular and hyperkeratotic hand eczema). However, they conclude with the notion that "there is no simple translation from morphology to subdiagnoses of hand eczema", which excludes their research from our investigation.

2.1.2 Photographic scales

Photographic scales are generally less specific, placing the severity of the hand in one of these classes which are represented by a number of different example photographs. The downside of photographic scales is that they do not allow for the inclusion of pruritus or pain, and cannot display all symptoms of the disease as clearly. However, photographic standards have been shown to perform better than descriptive ones [5].

The photographic guide [7] selected 50 representative photos, providing a mixture of dorsal and palmar views, male and female hands and level of severity. A guide was created with five levels of severity, where experts chose the four most representative photos for each level. Hand eczema severity assessment is performed by resembling the patient's hands to the photographs, selecting the image with the closest resemblance and its corresponding classification.

2.1.3 Other scales

According to [6], both methods have their drawbacks. A photographic scale has the drawback that it cannot consider the patient's perception of pain and the impact of the disease on the patient's quality of life. Written scales have the drawback that they are not ideal for the integration of multiple clinical signs into a single severity score. To counter the drawbacks of both methods, Coenraads et al. [6] proposed a combined scale with additional features like the HECSI scale has. They base their scale on the photographic guide, but add descriptions for symptom severity for each scale, and add additional disease symptoms that cannot be measured from a photograph, being pruritus and pain. After scoring all seven symptoms on a scale from $0-3$ (absent, mild, moderate, severe), the severity level, from $0-4$ inclusive, is shown in a table, called the Overall PGA Severity Rating. Four of the seven symptoms are given more primary importance, since they are "especially bothersome"[1].

2.1.4 Comparison

It is hard to near impossible to compare the results of the different severity scores, because (almost) each paper proposes another scale to measure hand eczema severity, and use different statistics to back up their method.
Charman and English [5] review both the photographic guide and HECSI scale. they argue that both scales were tested on reliability rather than validity. Overall, the photographic guide was more insightful and produces better reproducibility numbers. Also, both methods focus more on disease symptoms than the quality of life of the patient. Clearly, the combined scale solved this partially by adding pruritus or pain to their scoring form. There is no review of the combined scale, though it was used successfully in practice [21].

[1] The features fissures, vesiculation, oedema and pruritus/pain were given primary importance over erythema, hyperkeratosis and desquamation.

2.2 Automatic hand eczema recognition

In the image processing field, there have been some researches as to recognise hand eczema from digital photographs, but most research limits itself to finding regions that can be classified as having hand eczema as opposed to clean skin in a close-up of the skin [11, 24]. This in contradiction to our current problem, which also has non-skin pixels (i.e. background). Therefore, these studies are not applicable to our current problem. Furthermore, the area we are researching contains a whole hand, including nails, hair, etc. This requires a radical different approach than segmenting a lesion from a skin image.

Besides this, as told before, the dermatologists are specifically interested in the features of the hand which are most prominently used in the automatic classification. We were told explicitly not to mimic the dermatologist, looking for the signs and symptoms described in section 1.1.

Chapter 3

Materials and methods

For this research, the photographs of the previous research could be used, as well as additional photographs with corresponding classifications, gathered by the University Medical Centre Groningen (UMCG). According to [26], the images are distributed over class I as presented in table 3.1(a). However, the photographic guide used to rate the images has only five classes (see [7]), instead of the six ones mentioned. We therefore concluded that the single image in series I, class 1 should belong in one of the other classes. We decided to shift all images with classes 2 and higher one class number downward. With the additional images provided by the UMCG shown as series III, including some effort made by the author to obtain more class 0 (clean) hands, the distribution becomes as in table 3.1(b).

(a) The original distribution of images in series I

Class	0	1	2	3	4	5
Nr of images	8	1	19	5	8	9

(b) The current distribution of images

Class	**Series**			**Total**	**A priori chance**
	I	**II**	**III**		
0 (clear)	8	14	35	57	23.75 %
1 (mild)	20	62	26	108	45.00 %
2 (moderate)	5	7	30	42	17.50 %
3 (severe)	8	0	14	22	9.17 %
4 (very severe)	9	0	2	11	4.58 %
Total:	50	83	107	240	100.00 %

Table 3.1: The original and current distribution of images over the different classes.

3.1 Extensions to the previous work

3.1.1 Preprocessing

The preprocessing used in practice proved to be different from the preprocessing mentioned in [26]. We modified the image segmentation slightly, because the k-Means algorithm starting points proved to give incorrect segmentations (while segmenting series III, 6.9 % was segmented incorrectly).

The segmentation of the hand from the background is done by first defining a typical skin pixel colour, and finding the pixel closest to that colour. This pixel is used as the key point for the hand component. For the background, we assume there is a relative smooth, uniformly coloured background. Therefore, we decided to take one of the corner points as the key background point. Because it is possible that the hand intersects with one of the corners, we calculate the most deviating corner, and take the opposite one. These two key points are handed to the k-Means clustering algorithm [18], which in our case has two means, which separates all image pixels based on their resemblance to the colours of the two key points. After an iteration, the k-Means clustering calculates the average colour of both components, and takes these colours again as starting points for clustering. Eight iterations are performed. The holes in the resulting binary mask are filled up. This discards all components that are not 4-connected to the hand component, but is necessary because the program is not yet capable of handling masks containing multiple components. Next the hand component is opened with a disc of radius five. The original image is segmented using this final mask, and trimmed to the hand component. Figure 1.2 shows the complete preprocessing phase.

3.1.2 Feature extraction

A lot of features were calculated for each image in the previous research. First of all, the image histogram is calculated and binned into 7 bins. The other features are the Haralick features and the area/shape spectra.

Haralick features

[13] proposes a system in which several features, called the Haralick features, are calculated over Gray Level Co-occurrence Matrices (GLCMs) of an image. The GLCMs contain information about the distribution of the gray values in an image.

For example, the 1-distance horizontal GLCM updates for each pixel pair (p_1, p_2) which are horizontally next to each other. So they fulfil the conditions $|p_1.i - p_2.i| = 1$ and $p_1.j = p_2.j$. Then the values of $glcm(f(p_1), f(p_2))$ and $glcm(f(p_2), f(p_1))$ are increased by one, because the horizontal direction operates both left-to-right and right-to-left. Here, the $f(p)$ operation returns the gray value of pixel p.

This creates a matrix $glcm$ of size $N \times N$, where N is the number of gray levels, 256 in our case. These GLCMs are created for each of the distances 1,

0	0	1	1
0	0	1	1
0	2	2	2
2	2	3	3

(a)

		Gray value			
		0	1	2	3
Gray value	0	$\#(0,0)$	$\#(0,1)$	$\#(0,2)$	$\#(0,3)$
	1	$\#(1,0)$	$\#(1,1)$	$\#(1,2)$	$\#(1,3)$
	2	$\#(2,0)$	$\#(2,1)$	$\#(2,2)$	$\#(2,3)$
	3	$\#(3,0)$	$\#(3,1)$	$\#(3,2)$	$\#(3,3)$

(b)

$$0^\circ: \quad P_H = \begin{pmatrix} 4 & 2 & 1 & 0 \\ 2 & 4 & 0 & 0 \\ 1 & 0 & 6 & 1 \\ 0 & 0 & 1 & 2 \end{pmatrix} \qquad 90^\circ: \quad P_V = \begin{pmatrix} 6 & 0 & 2 & 0 \\ 0 & 4 & 2 & 0 \\ 2 & 2 & 2 & 2 \\ 0 & 0 & 2 & 0 \end{pmatrix}$$

(c) (d)

$$135^\circ: \quad P_{LD} = \begin{pmatrix} 2 & 1 & 3 & 0 \\ 1 & 2 & 1 & 0 \\ 3 & 1 & 0 & 2 \\ 0 & 0 & 2 & 0 \end{pmatrix} \qquad 45^\circ: \quad P_{RD} = \begin{pmatrix} 4 & 1 & 0 & 0 \\ 1 & 2 & 2 & 0 \\ 0 & 2 & 4 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

(e) (f)

Figure 3.1: Gray Level Co-occurrence Matrix explanation taken from [13]. (a) A 4 by 4 image with 4 gray values. (b) General form of any gray-tone spatial dependence matrix for images with gray tones 0–3. $\#(i,j)$ stands for the number of times gray values i and j have been neighbors, for a specific direction. (c)-(f) Calculation of all four distance 1 gray-tone spatial dependence matrices. P_H, P_V, P_{LD} and P_{RD} are the horizontal, vertical, left diagonal and right diagonal GLCMs, respectively. In our research, these matrices are created for distances 1, 2 and 4 pixels, with 256 gray values. The Haralick features (see table 3.3) are calculated based on these matrices.

2 and 4, and each of the directions horizontal, vertical, right diagonal and left diagonal. See figure 3.1 for an example calculation of the different direction GLCMs.

Using these GLCMs, the Haralick features are calculated, see table 3.3 for an overview. We calculated the Haralick features over the 1-, 2- and 4-distance GLCMs, over all directions at once. The Haralick features measure several characteristics of the image, such as contrast, entropy, variance, etc. These are the values that are used in our feature vector.

Area/shape spectra

The area/shape spectra [25], a type of pattern spectra [19], are calculated over the Min- and Max-Tree [22] of the image. A Max-Tree is a tree representation of the image, where each node represents a connected component in which all

gray values are equal or higher:

$$\forall (i,j) \in C_h^k : f(i,j) >= h$$

Where C_h^k is the kth connected component in gray level h and $f(i,j)$ denotes the gray value at pixel (i,j).

As each tree node is a connected component containing gray levels that are higher than those of his parent node, the tree leaves represent connected components in which all gray values are equal, so-called flat zones. All pixels of a child node are therefore contained in its parent node. The parent nodes keep increasing in size towards the root of the tree, which contains the entire image – or, in our case, just the hand component. See figure 3.2 for an example Max-Tree. Additionally, a Min-Tree is a Max-Tree of the inverse image.

Because the size of the tree nodes is increasing when traversing towards the root of the tree, they are excellent for performing connected filtering using attribute openings and thinnings, also known as granulometries [19]. These capabilities of the Max-Tree are not used in this research.

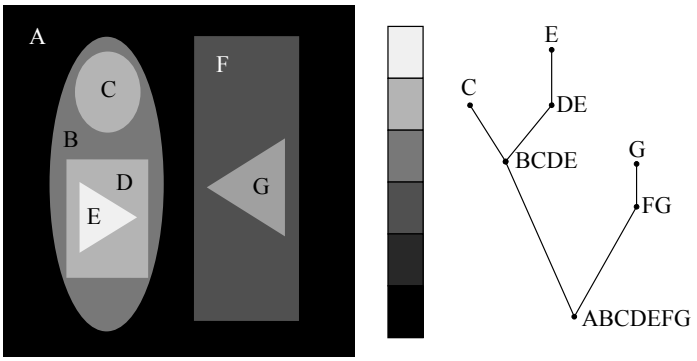


Figure 3.2: An example of a max tree. The left image is shown as a tree structure on the right, with corresponding colour spectrum. As shown in this image, the lighter colours are the leaves of the tree, while the other nodes represent a connected component of which the gray value is equal or higher.

Over these Min- and Max-Trees, the area/shape spectra were calculated. These spectra can be seen as a 2D histogram of the binned area versus the binned shape feature. They are constructed as follows. For each of these features, an 8 by 8 matrix is created, representing a spectrum of the area versus one of these shape features, as defined by [25]. For each node in the tree, the area and the features in table 3.2 are calculated. The scales used in these spectra are logarithmic. This emphasises smaller nodes and corresponding shape feature values. These are the interesting nodes, because they could contain shapes of blisters or fissures which could play an important role in classification, in contrast to the larger nodes which will contain larger sections of the hand. Figure 3.3 shows the difference between linear scales and logarithmic scales for a clean hand and an afflicted hand.

The scaling of the horizontal axis is defined by the smallest and largest possible areas and the scaling of the vertical axis is defined by the smallest and largest feature values. To be able to compare these area/shape spectra for different images, these minimal and maximal scales are first stored for each image, and then the optimal scaling that can contain all values is used for all images.

