Analysis of Mouse Joints

Examination of Osteoarthritis by Automatic Visual Inspection

Master's thesis

by

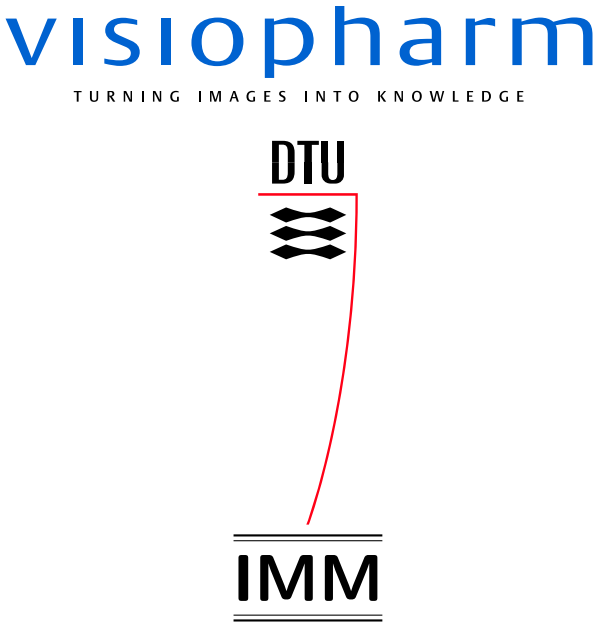
Jesper Skjerning

Supervised by

Bjarne Kjær Ersbøll

and

Michael Grunkin



Lyngby 2003

IMM-Thesis-2003-58

Preface

This Master's thesis is carried out by student Jesper Skjerning (s952662) at Dept. of Informatics and Mathematical Modelling (IMM) at the Technical University of Denmark (DTU), Lyngby, Denmark. The work is realized from April to October 2003 at Visiopharm Aps, Hørsholm, Denmark. The used data for this project is recorded and provided by Aventis Pharma GmbH (Germany).

Thesis supervisor is Associate Professor Bjarne Kjær Ersbøll, IMM, DTU.
Assistant supervisor is Ph.D Michael Grunkin, Managing Director of Visiopharm Aps.

The reader would most probably benefit from having a basic knowledge of mathematics, statistics and digital image analysis.

The software used for the project is:
Microsoft, Visual C++, for programming and imaging analysis.
Visiopharm, Imaging Utilities, for imaging analysis.
Insightful, S-Plus, for statistics and plots.
SAS Institute, SAS, for statistics and plots.
MathWorks, Matlab, for plots.
Winedt and Miktex for writing and compiling the report in Latex.
Adobe, Photoshop, Jasc Software, Paint Shop Pro, Microsoft, Imaging, ACD Systems, ACDSEE, etc. for image handling, visualization and manual processing.

The developed programs, functions and scripts can be found on the CD in the back of the report and a list of them can be found in Appendix H - J, for MS C++, S-Plus and SAS, respectively.

Keywords
Image analysis, Classification, Clustering, Clustering survey, Color transformation, Osteoarthritis.

Lyngby, 17 October 2003

Jesper Skjerning

Acknowledgements

I would like to express my gratitude to my supervisors during this project.
Thanks to Associate Professor Bjarne Kjær Ersbøll at IMM, DTU, for patiently answering good as well as less good questions and for providing and discussing ideas.
Thanks to Ph.D Michael Grunkin, Managing Director of Visiopharm Aps, for letting me work at Visiopharm with a genuine project, for correcting some of my weaknesses and also for providing and discussing ideas.

Thanks goes to Aventis Pharma without whose data and permission to use it, this project would not be.

Thanks to the nice colleagues at Visiopharm, especially Steen Tofthøj Rasmussen and Lars Pedersen for daily support and discussions.

I would like to thank Jeanette Gylling, Pia Stenberg and Toke Koldborg Jensen for revising the report.

Finally, I would like to express my gratitude to my fiancee Malene Lyngesen who has patiently supported me and made sure I did not starve during these six months!

Abstract

This project examines the use of automatic image processing to test for osteoarthritis (OA) in laboratory mice. By using mice that are predisposed for developing osteoarthritis, there should be a high correlation between their age and their osteoarthritis stage. The purpose of this project is to generate an automatic OA measure of the osteoarthritis stage of each mouse.
Visiopharm has earlier worked on the project and would like to find out if the results can be improved by other approaches than the ones they have chosen. They mainly calculate features from histograms of RGB and other color representations. Their obtained OA measure has a highly significant correlation with age at 0.54.

The data for this project consists of images of half of the right knee (the medial side of the tibia seen from above) of the mice at different ages. The mice are sacrificed and their tibia removed and stained with a blue dye. The dye binds to proteoglycan (an active part of the cartilage) that is destroyed by osteoarthritis. It is the belief that the healthy parts of the tibia appear blue in the images, the first sign of lesion appears purple and the worse lesion areas are bright / white. By automatic image processing, it is attempted to identify the different types of areas and their relative amounts are used to calculate an OA measure. This OA measure is tested against age and other measures known about the mice.

There is generated a probability map of the lesions and it is thereby shown that the lesions mainly emerge on or near the upper border of the tibia. Average images of the tibias at different ages are generated to see how the different lesion types behave according to age in different positions of the tibia. The purple areas are shown to be more concentrated near the bright lesions and hence are likely to represent an early lesion stage.

Samples from the images reveal that the different types of areas are not grouped but are mixed to an oblong point cloud in feature space. There is therefore not a sharp border between the different types of areas and that makes their separation and identification more difficult.
Eight types of areas are defined and labelled in the images and samples from these are used throughout the report. It is the basic types of areas (blue, purple and white) and subdefinitions of these that give the eight classes.
Examination of the classes shows that by looking at the images separately, the classes are pretty much separable but merging the class' centers across the images results in overlapping classes. Thus there is a difference between the images which is found to be somewhat systematic for all the classes in the images and therefore it might be possible to reduce it. Trials with different color transformations cannot remove the difference but only change and improve it for some classes and worsen it for others, depending on which color transformation is used. Among others, trichromatic colors and the IHS transformation are used.

Due to non-separable groups in feature space, clustering is not believed to be the perfect solution and likewise with classification due to the overlapping classes. Both methods are tried. Classification is tried first, to learn more about the classes and the difficulties in the images, and hence improve the knowledge for a clustering solution.

The classification of the images uses different combinations of input and output, to see the effect of the different classes. The classification is tried with each of the colorbands separately,

and up to three of them simultaneously.
From the implemented Bayes classifier, using Mahalanobis distance, the highest correlation of the OA measure with age is 0.58. Due to possible collinearity, a more reliable result is 0.56, obtained using RGB as input. The results are based on all eight defined classes. Merging the classes to the three basis classes, results in an OA measure that correlates with age at 0.57, also with RGB as input.
A set of manually defined decision rules are tried in order to classify the three basis classes. Here, the OA measure's correlation with age is 0.44 and hence not an improvement.

A couple of clustering approaches are suggested and a clustering survey is carried out. Trials do not separate the point cloud in feature space into directly usable clusters. Blue and white overlap and likewise with blue and purple. Strong tendencies according to age are not found and further actions to use the obtained clusters would be similar to the manually defined decision rules, tried in classification. Clustering is therefore not believed to be a usable approach for these images.

The classifier is tried again and improved by using a priori knowledge about the position of the bright lesions. Further improvements are removal of collinearity and noise reduction, among others. The noise is reduced using a median filter and this improves the classification. The obtained correlation is 0.60, using RGB as input. The addition of position information results in an improved correlation at 0.66, which is the maximal obtained result during this project. The interpretation of the last optimization is somewhat unclear hence the 0.60 is the most correct result.

In this project the purple areas are found to be an early lesion stage and the bright / white areas are found to mainly emerge near the upper border of the tibia. It is shown that automatic image processing can be used to establish a reliable OA measure. The classifier approach obtains results at least as good as Visiopharm's earlier solution by measures on histograms. It is believed that the classifier solution can be optimized even further.

Resume

I dette projekt undersøges muligheden for brug af automatisk billedgenkendelse til test af slidgigt (OA) i laboratoriemus. Ved at benytte mus, der er prædisponerede for at udvikle OA, bør der være et stort sammenfald mellem deres alder og deres OA stadie. Formålet med dette projekt er at generere et automatisk OA mål for OA stadiet af hver mus.
Visiopharm har tidligere arbejdet på projektet og vil gerne have undersøgt, om resultaterne kan forbedres ved brug af andre metoder. De har hovedsageligt beregnet egenskaber fra histogrammer af RGB og andre farverepræsentationer. Deres OA mål har en markant korrelation til alder på 0.54. Visiopharms nye løsningsforslag er anvendelse af clustering.

Data til projektet består af billeder af halvdelen af højre knæ (indersiden af skinnebenet, set fra oven) fra mus i forskellige aldre. Musene er slået ihjel, deres skinneben er fjernet og herefter farvet med en blå farve. Denne farve binder sig til proteoglykan (en aktiv del af brusken), der destrueres af slidgigt. Det er opfattelsen, at den sunde del af skinnebenet optræder blå i billederne, det første læsionsstadie er lilla, mens de værre læsionsstadier optræder lyse / hvide. Med automatisk billedgenkendelse er det forsøgt at genkende de forskellige typer af områder og deres relative areal bruges til at udregne OA målet. Dette OA mål testes i forhold til musenes alder og andre målinger fra musene.

Der er genereret et sandsynlighedskort over læsionerne og dermed vist, at de hovedsageligt opstår på eller nær den øverste kant af skinnebenet. Der er genereret gennemsnitsbilleder af skinnebenene ved forskellige aldre for at se, hvordan de forskellige læsionstyper opfører sig i forhold til alder på forskellige positioner af skinnebenet. De lilla områder af skinnebenet er fundet mere koncentreret omkring de lyse læsioner og er derfor sandsynligvis et tidligt læsionsstadie.

Udtag fra billederne afslører, at de forskellige typer af områder ikke er grupperet men overlapper til en aflang punktsky i featurerummet. Der er derfor ikke en skarp overgang mellem områdetyperne og det gør adskillelse og identifikation vanskeligere.
Otte områdetyper (klasser) er defineret og markeret i billederne og udtag fra disse bruges gennem rapporten. Det er de grundlæggende områdetyper (blå, lilla og hvid) og underopdelinger af disse, der giver de otte klasser.
Undersøgelser af klasserne viser, at når billederne undersøges enkeltvis, er klasserne tæt på at kunne adskilles, men når klassernes middelværdier sammenlægges på tværs af billederne, er resultatet overlappende klasser. Mellem billederne er der derfor en forskel og den er fundet til at være delvis systematisk for alle klasserne i alle billederne og derfor kan den muligvis fjernes.
Forsøg med forskellige farvetransformationer kan ikke fjerne forskellen, men ændre den og forbedre den for nogle klasser og forværre den for andre, afhængig af hvilken farvetransformation der benyttes. Bland andre er *trichromatic colors* og IHS transformationen benyttet.
På grund af de ikke adskilte grupper i featurerummet forventes *clustering* ikke at være den perfekte løsning og ej heller en klassifikator løsning, dette pga. overlappende klasser. Begge metoder afprøves. Klassifikatoren er afprøvet først for at lære mere om klasserne og de udfordringer, der evt. er i billederne. Dette for at blive klogere inden en clustering løsning.

Klassifikationen af billederne benytter forskellige kombinationer af in- og output for at se effekten af de forskellige klasser. Klassifikationen er baseret på farvebåndene enkeltvis og op til tre af dem ad gangen.

Med den implementerede Bayes klassifikator, der benytter Mahalanobises afstand, er den højeste korrelation af OA målet med alder 0.58. På grund af muligheden for multicollinearitet, er et mere troværdigt mål 0.56, som er opnået med RGB som input. Resultatet er baseret på alle otte definerede klasser. Sammenlægning af klasserne til de tre grundlæggende klasser resulterer i et OA mål, der med RGB som input, korrelerer til alder med 0.57.
Et sæt manuelt definerede beslutningsgrænser er forsøgt for at klassificere de tre grundlæggende klasser. Her opnås et OA mål, hvis korrelation til alder er 0.44 og dermed ikke en forbedring.

Der er foreslået et par clustering fremgangsmåder og lavet en clustering oversigt. Forskellige forsøg adskiller ikke punktskyen i featurerummet i direkte brugbare clusters / grupper. Blå og hvide grupper overlapper og ligeledes gør blå og lilla grupper. Grupperne viser ingen overbevisende tendenser til alder og videre udvikling af rutinen vil være i lighed med de manuelt definerede beslutningsgrænser, der er afprøvet under klassifikationen. Der regnes derfor ikke med, at en clustering løsning er en brugbar fremgangsmåde til disse billeder.

Klassifikatoren prøves igen og forbedres ved at benytte forhåndsviden omkring læsionernes position. Yderligere forbedringer er bla. fjernelse af multicollinearitet og støjreduktion. Sidst nævnte udføres med et median filter, hvilket forbedrer klassifikationen. Herved opnås en korrelation på 0.60, igen med RGB som input. Tilføjelsen af positions information, resulterer i en forbedring af korrelationen til 0.66, hvilket er den højeste korrelation, der opnås i projektet. Fortolkningen af den sidste optimering er dog ikke klarlagt og derfor er den tidligere opnåede korrelation på 0.60 det mest korrekte resultat.

I projektet er det sandsynliggjort at de lilla områder er et tidligt læsionsstadie og at de lyse / hvide læsioner hovedsageligt opstår nær den øvre kant af skinnebenet. Det er vist at automatisk billedegenkendelse kan benyttes til at etablere et pålideligt OA mål. Ved brug af en klassifikator kan det gøres mindst lige så godt som Visiopharms tidligere løsning med beregninger på fordelingen af farvebåndene. Klassifikatorløsningen menes endog at kunne optimeres yderligere.

Contents

Chapter 1

Introduction

This chapter is an introduction to the project. The first part is about osteoarthritis and the corresponding medicine development and treatment tests. After an introduction of the data, the overall problem and corresponding hypotheses are formulated. A literature search is carried out and suggestions to solutions of the problems are presented.

1.1 Osteoarthritis

Osteoarthritis (OA) is a degenerative joint disease [4, 5] where the cartilage in the joint between two bones is weakening and in a more progressed stage there is almost nothing left. The function of the cartilage is to absorb shock and to some extent also guide the movements of the joint. The cartilage contains proteoglycan (and other kinds of sugar molecules) which have a formidable ability to hold water. The water is pumped in and out when moving the knee in order to absorb physical stress. The knee is at the same time greased, so the friction is extremely low.
In all human beings there is a constant process of tearing down and building up the cartilage (just like the process in bones) which is not a problem when balanced. When suffering from osteoarthritis this balance does not exist and the tearing down (catabolic process) is faster than the building up (anabolic process).
The anabolic process weakens with age and hence the chance of suffering from OA increases the older a person gets. Approximately 50 % of the people turning the age of 50 have some sort of OA and in the seventies approx. 85 % have OA. Due to the increasing lifespan the problem is growing.
Farmers, ballet dancers, football players, mine workers etc. have an increased chance of developing OA due to hard physical labor and it is in some cases inherited but otherwise it is impossible to predict who will suffer from OA.
OA is independent of race and sex when looking at who develops it, but there is a difference between where in the body men and women have it.
At the beginning, the joints start to stiffen and is only painful in cases of overload. Later the pain is more constant but movement is still possible. In much progressed cases the pain is unbearable also when the joint is not used. There are no nerves in the cartilage itself so the pain comes from the surrounding bone and tissue. The pain from bone is due to changed blood streams and growth to withstand the increased stress. There is also pain from the surrounding tissue due to swelling and inflammation in and around the cartilage.
Most OA cases are visible on X-ray, but sometimes when OA is predicted from X-ray, the patient

has no pain nor other indication of OA. MRI scanning is a better way to diagnose OA but is still too expensive for routine scanning. Diagnosing OA is therefore mainly based on the patient's symptoms.

In Appendix A there is a more thorough explanation of the cartilage and osteoarthritis.

1.2 Medicine Development and Treatment Test

There are many types of medicine for treating osteoarthritis, but many of them are only dealing with the inflammation and hence reducing the pain. The treatments dealing with the problem have many unwanted side effects and can only slow down or stop the degeneration. It is therefore important to diagnose and start treating OA as early as possible.
Increasing lifespan, improved economics in the third world and the fact that the present medicine is not flawless mean that large sums can be earned by developing better OA medicine, which is why there is a great deal of focus in this field.
One of the early test stages in the medicine development is to use mice. A genetic altered mouse (STR1N) is predisposed to develop OA and in this way the OA stage should have a large correlation with the age of the mouse.
At present the test result is based on histology which consists of thin slices of cartilage taken after the mouse has been killed. The slices are manually graded by looking mainly for cracks and dents in the cartilage layer. This is a time consuming process and not objective.
Another measure is a biomarker (a urine sample) taken just before killing the mouse. The measure only indicates the current degeneration speed and not the absolute stage. The reason why it could be interesting to include the biomarker is because the mice are examined at a relatively young age where the degeneration has not stopped, hence some correlation with age is expected.
A new kind of data is acquired with the purpose of automating the grading process and obtaining objectivity. The right tibia (the bone below and including the lower part of the knee) is stained with a blue dye (Alcian Blue) and photographed through a microscope. The dye binds to the active part of the cartilage (the sugar molecule "proteoglycan"), and therefore areas with less staining are assumed to be lesions caused by OA.

Appendix A has a more thorough explanation regarding medicine development.

In the field of medicine, a level of significance at 0.05 is normally used. In this project a test value is defined as significant at a level of 0.10 and highly significant at a level of 0.01.