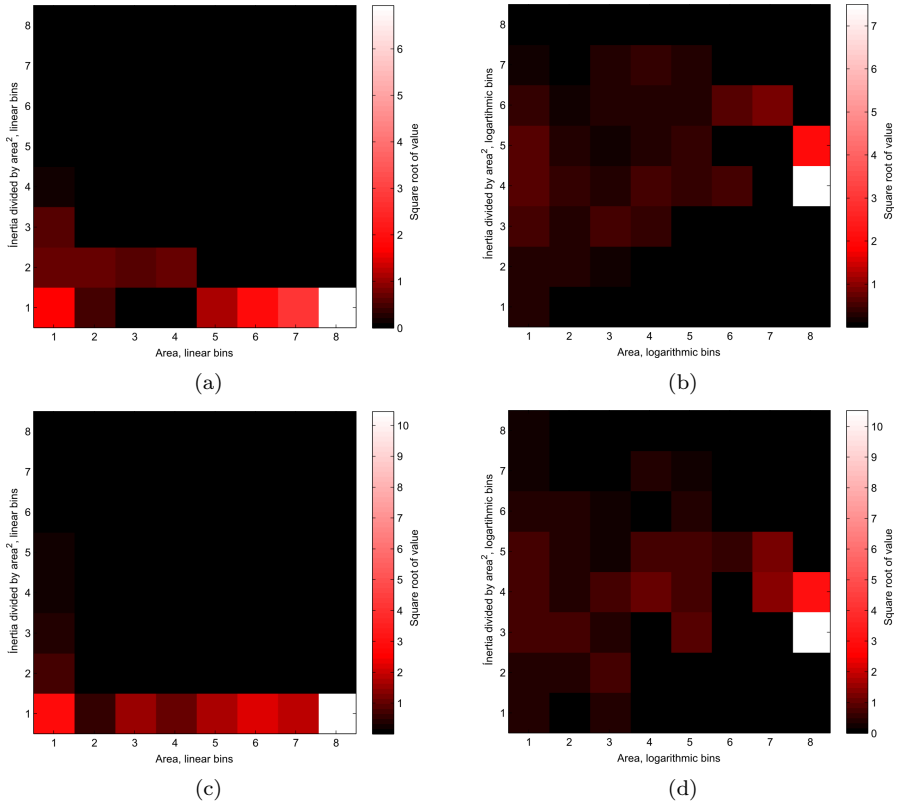


Figure 3.3: Area/shape spectra with linear (a,c) and logarithmic (b,d) scales. The top row shows a clean hand (class 0) and the bottom row shows a diseased hand (class 4). In the spectrum with linear scales, the size of the bins is defined by the outliers and the very large components. The logarithmic scale emphasises the smaller nodes.

3.1.3 Feature selection

The feature extraction results in a total feature vector of 811 features. We will assess the quality of these features again by redoing the feature selection.

The number of features mentioned above is for just a single colour band. When we would use all three colour bands, this number would be multiplied

by three. If we use the different scales – calculate all features for the original, half and quarter size image – this number would again be multiplied by three. Therefore, we decide to skip the three scales, and instead take a good look at the feature scalability. We will try to look at the three different colour bands at the same time.

A priori, we have the following considerations about the feature vectors:

Feature validity The features we chose were chosen with some care, but it is reasonable to imagine that there are features that have no additional value for the classification, which are highly correlated with other features or which are just noise.

Feature vector length Because of this large number of dimensions in the feature vectors, it is argued that we run the risk of "the curse of dimensionality" [1], which states that, with a large number of dimensions as opposed to the number of samples, there is always some correlation. More generalized data is favourable, because this would predict future data better, the classifier is not fitted to the training data.

For above reasons, we try to reduce the dimensionality of the feature vector by:

Principal Component Analysis Principal Component Analysis (PCA) [20] reduces pairs of two features by calculating the maximum variance and its axis, and mapping the data onto that axis, reducing the 2D problem to a 1D one. In practice, this means that a data set with an arbitrary number of features can be reduced to a lower-dimension problem by preserving the n components with the highest variance. The downside of this is that it works best on Gaussian (bell-shaped) data.

Self-Organising Maps Self-Organising Maps (SOMs) [16] define a grid of prototypes for the samples. This grid is trained to cover the entire sample space. The key issue is that SOMs do not use the label information. Instead, they try to find the inherent structure in the data. Therefore, the prototypes of the SOM are labelled afterwards (for example by the majority vote of the closest sample labels). SOMs are extremely useful for testing if a data set is separable in clusters or not.

Learning Vector Quantization Learning Vector Quantization (LVQ) [3] trains several prototypes that represent the data. In contrast to SOMs, they do use the labels for training. Each class is represented by one or more prototypes. Samples are labelled based on the closest prototype. LVQ is a simple and efficient algorithm for data classification. We adapt the distances to the prototypes to also yield a regression score.

3.2 Classification versus regression

As stated in the Related work chapter (Chapter 2), several classification schemes were developed. Each has its own scale, varying in precision. When the classification problem would be transformed into a regression problem, this might give the possibility to transpose each of the classification schemes to this new regression scale. Current classification schemes all classify on a natural number scale, while we would like to make this a real number scale, to enhance precision and allow mapping of one classification scheme to the other.

A more precise scale would be valuable, because the classification scheme utilised in the research used for this paper scaled from 0 to 4, and one can imagine that the dermatologists can be at a loss sometimes when choosing between two classes. Also, when monitoring a patient's progress, it could prove worthwhile when the severity assessment is more specific, to see that it is improving or decreasing. It could take some time to jump into another class in a five-class system.

	Name	**Function**
f_{15}	Moment of inertia	$\sum x^2 + \sum y^2 - \frac{(\sum x)^2 + (\sum y)^2}{A} + \frac{A}{6}$
f_{16}	Inertia divided by area squared	$\frac{f_{15}}{A^2}$
f_{17}	Compactness	$\frac{4\pi A}{P^2}$
f_{18}	Jaggedness	$\frac{AP^2}{8\pi^2 f_{15}}$
f_{19}	Entropy	$-\sum z(i)\log(z(i))$
f_{20}	Lambda max	Maximum child gray level minus current gray level

Where:

A = the area of a node, in pixels
P = the perimeter of a node, in pixels
$z(i)$ = the value of the histogram at gray level i

Table 3.2: The shape features that are calculated over all nodes of a tree and binned versus their area into an 8 by 8 matrix with log scales (the area/shape spectra).

	Name	Function
f_1	Angular second moment	$\sum_i \sum_j p(i,j)^2$
f_2	Contrast	$\sum_{n=0}^{N_\theta - 1} n^2 \left(\sum_{i=0}^{N_\theta} \sum_{j=1 \wedge \lvert i-j \rvert = n}^{N_\theta} p(i,j) \right)$
f_3	Correlation	$\frac{\sum_i \sum_j (ij) p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$
f_4	Sums of squares: variance	$\sum_i \sum_j (i-\mu)^2 p(i,j)$
f_5	Inverse difference moment	$\sum_{i}^{2N_\theta} \sum_{j} \frac{1}{1+(i-j)^2} p(i,j)$
f_6	Sum average	$\sum_{i=2}^{2N_\theta} i p_{x+y}(i)$
f_7	Sum variance	$\sum_{i=2}^{2N_\theta} (i - f_8)^2 p_{x+y}(i)$
f_8	Sum entropy	$-\sum_{i=2}^{2N_\theta} p_{x+y}(i) \log(p_{x+y}(i))$
f_9	Entropy	$-\sum_i \sum_j p(i,j) \log(p(i,j))$
f_{10}	Difference variance	$\sum_{i=0}^{N_\theta - 1} i^2 p_{x-y}(i)$
f_{11}	Difference entropy	$-\sum_{i=0}^{N_\theta - 1} i^2 p_{x-y}(i) \log(p_{x-y}(i))$
f_{12}	Information correlation 1	$\frac{HXY - HXY_1}{\max(HX, HY)}$
f_{13}	Information correlation 2	$\sqrt{1 - \exp(-2(HXY_2 - HXY))}$
f_{14}	Max. correlation coefficient	$\sqrt{\text{second largest eigenvalue of } \mathbf{Q}}$

Where:

N_θ = the number of gray levels, equal to the height and width of the GLCM

$p(i,j)$ = the value of the GLCM at (i,j)

p_x, p_y = the partial probability density functions

μ_x, μ_y = the means of p_x and p_y

σ_x, σ_y = the standard deviations of p_x and p_y

p_{x+y}, p_{x-y} = the probability of all GLCM coordinates summing to $x+y$ and $x-y$

HX, HY = the entropies of p_x and p_y

$HXY = -\sum_i \sum_j p(i,j) \log(p(i,j))$

$HXY_1 = -\sum_i \sum_j p(i,j) \log(p_x(i) p_y(j))$

$HXY_2 = -\sum_i \sum_j p_x(i) p_y(j) \log(p_x(i) p_y(j))$

$\mathbf{Q}(i,j) = \sum_{k=0}^{N_\theta} \frac{p(i,k)p(j,k)}{p_x(i)p_y(k)}$

Table 3.3: The Haralick features, as described in [13].

Chapter 4

Experiments

4.1 Introduction

In this section, the results of the experiments mentioned in the materials and methods chapter (chapter 3) are shown. When results required extra investigation and further experimentation, this will be clearly indicated.

First, a scale-invariance test was performed. This test had two purposes: first to check if all the used features were scale-invariant enough, and second, to check if the system would still give acceptable results in this smaller scale. This second purpose has a more practical goal: if the system is robust to scaling of the image down to the dimensions that are acquired by using a regular web cam or mobile phone camera, this provides solid ground for further research to let patients take their own images instead of requiring them to go to the hospital each time.

The second test performed was to see if the problem could be extended to a regression problem. As explained, the photographs are rated by the dermatologist on a discrete scale from 0 to 4. However, one can imagine that the disease severity is a continuous scale, rather than a classification problem. So, we would like to rate the image on a linear scale from 0 to 4. Thus, we want the system to be able to rate the severity continuous. Because we can assume that it is sometimes hard for the dermatologists to place a hand in one category or another, and we have several example data where multiple different classifications are given, we would like to investigate if extending the problem to a regression problem would prove valuable.

For this second test, we take a look at how a Self-Organising Map (well summarised in [15]) would cluster our images. This method tries to map the multi-dimensional feature space to a lower-dimensional space. This low-dimensional space is usually 2D, for visualisation purposes. In mapping the data to a lower dimension, the SOM algorithm works unsupervised, it has no knowledge of the labels of the data. Therefore, if the outcome of the SOM is clearly divided into the five different classes, we can see that it is a classification problem. If

the outcome is more of a smooth curve along which the classes lie, or no clear clusters are visible, the problem is more likely a regression one.

Finally, as third test, we will have to assess the outcomes of the classifier or regression function. It depends upon the previous experiment which of the two we will have to evaluate. If it is a classification problem, we can resemble our results to the results of the previous research. This will be done by testing the performance of different classifiers, and improving the performance of the best classifiers in the previous research. Especially the research into Learning Vector Quantization (LVQ) has made great advances [3], and will be a key focus classifier. In the case of regression, we will also look at the LVQ method, because it also has regression capabilities.

We hope to come to good enough assessment to match against the dermatologist's assessment. This would provide a good foundation for further research into practical use.

4.2 Scale-invariance

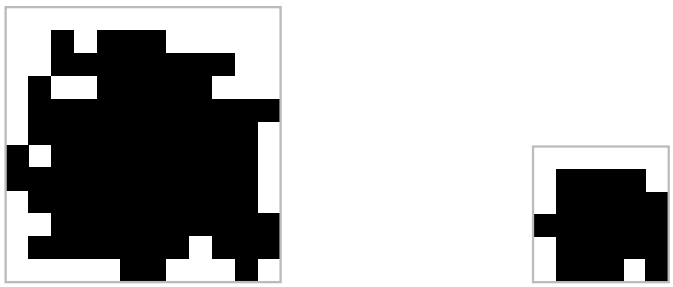
First of all, the features used in the previous research [26] were reviewed. The area in the area/shape spectra was still measured in pixels (because the software did not allow for different size images), which was divided by the total hand area to make the horizontal axis of the area/shape spectra matrix scale-invariant. As the area was measured in pixels everywhere, this was modified in the features that use the area (see table 3.2) to make them scale-invariant. For example, the inertia was divided by the total hand area in pixels squared to make it scale-invariant.

This scale-invariance conversions were done because of the different photo sizes in the latest data set and the differences in photo size with the previous series. Converting the hand area representation such that it would be scale-invariant enabled us to work specifically on the relative area size, independent of image size. The image size varies in the data set by factors like the manner in which the photo's are taken and the spread of the fingers, as the images are cropped to the hand component.

During this process, all features were scrutinised on their scalability and made scale-invariant where possible. Two of the features that were used as shape features in the area/shape spectra proved to be less scale-invariant than their definition suggests. The compactness and jaggedness attribute, both dependent on the perimeter, showed fluctuating behaviour that was not easily removable by changing one of the scale factors. Theoretically, the perimeter scales linearly with the image. The only problem lies in the loss of detail by downscaling. In this case, the scale factor of the perimeter is larger than the scale factor of the image, i.e. the perimeter becomes shorter in comparison. [26, Section 5.4] states: "These attributes are theoretically shape and rotation invariant. In practice interpolation and difficulties in measuring perimeter lengths due to sampling do affect the values when an image is scaled or rotated." More advanced perimeter calculation methods exist [23], but this is beyond the scope

of our research, and would increase our already significant computation time.

Figure 4.1 shows a four-connected shape and its scaled down version. The original shape (a) has a perimeter of 68 pixels, while (b) has a perimeter of 24. This is no factor 2, which it should have as a shape feature, while the area scale factor is roughly the factor 4 that it ought to be. As can be seen, a lot of detail is lost when downscaling a shape. This results in the irregular nature of the perimeter dependent features.



(a) Figure with area 93 and perimeter 68. (b) Figure with area 24 and perimeter 24

Figure 4.1: (a) A shape and (b) its down sampled counterpart of half the size, illustrating the loss of precision and the change in size and perimeter length.

The graphs in figure 4.2 show the irregularity of these attributes. In each graph, a hand image is scaled, as indicated on the horizontal axis. Scaling is done using the Lanczos filter [9]. The vertical axis represents the maximum attribute (compactness or jaggedness) value of the min- and max-trees. So, each image is converted to a max-tree and a min-tree (max-tree of the inverse), for each node the attribute values are calculated, and the graphs show the maximum values of the compactness and jaggedness attribute in the min- and max-tree of the image, at different scales. For reference, the image at scale 1 is about three times as large in graph 4.2(a) than in graph 4.2(b).

As can be seen, the attributes are very irregular when scaled, sometimes jumping towards (almost) zero. Some of this behaviour has to be attributed to the scaling algorithm, although it was chosen with care. However, when scaling an image to a larger scale, sharp edges will soften. Some number overflows have probably occurred here. Ignoring the jumps towards zero, observing only the scales smaller or equal to one, we still see that the graphs have a fluctuating nature, and no real estimation of a function that approximates them can be made. However, the area/shape spectra of the compactness and jaggedness still show similarities when calculated over the same image at different scales. Therefore we will try to calculate the feature vector with and without these area/shape spectra. Afterwards, we will evaluate if the removal of these features influences the classification, and in what way.

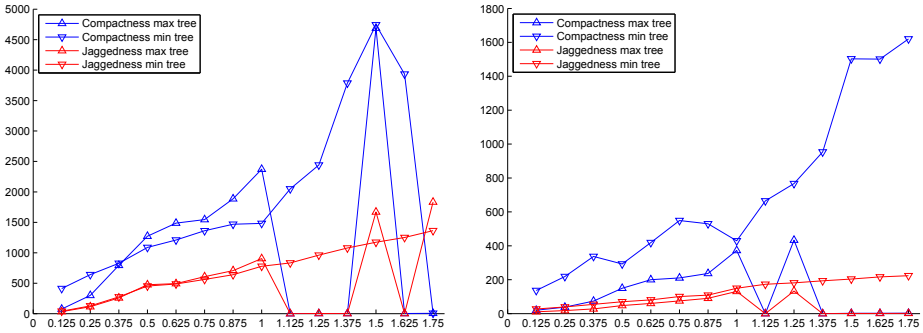


Figure 4.2: The maximum compactness and jaggedness values for the max- and min-tree against the scale of the image, for two example images.

4.3 Determining colour space and feature space

Thus, the feature space that we are assessing is the feature space with and without the perimeter dependent features. As the removal of the perimeter dependent features (in the following sections sometimes abbreviated as pdf) causes a feature space reduction from 2433 to 1655 features, this is a key focus when trying to reduce the feature space.

Several colour spaces were proposed in the previous research (see [26, appendix A]). Eventually only the red colour band of the RGB colour space was used. Because we are interested in how the other colour spaces would perform, we chose the most promising ones, being:

- **RGB** (Red, Green, Blue) is the most obvious choice, because the images we receive are in the RGB format. Using the RGB colour space would mean no colour space transformations are necessary, which would reduce calculation time.
- **HSB** (Hue, Saturation, Brightness) might be a more appropriate colour scheme for our purposes, because increases in saturation are prominently present in hands affected with eczema.
- **RSB** (Red, Saturation, Blue) is a combination of the two above colour spaces. It drops the green component, which does not contain much additional information in skin images, in favour of the saturation component, for the same reasons as above.
- **YCbCr** is a variation of the YIQ colour space, which tries to mimic the human perception of colours. As our results will be resembled to the observations made by the dermatologists, this might be an interesting colour space. We choose YCbCr as alternative for YIQ, because its values lie inside the [0,255] range of the RGB colour space, which means that no additional scaling is required.

- **CIE XYZ** (abbreviated as XYZ) mimics the sensitivity to colour for the cones and rods in the retina, known as the tristimulus model. It mimics the nonlinearities in human colour perception. Like the YCbCr colour space, this could give interesting results.
- **CIE L*a*b*** (abbreviated as L*a*b*) is a modification of the XYZ colour space in which the Euclidean distance between colours is equal to the perceived difference in colours, which is known as the perceptual model.

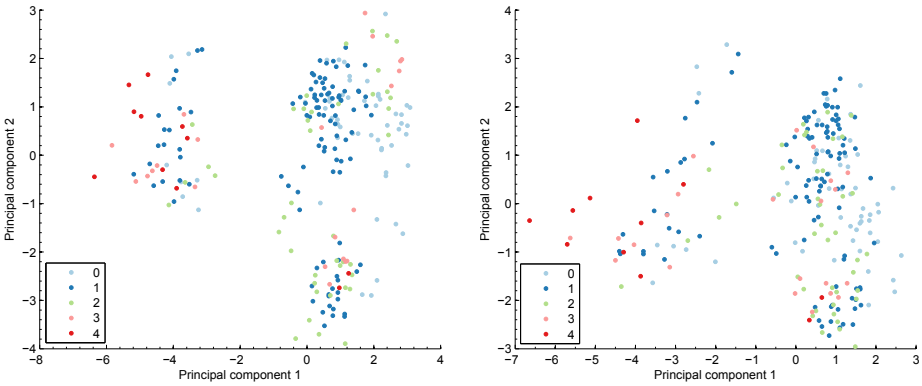
In initial tests, the colour spaces produce similar results, so the images shown as examples are from the L*a*b* colour space, which was arbitrarily chosen.

4.3.1 Principal Component Analysis

When trying to find a clustering of the data, it is wise to start searching for such a clustering using the easiest methods. The Principal Component Analysis is a simple yet intuitive and powerful feature space reduction method, which tends to give fairly good results.

Principal Component Analysis finds the linear combination of features with the largest variance, called a principal component. Each principal component is orthogonal to the others. When the first few principal components are mapped against each other, they show the largest part of the variance present in the data. In our case, the feature space was reduced to only two dimensions, for visualisation purposes.

The results are shown in figure 4.3. As can be seen, there is some clustering present. The PCA of the feature space including the perimeter dependent features clearly shows three clouds of samples, yet all three groups contain samples from different classes, so this clustering has no additional value.



(a) (b)

Figure 4.3: The Principal Component Analysis of the L*a*b* feature space. Shown is the PCA reduction of the feature space (a) including and (b) excluding the perimeter dependent features.

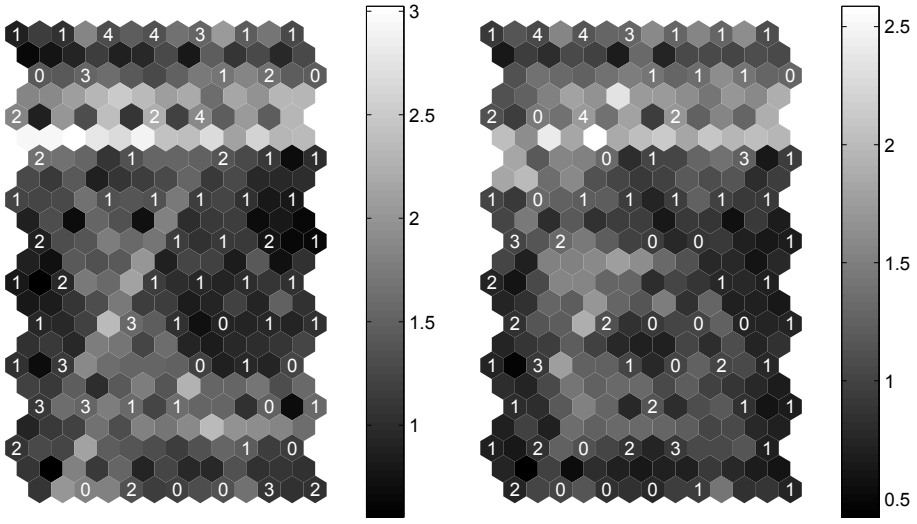
4.3.2 Self-Organising Maps

We first try to cluster the data into the inherent structure in it (without using the label information) using Self-Organising Maps [15].

Self-Organising Maps use unsupervised learning to train a grid of prototypes, which try to represent the data. The labels of the prototypes are added later, for example by a majority vote between the closest samples. During training, this grid adjusts itself to the structure present in the data, such that all data points are close to a prototype in the grid. The dimensionality of this grid defines the mapping the the SOM makes. For example, if we use a $M \times N$ grid, the SOM makes a 2D mapping. If there is any inherent clustering present in the data, the SOM will find it. If there is no clustering visible in the SOM, the data is probably not separable into different classes.

As the SOM has no visualisation method itself, this grid is visualised using a U-Matrix (see [15, section 2.4.2], mentioned as a 'display'). A U-Matrix is a hexagonal or rectangular grid of the SOM prototype grid. It visualises the distances between these prototypes using colour information. The brighter the color in the U-Matrix, the further apart the two neighbouring prototypes are. The colour of the nodes themselves is the mean value of all distance hexagons surrounding the node. The U-Matrix is a 12 by 7 hexagonal grid in our case. This size was automatically determined by the algorithm as the best size. The labels shown in the U-Matrix are the majority vote labels of the samples closest to that prototype in the map.

A Self-Organising Map is created for the data including and excluding the



(a) (b)

Figure 4.4: U-Matrix with majority vote labels for the L*a*b* features (a) including and (b) excluding the perimeter dependent features.

perimeter dependent features, to see which one has better clustering. Because no colour space has been selected, we create these two for all colour spaces. See figure 4.4(a) and (b) for an example result showing the U-Matrix and the labels after voting.

As can be seen from figure 4.4, there is no obvious clustering in the data. The SOM algorithm tries to group the data into two major clusters, as can be seen from the light horizontal 'line', which indicates a more distinct difference between that nodes. However, when we compare this segmentation to the over layered labels, this is not a boundary that really separates prominent clusters of classes. This irregularity in class labels gives a hint that this problem is not really a classification problem at all, and we should rather search in the direction of a regression problem.

When we compare figure 4.4(a) to figure 4.4(b), we see that this hard decision boundary is somewhat softened. When the softened decision boundary is seen as an argument for the regression problem, this could be read as a vote in favour of the feature space without the perimeter dependent features.