1.3 Data

Aventis Pharma develops and tests osteoarthritis medicine, and has provided the data for this project. Besides the histology and biomarker data, there are approx. 450 images of mice-tibia with different kinds and doses of treatment from 10 different studies. The mice were between the age of 40 and 154 days when sacrificed. Approx. 2/3 of the mice are 84 days old (12 weeks). All images are followed by a mask, marking which areas are usable. Areas too far from the focal-plane or with artifacts are excluded. The tibia is submerged into water when photographed, to avoid reflections in the focal-plane. Due to numerous image acquisitions of the same knees (during improvement of the process etc.), the oldest studies have been bleached and are excluded

from this project. Of the eight usable studies, 126 of the images belong to untreated mice and their OA stage should thus have a large correlation with age. These images, of only the medial side of the right tibia seen from above, constitute the main data set for the project.

The appearance of the stages of OA varies from image to image and / or from study to study. The large variation in the data will probably not give good results when grading a single knee (or image), but the mean of each study is expected to have a large correlation with age, but with a lower level of significance due to only eight studies compared to 126 images.

1.4 Hypotheses

The blue staining binds to the proteoglycan in the cartilage. It is the proteoglycan that is reduced when suffering of OA, which is expected to cause visual changes in the images of the tibias. There are blue areas which are assumed to represent healthy cartilage, while bright / white areas are assumed to represent lesions.
The mice are genetically altered so they will develop OA, hence the older they are the larger their probabilities are of having OA and the larger the extent is expected to be. This is the background for the main hypothesis of this project: **"The relative amount of the lesion area increases with age and the relative amount of healthy area decreases with age"**.

Besides age there are also biomarker (urine sample) and histology (slices of cartilage) measures for some of the image data. These lead to the hypothesis: **"Biomarker and histology are reliable measures of osteoarthritis"**.

The positions of the bright lesions do not seem to be random and have lead to the hypothesis: **"The bright / white lesions are more likely to appear near the border of the tibia than in the middle of it"**.

There are some purple areas in the images, which were said to be healthy cartilage with staining variation or failure. During the project, observations and test results changed this assumption and resulted in the following hypothesis: **"The purple areas are more likely to appear next to bright / white lesions than in the other positions of the tibia"** and hence **"The purple areas in the images represent an early lesion stage"**.

The appearance of the same type of area varies between the images but it seems to follow some sort of structure. It will be examined by the hypothesis: **"The diverse appearances of the same type of area from image to image is mainly due to an intensity variation between the images"**.

Besides rendered probable these hypotheses, the quality of the images should be examined and likewise with possible differences between the studies. Preprocessing of the images should be carried out, if needed.

If time permits it, the developed routines should be used on images from treated mice to see if it is possible to measure the same treatment effects as Aventis Pharma has found using other methods.

1.5 Problem Analysis

To solve the problems properly, information about osteoarthritis, medicine development and tests are found. Visiopharm's reports are studied and a literature search is carried out on tissue segmentation, color transformation, classification and clustering.
Based on this knowledge and general experience a short solution suggestion / summary is given for each hypothesis. In the report they are in order of appearance.

1.5.1 Previous Work

Visiopharm has earlier prepared three reports for Aventis [1, 2, 3] regarding the same problem. The first two, directly deal with developing algorithms to find connections between the images and age, biomarker and histology data. The third report is the work of Bachelor student A. K. Poulsen and here the focus is an exhaustive examination of color transformations, to improve the results of the earlier developed algorithms.

The first project [1] consists of a joint degeneration index, resulting in a highly significant correlation with age at 0.54 and with biomarker at 0.43. There was no correlation with histology. The entropy of trichromatic green leads to these results. Standard deviation and inverse mode density level on the trichromatic green also gave good results. These were the results when choosing between several measures on histograms and different color transformations.
The algorithms have later shown not to produce equally good results on new studies.

The second project [2] resulted in a measure combined by trichromatic blue and the entropy of trichromatic green, giving an upside down U-shaped curve for the development from healthy to degenerated cartilage.
Again the algorithms have later failed to generalize to new study data.

The third project [3] does not match the earlier obtained results, hence no better color transformations are found than those used in the first projects. One interesting thing can though be used; the results are very dependent on the angle of the Hue in the IHS transformation. The best angle depends on the used images, but this project uses more or less the same data and the obtained angle ($\theta = \frac{\pi}{3}$) will therefore be tested in this project. Only four angles have been tested and hence it could be more exhaustive.

More fulfilling resumes are to be found in Appendix B.

1.5.2 The Main Problem

The problem at hand is quite unique. Sure others are working with blue staining of the tibia but the found publications mostly work with stained slices of cartilage or stained slices of the bone, to examine its growth changes. Others use a much larger part of the bone and hence the details searched for, in this project are not an issue.
A second trial is tissue identification. This mainly results in finding publications on birthmarks and whether or not they are abnormal. The general approach is to find the edges. Some also use clustering but this is not directly usable for this project. One thing is interesting though; when clustering structures, that are not separated, then fuzzy clustering is preferred instead of

crisp clustering.

1.5.3 Clustering

The first clustering routine was published in the year 1965 by E. Forgy [11] and is the basis for the well known k-means clustering by J. MacQueen in 1967 [12].
Since then, many types and improvements of clustering have been published and the literature search in this area has developed to something like a survey, which can be found in Appendix K.
Clustering is quite simple to implement and alter to specific cases but it is not so easy to control and adjust the parameters. A disadvantage is that spectral areas do not always correspond to informational classes. K-means clustering should be tried and maybe some density based or nearest neighbor clustering.

1.5.4 Classification

Classification is used more often than clustering but there are not as many alternatives. There are some basic ones and then publications are more concerned with add-ons like using spatial features. A classifier can, besides class centers and covariance, use a priori knowledge where the relative amount of each class is included and losses can be defined for each combination of original and resulting classes.

The main classification routines are

- Minimum distance. Classifies a pixel, as the class, for which it has the shortest distance to.
- Parallel piped. Defines a set of decision boundaries for each class. Here a pixel can be classified as undefined and as overlapping, besides the normal classes.
- Maximum likelihood. Uses the mean and covariance of the classes, to find the class with maximum probability for each pixel. Often a Bayesian classifier is used.

Normally a Maximum likelihood classifier gives the best results because it uses most a priori knowledge.

1.5.5 Color Transformations

There are several types of color transformations but they are not always followed by a description of their advantages and disadvantages.
Two color transformations are found interesting; the first, trichromatic colors [6] normalize the images with respect to the intensity by dividing each band by the sum of the bands. Trichromatic red, green and blue thus represent the relative intensities of their respective bands. The second is Intensity, Hue and Saturation (IHS) [6] where Hue and Saturation are independent of the mean intensity (Intensity).
A. K. Poulsen [3] has gathered a lot of color transformations and some of them also normalize the color space. Among the color transformations trichromatic and IHS are found. Some of the other transformations also separate the intensity and are much like the IHS. They define the intensity differently and some of them are based on the human perception of colors, which is probably not a good idea for this project.
It is possible to make other transformations, e.g. subtracting the mean intensity from the bands.

1.5.6 Solution Suggestions

Here is a short solution suggestion / summary for each of the hypotheses.

"Examine the quality of the images and possible differences between the studies. Preprocessing of the images should be carried out, if needed"
Visiopharm has, together with Aventis, improved the image acquisition and removed a lot of artifacts and variation between the images, see Appendix B. The images are downsampled before they are used in this project hence the noise is reduced even more.
Due to the improved image acquisition and the developed protocol for it, reproducibility should be obtained and hence the possibility of comparison across the studies. The images are thus prepared for image processing and a thorough examination of them is out of this project's scope.

Three problems are though found during the project

1. The images can be saturated.
   There is no need for moving or scaling the intensities to obtain a better dynamic of the highest intensities.

2. Analyses have shown that there is an intensity variation between the images. This is a problem for the analyses.
   The histograms of each study can be examined to see, if the studies of the same age have alike distributions and mean values. The studies of the older mice are expected to be brighter than the studies of the younger mice due to more and larger bright lesions with age. For the same age, the studies should be alike.
   Color transformations or intensity invariant measures should be used to avoid this problem.

3. In a few images there are pixels for which the red band is 0. This occurs in the left side of the images, near the border of the tibia. They are mostly single pixels and thus not very visible to the human eye. There are in the order of 100 pixels pr. affected image. No pixel values in these images should be 0 and their neighbors have intensities of 10,000+.
   They are easy to find automatically and can be replaced with the mean of their neighbors. A median filter can also be used to remove them. The pixels are discovered late in the project and are hence not corrected for most of the project. Their part of the usable area in the respective images is less than 1 ‰, and will not affect the results notably.

Due to lack of generalization of Visiopharm's solutions, a large effort should in this project be focused on thoroughly examining the data in order to reveal new information and structures.

"Biomarker and histology are reliable measures of osteoarthritis"
There are only biomarker and histology data for a few studies, all with mice in the same age, which reduces the test possibilities. Biomarker and histology from the same mice can be correlated to find out if and how they are connected. For histology data from different studies, the probability of whether or not they come from the same distribution can be examined.

"The bright / white lesions are more likely to appear near the border of the tibia than in the middle of it"
The tibia is not positioned exactly in the same place from image to image, therefore the images should be aligned (scaled, rotated and translated) to a common shape. This allows comparison of the images for each position (pixel) of the tibia. Different features (like the mean value) can then be extracted from different positions of the tibia and can be compared to each other, using analysis of variance, to obtain statistical proof of the differences. The mean values of the different areas can then be estimated to calculate the change in appearance between these areas. The lesions can also be manually marked to generate a probability map. This reveals the probability of symptoms (bright lesions) emerging at each position of the tibia.
Both solutions should be tried to obtain both strong visual results (the probability map) and strong mathematical results (analysis of variance).

"The purple areas are more likely to appear next to bright / white lesions than in the other positions of the tibia"
This hypothesis can be tested using the manually marked lesions, suggested above. A suitable purple measure can be calculated for different distances to the nearest lesion. It should show a falling concentration of purple as the distance to the lesions increases.

"The purple areas in the images represent an early lesion stage"
If the above hypothesis is rendered probable so is this hypothesis. Second, partitioning and identifying different areas in the images can be used. If the hypothesis is true, then the amount of purple area will increase with age, unless the higher lesion stages totally dominate the images with time.

"The diverse appearances of the same type of area from image to image is mainly due to an intensity variation between the images"
Due to variation of the tibias and their various lesion stages, the offset (average intensity of the image) cannot be subtracted with the purpose of normalizing them.
The average over entire studies could be compared to each other with respect to age. There are ages for which there exists only one study and this can therefore not be normalized and besides, the variation is believed to be between images mostly.
Samples from the different types of areas and from different images can be used to show whether the intensity variation is systematic or at random.
Different color transformations, like trichromatic and IHS, should be tried and also inter band relations, to find out if the intensity variation can be reduced and if new facts of the data are revealed.

"The relative amount of the lesion area increases with age and the relative amount of healthy area decreases with age"
A simple solution is to calculate the percentage of the area of the manually marked lesions out of the total usable area for each image and correlate it with age. This is functional for an initial test, but with respect to the purpose of the project, the areas should be found automatically and objectively.
The hypothesis can be tested by partitioning the images in different types of areas, each with similar appearances. The found areas will be identified (as healthy or lesion, or more detailed)

and used in order to calculate the amount of lesion area compared to the total used area of the tibia. This is a suggestion for an OA measure and will be tested against age, biomarker and histology.
There are four obvious partitioning types to look into; texture analysis, segmentation, clustering and classification.
Texture analysis uses the neighborhood of a pixel to describe patterns.
Segmentation is mainly about finding borders and other dissimilarities in the images.
Clustering groups similar pixels / areas together.
Classification finds, for each pixel, the class for which it has the largest probability of belonging to.
There does not seem to be a strong difference between the patterns in all the different types of areas in the images. There are patterns in the images, but besides the fibrillated type they are not related to a specific type of area. Texture analysis is therefore not considered for this project.
Due to the nature of the images (soft and non-definable edges) segmentation will probably not do, at least not alone.
Due to non-generalizable appearances of the different areas, clustering will probably do better than a classifier. After a color transformation of the data, classification might work and the transformation might also improve the clustering results.
Clustering might seem to be the best solution, but it can be difficult to control and when there are no separate / natural groups in the data, it might result in less usable clusters. The identification of the found clusters is not without problems. Classification is easier to control and is more objective (less choices). Both, classification and clustering, are tried in order to learn as much about the data as possible and to obtain the best results.

1.6 Schematic Approach

With the classifier as basis, the main results and working approach are sketched in the following figures.

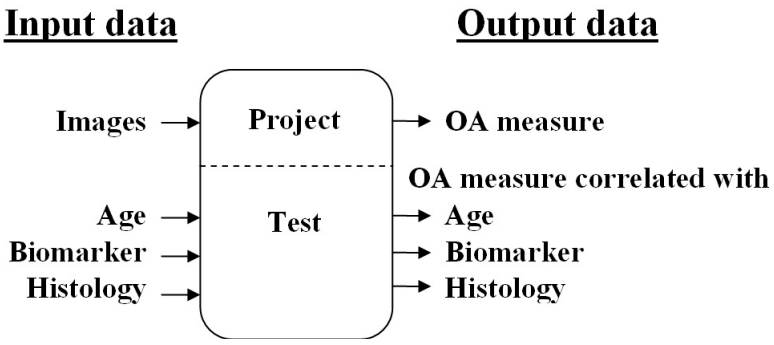


Figure 1.1: The overall purpose of the project.

Figure 1.1 shows the overall purpose of the project; to generate an automatic osteoarthritis measure of each image that has a high correspondence with age, biomarker and histology.

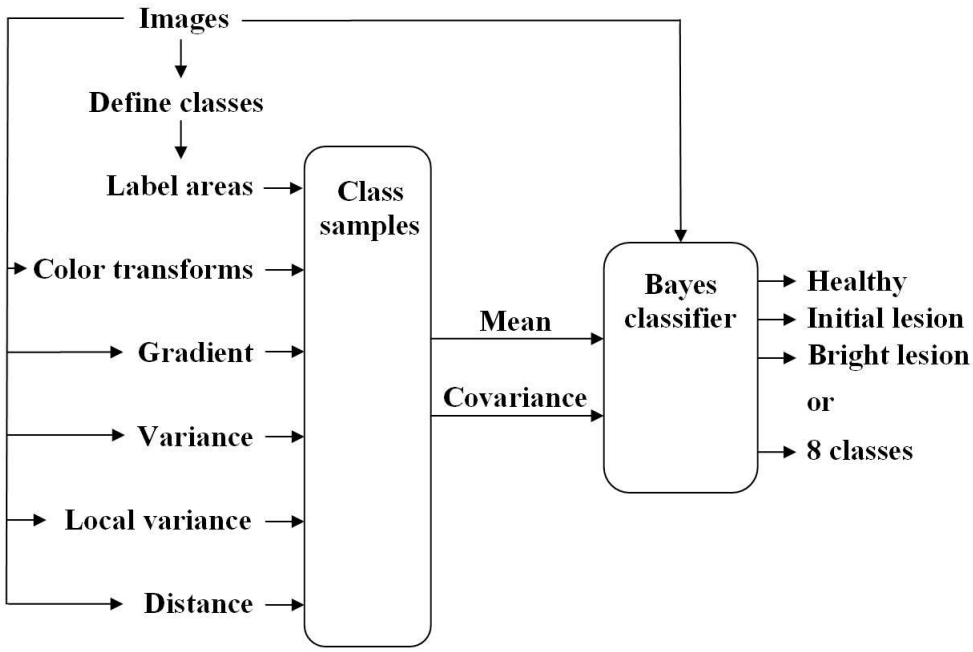


Figure 1.2: Input data, preparation and classification.

Figure 1.2 shows, based on the images, that different classes are defined and areas in some of the images are labelled according to them. The images are color transformed by different methods and samples are extracted. For the same images and pixels the gradient, variance, local variance and distance are extracted and included in the sample. The distance is defined as the distance from the upper border of the usable area in each image. The mean and covariance of each class in the entire sample are calculated for each feature / color representation and used to classify the images. The classified images contain either three or eight classes.

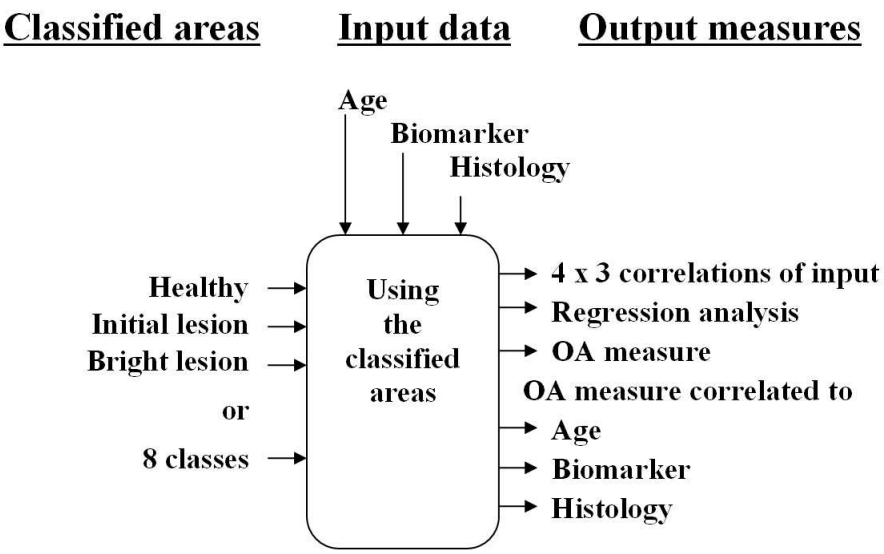


Figure 1.3: The classification results are used to generate the OA measure which is tested against age, biomarker and histology.

Figure 1.3 shows that the relative amount of each class in each image is correlated with age, histology and biomarker. It is also used in a regression analysis to find the significant classes and their coefficient for the OA measure. This OA measure's correspondence with age, histology and biomarker, is tested.