Error measure

Because the results are so similar, we take a look at the error measures given by the SOM mapping algorithm. These values are presented in table 4.1. Here, the topographic error is defined as the proportion of data points for which the closest and second-closest weight neurons are not adjacent on the neuron lattice. The quantization error is defined as $E_q(x) = x - m_{c(x)}$ where $c(x)$ indicates the best-matching unit for x, and $m_{c(x)}$ is its location. The quantization error shown is the average quantization error over all samples. Looking at these error

Colour space	**Quantization error**
RSB excl	3.251
HSB excl	3.257
XYZ excl	3.269
RGB excl	3.277
L*a*b* excl	3.696
YCbCr excl	3.757
HSB incl	4.620
RSB incl	4.689
XYZ incl	4.707
RGB incl	4.730
L*a*b* incl	5.082
YCbCr incl	5.118

Colour space	**Topographical error**
L*a*b* incl	0.000
YCbCr incl	0.000
XYZ incl	0.004
XYZ excl	0.004
RSB incl	0.008
RSB excl	0.008
L*a*b* excl	0.013
RGB excl	0.013
YCbCr excl	0.013
HSB incl	0.017
HSB excl	0.017
RGB incl	0.017

Table 4.1: The errors made in clustering by the different colour models. The notation 'incl' and 'excl' denotes if the error belongs to the feature space inclusive or exclusive the perimeter dependent features, respectively.

measures, it is obvious that the colour spaces without the perimeter dependent features perform better on the Quantization error, because less features means less distance. Therefore, the error between the feature spaces with and without the perimeter dependent features is not comparable, but we can see that the results for both cases are fairly similar: in both cases, the RSB and HSB colour space perform best. Further conclusions are not possible, because colour spaces that have a low quantization error have a (relative) high topographical error, and vice versa. Furthermore, these errors are from the U-matrices in figure 4.4, where we argued that thy do not give a good clustering of the data. It would be unwise to select a colour space based on an ill-classified system.

4.3.3 Density plots

When plotting the density of the data points, using the mapping by the two principal components and the labels from the SOM, in figures 4.5(a) and (b), we would like to see the densities of the different classes in different sub parts of the image. However, the densities overlap a lot, so there is no way in which we can separate the area to point out a single class.

We see that the density is much more spread out in the feature space excluding the perimeter dependent features, suggesting that the final regression of the data may become more robust when using the feature space excluding the perimeter dependent features. However, no conclusion can be drawn based on these plots, so we proceed with the next classification and regression methods.

(a)

(b)

Figure 4.5: The density of these same SOMs of figure 4.4 for each class (0–4, horizontally), mapped on the two principal components, (a) including and (b) excluding the perimeter dependent features.

4.4 Classification or regression

To choose between classification or regression, we use trees that can perform both methods [2]. Because no error margin was substantial enough, we perform the classification tree and the regression tree with all colour models, and with and without the perimeter-dependent features. As our sample space is limited, we use leave-one-out cross validation for this.

4.4.1 About classification and regression trees

We use the classification and regression trees from the statistics toolbox in Matlab. See Appendix B for the function calls.

Using a classification tree, the data is labelled according to the decisions made in the tree. At each tree node, the data samples are separated into two groups based on the feature that segments most of the data, using Gini-Simpsons diversity index:

$$I_{GS}(p) = 1 - \sum_{c=1}^{C} p_c^2$$

Where p is the distribution of the samples at the current node, C is the number of classes, and p_c is the probability of class c. For splitting parent distribution p_p into child nodes p_1 and p_2, the Gini improvement measure is than defined as:

$$I_{GS}(p_p) - \left(\frac{\text{size}(p_1)}{\text{size}(p_p)} I_{GS}(p_1) + \frac{\text{size}(p_2)}{\text{size}(p_p)} I_{GS}(p_2) \right)$$

Which defines the improvement as the parent diversity index minus the sum of the child diversity indices times their fraction of the samples. The split is chosen which maximises this improvement measure.

This splitting continues until all sample labels in a tree node are the same, or the number of samples in a tree node is too small for further splitting. In our trees, an impure node, containing samples with multiple labels, is split when the number of samples is greater or equal to 10.

After the tree has been created, leaves of the tree are merged when the sum of their risk values is greater or equal to the risk value of their parent node. For a classification tree, this is the misclassification cost, and for a regression tree, this is the mean squared error for all samples in that leaf node.

In a regression tree, the leaf node labels are defined as the mean value of the labels of the samples that are represented in that node. Therefore, a regression tree is capable of returning continuous values as its prediction score.

4.4.2 Error measures

The result of the classification and regression trees is a severity assessment for each of the samples. To be able to compare these trees, we have to use an error measure to indicate the quality of the severity assessment given by a tree.

Baseline

If we want to interpret these error measures, we have to compare the results to a baseline. The baseline we would like to use is the same error measure between severity assessments made by dermatologists.

Because we have only a single classification for our photographs, we cannot compute the dermatologist's error over that. However, we do have a table of different classifications, provided by the UMCG, in which the severity assessments are given in the same scale that we use, the classifications provided are the results from [7]. This classifications are severity assessments made by dermatologist using the photographic guide on present patients. This means that the dermatologists had the advantage of being able to touch the afflicted hand and view it from all sides. Therefore the error that they make should be less than the error when only photographs are available, as in the computer vision approach.

In the following subsections, we define two error measures that give an indication of the error made by the dermatologists.

Standard error measure

We would like to make large errors weigh more heavily than small errors, a classification that is one class off is less bad than a classification that is further from the truth.

The first and most logical error measure which came to mind is the root mean squared error, also known as the standard error:

$$\begin{aligned} \varepsilon_{rmse}(n) &= \sqrt{\frac{1}{D-1}\sum_{d=1}^{D}(c_d(n)-\mu(n))^2} \\ \text{where} \qquad \mu(n) &= \frac{1}{D}\sum_{d=1}^{D} c_d(n) \end{aligned} \tag{4.1}$$

Here, $\varepsilon_{rmse}(n)$ is the root mean squared error for sample n, D is the number of dermatologists and $c_d(n)$ is the classification of sample n by dermatologist d.

Classification error

However, when we use this error measure for a classification problem, it is as if we are saying: the mean value of all classifications is the optimal classification. But the dermatologists classify on a natural number scale from 0 to 4, they cannot classify a sample as a real number, which this mean will probably be most of the time.

For example, when we apply the mean squared error to a sample which ten out of the eleven dermatologists have classified as 1 (mild), and one as 0 (clear), the mean squared error would first calculate the mean classification, and then say that ten of the eleven dermatologists are slightly wrong, and one of them

is somewhat more wrong, because the true classification is at $\frac{10}{11}$ (or 0.9091). Because it is not possible to give such a classification, the dermatologists will practically always be somewhat wrong, except in cases in which they all agree, or the spread of the classifications is even. One could say that when ten out of the eleven dermatologists agree, they must be right. Thus then only the one dermatologist with an alternative classification would be wrong. This seems much more sensible. Therefore, we propose a new classification error, based on the mean squared error, but with the rounded mean as its mean value:

$$\begin{aligned} \varepsilon_{class}(n) &= \sqrt{\frac{1}{D-1}\sum_{d=1}^{D}(c_d(n) - c_{rm}(n))^2} \\ \text{where} \quad c_{rm}(n) &= \text{round}(\mu(n)) \end{aligned} \tag{4.2}$$

Using the same terminology as in equation 4.1.

In table 4.2, the distribution of the dermatologist classifications per sample is given for 28 samples. Also, two examples are given in which the error measure ε_{rmse} would give a fair mean value in a classification problem.

Computer severity assessment error

As said before, the result of the classification and regression trees is a severity assessment for each of the samples. To be able to resemble the error of these results to the error made by the dermatologists, a similar error measure is used. However, because all samples already have a label, a single classification given by a dermatologist, we do not have to take an average value. Thus, the error made by the computer severity assessment is:

$$\varepsilon_c = \frac{1}{N}\sum_{n=1}^{N}(a(n) - d(n))^2 \tag{4.3}$$

Where $a(n)$ is our severity assessment of sample n (out of N samples), and $d(n)$ is the dermatologist's classification of sample n. Note the difference in terminology: the severity assessment by the dermatologist is a classification, i.e. it is a natural number between 0 and 4, and our severity assessment can also be a real number, yet in the same range. Because the classification made by the dermatologist is used to test both the classification and the regression methods, there is no need to distinguish between them in this error measure. As this error measure is based on the mean squared error, it will be referenced to as the mean squared error occasionally.

Error margins in the computer error measure The error ε_c made by the computer severity assessment does not incorporate the variances in the severity assessments. As can be seen in table 4.2, the error made by the dermatologists is not to be neglected, and has to be incorporated in our calculations.

Classification					μ	ε_{rmse}	c_{rm}	ε_{class}
0	1	2	3	4				
0	0	5	17	1	2.8261	0.4910	3	0.5222
1	20	2	0	0	1.0435	0.3666	1	0.3693
2	11	10	0	0	1.3478	0.6473	1	0.7385
1	9	8	5	0	1.7391	0.8643	2	0.9045
0	0	4	15	4	3.0000	0.6030	3	0.6030
22	1	0	0	0	0.0435	0.2085	0	0.2132
0	8	15	0	0	1.6522	0.4870	2	0.6030
0	7	7	9	0	2.0870	0.8482	2	0.8528
0	1	21	1	0	2.0000	0.3015	2	0.3015
0	0	5	17	1	2.8261	0.4910	3	0.5222
0	0	0	2	20	3.9091	0.2942	4	0.3086
0	11	12	0	0	1.5217	0.5108	2	0.7071
0	6	16	1	0	1.7826	0.5184	2	0.5641
0	0	0	7	16	3.6957	0.4705	4	0.5641
0	20	3	0	0	1.1304	0.3444	1	0.3693
0	0	0	17	6	3.2609	0.4490	3	0.5222
0	0	5	16	2	2.8696	0.5481	3	0.5641
0	0	16	6	1	2.3478	0.5728	2	0.6742
2	21	0	0	0	0.9130	0.2881	1	0.3015
0	0	11	12	0	2.5217	0.5108	3	0.7071
0	2	2	18	1	2.7826	0.6713	3	0.7071
0	1	17	5	0	2.1739	0.4910	2	0.5222
0	13	10	0	0	1.4348	0.5069	1	0.6742
0	0	4	19	0	2.8261	0.3876	3	0.4264
0	2	20	1	0	1.9565	0.3666	2	0.3693
3	10	8	2	0	1.3913	0.8388	1	0.9293
0	0	2	20	1	2.9565	0.3666	3	0.3693
0	1	6	14	2	2.7391	0.6887	3	0.7385
					Mean:	0.5047	*Mean:*	0.5589

0	0	4	15	4	3.0000	0.6030	3	0.6030
0	1	21	1	0	2.0000	0.3015	2	0.3015

Table 4.2: Above: The different dermatologist classifications for 12 dermatologists (eleven gave ratings on two consecutive evenings) and 28 hands. Shown is the mean classification and corresponding error (the root mean squared error ε_{rmse}). Besides that, the rounded mean c_{rm} and the classification error ε_{class} is shown. The classification error is defined as the mean squared error, only with the rounded mean as its mean value. This is a more honest error measure in a classification problem.
Below: Two examples taken from the above table where the mean of the dermatologist's classifications does correspond to a valid classification. Notice that the error measures are identical in these cases.

We could write the error in the severity assessment for a single sample as follows:

$$\begin{aligned} \varepsilon_c(n) &= (m_{ts}(n) + \delta_a(n) - (m_{ts}(n) + \delta_d(n)))^2 \\ &= (\delta_a(n) - \delta_d(n))^2 \\ &= \delta_a^2(n) - 2\delta_a(n)\delta_d(n) + \delta_d^2(n) \\ \text{thus,} \quad \varepsilon_c &= \frac{1}{N}\sum_{n=1}^{N}\delta_a^2(n) + \frac{1}{N}\sum_{n=1}^{N}\delta_d^2(n) \\ &= \delta_a^2 + \delta_d^2 \end{aligned} \tag{4.4}$$

Where $m_{ts}(n)$ is the – unknown – true severity of sample n and $\delta_a(n)$ and $\delta_d(n)$ are the deviations from $m_{ts}(n)$ by $a(n)$ and $d(n)$ (see equation 4.3), respectively. In the second last step, $2\delta_a(n)\delta_d(n)$ can be ignored, because these deviations of the mean will sum to zero over all examples, because they are uncorrelated.

As we know ε_c, which is the result of our calculations, and we have an estimate of δ_d^2 for the classification and regression case: ε_{class} and ε_{rmse}, respectively. These values are shown as the means in table 4.2. They have to be squared to represent δ_d^2. This gives us the following definitions:

$$\varepsilon_{csa} = \sqrt{\varepsilon_c - \varepsilon_{class}^2} \tag{4.5}$$

$$\text{and} \quad \varepsilon_{rsa} = \sqrt{\varepsilon_c - \varepsilon_{rmse}^2} \tag{4.6}$$

Where ε_{csa} defines the error for a classification severity assessment, and ε_{rsa} defines the error for a regression severity assessment. They are representatives of the error $\sqrt{\delta_a^2}$, which is the root mean squared error made by the computer assessment.