Chapter 2

Data and the Image Acquisition

This chapter describes the images with respect to staining, physical data and region of interest. The image acquisition is shortly described while the acquisition of biomarker and histology are not dealt with in accordance with the main focus of the project. The data set and study images for the project are defined. Examples of the data are given and finally the ground truth information (age, histology and biomarker) is described.

The data for this project is rather comprehensive. It consists of a series of images from each of eight used studies. Each image contains half a knee joint from a mouse, captured through a microscope. Postmortem, the knees are removed, cleaned and stained with a blue dye that binds to the proteoglycan in the cartilage (which degenerates during osteoarthritis). It is only the right tibia's medial side (the side towards the other knee) that is used. It is believed that this side is worn the most. Throughout the report this half of the right tibia will be referred to as "the tibia".

2.1 Blue Staining

After dissection, cleaning and drying, the knees are stained with Alcian Blue. It is carried out by dipping the cartilage in the dye for 30 seconds and afterwards washing it in a salt resolution. The dye binds to the proteoglycan in the cartilage, which results in healthy areas obtaining a deep staining (dark blue) and the worse the osteoarthritis, the brighter the blue. The worst areas do not get colored at all and result in an almost white or natural cartilage color. Some local areas with OA can be colored dark blue due to the hyperactivity of the cartilage trying to "fight" the degeneration.
Some areas can also end up being purple. This was said to be some sort of missing staining or staining failure, but is in this report rendered probable to be an early lesion stage.

2.2 Imaging

The used data only consists of the medial part of the right tibia captured from above through a microscope. The tibia is positioned in clay and submerged into water to move reflections away from the focal plane, and hence remove reflections from the images. The light source is warmed up for at least 15 minutes before capturing the first image to insure constant illumination. Aventis Pharma and Visiopharm have improved the protocol for the capturing process

hence reproducible images should be the result as long as the tibia only is submerged a few times to avoid bleaching.

Figure 2.1 shows the tibia from the side and the whole tibia from above to get a better understanding of the images used.

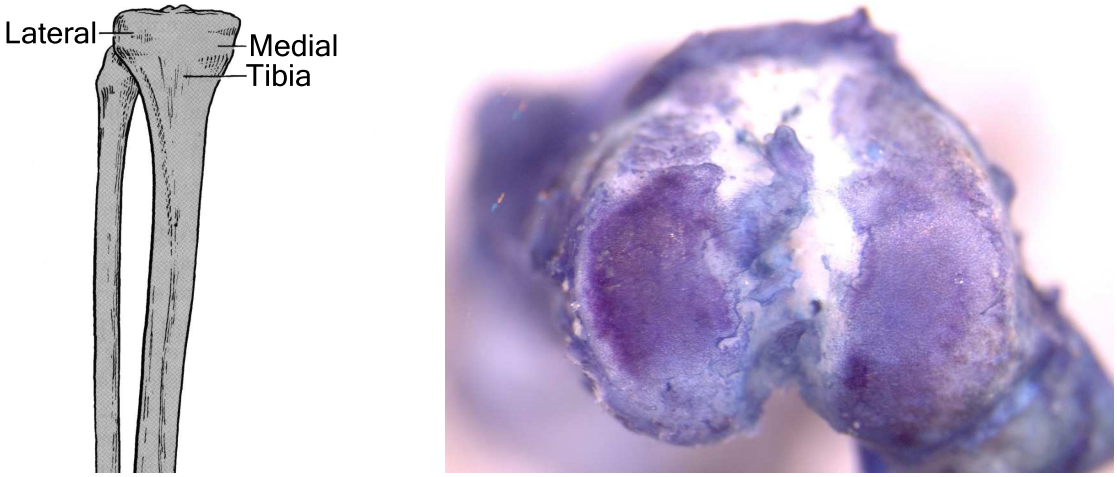


Figure 2.1: The tibia. Left: The tibia from the side. Right: The whole tibia captured from above. It is the left side of the tibia, turned 90 degrees clockwise, that is captured in the images.

Calibration was developed by Visiopharm for the image acquisition but due to problems it has not been implemented. The main problem is the large magnification in the microscope. The result is that normal calibration schemes are too rough to calibrate from. Small cracks, dents and variation in color become visible and dominating in the images.
This project discovered an intensity variation between the images that probably could have been reduced if the calibration was implemented.

2.3 Region of Interest

For each image a Region of Interest (ROI) has been defined. It is a hand drawn mask (by an operator), describing the border of the ROI, which is stored in a separate file, one for each image. The ROI is defined from the focal plane of the microscope hence areas outside have height distances too far from the focal plane and are too blurred to be used. Artifacts and other disturbances (like manually torn cartilage and air bubbles) and other areas that should have no effect on the OA grading are also excluded from the ROI.
The border of the ROI does not have a sharp edge but is a manual decision all the way, so the precision and definition may vary from image to image and from operator to operator. Figure 2.2 shows a schematic drawing of the camera, the microscope and the submerged tibia. The ring light assures monotone light conditions.

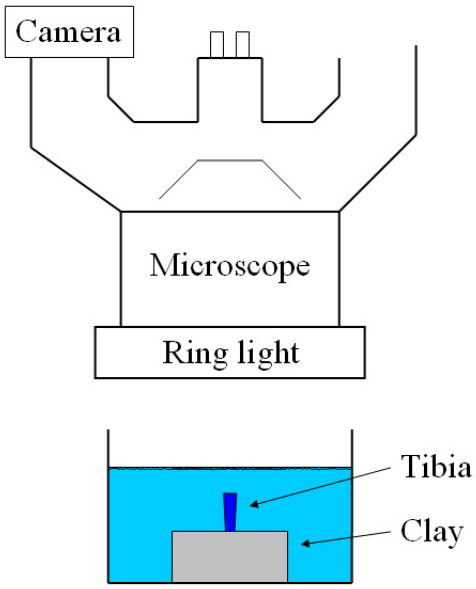


Figure 2.2: Schematic drawing of the image acquisition.

2.4 Protocol and Human Variation

The tibia is fixated in clay and hence the angle of the tibia, compared to vertical, can thus vary in two directions. This will hardly be directly visible in the images, but might affect the appearance of the tibia and increase the variation between the images. The distance from the lens to the tibia and the focus may also vary from image to image. The variance due to the last part is reduced by defining the ROI.

Different operators or users might do things a little differently compared to each other and from time to time (inter and intra variation). Visiopharm has found that the ROI drawing of the same images varies significantly between operators.

2.5 Physical Data

The images are reduced in size from the original which is double the size in both directions. The reduction is carried out by cubic interpolation. The image size is now 1044 $\times$ 775 pixels and in RGB format with 16 bits pr. channel, which is 4.85 MB information each. They are stored as LZW-compressed TIFF files (lossless) which results in a mean size of 6 MB each. This is approx. 20 % more than if they were not compressed, which is strange!
There is one frame per image and the active bit depth has been measured to be only 9 to 11 bits.
The physical size of the pixels vary from image to image due to the variable distance between the tibias and the lens. Variation within each image is also rounded due to the fact that the tibia is somewhat spherical. A mouse knee is in the order of 1 mm in diameter and therefore a qualified guess of a pixel size is in the order of 1 $\times$ 1 μm. The ratio between vertical and horizontal size may vary.

2.6 The Study Data

The images for this project basically comes from two kinds of experiments; one in which the mice are dissected at different ages in order to examine the development of the disease and one in which the mice are examined at the same age, now with different doses of medicine (*high, middle, low* or *no treatment*) in order to examine the treatment effect. There are two age dependent studies, five treatment studies and two studies with no treatment and equally aged mice. The information about the studies is organized in Table 2.1.

	Age	Treatment							Ground
Study	[weeks]	No	Low	Middle	High	Effective	Bleached	Blinded	truth
ADEP	$6-21$	15				–	*Yes*	*Yes*	
STR1N-04	12	25	25		25	*Yes*	*Yes*	*Yes*	*B&H*
STR1N-05	12	18	25			*No*	*Yes*	*No*	
STR1N-09	12	25	25		25	*No*	*No*	*No*	*B&H*
STR1N-12	12	24	24		24	*No*	*No*	*Yes*	*H*
STR1N-13	6	24				–	*No*	–	
STR1N-14a	6	9				–	*No*	–	
STR1N-14b	12	9				–	*No*	–	
STR1N-15	12	22	24	25	26	–	*No*	*Yes*	
STR1N-40	5.71	9				–	*No*	–	
STR1N-154	22	9				–	*No*	–	

Table 2.1: Overview of the entire study data from Aventis.

in which

- *Effective* means that the treatment had a (positive) effect on the mice.
- *Bleached* means that the tibias are believed to have lost some of the blue staining.
- *Blinded* means that the operator was unaware of the treatment given to the mice when drawing the mask and grading the OA.
- *Ground truth* means that Biomarker (B) and / or Histology data (H) is available.
- The ADEP and STR1N-04 are from year 2001 and the rest from year 2002.
- Biomarkers from STR1N-04 and STR1N-09 are not directly comparable.

Missing images or masks etc. are not included in the count. The images accounted for can still miss biomarker and / or histology data, but will be included due to the overall measure - correlation with age.

The age dependent study (ADAP) is bleached so much that it can be excluded without testing. In studies STR1N-04 and STR1N-05 the bleaching is more doubtful. A hypothesis [1] is used as a simple test (carried out in Chapter 4.4) but it does not separate these studies from the rest. They are older than the rest of the studies and could be weakened either by time or by being submerged into water so many times that the staining is washed out. To make sure that they do not disturb the algorithm development, they are left out of this project.
Images in which the left tibia was captured (if the right was unusable) are also left out and likewise if the mask (that defines the ROI) is too different from the others.

The final data set, consisting of 126 images, can be seen in Table 2.2.
Study STR1N-14 is the only used study with mice of different ages. In order to make compari-

Study	Age [days]	Age [weeks]	Images	Bio-markers	Histo-logy	Left tibia	Abnormal mask
STR1N-09	84	12	23	23	25	$123LT$	$117RT$
STR1N-12	84	12	24		24		
STR1N-13	42	6	24				
STR1N-14a	42	6	9				
STR1N-14b	84	12	9				
STR1N-15	84	12	19			$120LT$	$105RT, 111RT$
STR1N-40	40	5.71	9				
STR1N-154	154	22	9				

Table 2.2: Overview of the final data used in this project.

son to age possible, it is divided in the studies STR1N-14a and STR1N-14b which contain the images of the 6 and 12 week old mice, respectively.

During image acquisition, grading etc. some of the studies were blinded (the tibias and / or images were in random order with a random number) but they are all unblinded before they are used in this project. The images in each study now have a two or three digit number in ascending order. There might be missing values due to abnormal masks etc. as mentioned above. In the report there are references to single images by using the following name format: "STR1N-09-110". It refers to image 110 in study STR1N-09. Because of the fact that only the Right Tibia is used, the original extension (RT or LT) is left out of the name.

2.7 Data Examples

The variation in the data set is rather large, hence showing a few examples will only give a vague idea of the different cases. Figure 2.3 shows four examples and Figure 2.4 shows another four with their respective masks.
Image STR1N-14a-203 contains healthy varying blue, lesions and some brownish areas. The circle at the bottom part of the image is an artifact from an air bubble. Image STR1N-12-124 contains purple areas, lesions and some healthy varying blue. Image STR1N-14b-104 contains healthy blue material, lesions and in the top left part some dark spots. Image STR1N-14b-106 contains healthy varying blue, some brownish areas and some dark spots. The different types of areas will be defined in Chapter 6.

Figure 2.4 shows that image STR1N-40-03 contains healthy blue areas and some small bright spots and image STR1N-40-03 is mainly purple with lesion looking areas within. Looking at the color of the bright areas shows that they are blue areas, but looking at the entire image the bright areas look like lesions. Image STR1N-154-03 contains bright healthy varying blue, a little purple and some lesions. Image STR1N-154-09 probably represents the tibia with the worst case of osteoarthritis in the data set. The lesions are bright and with some brownish or yellow inside. Besides lesions it contains blue and purple areas. These images are examples of the youngest and oldest mice but in the same studies almost the opposite appearance can be found, hence a large variation exists in the data.

The general ROI shape is somewhat like the one in STR1N-40-03 (the top left image in Figure

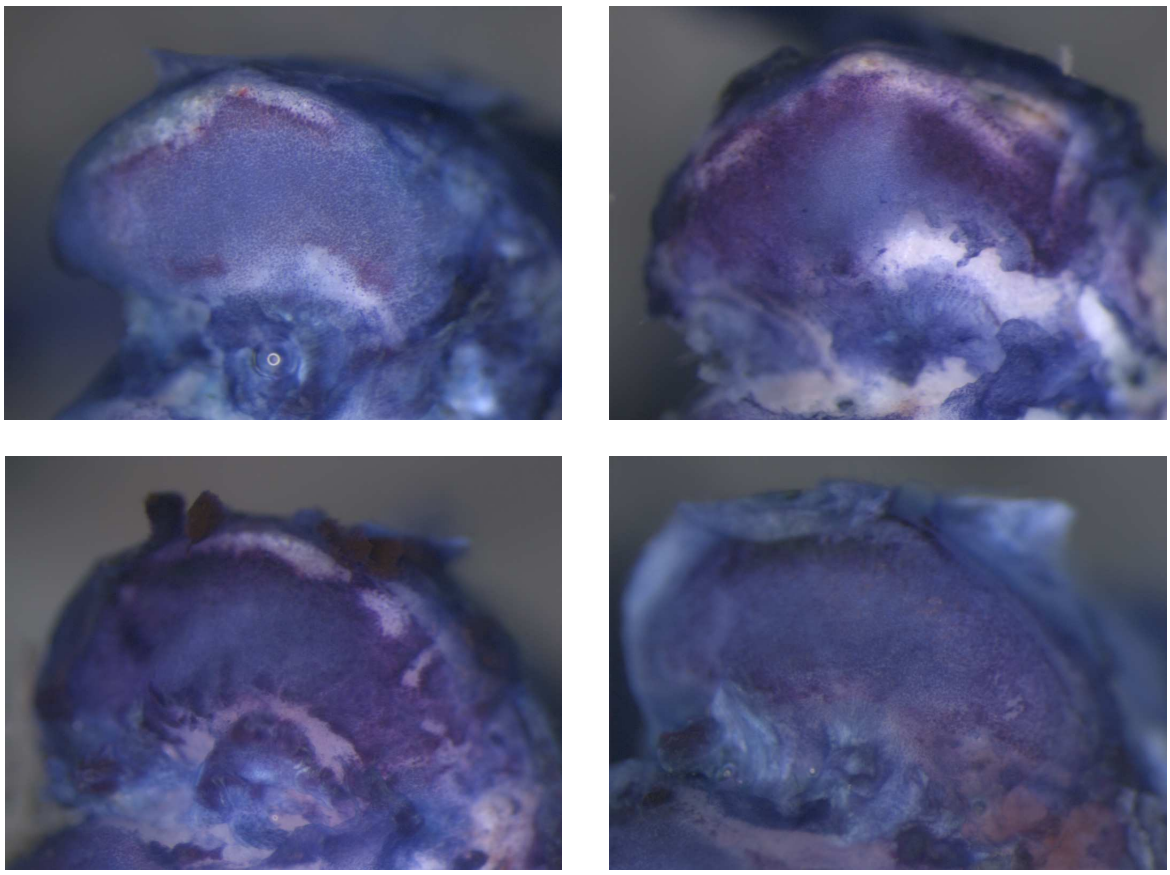


Figure 2.3: Examples of the images. Top row: STR1N-14a-203 and STR1N-12-124 (6 and 12 weeks old). Bottom row: STR1N-14b-104 and STR1N-14b-106 (both 12 weeks old).

2.4). The bright area, just outside and below the middle part of the ROI, is below the focal plane and is always left out of the ROI. It should not be considered a lesion.

2.8 Ground Truth

For this project there are three indications of the OA stage in the images.