Definition of ε_c (see equation 4.4) implies that our proposed assessment performs worse when it is larger than the corresponding dermatologist's error, and better than the dermatologists when it is smaller.

4.4.3 Cross validation on trees

Cross validation method

Preliminary test using k-fold cross validation proved that the results were not reliable. The precision of the outcome was not even guaranteed to one decimal digit, even with a large number of folds and therefore a low number of samples per fold. Stratifying the folds, which means that each fold contains roughly the distribution of the samples, improved this slightly, but not good enough for scientific results. Therefore, and because of the relatively low number of samples, it was decided that the trees would be evaluated using leave-one-out cross validation.

The leave-one-out cross validation method takes one sample out of the sample space. The tree is trained on the rest of the samples with their corresponding

labels. Then the one removed sample is presented to the tree for a severity assessment. This outcome is stored, and then a tree is trained, leaving out the next sample. Continuing in this way, all samples are assessed using a tree trained on all samples but themselves.

Results

The resulting assessments are compared to the classification given to each sample by the dermatologist, as shown in equation 4.3. This is done for all colour bands, including and excluding the perimeter dependent features, and using a classification tree and a regression tree. The results are shown in figure 4.6.

Overall, the regression tree seems to perform better. There are some exceptions, i.e. in the case of the YCbCr and L*a*b* colour spaces. It is not at all clear which feature space performs better, e.g. with the RSB colour space the feature space without the perimeter dependent features performs better, and on other cases, the feature space including the perimeter dependent features performs better.

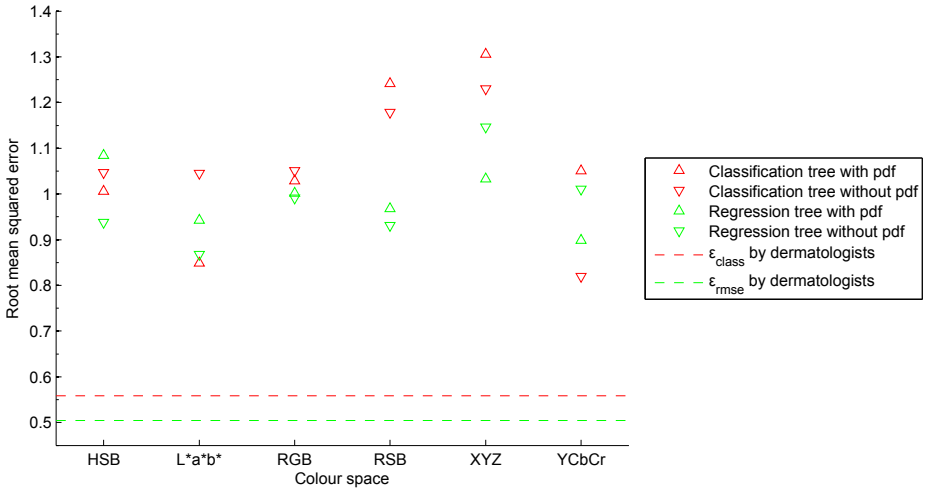


Figure 4.6: The mean square error scores for the classification tree (ε_{csa}) and regression tree (ε_{rsa}), for the six different colour spaces and with and without the perimeter dependent features (pdf).

4.4.4 Tree bagging

A single tree might not give the best possible severity assessment, the given severity assessment is limited to the decisions made in the tree nodes. Besides that, a tree is considered a weak learner. Therefore, we investigated the use of multiple trees for classification, called tree bagging. Bagging (also known as bootstrap aggregation) combines several weak learners into a strong learner.

This is enabled by training each tree on a separate bootstrap of the data. A bootstrap of the data implies that a data set is created with as many samples as the input data set, however some of the entries are duplicated, and thus, some are left out. This effectively averages out the weak points of the trees – training on outliers – over the bags, and makes the classification vote between the trees more robust.

Example results for the reduction of the root mean squared error are shown in figure 4.7. As can be seen, the error rapidly decreases at first, and comes to a slower descent about half-way of the 50 tree bags, especially for the regression trees. We decided to use 30 tree bags as a tradeoff between good severity assessment and computational costs. As the error rapidly decreases with the number of bags, we performed all tests from the previous section (all colour spaces, including and excluding perimeter dependent features, classification and regression tree) again, in the expectation of more robust and thus comparable results. The results of these tests, the mean square errors made by the trees, are plotted in figure 4.8.

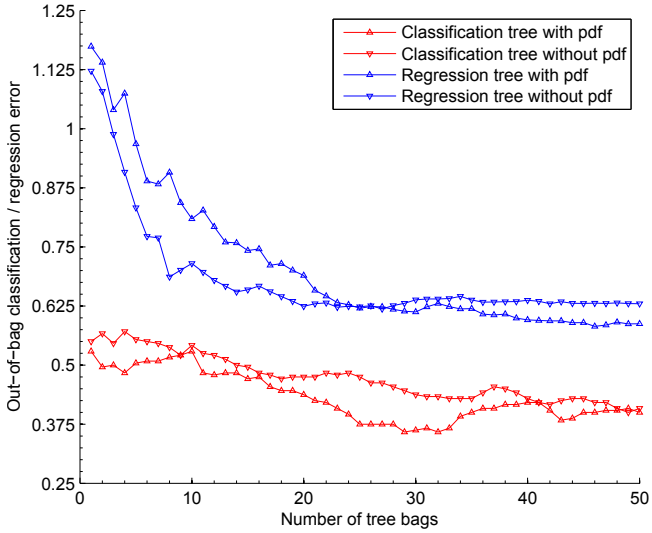


Figure 4.7: The error reduction with increasing number of tree bags, for the classification and regression tree of the L*a*b* colour space, including and excluding the perimeter dependent features (pdf).

4.4.5 Results

Figure 4.8 shows the results made by the bagged trees. The horizontal lines in the graph represent the root mean squared error of the dermatologist's classifications for both the classification (ε_{class}, equation 4.2) and the regression (ε_{rmse}, equation 4.1), whose values are the means from table 4.2.

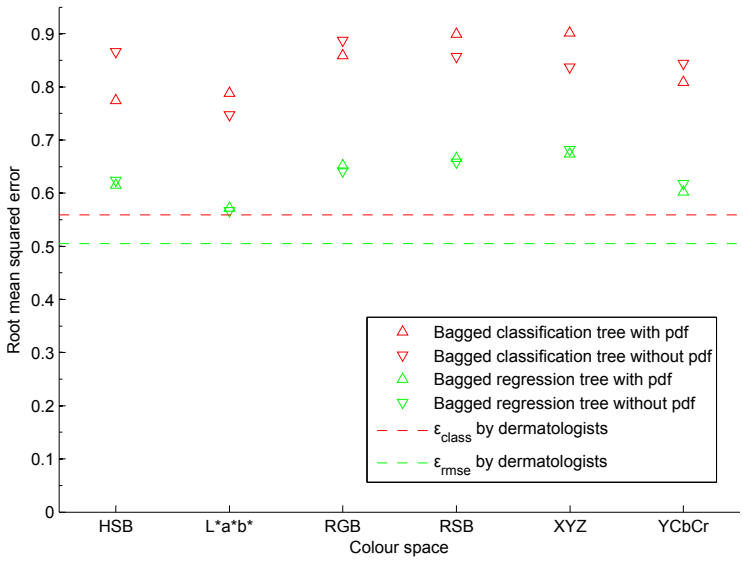


Figure 4.8: The root mean squared error scores of the bagged classification and regression trees, ε_{csa} and ε_{rsa}, respectively. The two horizontal lines show the ε_{rmse} and ε_{class} by the dermatologists as shown in table 4.2.

We see here that the regression trees definitely have an advantage over the classification trees, all regression trees perform better than all of the classification trees. Note that the error values of the classification and regression trees cannot be resembled to each other, because they are based on a different error measure. The regression trees perform better because they are relatively closer to the corresponding baseline set by the dermatologists than the classification trees to their corresponding baseline.

Furthermore, we see that the absence or presence of the perimeter dependent features does make a difference in the classification trees, but hardly makes a difference in the regression trees.

For the classification trees, this difference in performance is fluctuating: for half of the colour spaces, including the perimeter dependent features increases performance, and for the other half it decreases performance. For the regression trees, the difference in performance is marginal. But the decrease in feature space without the perimeter dependent features is a reduction from 2433 features to 1665 features, which is also to our liking.

These results lead us to several conclusions:

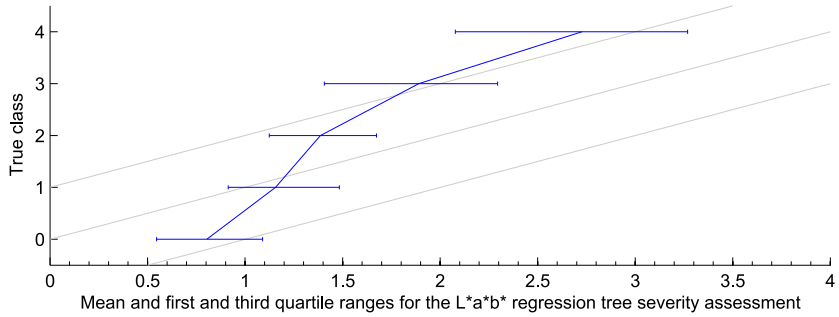
1. The L*a*b* colour space performs the best.
2. The feature space without the perimeter dependent features has our favour.
3. The regression trees clearly perform better than the classification trees.

		True class				
		0	1	2	3	4
Pred. class	0	38	7	4	4	0
	1	18	97	3	1	0
	2	0	3	11	4	0
	3	1	1	3	4	3
	4	0	0	0	1	3

(a)

		True class				
		0	1	2	3	4
Pred. class	0	9	0	0	0	0
	1	47	88	32	5	0
	2	0	20	9	15	5
	3	1	0	1	2	5
	4	0	0	0	0	1

(b)



(c)

Figure 4.9: For the L*a*b* colour space without perimeter dependent features: (a) Confusion matrix of the true class versus the predicted class of the classification tree. (b) Same confusion matrix for the rounded prediction of the regression tree. (c) The mean, first and third quartile ranges of the regression tree prediction. Shown are the optimal diagonal and the 1-class variances from that diagonal.

Classification versus regression

Figure 4.9(a) and 4.9(b) show the confusion matrix of the best classification and regression tree, respectively. This are the confusion matrices of those trees over the L*a*b* colour space without the perimeter dependent features. However, since the regression tree does not (only) return natural numbers, the severity assessment made by the regression tree was rounded to the nearest natural number. This of course does affect the quality of the assessment, but is shown here in comparison to the confusion matrix of the classification tree. Figure 4.9(c) shows a plot of a confusion matrix for the regression tree, with the mean and first and third quartile ranges of the regression values against the natural number dermatologist's classifications.

As can be seen in figure 4.9(c) and the confusion matrix in figure 4.9(b), the diagonal with the mean prediction is somewhat skewed from the optimal diagonal (which is from the top left towards the bottom right in the confusion matrix of figure 4.9(b)). A regression algorithm will try to make a function through the data such that the error is as low as possible. In the case of the classes 1 to 3, this is possible, because examples will exist at both sides of

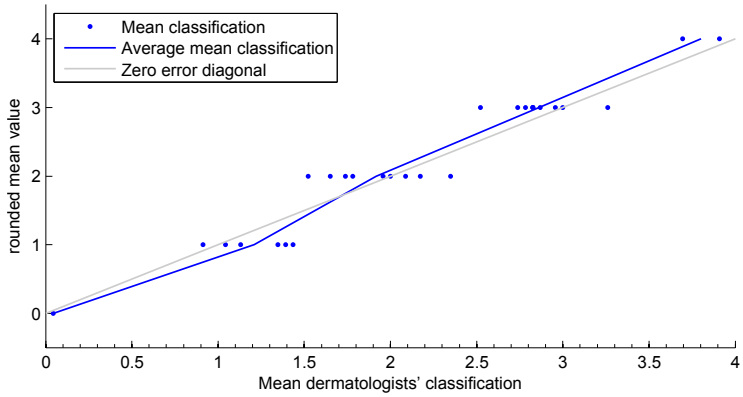


Figure 4.10: The natural error made by the average classification. Shown are the mean dermatologists' ratings on the x axis, with on the y axis their rounded mean classification. The blue graph shows the mean of the ratings.

the classification. However, because class 0 only has examples on or above its classification, the regression algorithm will naturally predict a higher score than zero. For class 4, this is the same, but in the other direction. See figure 4.10 for an example. In our case, one would predict that class 2 (being the centre class) would be predicted the best. However, because class 1 has many more samples, it is predicted better than class 2.

Methods exist that correct this natural behaviour of the regression predictors, for example by forcing the regression function to pass through the origin. This is an issue that we did not further investigate, but is worth more investigation in the future to increase the performance of the regression severity assessment.

Comparison with previous research

When we compare our results to the best result from the previous research [26, Section 6.6.7], we observe the mean squared error values and prediction accuracy in table 4.3. Note, however, that some results are not inherently comparable. The values for the best method in the previous research are hard, if not impossible to resemble to the other values, because, as explained in the introduction of Chapter 3, it is based on a six class system instead of the five class system used in the current research. Furthermore, the correct and within one class prediction of the regression tree are based on a rounding of the predicted value to the nearest natural number.

Interpreting these results, we observe that the best classification tree ensemble proves slightly better than the Bayesian net. The regression tree shows a disappointing 45.42 % accuracy, but has a much better accuracy of 95.42 % within one class.

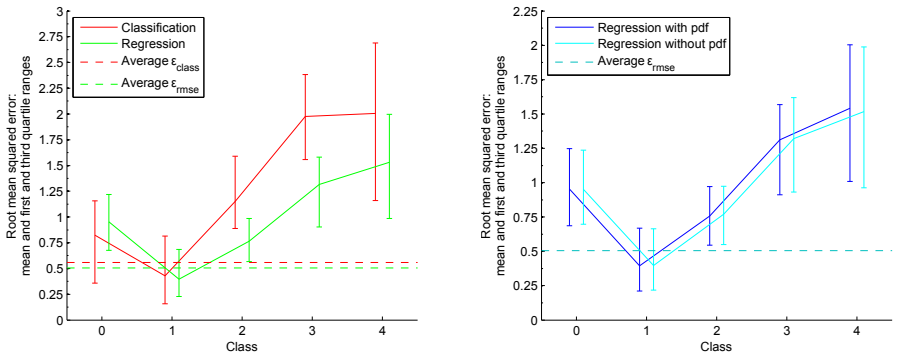
Property	**Best from [26]**	**Best classification**	**Best regression**
Method	Boosted Complement Naive Bayes	Bagged classification tree	Bagged regression tree
Colour band/space	Red	L*a*b*	L*a*b*
Perimeter dependent features	Yes	No	No
Number of features	811	1665	1665
Correct prediction	65.91 %	63.75 %	45.42 %
- within one class	87.97 %	90.00 %	95.42 %
Class. error ε_{csa}	0.6686	0.7473	0.5137

Table 4.3: Comparison between the results of the previous research and the best current results.

Errors per class

This results asked for further investigation into them, because a lot can be said about it. A particular interest of ours was to compare the errors made per class by the classification and regression trees. This gives a good indication as to which severity classes are hard to recognise. Figure 4.11(a) shows the root mean squared error for the classification and regression trees per class, its mean and first and third quartile ranges.

We can see that the error of the severity classes corresponds roughly with the number of images available in our database (see table 3.1(b)): the more images we have of a certain severity class, the better it is assessed by the classification



(a) (b)

Figure 4.11: The average mean squared error over all colour spaces (a) for the classification and regression tree and (b) for the regression trees with and without the perimeter dependent features (pdf).

and regression trees.

Furthermore, when we compare these scores with the mean squared error made by the dermatologists, we see that the trees are capable of classifying severity class 1 with a lower error margin than them. Unfortunately, the numbers of the classifications provided by the dermatologists are so few that they do not allow us to plot them against the severity classes. But the conclusion we can draw is that, when provided with enough examples, the classification and regression trees can outperform the dermatologists intraobserver error: the error that the dermatologists make when classifying the same hand more than once at different times.

As the differences between the regression tree including and excluding the perimeter dependent features are so small, it is interesting to see if this difference becomes clearer when averaged over all colour spaces, but split out for the different severity classes. Figure 4.11(b) shows this graph. However, the changes are very small. To calculate the significance, we use Welch's t-test [27], because both assessment distributions have different variances. Table 4.4 shows Welch's t-test, the degrees of freedom and the confidence interval per class. As can be seen, most of the differences in the distributions are not statistically significant, but the regression tree is significantly better at predicting classes 2 and 3.

Class	0	1	2	3	4
Classification versus regression					
Welch's t	0.6452	0.4293	2.1461	1.7643	0.6949
D.o.f.	108.6205	207.9462	69.9629	37.0376	18.0313
Confidence interval	52.02%	66.81%	3.54%	8.59%	49.60%
Regression including pdf versus excluding pdf					
Welch's t	0.0118	0.0156	0.0950	0.0304	0.0461
D.o.f.	111.8223	213.9027	81.9973	41.9064	19.9996
Confidence interval	99.06%	98.76%	92.45%	97.59%	96.37%

Table 4.4: Welch's t, the degrees of freedom (d.o.f.) and the confidence interval that the two different distributions that are resembled actually are the same for the classification versus regression and regression including the perimeter dependent features (pdf) versus excluding them.

Prominent features

The advantage of using regression and decision trees is that the decisions made in the tree are viewable. This gives us the opportunity to see which of our features is most prominently present in the trees. Thus, which feature is of great importance to the decision making process in the trees. We summed the variable importance over all bagged trees for the L*a*b* colour space, and did so for the classification trees and the regression trees. As no testing was necessary,

the trees were trained on all of the data. The resulting eight features with the largest variable importance are shown in table 4.5.

Feature	**Tree**	**Index**	**Colour**	**Var. imp.**
Bin	-	3	L*	0.0034
Entropy	max	(0,2)	b*	0.0031
Cluster prominence	-	2	L*	0.0027
Inertia / area^2	min	(0,3)	L*	0.0026
Entropy	max	(0,0)	b*	0.0022
Entropy	min	(0,0)	a*	0.0021
Std.dev.	-	1	L*	0.0018
Cluster prominence	-	4	L*	0.0018

Feature	**Tree**	**Index**	**Colour**	**Var. imp.**
Inertia	min	(7,6)	L*	0.0382
Local homogeneity	-	1	a*	0.0291
Entropy	min	(0,4)	a*	0.0246
Lambda max	max	(4,5)	L*	0.0229
Lambda max	min	(0,4)	a*	0.0217
Entropy	min	(0,0)	a*	0.0183
Entropy	max	(0,0)	L*	0.0179
Entropy	max	(0,7)	a*	0.0174

Table 4.5: The most important variables used by the classification trees (above) and regression trees (below). *Colour* indicates the colour band and *Var. imp.* indicates the variable importance (sorted from high to low).

Frequently present in this table are the features based on the entropy, which indicates that this is an important feature to determine the severity of a sample. We also observe that the (0,0) index occurs multiple times for the entropy attribute. We conclude from this that samples with a low entropy and area are important, in more colour bands and in both the Min- and Max-Tree. This is an indication us that the segmentation of the hand images into three different colour bands and concatenating the three different feature vectors might not be the most intuitive approach. Because if the entropy (0,0) index is important for the separate colour bands, how much more important will it not be when the colour bands are processed at once? Future research should therefore investigate the option of using a tree which is capable of handling multiple colour bands at a time, such as a binary partition tree [22].

4.5 Learning Vector Quantization

Trees are essentially weak learners, whose accuracy can be boosted by bagging the trees, as was done in section 4.4. Using the trees, we deduced that the

L*a*b* colour space without the perimeter dependent features performed the best, and with this feature space, we investigate into another classifier: Learning Vector Quantization (LVQ) [17]. Or specifically, we take a look at a recent method called Limited Rank Generalisation Matrix LVQ (Limited Rank GMLVQ) [3]. By reducing the number of output dimensions, they enable easier adaptation of the variables and thus better adaptation to the data and visualisation of the data.

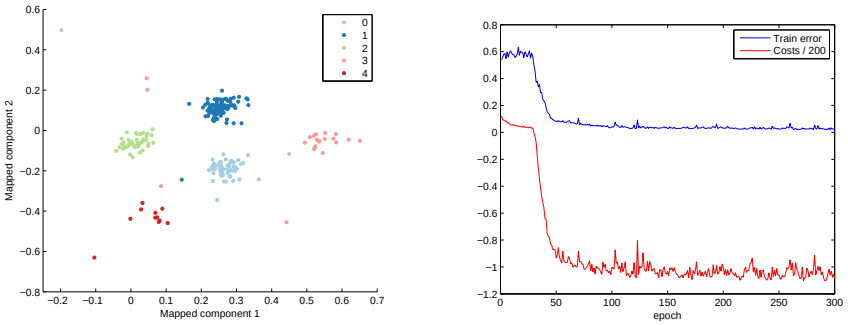
The Limited Rank GMLVQ system trains prototypes based on the training data. When presented with a new sample, it proposes a classification by measuring the distances of the new sample to all of the class prototypes which define a class. The class of the closest prototype is the prediction that is given to the sample. We decided that this was easily adaptable to a regression severity assessment. Therefore, we propose a conversion to a regression severity assessment. First, we take the negative square of all distances to the classes, which transforms these values to a distance measure which is greatest for the closest samples. These distance measures were divided by their sum, such that their distance measure becomes the influence fraction that prototype p has in the severity assessment. These values were multiplied with the class labels the prototypes represent. This is possible because the labels are correlated: the labels left and right of the current label are one step less or more severe, respectively. The sum of these influence values multiplied by the prototype labels gives our regression severity assessment. An example of this calculation is given in table 4.6.

	Prototype p					**Sum**
	0	1	2	3	4	
distance d	0.5393	0.1906	0.2774	0.0468	0.0327	
d^{-2}	3.4385	27.5241	12.9932	456.3026	937.1802	
$d^{-2}/\sum d^{-2}$	0.0024	0.0191	0.0090	0.3179	0.6515	1.0000
$(d^{-2}/\sum d^{-2})p$	0.0000	0.0191	0.0181	0.9538	2.6060	3.5970

Table 4.6: The calculation steps for determining our regression score after application of the Limited Rank GMLVQ algorithm. Shown is an example which regression score is influenced for 65% by class 4 and for 32% by class 3, with some minor influence by classes 0 to 2 (see rows three and four).

Preliminary tests using the Limited Rank GMLVQ proved to segment the data quite well on visual inspection, apart from an occasional outlier. Remarkably, the order of the classes in the image is not on a line or simple curve, as might be expected. Figure 4.12(a) shows a preliminary test based on 239 out of the 240 samples, using 300 epochs to train the mapping. Figure 4.12(b) shows the training error and the cost function against the current epoch. The training error drops in the first quarter of the training, and not much improvement is seen after that. Therefore, we reduce the number of epochs to 100 for further tests.

Because of time restrictions, we performed 40-fold cross validation on the



(a) (b)

Figure 4.12: Using the Limited Rank GMLVQ algorithm over 300 epochs results in (a) a mapping of the data to 2D and (b) error measures for each of the learning epochs.

Limited Rank GMLVQ algorithm, which indicates 6 samples per fold. The resulting errors ε_{csa} (see equation 4.5) and ε_{rsa} (see equation 4.6) are shown in table 4.7. The regression RMSE score is based on the regression prediction explained above.

Run	**1**	**2**	**3**
Folds	40	10	10
- Samples/fold	6	24	24
Epochs	100	150	150
Dimensions	2	2	4
Classification RMSE score ε_{csa}	0.9220	0.8898	0.7720
Regression RMSE score ε_{rsa}	0.8405	0.7620	0.6531

Table 4.7: The results of running the Limited Rank GMLVQ algorithm.

The results of the first run were quite disappointing, when resembled to the results from the bagged trees. However, the data is transformed to just two dimensions in the first run, which is a feature space reduction of three orders of magnitude. With this in mind, these first results are remarkably good. To improve the results, fearing that our epoch reduction was too radical, we increased the number of epochs to 150 and lowered the number of folds to 10 (24 samples per fold). This made the classification RMSE score drop slightly, but the regression RMSE score was lowered considerably. This proves that there is some improvement to gain. However, the classification RMSE score is worse than most of the bagged classification trees, and the regression RMSE score is far worse than all of the bagged regression trees.