- Age: Because the mice are going to develop OA, then the older they are, the more progressed should their OA be. The time of onset and the speed of degeneration varies from mouse to mouse, hence correlation with age is expected to be relatively high using the mean of each study and somewhat lower when correlating age to the OA measure of the single mouse.
- Biomarker: The urine samples only show the speed of degeneration. The optimal measure would be the absolute stage and hence the summed degeneration. If the turnover had not reached its maximum at the age of the oldest mice it could be a reasonable measure, but it is believed that the turnover peaks at an age of 12 weeks for an untreated mouse. So the usability of this measure is somewhat doubtful.
- Histology: It is only the right knee that is stained and imaged. The left knee is used to

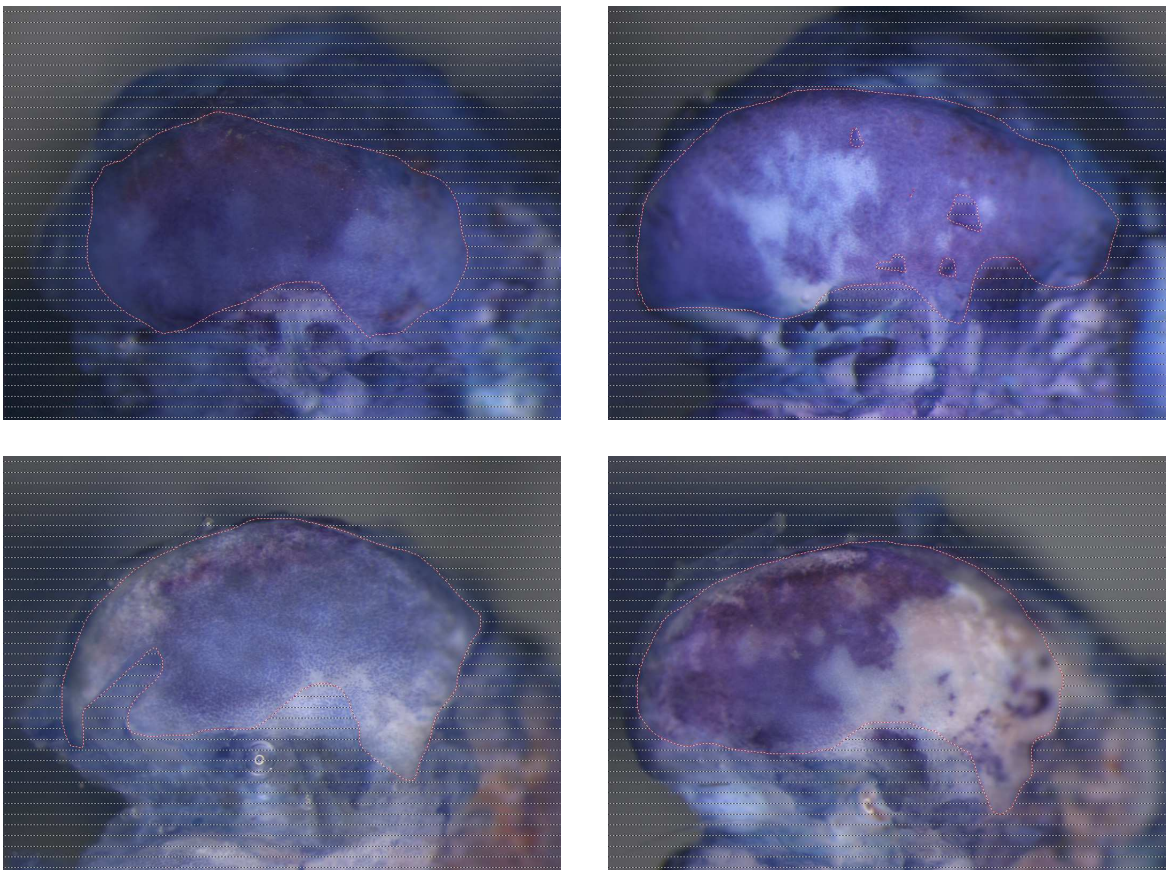


Figure 2.4: Examples of the images with their ROI mask. Top: STR1N-40-03 and STR1N-40-04 (both 5.71 weeks old). Bottom: STR1N-154-03 and STR1N-154-09 (both 22 weeks old).

extract histological information. A slice (or several) of the cartilage is examined and the extent of cracks and dents in the cartilage is a measure of the osteoarthritis' stage. This is Aventis' current gold standard for measuring the treatment effect of the OA medicine. There is of course an uncertainty here, because the two knees of a mouse do not have to be equally struck by OA and due to subjective grading.

Chapter 3

Biomarker and Histology

This chapter tests the validity of the biomarker (urine sample) and the histology data (manually graded slices of cartilage). The hypothesis is that "*Biomarker and histology are reliable measures of osteoarthritis*".

There is histology information for study STR1N-09 and STR1N-12, but only biomarker data for study STR1N-09. The biomarker data for STR1N-12 are found identical to the biomarker data for study STR1N-09 where it is believed to belong.

The data makes it possible to test the correlation between biomarker and histology for the STR1N-09 study and to test for equal distribution of the histology data for the two studies. Both studies are of 12 week old mice hence no age correspondence test can be carried out.

3.1 Comparison of Biomarker and Histology

Summary statistics, correlation and scatterplot of the biomarker and histology data for study STR1N-09 are shown in Figure 3.1.

```
*** Summary Statistics for data in: Bio.Histo.09 ***

numeric matrix: 6 rows, 2 columns.
          Biomarker Histology
     Min:  49.19140 16.000000
    Mean:  69.42102 20.760000
  Median:  68.86772 19.000000
     Max:  98.43544 27.000000
 Total N:  25.00000 25.000000
Std Dev.:  12.13592  3.992493

*** Correlations for data in: Bio.Histo.09 ***

numeric matrix: 2 rows, 2 columns.
          Biomarker Histology
Biomarker 1.0000000 0.1726878
Histology 0.1726878 1.0000000
```

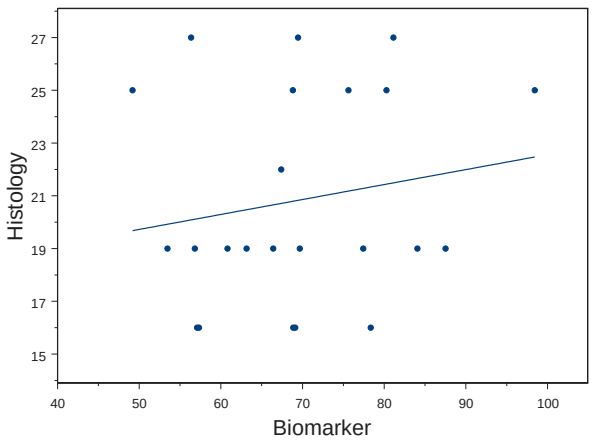


Figure 3.1: Correlation test between biomarker and histology data. Left: Summary statistics and correlation. Right: Scatterplot with a linear regression line.

The correlation is only based on 23 data pairs because two animals are without biomarker and hence excluded here. The correlation coefficient equals 0.17 with a t-value of 0.80. This means that the rather low correlation is not at all significant. The calculation of correlation and level of significance is specified in the Theory chapter, p. 105.
Visual inspection of the scatterplot shows no sign of correlation and the histology measures are concentrated on a few values and within these the spread of the biomarker is large.

3.2 Comparison of Histology Data from Different Studies

The two sets of histology data are independent because they come from different animals, hence a correlation test is meaningless, but a test of whether or not they come from the same distribution is carried out, see Figure 3.2.

```
*** Summary Statistics for data in: Histology ***

numeric matrix: 7 rows, 2 columns.
           STR1N-09  STR1N-12
     Min: 16.000000  9.000000
    Mean: 20.760000 18.583333
  Median: 19.000000 16.000000
     Max: 27.000000 27.000000
 Total N: 25.000000 25.000000
   NA's :  0.000000  1.000000
Std Dev.:  3.992493  6.240935


    Two-Sample Kolmogorov-Smirnov Test

data:  x: STR1N-09 in Histology ,
   and y: STR1N-12 in Histology
ks = 0.3433, p-value = 0.0896 alternative hypothesis:
 cdf of x: STR1N-09 in Histology does not equal the
cdf of y: STR1N-12 in Histology for at least one
 sample point.
```

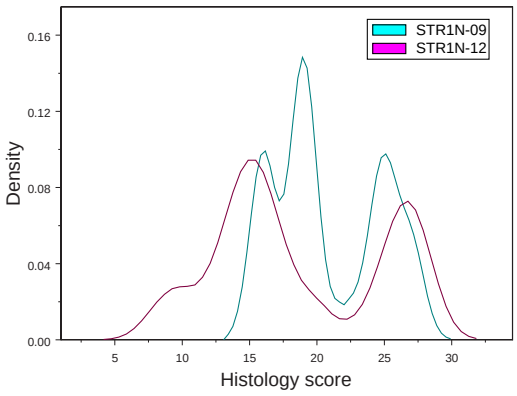


Figure 3.2: Comparison of histology data. Left: Summary statistics and test for equal distributions. Right: Density plot (smoothed) of the histological data.

The Two-Sample Kolmogorov-Smirnov Test gives a p-value of 0.0896. Using a level of significance at 0.05, which is normally used for this test, means that the distributions of the data sets do not differ significantly. Therefore it can be assumed that they come from the same distribution.

3.3 Conclusion

There is no correlation, either significant nor visual, between biomarker and histology data for study STR1N-09. Hence at least one of the grading methods must be considered a bad measure (either error prone, too large variation, maybe the measure has no relation to osteoarthritis or the histology grading is too subjective).

Comparison of the histology data shows that the density functions are not completely identical nor normally distributed, but shows some united shape. A Kolmogorov-Smirnov test do not show difference at a 0.05 level of significance and therefore it can be assumed that they come

from the same distribution, which indicates their usability.

There is not much to conclude from, but it seems like the histology data is the most trustworthy measure, if any. This is supported by a reason to believe that the OA turnover, measured by the biomarker, is at its maximum around the 12^{th} week of the STR1N mouse's life.
The hypothesis "*Biomarker and histology are reliable measures of osteoarthritis*" is not rendered probable, but it cannot be disproved that histology is a reliable measure of osteoarthritis.

The extent of biomarker and histology data is rather small but the corresponding automatic OA graded images can be correlated to them as a test of functionality.

Chapter 4

Initial Analysis of the Images

In this chapter the images and studies will be initially examined in order to familiarize with the data. Image examples, scatterplots, boxplots, the gradient and variance will be looked upon.

The used images, shown as examples and used for calculation through the report, are not the same from time to time. This is done to show some of the very different types of images. It should be noted that the printed images tends to favor / enhance the dominating color, so differences appear larger, than they are.

4.1 Examples and Scatterplots

To begin with, four images and their respective scatterplots are shown in Figure 4.1 and 4.2. The scatterplots show a random sample of a 1000 pixels from each image ROI.

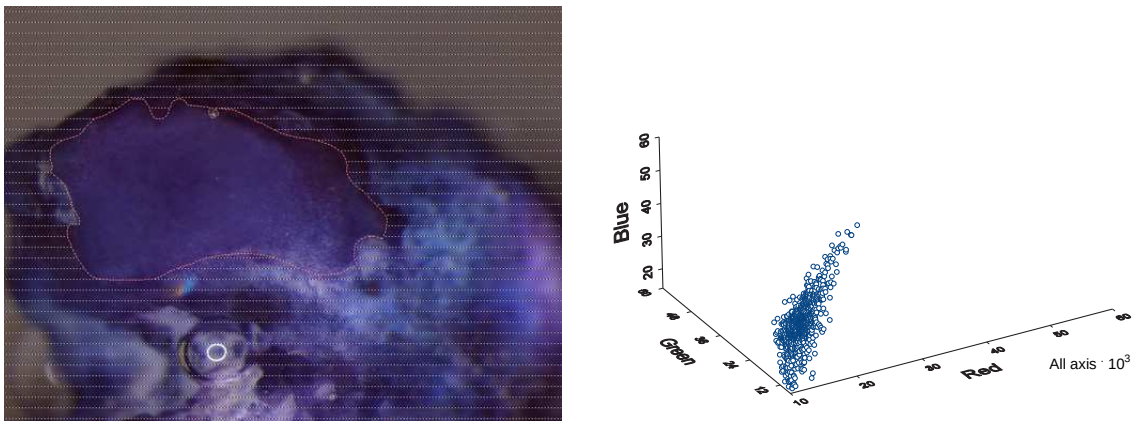


Figure 4.1: Image STR1N-40-10 and its corresponding scatterplot (a 1000 pixels sample from the ROI).

Figure 4.1 and 4.2 show that the images with lesions have pixels with higher intensities than the healthy one (in Figure 4.1). This is not surprising, but on the other hand it is surprising that there is no grouping tendencies for the different areas (blue, purple, white, etc.). Especially is STR1N-154-09 expected to show some grouping.

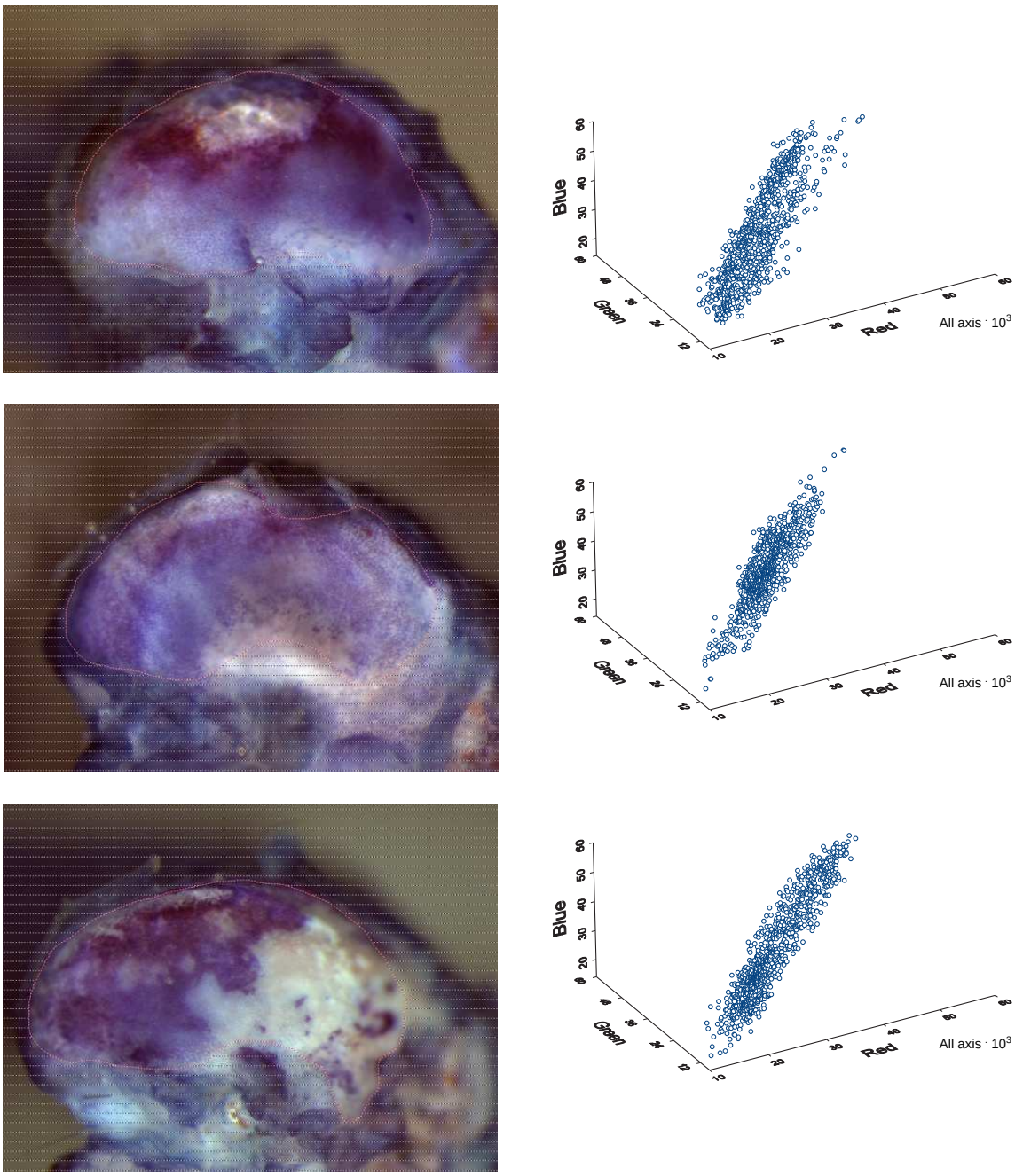


Figure 4.2: Images and their corresponding scatterplots (a 1000 pixels sample from the ROI). Top: Image STR1N-09-101. Middle: Image STR1N-15-102. Bottom: Image STR1N-154-09.

Note how different the blue areas are from image to image. They can be dark or light blue and the areas can be fast varying / fibrillated.

4.2 Histograms of the Colorbands

In order to examine the color intensities in different images and studies, histograms of the three colorbands are calculated by using all pixels in the ROI. Figure 4.3 shows the histograms for two single images. The histograms are normalized (by the largest value in all the bands in each

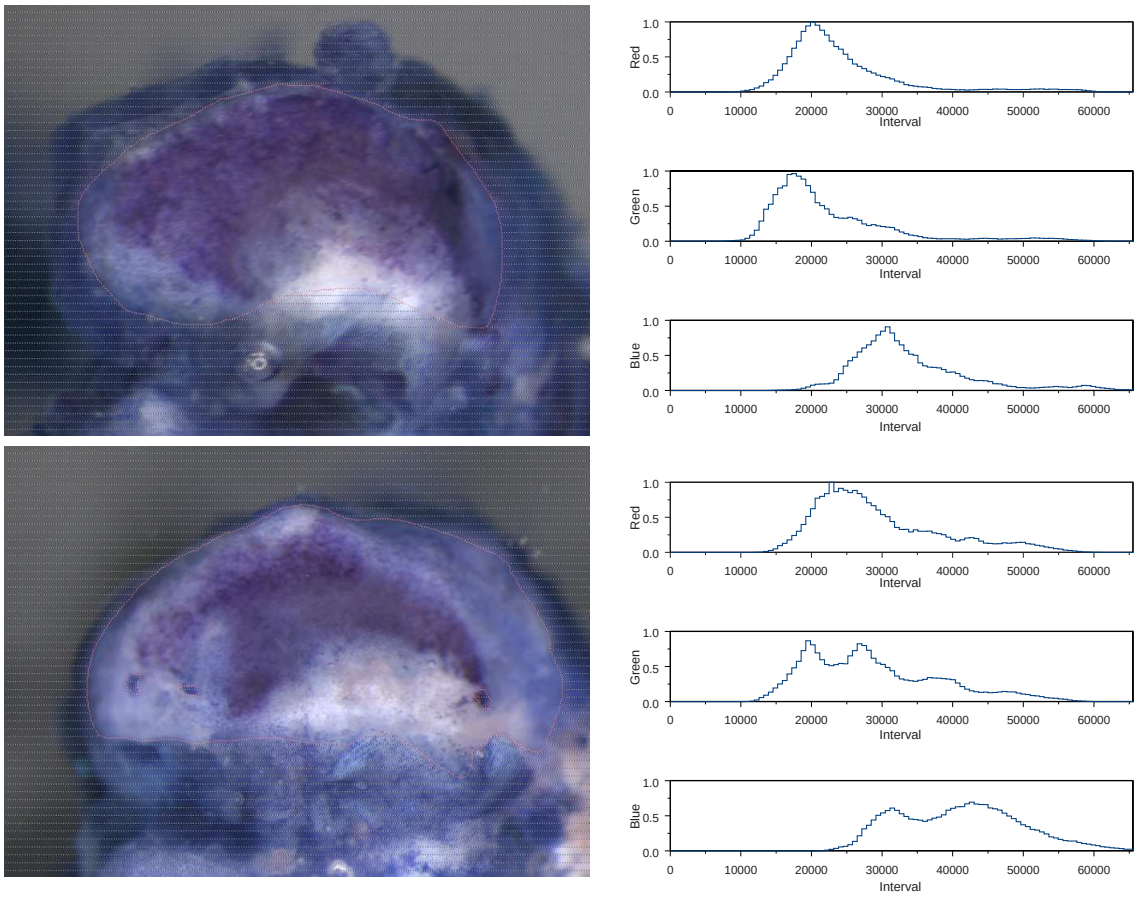


Figure 4.3: Histograms, with 100 bins, of the three colorbands, normalized between 0 and 1. Top row: Image STR1N-09-108 and its histograms. Bottom row: Image STR1N-15-110 and its histograms.

image), hence comparisons between the two images should be carried out only by shape and position of intensities and not by their absolute values.
The first image is mostly purple with some blue and some white and yellow lesion. The intensities are as expected, except that a flat tail goes to the top of the dynamic range as a result of the lesions (a peak is expected).
The second image has it all; purple, blue, bright blue and white and yellow lesions. This corresponds well with the green histogram which has two main peaks (purple and blue) an some smaller ones (yellow and white). The intensities have a larger variance and the mean values are higher (brighter) than the other image. It is a little surprising that the lesions do not result in larger peaks. Hardly any values are below 10,000 while the range in the high end is used by all the bands, especially the blue.