As a final test, the number of used dimensions to represent the data was doubled from two to four. The original two dimensions are great for visualisation, but [3, Section 3.2] shows that the performance rises quickly when the dimensionality is enlarged. So, this is our final test to see if the algorithm can prove to perform even better. The results are shown in table 4.7. As additional information, the confusion matrices of the classifications made by the trained prototypes for each run are shown in table 4.8.

As shown in figure 4.13, this RMSE scores come even closer to the optimal scores by the L*a*b* classification and regression trees. This proves that the data segmentation provided by the Limited Rank GMLVQ method is worth further investigation, but this is beyond the scope of our investigation.

(a) Run 1

		True class				
		0	1	2	3	4
Pred. class	0	30	15	5	3	1
	1	17	67	23	9	2
	2	9	18	13	4	0
	3	1	7	1	6	2
	4	0	1	0	0	6

(b) Run 2

		True class				
		0	1	2	3	4
Pred. class	0	33	18	9	5	2
	1	17	67	17	3	0
	2	6	18	13	2	1
	3	1	5	2	11	4
	4	0	0	1	1	4

(c) Run 3

		True class				
		0	1	2	3	4
Pred. class	0	34	22	7	2	0
	1	19	66	19	6	2
	2	3	17	14	5	0
	3	0	3	2	6	3
	4	1	0	0	3	6

Table 4.8: Confusion matrices of the true classification against the classification predicted by the Limited Rank GMLVQ algorithm using k-fold cross validation. See table 4.7 for the information about each run.

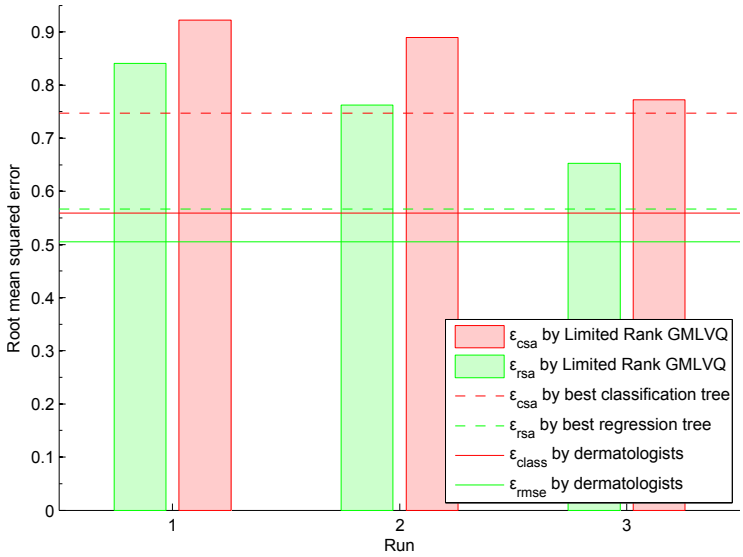


Figure 4.13: The error made by the Limited Rank GMLVQ prediction, showing the best classification and regression tree, the bagged tree over the L*a*b* colour space, and the classification and regression error made by the dermatologists (ε_{class} and ε_{rmse}, respectively).

Chapter 5

Conclusion

This master theses focuses on the recognition of hand eczema from photographs. It is a continuation of the work by Van de Wal [26], entitled "Automatic Classification of Hand Dermatitis". As this research tried, amongst others, to convert the classification problem to a regression problem, it is entitled "Automatic Severity Assessment of Hand Dermatitis", to emphasise the difference in focus with the previous research.

5.1 Preprocessing

The preprocessing was slightly adjusted but mostly reused. As the previous work focused on the red colour band, we retried the experiments with a broader range of colour spaces. Extra images were acquired, especially images of 'clean' hands, that is, hands without eczema. This broads our sample space, and diminishes the influence of the relatively high number of samples in class 1 (see table 3.1(b)). The preprocessing was adapted, improving the segmentation of the photographs into a hand and background component. For further information, see Section 3.1.1. As features, the histogram, the Haralick features over the Gray Level Co-occurrence Matrices and the area/shape spectra of several shape features for the min- and max-tree were calculated. This resulted in a total feature vector of 811 features, explained in detail in Section 3.1.2.

5.2 Experiments and results

The features were reviewed, and it was discovered that the perimeter dependent features of the area/shape spectra features were not scalable, nor easily adapted to be scalable. It was decided to try and perform the experiments with and without these features. Removing the perimeter dependent features would mean a feature space reduction from 811 to 555 features.

The tests were run on the complete colour space instead of just a single colour band. This was done by concatenating the feature vectors of the different

colour bands. This increased the number of features to 2433 (or 1655 without the perimeter dependent features). Several colour spaces were selected, being the CIE L*a*b*, CIE XYZ, CIE YCbCr, HSB, RGB and RSB colour spaces. See Section 4.3 for the argumentation of this choice.

Preliminary tests with Principal Component Analysis proved that the data was not easily separable into the five different classes (Section 4.3.1). Further tests using Self-Organising Maps, to see if there was any inherent structure in the data, proved futile (Section 4.3.2). Plotting the densities of these maps showed no visible structure either (Section 4.3.3).

23 dermatologist's ratings for 28 hands were obtained, from a clinical experiment. From these ratings, we calculated the root mean squared error ε_{rmse} as baseline to compare against regression errors, and defined the classification error ε_{class} as the root mean squared error with the rounded mean as a baseline for classification errors. We defined the error made by the computer for both the classification severity assessment (ε_{csa}) and the regression severity assessment (ε_{rsa}). See Section 4.4.2 for further details.

Severity assessment using leave-one-out cross validation using trees proved mixed results between the different feature spaces, including and excluding the perimeter dependent features, and between the different colour spaces (see Section 4.4.3). The regression trees seemed to perform better, but the classification tree was better at some points (see figure 4.6). The feature spaces with and without the perimeter dependent features varied in prediction error, so no conclusion could be drawn.

As a single tree is a weak learner, we used bootstrap aggregation (bagging) to produce a more robust outcome. These results are shown in figure 4.8. Bagged regression trees are clearly better than bagged classification trees, with the bagged trees over the L*a*b* colour space as best. The presence of the perimeter dependent features gave mixed results, statistically insignificant between the bagged regression trees. We therefore discarded them, as this reduces our feature space significantly. For details, see Section 4.4.5.

When viewing the error measures for the different classes, we observed that class 1 was actually better recognised by the bagged trees than by the dermatologists (see Section 4.4.5). This is also the class with the most samples, therefore it does not surprise us that it is well recognised, but it also encourages us to invest in finding additional photographs for the other severity classes.

Key features for decision making in the bagged classification and regression trees are primarily the entropy area/shape spectra. Results direct us to use all colour bands when creating the trees. See also Section 4.4.5.

Another approach which was tested is Learning Vector Quantization, as this is a promising field of interest for classification problems. We modified the distances given by the Limited Rank GMLVQ system to produce a regression score (see Section 4.5). During experiments, Limited Rank GMLVQ proved very good and finally came very close to the best performances by the tree structures. The main results are listed in table 5.1.

Error by dermatologists	**Classification** ε_{class}: 0.5589		**Regression** ε_{rmse}: 0.5047	
Error by the computer	**Classification** ε_{csa}	*colour(s)*	**Regression** ε_{rsa}	*colour(s)*
Comp. Naive Bayes	0.6686	Red	-	-
Best tree	0.8191	YCbCr	0.8678	L*a*b*
Best bagged tree	0.7473	L*a*b*	0.5671	L*a*b*
Best Limited Rank GMLVQ result	0.7720	L*a*b*	0.6531	L*a*b*

Table 5.1: The results of this thesis, summed in one table. All best computer assessment results use the feature space without the perimeter dependent features except the Complement Naive Bayes, which is the best result from [26].

5.3 Future work

We were fortunate enough to acquire different classifications by a group of dermatologists over the same set of hands, which gives us an indication of the error that exists between the dermatologists, and gives us a good baseline to resemble our results to. Unfortunately, the results of the best method, the bagged regression tree for the L*a*b* colour space excluding the perimeter dependent features, proved overall slightly worse than the error made by the dermatologists. However, the classifications made by the dermatologists came from a clinical trial, where the dermatologists were capable of touching the hand and viewing it from all sides. It is therefore remarkably that our best severity assessment error comes very close to the error made by the dermatologists. This results call for additional research into this field, to improve the computer severity assessment even further. Especially the LVQ method deserves further research, by increasing the dimensionality of Limited Rank GMLVQ, or by switching to true regression LVQ [12].

After all, when the performance of our automatic classification would surpass that of the dermatologist classification, we would have created an objective, repeatable measure, which would become a great help in reliable severity assessments and as a baseline for pharmaceutical experiments. However, to truly improve beyond the error by the dermatologists, we would have to stop viewing the dermatologist's assessments as crisp labels, and move into the terrain of regression analysis.

Experiments showed that class 1 is assessed better by the bagged trees than by the dermatologists. When class 1 is recognised well because of its number of samples, other classes are surely better classified when the number of samples increases. However, it is relatively easy to acquire additional photographs for class 0, because one can round up some people without hand dermatitis, but it is much harder to acquire additional photographs for the other severity classes. We propose that, in cooperation with the UMCG, patients with eczema might be asked to let their hands be photographed more often. This is a small additional

burden for them, but gives us the extra data that is needed to improve the automatic severity assessment of hand dermatitis.

Another issue worth some further research is that the feature vectors are now a concatenation of the feature vectors of the different colour bands. Therefore practically all information that comes from the union of these colour bands is lost. We propose that future work includes research into binary partition trees [22], which are capable of creating a tree which consists of all three colour bands.

Now that the feature space is known and scale invariant, additional research could explore if the feature vectors are scale invariant enough to let people at home take pictures of their own hands. One could think about pictures taken with a cell phone camera. These cameras become better and better, and might prove good enough to give reliable severity assessments on. Experiments in this direction would have to deal with scale invariance tests – how good is the severity assessment for a given size photograph – and issues like lighting conditions and blurry images.

In conclusion, the first steps are taken towards our ultimate goals. First, to reduce the burden for patients with hand eczema to travel to the hospital for frequent check-ups and enable hand eczema severity assessment from your home. And second, to create a severity measure that can be used by dermatologists and pharmaceutical research anywhere as a reliable and objective severity assessment measure.

Bibliography

[1] R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[2] L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, California, 1984.

[3] K. Bunte, P. Schneider, B. Hammer, F.-M. Schleif, T. Villmann, and M. Biehl. Discriminative visualization by limited rank matrix learning. Technical Report MLR-03-2008, Leipzig University, 2008.

[4] C. Charman, C. Chambers, and H. Williams. Measuring atopic dermatitis severity in randomized controlled clinical trials: What exactly are we measuring? In *Journal of Investigative Dermatology; 120*, pages 932–941, 2003.

[5] C.R. Charman and J.S. English. Getting to grips with hand eczema: measuring skin disease severity objectively. *British Journal of Dermatology*, 152(2):296–301, February 2005.

[6] P.J. Coenraads, T. Ruzicka, B. Dreno, and J. Maares. Combined written and photographic guide for assessing severity of chronic hand eczema.

[7] P.J. Coenraads, H. van der Walle, K. Thestrup-Pedersen, T. Ruzicka, B. Dreno, C. de La Loge, M. Viala, S. Querner, T. Brown, and M. Zultak. Construction and validation of a photographic guide for assessing severity of chronic hand dermatitis. *British Journal of Dermatology*, 152(2):199–201, February 2005.

[8] T.L. Diepgen, K.E. Andersen, F.M. Brandon, M. Bruze, D.P. Bruynzeel, P. Frosch, M. Gonçalo, A. Goossens, C.J. le Coz, T. Rustemeyer, I.R. White, and T. Agner. Hand eczema classification: a cross-sectional, multicentre study of the aetiology and morphology of hand eczema. *British Journal of Dermatology*, 160(2):353–358, October 2008.

[9] C.E. Duchon. Lanczos filtering in one and two dimensions. In *Journal of Applied Meteorology*, volume 18, pages 1016–1022, May 1979.

[10] M.H.F. Wilkinson F. Tushabe. Content-based image retrieval using combined 2d attribute pattern spectra. In *Lecture Notes in Computer Science*, pages 554–561, 2008.