Histograms for all the ROIs and merged by study are calculated and shown in Figure 4.4. The youngest and oldest mice (STR1N-40 and STR1N-154, respectively) are highlighted. Figure 4.4

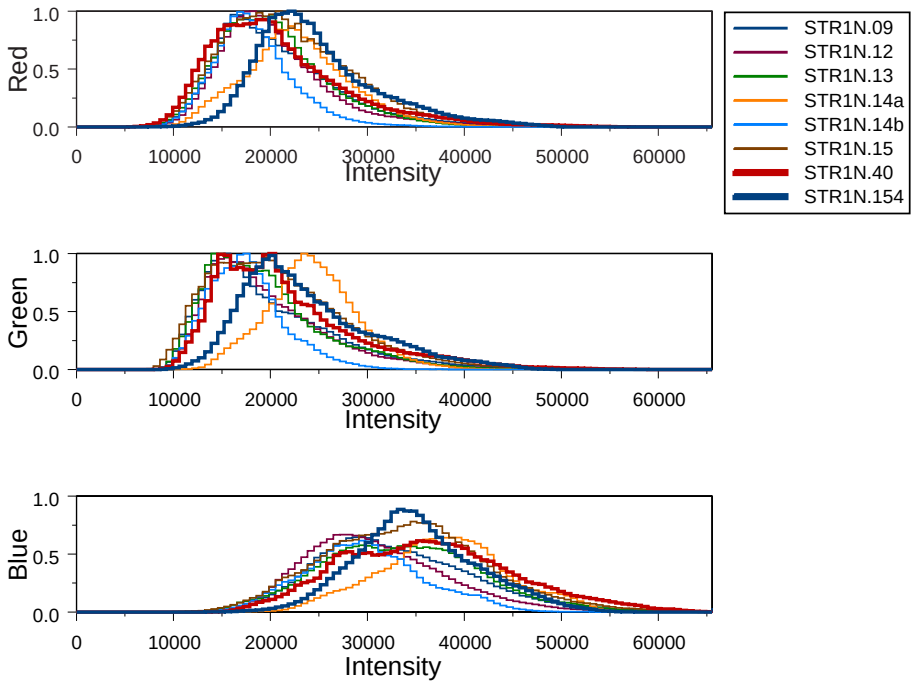


Figure 4.4: Histograms of the three colorbands merged by study. The youngest and oldest studies (STR1N-40 and STR1N-154) are shown by thicker lines. The histograms are normalized so the highest amplitude of all the bands for each study equals one.

shows that by looking only at the two highlighted studies, the images of the older mice have higher red and green intensities than the young ones, which could be due to more and / or larger lesions. The blue intensity is more concentrated for the older mice, and the blue intensities are more symmetric and normally distributed than the red and green band, which look more like a Rayleigh distribution. The increased red and green and somewhat equal blue results in brighter and more grayish colors.

Looking at all the studies, the youngest and oldest mice are not the extremas which could be expected. It is hard to see a general age dependency. In the green band STR1N-154 (22 week old mice) peaks between studies STR1N-40 and STR1N-14a (5.71 and 6 week old mice, respectively).

The histograms are also examined in order to find out if it reveals anything about the comparability across studies. Study STR1N-154 and STR1N-14a differs from the other studies. Study STR1N-154 is brighter than the others, but represent the oldest mice hence it is expected. Study STR1N-14a contains 6 week old mice and is expected to be darker than average, but that is not the case. It is brighter than average for all the bands and is the brightest for the green band.

A look at the images in study STR1N-14a reveals that they do look brighter than the images in the other studies. It is not large and bright lesions that dominate nor a general higher intensity of the entire image. In many of the blue areas there are bright pixels within. It varies with a high frequency and thus have a coarse but small structure. An example of this fibrillated appearance is shown in Section 2.7, p. 16 (the upper left image). Images from the other studies can contain the same type of areas but they do not have it to the same extent. For this reason the study cannot be determined to be abnormal and is kept in the analyses.

4.3 Summary of the Studies

Table 4.1 shows the mean and standard deviation calculated for all bands in each study.

	Age	Red		Green		Blue	
Study	[weeks]	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
STR1N-09	12	22266	7867	21065	8197	33448	8175
STR1N-12	12	19907	6264	18916	6303	30319	7596
STR1N-13	6	21289	5712	20551	6380	35278	8031
STR1N-14a	6	22529	5536	24051	4692	37469	6584
STR1N-14b	12	18063	4317	17312	3979	29456	6201
STR1N-15	12	22561	7384	21577	7729	34078	8140
STR1N-40	5.71	20707	7373	21522	7640	36014	8886
STR1N-154	22	24551	6407	23500	6769	34863	6322

Table 4.1: Mean and standard deviation for the colorbands in all the studies.

Table 4.1 shows that STR1N-14b and STR1N-12 (both contain 12 week old mice) together have the lowest and second lowest mean intensity for all the bands. It could be expected that mice's tibias at this age were brighter than average. STR1N-14a (6 week old mice) has the highest intensities for the green and blue band while STR1N-154 has the highest intensities for the red band. The mean values of the blue band is largest for the 5.71 and 6 week old mice, but the 22 week old mice are not much lower. The mean values of the 12 week old mice vary and their value are mixed with those from the other ages.
The standard deviation shows no relationship to age.

4.4 Study Tendencies and Bleaching

For further examination of the intensities of each study, the mean value for each colorband is plotted in Figure 4.5. Here the excluded studies STR1N-04 and STR1N-05 are shown as well, to see if they are bleached.
Figure 4.5 shows that the intensities vary a lot, and that the red band always has higher intensity than the green band for the 12 and 22 week old mice, while their blue bands are generally lower than average. For the 5.71 and 6 week old mice, two out of three have larger green intensities than the red. For the relative intensities these observations are still valid and the values have less variance. Relative intensity is the same as trichromatic colors where each band is divided by the sum of the three bands. This will be explained further in Chapter 7.1.1, p. 59.
The studies STR1N-04 and STR1N-05 have intensities above average but are not different from the other studies. The bleaching hypothesis [1] says that if the green intensity is larger than the red the study is bleached. This is not the case for STR1N-05. It is the case for STR1N-04

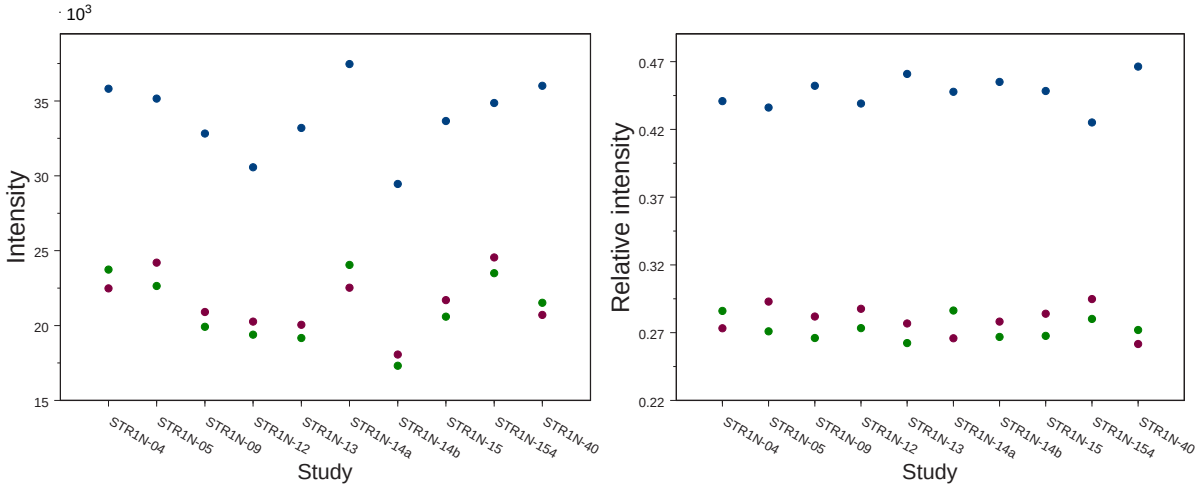


Figure 4.5: Mean intensity of each colorband for each study. Left: The intensities against study. Right: The relative intensities against study.

but also for the studies STR1N-14a and STR1N-40 (which on the other hand are younger). The hypothesis is therefore not believed to be true, but the studies are still left out, as explained earlier.
There is not much consistency from study to study. An odd observation is that the distance between the red and green band is almost equal from study to study even though it differs which are is bigger. It is probably just a coincident.

4.5 Intensities by Image

Here the behavior of the colorbands according to age is examined by plotting the mean of each colorband for each image ROI against age. The result can be seen in Figure 4.6. It shows that the intensity of the red and green colorbands increases with age while it decreases for the blue band. The red band increases the most. The images are brightened and / or are more purple the older the mice. Looking at the relative intensities, the blue decreases more visibly here and the red increases more. The difference from age to age looks convincing for the blue band, but not enough to separate between the ages.

Correlation with Age
The correlation between the colorbands and age are calculated for both the absolute and relative colors and both by each image and by each study. The corresponding level of significance is also calculated, see Table 4.2. The correlation and calculation of level of significance are defined in Chapter 11, p. 105.
Table 4.2 shows that the relative red and blue, for each image, correlates highly significantly to age at 0.55 and -0.48, respectively. The observations merged by study correlate better to age, 0.78 for red and -0.81 for blue. With only eight studies, the level of significance are 0.0185 and 0.0117 which is significant, but not highly significant.
The absolute values have lower correlation with age and hence a lower level of significance.
The green band has lower correlation coefficients than the other bands.

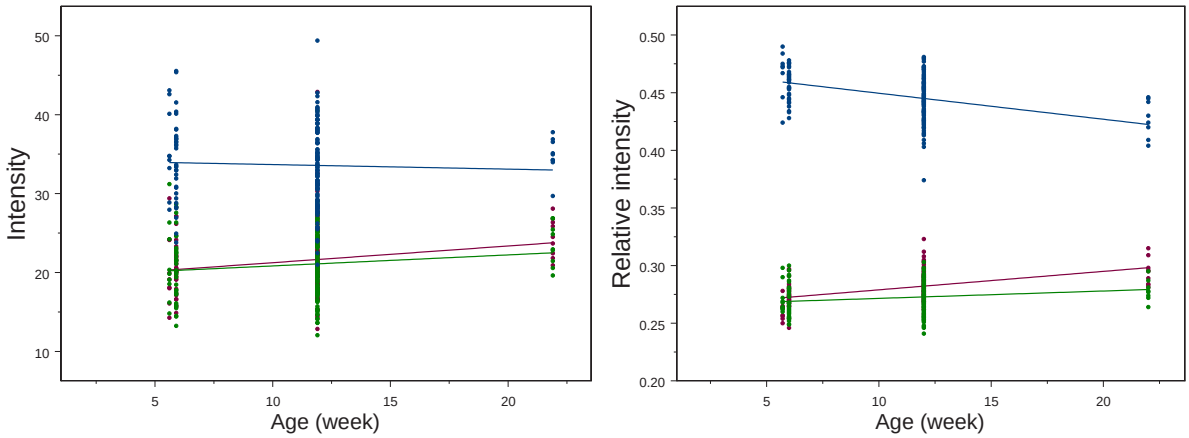


Figure 4.6: Mean of red, green and blue for each image ROI plotted against age and added a linear regression line. Left: The intensities against age. Right: The relative intensities against age. Note that red points are drawn first and almost covered by the other bands.

		Red		Green		Blue	
Data	Samples	Corr.	Signif.	Corr.	Signif.	Corr.	Signif.
Absolute, each image	126	0.20	$2.47 \cdot 10^{-2}$	0.09	$3.16 \cdot 10^{-1}$	−0.09	$3.16 \cdot 10^{-1}$
Relative, each image	126	0.55	$2.40 \cdot 10^{-9}$	0.16	$7.35 \cdot 10^{-2}$	−0.48	$1.27 \cdot 10^{-7}$
Absolute, each study	8	0.45	$2.57 \cdot 10^{-1}$	0.16	$7.03 \cdot 10^{-1}$	−0.19	$6.50 \cdot 10^{-1}$
Relative, each study	8	0.78	$1.85 \cdot 10^{-2}$	0.18	$6.68 \cdot 10^{-1}$	−0.81	$1.17 \cdot 10^{-2}$

Table 4.2: Correlation with age of the averaged colorbands over image and over study, including level of significance.

4.6 Gradient of the Colorbands

The gradient of the colorbands is calculated to see if it reveals anything about the different types of areas, see Figure 4.7. It is calculated by filtering the image with a 1 × 2 filter (+1 , -1) in the x-direction and the transposed filter in the y-direction, and adding the results by root-mean-square (RMS). The gradient images in Figure 4.7 show that red generally varies more than the other bands. This is not the case when looking at the standard deviation in Table 4.1, page 27, so the red info must vary with a higher frequency (here from pixel to pixel) than the other bands. In the bright lesion area, red and green show equal amounts of variation (which makes it yellow). Many of the steepest changes are alike for all bands (the white gradients).
The summed gradient image shows the same. It is surprising that there are large gradients within the bright / white areas, but the earlier shown histograms revealed that bright lesions did not result in a peak but have a large dynamic.
The logarithm of each colorbands gradient reveals that the image is saturated. The yellow and green areas are the troubled areas. The gradient is of course zero in the saturated band(s). For the given image, parts of the blue band (yellow areas) and the blue and red band (green areas) are saturated.
Some bright lesion areas give a yellow gradient, others give a green gradient and some areas are mostly marked on the border. After mean filtering it might be used as an indication of the different types of areas.

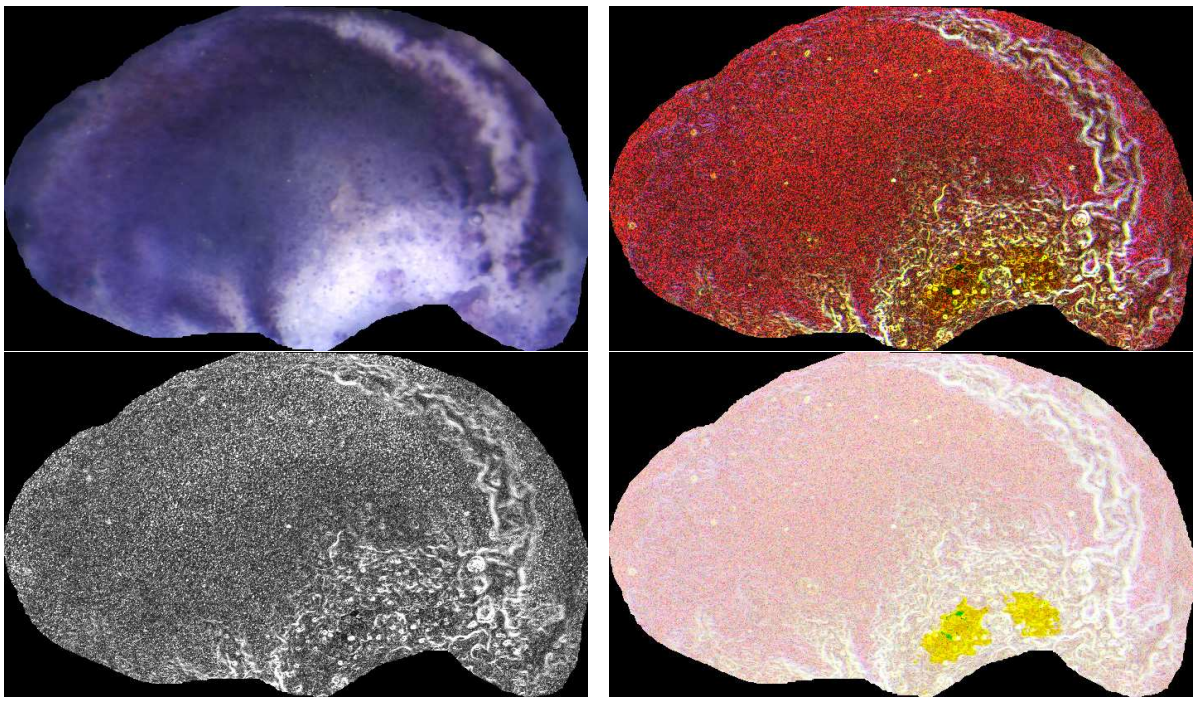


Figure 4.7: Gradient of a ROI (STR1N-09-118). Top row: The original ROI and the gradient of each colorband. Bottom row: The summed gradients (by RMS) and the logarithm-scaled gradient of each colorband.

4.7 Variance of the Colorbands

The variance of the colorbands is calculated to see if the different areas behave differently, see Figure 4.8. The same image as for the gradient is used. The variance is calculated for a 10×10 pixels window. Different window sizes are tried (5×5, 10×10, 15×15 and 30×30 pixels) and the 10×10 pixels is chosen because it prevents local peaks from being too dominating and at the same time it does not blur the result too much.

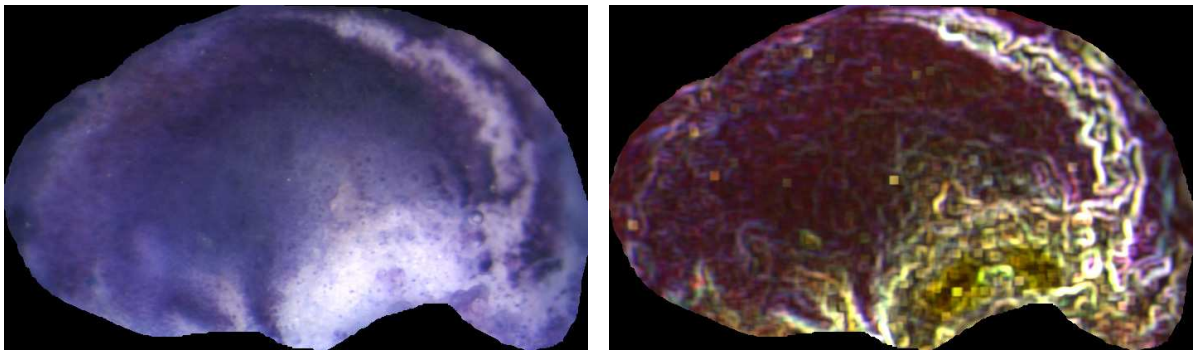


Figure 4.8: Variance of a ROI (STR1N-09-118). Left: The original ROI. Right: The variance of each colorband.

Figure 4.8 shows large resembles with the gradient image, but here the domination of red is more purple (varying blue and red).

The bright lesion areas generally have higher variance than the other types of areas, either high in all the bands or just in the green and red bands. In the first case, it is mostly the borders that have the high variance and hence a mean filter should be applied before using it to describe the different areas.

4.8 Conclusion

There is a large variation from image to image, and the shown images contain both easily identifiable areas and some rather diffuse ones. Scatterplots of the image ROIs show that images with lesions have higher intensities but the only sign of grouping between the different areas is a little different concentration in a continuous point cloud in feature space.
The histograms do not show much peaking either, according to the different types of areas. The histograms of the studies show no consistent tendency according to age. STR1N-14a looks brighter than the other studies and could be a problem when comparing results or merging information across studies. An examination reveals that all images can contain fibrillated areas that are blue with a small bright structure within. The images in study STR1N-14a contain more of this area type than images from the other studies. It is hence not reasonable to remove study STR1N-14a.

The bleaching hypothesis [1] shows that the studies STR1N-04, STR1N-14a and STR1N-40 are bleached and the other studies, including STR1N-05, are not. The intensities do not show that the studies STR1N-04 and STR1N-05 are different than the other studies. The hypothesis is therefore believed not to be true, but study STR1N-04 and STR1N-05 are still left out of the project to be sure that they do not affect the results.

The mean intensities for each ROI show some relationship to age. The tendency is most obvious for the relative intensities where blue decreases with age, while red and green increase. The relative intensities of each image have a highly significant correlation with age at 0.55 and -0.48 for red and blue, respectively. The absolute intensities have no significant correlation with age. The large variation in intensity between images from the same study could lead to the conclusion that correlation with age should not be based on the single images but must be averaged for each study and then correlated with age.

The gradient of the images shows that the red band differs more between neighbor pixels than the other bands and that the images can be saturated. There is a large gradient for all bands in or around the bright lesion areas, but in large bright areas the gradient is largest at the border.

The variation of the images, in a 10×10 window, shows that the red and blue band vary more than the green band. Their variance is large for all bands in or around the bright lesion areas, but in large bright areas the variance is largest at the border.

It is generally hard to find tendencies according to age. Especially the 5.71 and 6 week old mice, which are two days apart, have large differences. Likewise do the studies which represent 12 week old mice differ.

Chapter 5

Examination of the Bright Lesions

In this chapter the bright lesions and the purple areas are examined.
The images are aligned and the lesions are manually marked and used to generate a probability map of the position of the lesions.
The appearance at different positions is examined visually and by analysis of variance. The mean values of the different positions are estimated and likewise with the significance of their differences.
The aligned images with bright lesions marked are also used to test whether or not purple is a lesion stage. This can be assumed if purple is more likely to be found around the bright lesions than further away.

5.1 Image Alignment

A directly comparison of corresponding positions on the tibias is not possible because the tibia is not positioned exactly in the same place from image to image. The tibia can be moved, tilted and rotated, and the distance to the microscope can also vary. The difference in the images is not large, but large enough to disturb a comparison. The images are therefore aligned, not only by moving them, but also by scaling and rotating them. A description of the alignment can be found in Appendix C.

The masks are not aligned and therefore they cannot be used with the aligned images. This means that the "border" of each tibia is unknown and likewise with artifacts and other areas that should be excluded. The results can therefore contain some noise, but most of the areas are good enough. Areas that should be excluded from the ROI due to the distance from the focal plane, still have approx. the correct color, but are more blurred, so this is a minor problem. The important problems are artifacts and whether or not the background is included in the used area. Air bubbles are normally located below the usable areas.
A common ROI is manually drawn in order to avoid some of the problems. The common ROI is larger than the ROI of the single images because it also includes some of the blurred areas. However it does not include the background.

There are 126 usable images in the eight used studies, so only some of the images are aligned. The studies STR1N-14a, STR1N-14b, STR1N-40 and STR1N-154 each contain nine images which are all used. For the rest of the studies (STR1N-09, STR1N-12, STR1N-13 and STR1N-15) a

sample of 9 images from each is used. The images are chosen without prior investigation by using every second or third approximately. A total of 72 images are aligned. An equal number of images from each study makes tendencies between studies more comparable. There is of course still an unequal amount of images pr. age.

5.2 Probability Map for the Bright Lesions

Here the following hypothesis is tested, "*the bright / white lesions are more likely to appear near the border of the tibia than in the middle of it*".
The white or bright lesion areas are manually marked in the aligned images. These markings are merged to show where the lesions generally emerge. They are merged by study, by age, and finally all of them are merged.
The manual marking of the lesions introduces errors. A pixel's relationship to either lesion or healthy areas are doubtful and there is no sharp border between them. The marking should be considered reasonable, but is carried out by an untrained person whose knowledge is obtained only by looking at and working with the images.
The 47 of the 72 aligned images contain lesions. Figure 5.1 shows the lesions merged by age.
There is surprisingly much lesion marked for the youngest mice, but going through the images again does not change the marked areas. There are relatively large differences between the images from mice at the age of 5.71 and 6 weeks.
The positions of the lesions are not entirely random. It might seem so for the 5.71 week old mice (which have fewest images with lesions) but for the rest, there is a tendency of more lesions near the border of the tibia, especially in the upper part, than in the middle of it.
Note that the bright areas at the bottom of the images should not be considered a lesion. It is the area just below the hand drawn ROI, shown in Figure 5.2. This part of the tibia is always bright and outside the ROIs. It is colored here because there is no sharp border between the lesions and this area.
In Figure 5.2 all the lesions are merged to a probability map and it clearly shows that the lesions are more likely to emerge near the top half of the border than in the middle of the tibia. At the bottom, to the left, there are few lesions while in the right side there are more lesions. The brightest area represents 15 lesions (approx. 1/3 of the images with lesion). In the middle of the tibia there are up to three lesions. Following the upper border of the ROI, there are between 7 and 15 lesions in each position with a few exceptions. Using the hand drawn ROI it is observed that the lesions appear to be very close to the border of the ROI and outside of it.
The lesions in the unaligned images can also occur outside their ROI so the above is not an indication of a bad alignment nor a too large hand drawn ROI.

Further use of the probability map
Besides using the probability map for visual inspection, it can be used in the partitioning of the images. The information can be used as a priori knowledge, so it takes less convincing to identify a pixel near the upper border as a lesion.
The partitioning is carried out on the original images (unaligned) though it demands that the outline of the tibia is identified in order to use this information. This could be done using Active Appearance Models (AAM) [9, 13] or the simple solution by using the ROIs which probably have to be smoothed first.
The probability map is based on only 47 images (all with lesions) which is a small sample for a

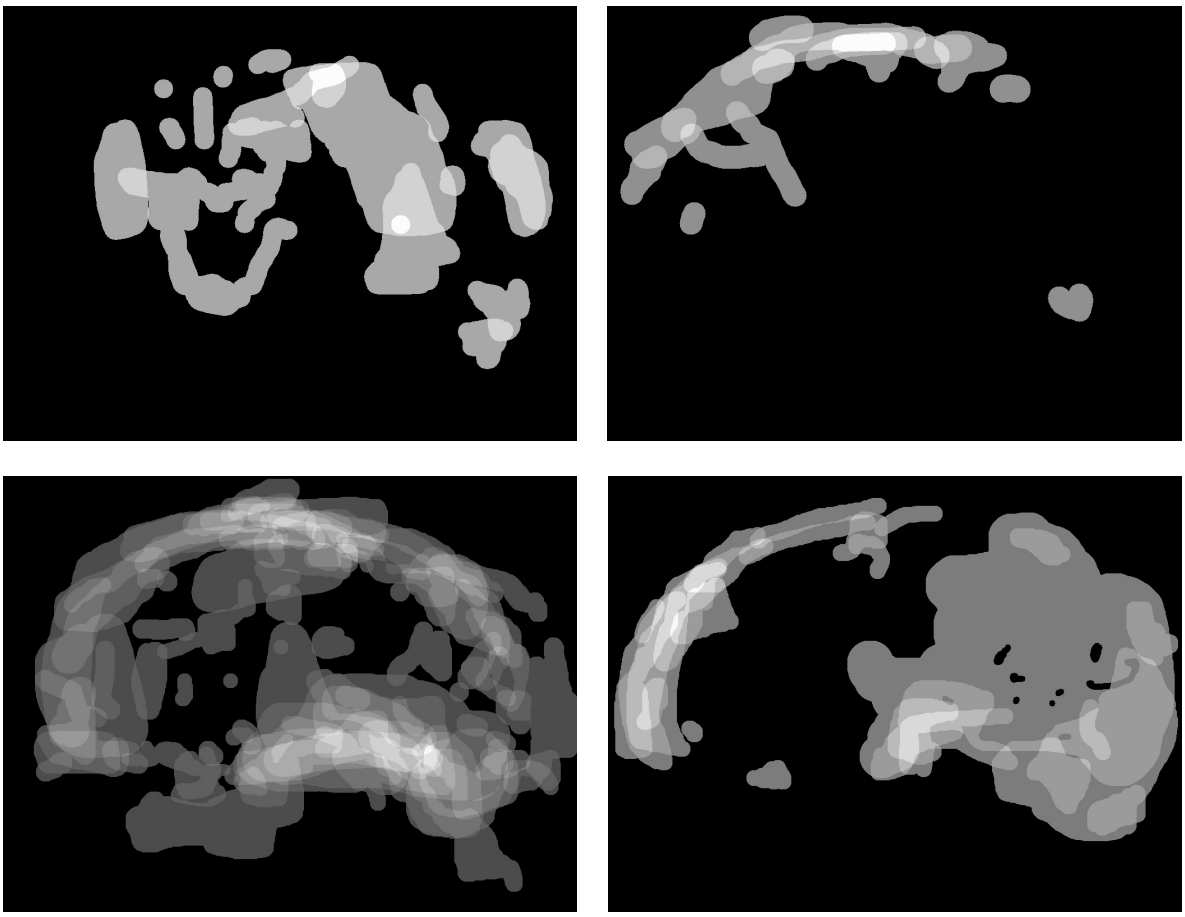


Figure 5.1: The manually marked lesions merged by age. Top row: Lesions for the 5.71 and 6 week old mice. Bottom row: Lesions for the 12 and 22 week old mice. Note that the images are scaled individually so the area with most lesions appears white.

good empirical basis of a priori knowledge.

Correlation with age

The area of the manually marked lesions is calculated for each aligned image. Both the lesion area within the hand drawn mask and the total marked lesion area are calculated and correlated with age.

For the areas within the mask, the result is a correlation at 0.27 with a level of significance at 0.063. For the total lesion area, the correlation with age is 0.36 with a level of significance at 0.014.

The correlations are relatively low, with respect to the main hypothesis of the project. The fact that the correlation is largest, when the entire area is used, must be because there is more lesion outside the ROI, the older the mice. The relatively large amount of bright lesions in the study, containing the youngest mice, is thus reduced.

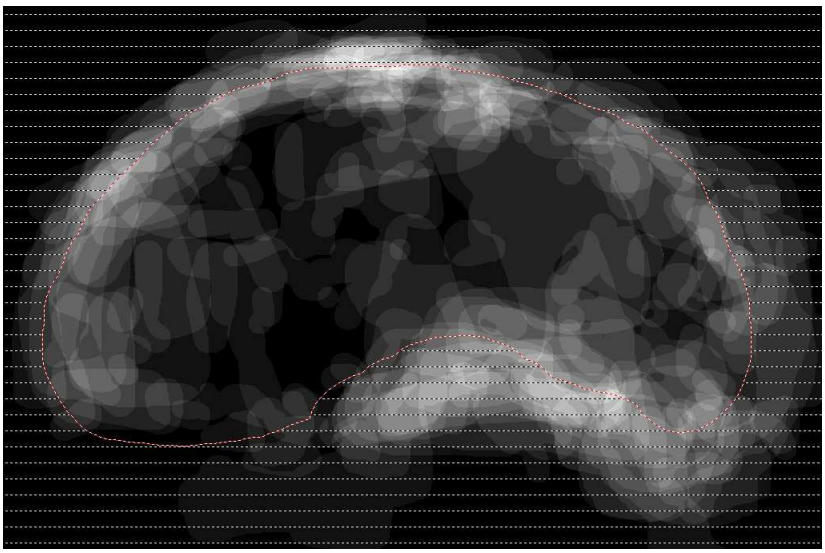


Figure 5.2: Probability map based on all the lesions in the aligned images. There are between 0 and 15 lesions in the same position (out of 47 images containing lesions). The image is scaled to full dynamics of the gray level scheme. Note that the bright area below the ROI should not be considered a lesion.

Correlation with biomarker and histology
The relative lesion area is also correlated with measures from biomarker and histology. For study STR1N-09 the correlation with histology is -0.39 and for study STR1N-12 it is 0.10 with levels of significance at 0.29 and 0.80, respectively. A negative correlation means that when the relative amount of lesion increases, the OA grading (using histology) decreases, which is worrying. The correlations are not significant and hence there is no problem.

For correlation, with the biomarker, of study STR1N-09 the result is -0.30 with a level of significance at 0.43. Here the correlation is allowed to be negative. This means that the active matter in the biomarker is decreasing, the worse the OA gets, e.g. if the turnover has topped and afterwards declines. The correlation is not significant hence nothing is concluded from it.

Scatterplots of the relative areas vs. the biomarker or histology (not shown) reveal no sign of correlation.

5.3 Examination of Different Tibia Areas

Here an alternative way is used to examine the hypothesis, "*the bright / white lesions are more likely to appear near the border of the tibia than in the middle of it*", and to estimate the mean value at different positions.

The aligned images are mean filtered with different window sizes (5×5, 10×10, 15×15 and 30×30 pixels). The 10×10 mean filter is chosen because it keeps a good contrast, but the images are less peaked than without filtering.

After alignment and filtering, the mean, minimum, maximum and variation are found for each position (pixel) in each study. The same four features are also found for each of the four ages (studies merged by age).

The Images Merged by Age
Figure 5.3 shows the four features merged for the 12 week old mice. See also Appendix D for the results of the other ages. The images for each study are not shown.