[11] H. Ganster, A. Pinz, R. Röhrer, E. Wildling, M. Binder, and H. Kittler. Automated melanoma recognition. *IEEE transactions on medical imaging*, 20(3):233–239, March 2001.

[12] M. Grbovic and S. Vucetic. Regression learning vector quantization. In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, pages 788–793. IEEE Computer Society, 2009.

[13] R.M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6):610–621, November 1973.

[14] E. Held, R. Skoet, J.D. Johansen, and T. Agner. The hand eczema severity index (hecsi): a scoring system for clinical assessment of hand eczema. a study of inter- and intraobserver reliability. *British Journal of Dermatology*, 152(2):302–307, February 2005.

[15] S. Kaski. Data exploration using self-organising maps. Master's thesis, Acta polytechnica Scandinavica, 1997.

[16] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, January 1982.

[17] T. Kohonen. Improved versions of learning vector quantization. In *International Joint Conference on Neural Networks*, volume 1, pages 545–550. IEEE Computer Computer Society Press, 1990.

[18] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley Symposium on Mathematics, Statistics and Probability*, volume 1, pages 281–297, 1967.

[19] P. Maragos. Pattern spectrum and multiscale shape representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):701–716, 1989.

[20] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, 2(11):559–572, November 1901.

[21] T. Ruzicka, C.W. Linde, G.B.E. Jemec, T. Diepgen, J. Berth-Jones, P.J. Coenraads, A. Kaszuba, R. Bissonnette, E. Varjonen, P. Holló, F. Cambazard, M. Lahfa, P. Elsner, F. Nyberg, A. Svensson, T.C. Brown, M. Harsch, and J. Maares. Efficacy and safety of oral alitretinoin (9-cis retinoic acid) in patients with severe chronic hand eczema refractory to topical corticosteroids: results of a randomized, double-blind, placebo-controlled, multicentre trial. *British Journal of Dermatology*, 158(4):808–817, 2008.

[22] P. Salembier, A. Oliveras, and L. Garrido. Anti-extensive connected operators for image and sequence processing. *IEEE Transactions on Image Processing*, 7(4):555–570, 1998.

[23] N. Sladoje, I. Nyström, and P. K. Saha. Measurements of digitized objects with fuzzy borders in 2d and 3d. *Image and Vision Computing*, 23:123–132, 2005.

[24] J.S. Taur. Neuro-fuzzy approach to the segmentation of psoriasis images. *Journal of VLSI Signal Processing 35*, pages 19–27, 2003.

[25] E.R. Urbach, J.B.T.M. Roerdink, and M.H.F. Wilkinson. Connected shape-size pattern spectra for rotation and scale-invariant classification of gray-scale images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):272–285, February 2007.

[26] B. van de Wal. Automatic classification of hand dermatitis. Master's thesis, University of Groningen, March 2007.

[27] B.L. Welch. The generalization of "student's" problem when several different population variances are involved. *Biometrika*, 44 (12):28–35, 1947.

Appendix A

Test results

This chapter is a summary of the results in this paper, showing the numbers behind the plots and referring to the numbers when presented in the thesis where applicable.

A.1 Perimeter dependent features

During testing, a test was run to observe if the features which depend on the perimeter (which is measured in pixels), the perimeter dependent features compactness and jaggedness, were scale-invariant enough. The results were shown in figure 4.2. The data is presented in table A.1.

A.2 Self-Organising Maps

Self-Organising Maps were used to observe if the data would cluster itself without knowledge of the labels. For each colour space and with and without the perimeter dependent features, a SOM was created. The resulting quantization error and topographic error are explained in Section 4.3.2 and shown in table 4.1.

A.3 Root Mean Squared Error measures

To interpret the misclassification made by the dermatologists, a separate error measure was proposed for classification and regression, in section 4.4.2. The calculation of these error measures is shown in table 4.2, including the data provided by the dermatologists, showing the spread in their classifications.

Scale	Image 1				Image 2			
	Compactness		Jaggedness		Compactness		Jaggedness	
Tree:	*min*	*max*	*min*	*max*	*min*	*max*	*min*	*max*
0.125	412.04	75.95	39.91	32.33	136.31	19.78	27.24	10.96
0.25	640.87	300.20	127.23	113.27	218.47	38.40	39.49	17.61
0.375	827.22	793.84	274.61	257.87	337.08	73.74	54.78	27.83
0.5	1087.83	1271.47	453.08	474.14	293.05	149.71	70.20	47.52
0.625	1209.25	1486.30	486.31	493.85	418.93	199.81	80.70	60.34
0.75	1362.38	1544.95	562.20	607.68	549.21	210.44	100.44	74.98
0.875	1468.00	1885.82	639.86	706.34	529.86	236.84	108.79	90.35
1	1481.84	2372.23	777.92	904.30	429.33	372.09	149.36	130.71
1.125	2049.77	*	832.54	*	664.81	*	173.16	*
1.25	2441.89	2.67	961.02	1.90	766.67	433.06	181.56	132.91
1.375	3786.87	1.99	1078.60	0.89	953.74	*	193.21	*
1.5	4744.24	4689.70	1173.29	1669.86	1502.96	1.99	203.85	0.89
1.625	3934.87	2.29	1250.08	1.33	1501.28	1.99	216.25	0.89
1.75	6300.18	5234.07	1363.28	1832.50	1620.19	2.29	223.74	1.48

Table A.1: Maximum values for the compactness and jaggedness of the min- and max-tree for two different images. See section 4.2 for more information. The asterisk denotes a number overflow (value $-1.79 \cdot 10^{308}$). Note that image 1 is about three times larger than image 2.

Colour space	Classification		Regression	
	pdf	*npdf*	*pdf*	*npdf*
Regular trees:				
HSB	1.0063	1.0469	1.0843	0.9375
L*a*b*	0.8491	1.0449	0.9423	0.8678
RGB	1.0288	1.0509	1.0018	0.9908
RSB	1.2417	1.1780	0.9680	0.9311
XYZ	1.3055	1.2299	1.0330	1.1460
YCbCr	1.0508	0.8191	0.8985	1.0102
Tree bagging with 30 bags:				
HSB	0.7747	0.8661	0.6155	0.6238
L*a*b*	0.7880	0.7473	0.5723	0.5671
RGB	0.8589	0.8875	0.6519	0.6417
RSB	0.8991	0.8564	0.6659	0.6582
XYZ	0.9015	0.8367	0.6742	0.6817
YCbCr	0.8089	0.8442	0.6026	0.6176

Table A.2: Table showing the root mean squared error of the proposed assessment by the different bagged trees. The abbreviation (n)pdf indicates (no) perimeter dependent features. All assessment values were obtained using leave-one-out cross validation.

A.4 Trees and bagged tree ensembles

Separate error measures for the computer error measure are defined in Section 4.4.2, which correct the error made by the computer using the error that the dermatologists have. See in particular equation 4.5 and 4.6.

A proposed severity assessment was made by the trees. First, this was tried using only a single tree. However, as a tree is essentially a weak learner, and the results proved inconclusive, we applied bootstrap aggregation to the trees, creating a tree ensemble that is a strong learner. The results of applying these trees and tree ensembles to the data of the different colour spaces, including and excluding the perimeter dependent features, is shown in figure 4.6 and 4.8. The data is presented in table A.2, which shows ε_{csa} for the classification trees and ε_{rsa} for the regression trees.

A.5 Average RMSE values

After noticing that the tree assessments were essentially worse than the error made by the dermatologists, we were interested in how this error was distributed over the different classes. The results are shown in figure 4.11. The data is shown in table A.3.

Class	**Classification**			**Regression**		
	Q1	**Mean**	**Q3**	**Q1**	**Mean**	**Q3**
0	0.3592	0.8245	1.1561	0.6761	0.9528	1.2184
1	0.1571	0.4294	0.8146	0.2276	0.3956	0.6860
2	0.8873	1.1521	1.5912	0.5660	0.7644	0.9862
3	1.5568	1.9752	2.3822	0.9020	1.3158	1.5799
4	1.1607	2.0057	2.6895	0.9860	1.5298	1.9956

Class	**Regression pdf**			**Regression npdf**		
	Q1	**Mean**	**Q3**	**Q1**	**Mean**	**Q3**
0	0.9538	0.6874	1.2475	0.9517	0.6964	1.2360
1	0.3951	0.2103	0.6684	0.3962	0.2173	0.6650
2	0.7578	0.5439	0.9724	0.7710	0.5496	0.9738
3	1.3113	0.9112	1.5685	1.3203	0.9314	1.6184
4	1.5427	1.0095	2.0033	1.5169	0.9620	1.9879

Table A.3: The mean and first and third inter quartile ranges for the Root Mean Squared Error values of the classification and regression tree (above) and of the regression tree with (no) perimeter dependent features ((n)pdf) (below). The first and third quartile ranges are the means of the samples below and above the mean, respectively.

A.6 Limited Rank GMLVQ results

The Limited Rank GMLVQ approach was tried because of the potential of Learning Vector Quantization methods, and it is currently one of the more recent improvements in this field. Table 4.7 shows the performed runs and their classification and regression RMSE. The confusion matrices of the runs are shown in table 4.8.

Appendix B

Matlab code

After the creation of the feature vectors, the rest of the computations and evaluations were performed in Matlab. We tried using several additional packages, but ultimately favoured the given Matlab packages above additional packages who were sometimes hard to handle and non-intuitive.
The used cross validation, tree building and testing methods were taken from the Matlab Statistics Toolbox™.

Leave-one-out cross validation was performed using the function `crossval`, which returns the mean squared error between the predicted class / value and the given label:

```
mse = crossval('mse',data,labels,'Predfun',fun,'leaveout',1);
```

As can be seen from the above function call, we have to indicate to `crossval` that we would like leave-one-out cross validation, as it is also capable of performing k-fold cross validation, for example. We also have to provide a function handle `fun`, which contains the training and testing method. We used several methods, which are presented in the following subsections.

Classification tree

```
fun = @(X_tr, C_tr, X_te) ( ...
    str2double( eval( ...
        classregtree( X_tr, C_tr, 'method', 'classification'), ...
        X_te ...
    ) ) ...
);
```

This function creates a classification tree (see line 3) using the training data `X_tr` and corresponding training labels `C_tr` and tests the created tree with the function `eval(tree, testdata)`, which returns the predicted labels for the samples in `X_te` as the function result. The results have to be converted from characters to doubles (using `str2double()`) to be able to compute the mean squared error. The code `...` indicates that the code continues on the next line.

Regression tree

```
fun = @(X_tr, C_tr, X_te) ( ...
    eval( ...
        classregtree(X_tr, C_tr, 'method', 'regression'), ...
        X_te ...
    ) ...
);
```

This function is similar to the one presented in the precious subsection, but creates a regression tree. Note that the function `str2double()` is not necessary in this case.

Bagged classification tree

```
fun = @(X_tr, C_tr, X_te) (
    str2double( predict( ...
        TreeBagger(30, X_tr, C_tr), ...
        X_te ...
    ) ) ...
);
```

This function creates a classification tree ensemble, which is used to classify the given samples. This function works slightly different from the single classification tree in the sense that the evaluation is done using a function `predict( treeEnsemble, testdata)` instead of `eval`. We also see the number 30 in line 3, referring to the fact that we used thirty tree bags in our experiments. Note finally that the function `str2double` again is necessary in this case.

Bagged regression tree

```
fun = @(X_tr, C_tr, X_te) ( ...
    predict( ...
        TreeBagger(30, X_tr, C_tr, 'Method', 'regression'),
        X_te ...
    ) ...
);
```

Creates a regression tree ensemble similar to the classification tree ensemble in the previous section. The difference with the previous function is the additional parameters specifying that this should be a `'regression'` tree ensemble, and the loss of the function `str2double`.

Naive Bayes classifier

```
fun = @(X_tr, C_tr, X_te) ( ...
    predict( ...
        NaiveBayes.fit(X_tr, C_tr, 'Distribution', 'kernel'),
        X_te ...
    ) ...
);
```

This function creates a Naive Bayes classifier which uses a so-called kernel distribution instead of the standard Gaussian distribution, which requires all data to have a bell-shaped distribution. The kernel distribution determines a function for the distribution of the data. The rest of the function handle is quite similar to the previous functions.