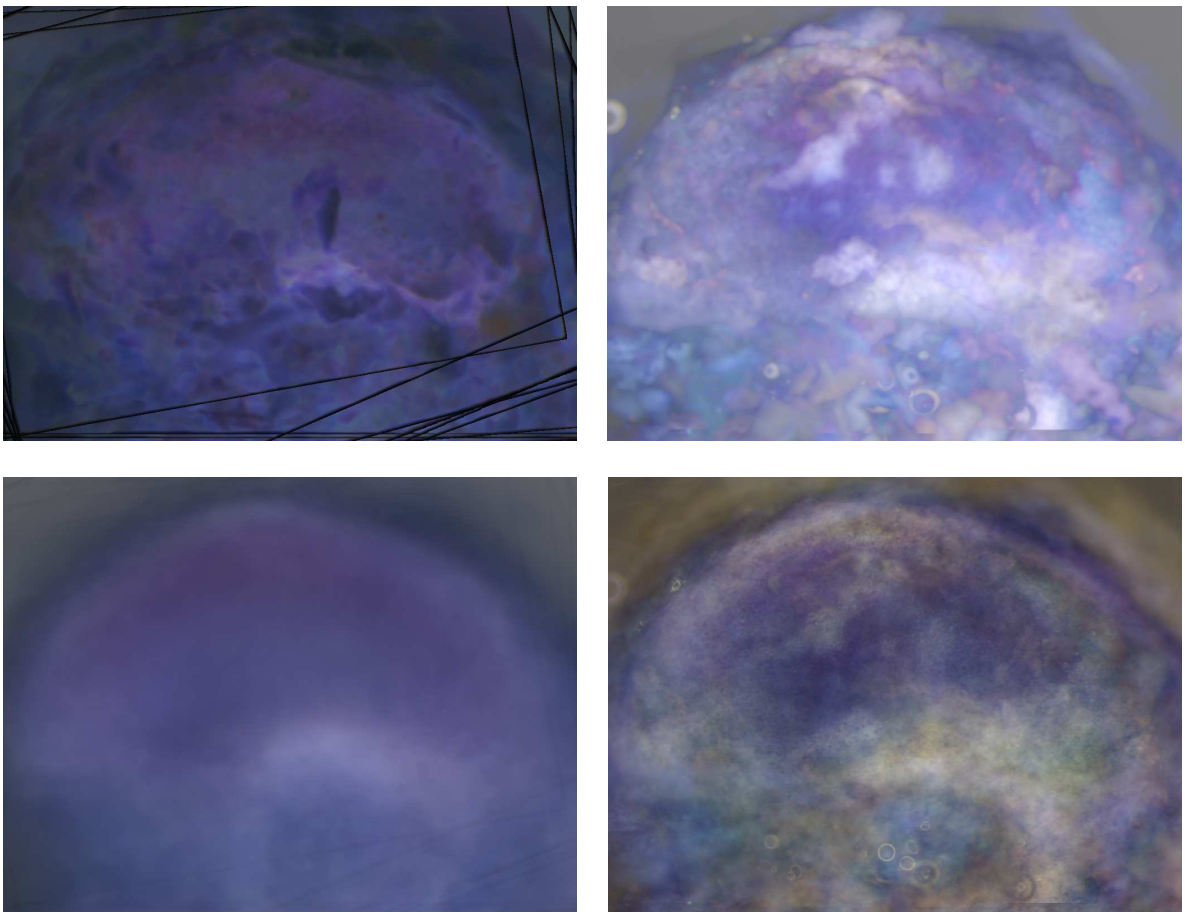


Figure 5.3: The aligned and merged images of the 12 week old mice. Top row: Minimum and Maximum image. Bottom row: Mean and Variation image. Note that the images are brightened individually for a better perception and that the lines in the images are due to an artifact of the alignment.

The Mean images get brighter, the older the mice are. Except for the youngest study, there is a tendency that the borders of the tibia are brighter or more purple than in the middle of it, especially in the top half of the tibia. There is little difference but it is more visible on the print out, than on the PC screen. The youngest study is brighter in the right side of the image due to the domination of a single image.

The Minimum images get brighter, the older the mice are. There are some dark spots in

the images, except for the 22 week old mice. The dark spots are close to the image border and could be from coarse cartilage, that is normally outside the ROI.

The Maximum image for the 12 week old mice is mostly bright on the top of the tibia (the bright spot on the bottom of the image should not be considered a lesion). The 6 week old mice's images have a tendency to be brighter or more purple around the border than in the middle. The 22 week old mice support the tendency, but to a less degree. For the 5.71 week old mice the tendency is, if any, the opposite.

The Variation images show that the 6 and 12 week old mice vary more near the border than in the middle and that all bands vary alike. In the middle, the images are more blue and purple (less variance in red and green or in green band, respectively). For the 5.71 and 22 week old mice, the bands variate unequally and cannot be generalized.

The tendency, that areas near the border become more purple with age, indicates that purple could be an early stage of OA.

The Merged Images by Study
Studies with 6 and 12 week old mice each appear similar to their respective merged images, except the tendencies are less obvious. Studies with 5.71 and 22 week old mice are the same as those merged by age because they each consist of only one study.

5.3.1 Point Development

The tendency of a brighter and / or more purple border compared to the middle, is tested here by extracting samples from different positions in the aligned images. The first three trials each have five sample positions, two are positioned in the middle of the ROI (to the left and to the right of the middle) and three on the border (on the left side, in the middle and the right side), see Figure 5.4.

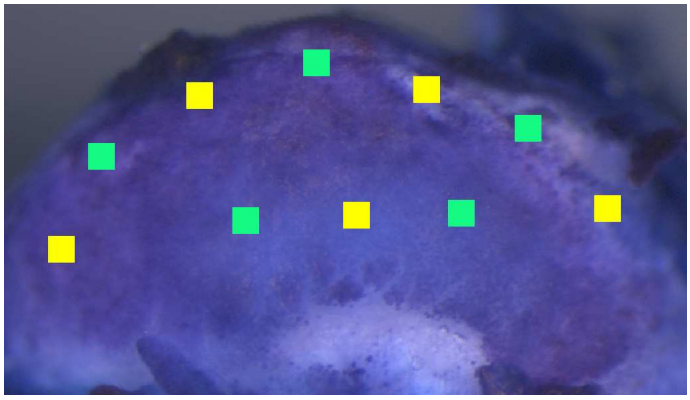


Figure 5.4: The sample positions. The green areas are the five initially used points and all the points together represent the points for the fourth trial.

The idea is to make an analysis of variance to obtain the simplest model, that describes the

development of the intensities. As a start, the sample values are shown developed by age and by position. The points are samples from the aligned images and are each the mean of a 40 × 40 pixels window, to avoid noise and local variation. Figure 5.5 and 5.6 show the samples, averaged by study, according to age and to position, respectively.

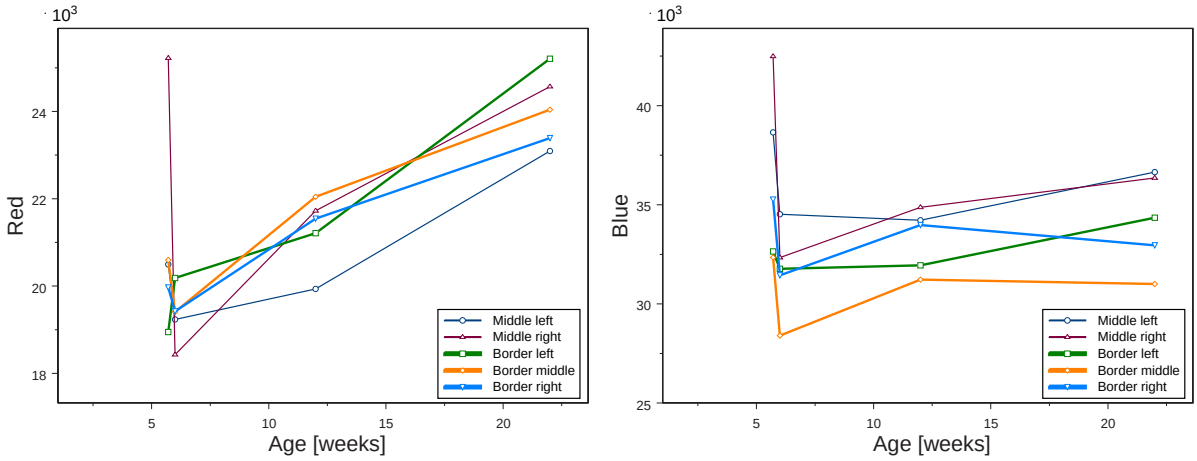


Figure 5.5: The samples shown according to age. Note that the y-axis' intensity is not the same from plot to plot.

Figure 5.5 shows that the intensities increase with age for the red band, except the fact that the 5.71 week old mice have some intensities higher than those from the 6 week old mice and a single value above all the others. The plotted values are in average over nine or more images, so a single outlier cannot affect the results that much.
For the blue band, there is a much smaller tendency of increasing intensities with age, where the 5.71 week old mice have higher intensities than all the other ages. The green band (not shown) behaves somewhat in between the two shown bands.

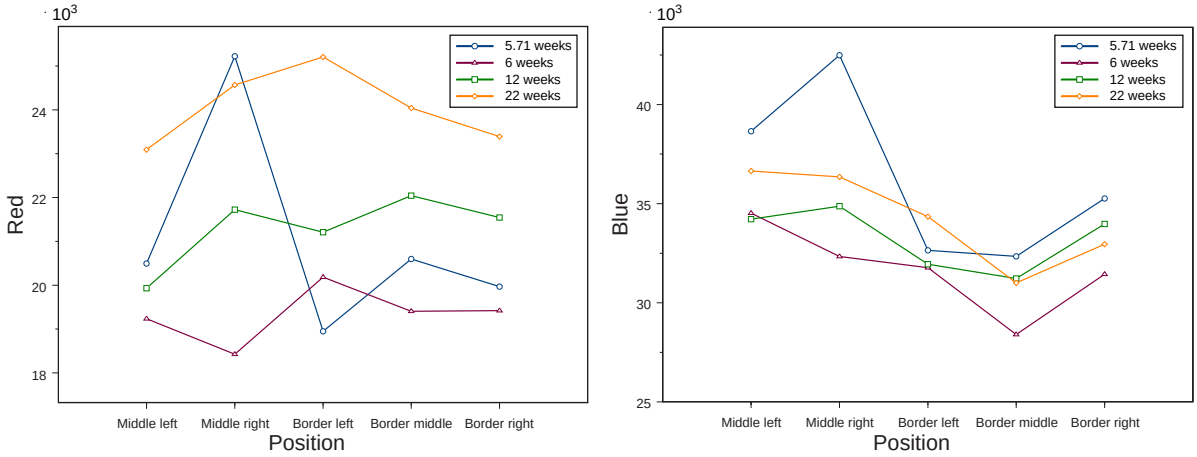


Figure 5.6: The samples shown by position. Note that the y-axis' intensity is not the same from plot to plot.

Figure 5.6 shows the samples according to position and thus it is expected to see a difference in the levels of the first two points compared to the last three points. The blue band shows this tendency, but the intensities on the border are generally darker than those in the middle of it, which is the opposite of the expected. For the red band, the 6 and 12 week old mice show an increasing intensity on the border, while the other ages do not. The plot of the green band (not shown) is like the blue band (but with lower intensities).
The results are surprising and two other sets of points are extracted to verify the results. The first new set is positioned just a little different than the first ones, and in the second new set, the positions near the border are moved a little towards the middle (away from the background). The results are basically the same for the three trials; green and blue decrease while red increases near the border for the 6 and 12 week old mice. The resulting color is more purple but not brighter near the border compared to the middle.
The bands all together result in falling intensities near the border. A closer look at the Mean images of each study shows that the area between the middle and the border is slightly darker than the other areas, probably making the border seem brighter, than the middle.

A fourth trial is carried out with 10 points, three in the middle and seven near the border. Their positions can be seen in Figure 5.4. The result, shown in Figure 5.7, is that the blue intensity differs depending on the different positions near the border. It has a lower intensity near the top of the tibia than on the sides of it. The red and the green band show more diffuse behavior. The points near the border might therefore not be pooled when analyzing the variance.

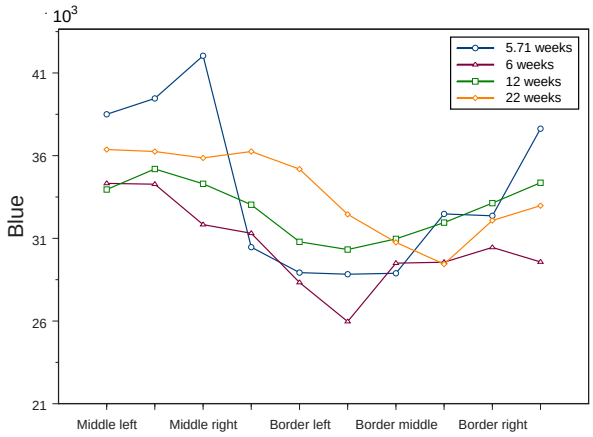


Figure 5.7: The samples of the fourth trial. The blue band is shown by position.

A fifth trial is carried out to be sure of the decreasing intensity near the upper border compared to the middle of the tibia. The sample positions are selected using the probability map generated in the previous section. The sample positions near the border are selected there where the largest concentrations of lesions are. In the middle, the sample positions are located where there are no lesions at all or only where there are a few ones, according to the probability map. The positions can be seen in Figure 5.8 and the resulting intensities of the green and blue bands are shown in Figure 5.9.
The test, with sample positions so near the border that it, for a few images, might tangent the border, does not change the fact that the images are brighter in the middle of the tibia than

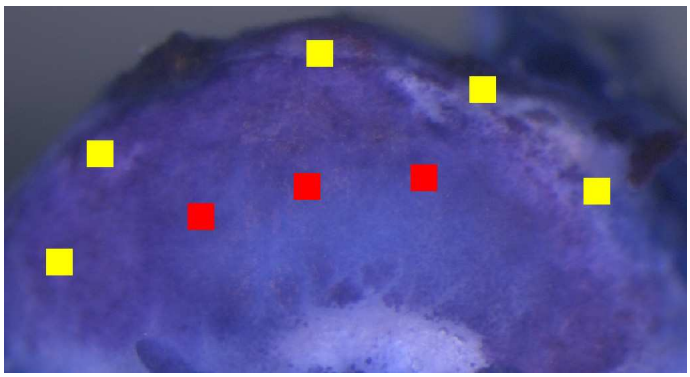


Figure 5.8: The sample positions for the fifth trial. The red areas mark the samples in the middle of the tibia and the yellow areas mark the samples near the border.

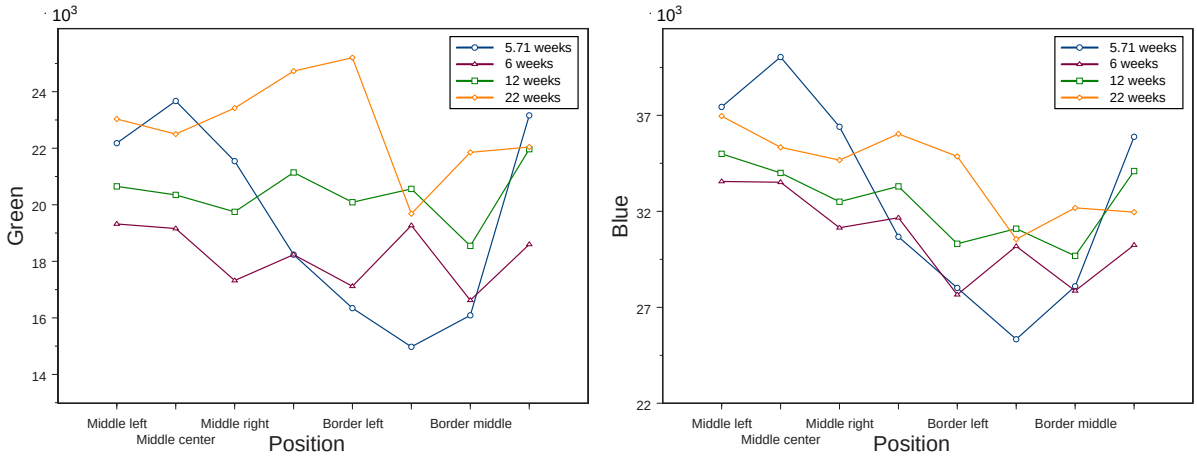


Figure 5.9: The samples of the fifth trial shown by position. Note that the y-axis' intensity is not the same from plot to plot.

close to the border.
The 5.71 week old mice support the tendency of darker samples in the middle of the border compared to the left and right side of the border. This tendency is not the case for the rest of the ages.

5.3.2 Modelling of the Intensities by Position, Area and Age

An analysis of variance is carried out in order to obtain a statistical foundation of the development of the intensities in the different positions of the tibia. The purpose is to find whether or not the intensities differ between the positions, between the areas and / or between the ages. The analysis of variance model includes the dependent variables Age, Study, Image, Area and Position and their mixed effects. Area is either the border or the center of the tibia, and Position represents the samples within these areas.

The model consists of the systematic effects (mixed effects) and random effects. The model, and

the call in SAS using PROC MIXED, can be seen in Chapter 11.2, p. 105.
Besides the level of significance of each systematic effect, the distance between the estimated mean values of Position is also calculated for every combination. The model reveals whether or not the samples near the border should be pooled.

The test is carried out for each of the three bands and for the first, the fourth and the fifth trial with five, ten and eight sample positions, respectively.

Analysis of Variance with 5 Positions
For the red band, there is no significant effect. The variable which is closest to significance, is Age with a level of significance at 0.53. Reducing the model, by repeatedly removing the least significant of the systematic effects, ends up with only Age and a level of significance of 0.52. Hence there is no significant development to the Area, the Position nor the Age etc., for the red band.
A look at the estimated mean values of each position reveals that the samples near the border have a small variance, but the samples from the middle have one value above and one value below them. The result is that the mean of each of the two areas are alike.

The green band shows that Area is the only significant variable with a level of significance at 0.071. It becomes more significant as the model is reduced and ends at 0.048. The estimated mean values for each position near the border vary more than the red band, but they are all below the estimated mean values from the middle of the tibia. The largest difference between the estimated mean values is between the "middle right" and "border middle".

For the blue band the Area is significant (0.019) and Position(Area) shows a tendency (0.11). Reducing the model results in keeping both Area and Position(Area) in the model with levels of significance at 0.0051 and 0.059, respectively. Thus there is a difference between the samples near the border and the samples in the middle of the tibia, and also between some of the samples within these areas.
Looking at the differences between the estimated mean values shows that the samples from near the border have two alike values and one further away, while the samples from the middle are alike. It must therefore be the samples from near the border that show significant difference. It is "border middle" that has a lower intensity compared to the other near-border samples. The difference between the estimated mean values is larger between the two areas than between the border samples.
Again, all the estimated mean values from near the border are below the ones in the middle.

Resume
The results for the different bands are gathered in Table 5.1. It reveals that Area is the most

Band	Effect	Level of Sig.	Effect	Level of Sig.	Mean middle	Mean border
Red					21,587	21,496
Green	*Area*	0.048			21,810	20,146
Blue	*Area*	0.0051	*Position*(*Area*)	0.059	36,261	32,746

Table 5.1: Summary of the analysis of variance for the first trial with five sample positions.

significant effect when explaining the development of green and blue. For the blue band, the samples near the border also show significant difference between themselves. It is the "border middle" that has the lowest intensity among them. This position is where there are most lesions according to the probability map.
It also shows that green is larger than red in the middle of the tibia and viceversa near the border, where the blue has also decreased. Near the border, the appearance is hence darker and more purple than in the middle.

A closer look at the Mean images, reveals that the tibia near and on the border becomes darker as it begins to curve. This tendency is also found for the unaligned images and it begins within the defined ROIs. Hence when there is no lesion, the images are darker here than the average image intensity. When added to images with lesions, these areas are still not brighter than the ones in the middle of the tibia.

Analysis of Variance with 10 Positions
The red and green bands have no significant systematic effects. For the blue band, Area and Position(Area) are significant (0.0009 and 0.079, respectively) when the model is reduced. This means that there is a highly significant difference between the middle and the near border and that there is a significant difference within one or both areas.

Analysis of Variance with 8 Positions
These sample positions are closer to the border and positioned where there are most lesions according to the probability map.

For the red and green bands, the differences between the two areas are small and non-significant due to the differences within each area (Position(Area)), which again are significant (0.09 and 0.08 for red and green, respectively). For the blue band, the difference between the middle and the border is significant (0.01) and the estimated mean values are 34,305 and 31,070, respectively.

Joint Resume
There is only one significant effect describing the development of the red band which is for the fifth trial and the positions within the different areas. The same is the case for the green band which also has Area as significant for the first trial. The blue band has Area as significant for all the trials and also Positions(Area) for the first and fourth trial.

All the trials show that the border is more purple and darker than in the middle which is more blueish. The border looks brighter than the middle, but it must be visual delusion.

The relationship to age is not near significance for any of the tests nor bands. Figure 5.5 showed an increasing intensity with age, but the values from the middle and from the border increase with approximately the same amount and hence there is no development between the two areas according to age. The fact that Age itself is not significant must be because the intensities for the 5.71 week old mice do not follow the trend.

5.4 Examination of Purple Lesions

One of the observations during the project is that the purple areas might not be healthy areas but an early lesion stage. The corresponding hypothesis is "*the purple areas in the images represent an early lesion stage*" and will be rendered probable by rendering probable the hypothesis "*the purple areas are more likely to appear next to bright / white lesions than in the other positions of the tibia*".

If the hypothesis is true, then in the near neighborhood of bright lesions, it should be more likely to find purple than further away.
Even if the purple area is an early lesion stage, it can also emerge in images without bright lesions or far from such one. It is then, these purple areas that will become the bright lesions with time. For these reasons, only images with bright lesions marked, are used for the test and mainly with focus on the local areas around the bright lesions. Only pixels within the hand drawn ROI are used.

The amount of purple pixels is calculated with respect to the distance to the bright lesions. The distance measure is calculated by using the manually marked bright lesions.
Purple is normally a mix of blue and red. In these images, blue always has the largest intensity and green and red are alike. This could mean that there are no purple pixels, but this is not the case. Samples have shown that a pixel appears purple if red is larger than green. This is of course based on the author's perception of purple. This red vs. green relationship can also occur in the bright lesions but their intensity is higher and easily distinguishable. Working with the marked lesions, removes the problem.
Purple is also defined from trichromatic red, which is the red intensity divided by the sum of intensities for each pixel. A more thorough description can be found in Chapter 7.1.1, p. 59. Purple is defined as "trichromatic red $>$ 0.27" meaning that 27 % of the total intensity is red. The threshold is based on samples from the images.

An euclidian distance map is calculated, for each image containing lesion, by using "The Signed Euclidean Distance Transformation" (SEDT) [7]. For each distance between 0 and 200 pixels, in bins of 10 pixels, the amount of purple pixels compared to the total amount of pixels in the bins is calculated.

The SEDT is calculated by recursive filtering with four filters. This is a much faster way of obtaining each pixel's distance to the nearest lesion pixel, than calculating its distance to each of the lesion's pixels or just to the pixels on the border of the lesions.
The mean of each purple measure in each bin is also calculated, and defined as the mean of "red - green" and as the mean of trichromatic red.
The different results are merged by study and shown in Figure 5.10. It shows that whether the measure uses the percentage of purple pixels or the mean of the purple measure, the results are alike for both purple measures.
The measure "red-green $>$ 0" increases with the first distance from the lesions, it is then more or less constant and then it decreases again with the distance from the lesions. The fact that the measure increases with the distance near bright lesions could be due to the fact that only "well defined" bright lesions are marked and hence there is a zone between the marked lesion and the more homogenies purple. The purple measure is above average for seven of the eight

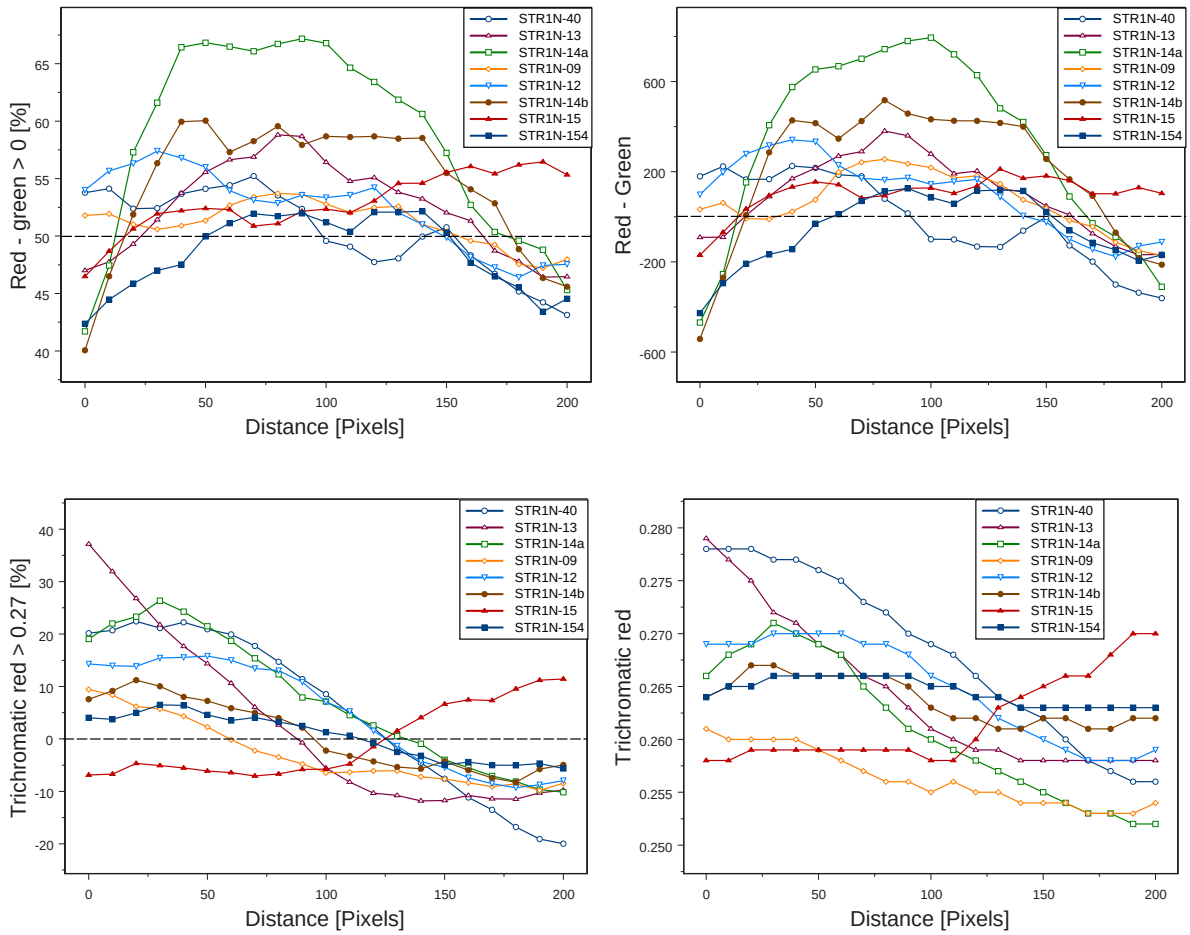


Figure 5.10: The purple measure against distance to the nearest bright lesion pixel. Top row: Purple defined as "red-green $>$ 0". Left: Percentage purple pixels in each distance bin. Right: The mean of the purple measure for all pixels in each bin. The dashed line in the plots is the mean purple of all non-lesion pixels. Bottom row: Purple defined as trichromatic red $>$ 0.27. Left: Percentage purple pixels over average in each distance bin. Right: The mean of the purple measure for all pixels in each bin.

studies within less than 25 pixels from the lesions.
For some studies, the trichromatic red decreases with distance from bright lesions and for other studies it increases a little first and then decreases. It looks even more convincing than the first purple measure. The mean amount of purple in the whole image is subtracted hence values larger than zero are above average.
Study STR1N-15 does not follow the tendencies for both measures. It is almost constant or increasing with the distance to the bright lesions.
Further from the bright lesions, more than 200 pixels away, the measures do not behave alike. Some increase and some decrease. At this distance from the bright lesions, the probability of purple is hence back to "random" / normal.

The amount of blue is also calculated with respect to the distance from the nearest lesion pixel. Here blue is defined as "trichromatic blue $>$ 0.48". The threshold is based on samples. The result can be observed in Figure 5.11.

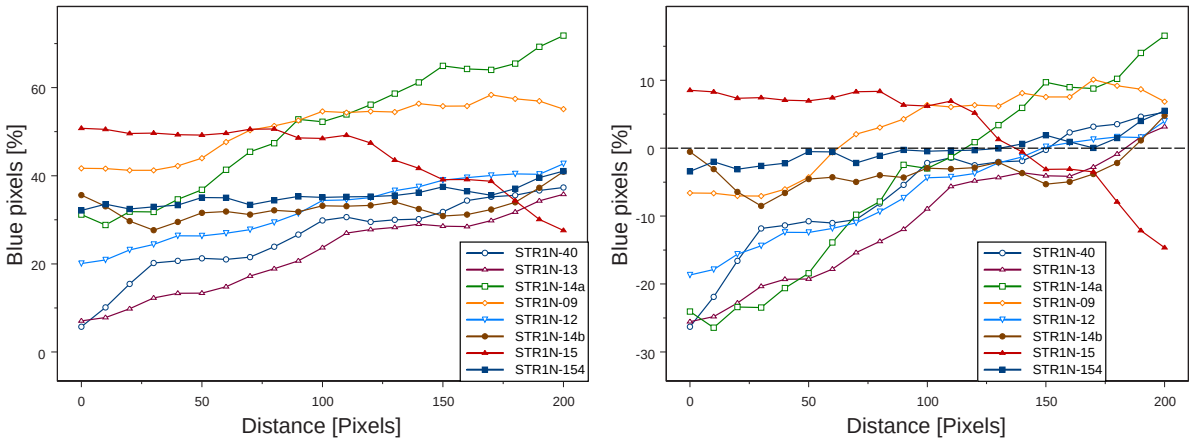


Figure 5.11: Blue measure, defined as trichromatic blue $>$ 0.48, against distance to the nearest bright lesion pixel. Left: Percentage blue pixels in each distance bin. Right: Percentage of blue pixels over average in each distance bin.

Figure 5.11 shows the percentage of blue pixels with respect to the distance to bright lesions. In the right plot the mean amount of blue pixels in each study is subtracted. The plots show that the blue, as it is defined here, increases with the distance from lesions in all the studies, except STR1N-15. Some of the other studies increase a bit while others are doubled or tripled, from next to the lesion and to positions 200 pixels away. They all start below their study average and end above it except, again, STR1N-15 which shows the opposite characteristics.
Blue increases with the distance and if it is assumed that there are only three types of pixels, and the bright lesions are marked (and left out), then the amount of purple must decrease when the amount of blue increases. This is not entirely true because the blue and purple definitions do not define / use all pixels.

Study STR1N-15 does not follow the tendency and a look at the images shows that they contain a lot of small areas, where the choice of whether or not they are lesions, is impossible. The bright lesions marked are the certain ones which for this study might leave to many unmarked lesion areas.

5.5 Conclusion

72 images, 9 from each study, are aligned and hence corresponding positions on the tibias can be studied. The ROIs masks are not rotated, therefore a common ROI is hand drawn.
The bright lesions in the aligned images are manually marked in order to examine their behavior according to position, age etc.

A probability map is generated describing the possibility of bright lesions emerging at each position of the tibia. It clearly shows that they are more likely to appear on and near the upper

border of the tibia, than in the middle of the tibia. The hypothesis "*the bright / white lesions are more likely to appear near the border of the tibia than in the middle of it*" is hence rendered probable.
The relative amount of bright lesions marked in each image is compared to age and this results in a correlation at 0.27 with a level of significance at 0.063. The entire lesion area (including the area outside the ROI) is also compared to age which gives a correlation at 0.36 with a level of significance at 0.014. These are rather low correlations but they are significant.

The aligned images are merged by age and show that the amount of purple near the border increases with age, which is the first indication that purple is not a healthy area but a lesion area in an early stage. The images also seem brighter near the border but the fact is, that the summed intensity is lower here than in the middle of the tibia.
Other tendencies according to age are hard to find and the 5.71 and 6 week old mice, which are two days apart, again show large differences.
Analysis of variation shows that there are no significant effects according to the red band, but that green and blue are significantly lower near the border of the tibia than in the middle of it (which gives the purple appearance). The blue band also shows that the different positions near the border of the tibia differ significantly.
The hypothesis "*the bright / white lesions are more likely to appear near the border of the tibia than in the middle of it*" could not be rendered probable using samples of the aligned images.

Examination of purple as a lesion stage results in decreasing purple with the distance to bright lesions, when purple is defined from trichromatic red. When purple is defined as red-green > 0, it first increases with the distance, then flattens out and at last it decreases with the distance. The amount of blue pixels increases with distance. Study STR1N-15 does not behave according to these observations.
The results show that purple is more likely to appear near bright lesions and hence might be an early lesion stage. The hypotheses "*the purple areas in the images represent an early lesion stage*" and "*the purple areas are more likely to appear next to bright / white lesions than in the other positions of the tibia*" are hence rendered probable.

Chapter 6

Defining, Labelling and Analysis of Classes

In this chapter the images, studies and different types of areas will be examined by defining eight different classes in the images. Data from the classes are used for comparing the classes to each other and between images and studies, using scatterplots, boxplots, statistics etc.

6.1 Data Classes

From the eight used studies there are 126 usable images. Twenty of these (two - three from each study) are partly labelled so that areas which can be identified are colored according to their class. See Figure 6.1 for an example. The RGB values are extracted according to these classes and constitute a data set. Due to S-Plus' instability concerning moderate amounts of data, only 100 randomly chosen pixels from each class in each image are used. The data set is the basis for most of the following tests and figures throughout the report.

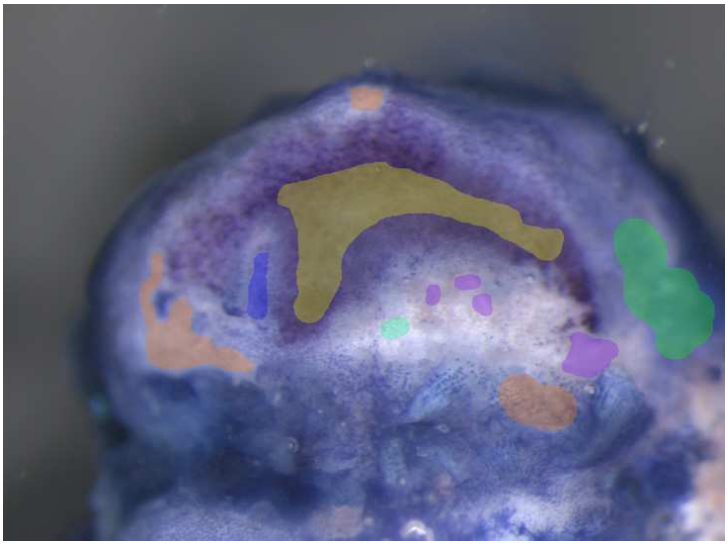


Figure 6.1: Labelled image example (STR1N-15-110). The label colors are transparent so that a bit of the underlying structure is visible.

The following eight classes are identified

1.	*Healthy blue*	Homogeneous dark or middle blue.
2.	*Healthy blue pattern*	Fast varying (fibrillated) between dark and middle blue.
3.	*Lesion purple*	Homogeneous dark or middle purple, often around a *Lesion white* or a *Lesion purple bright* area.
4.	*Lesion purple pattern*	Fast varying (fibrillated) between dark purple and middle purple, often around a *Lesion white* or a *Lesion purple bright* area.
5.	*Lesion purple bright*	Homogeneous bright purple area, often inside a purple area.
6.	*Lesion perhaps*	Areas that are varying between bright blue and white which have doubtful relations for the untrained eye.
7.	*Lesion white*	Homogeneous white or bright or light gray areas.
8.	*Lesion yellow*	Homogeneous yellow or brownish areas. Normally small areas within *Lesion white*.

The labelled images are not the same as for the bright lesions examinations in Chapter 5, though there might be repetitions.
The classes are hard to define precisely and the labelling are done iteratively by going back to see the interpretations in an earlier labelled image and maybe adjust it or the current one. Actually, the labelling has been changed twice before this stage.
The first time very little knowledge was at hand and after going through this chapter's tests and having several looks at the images, it resulted in a checking of the labels which led to a lot of changes.
This second trial removed a class (*Lesion pattern*) which was only present in one image. The labels were marked a little wider in comparison to the first where only "clean" classes were marked. After going through the tests again, it was discovered that the purple areas are not a part of the healthy classes, but an early stage of lesion. The purple areas are often found around a lesion area. This did not have to change the label definitions much, but in most of the images there are doubts about whether areas should belong to a purple class or a blue class. With the new knowledge these areas labels depend somewhat on the presence of lesions. At the same time the class *Lesion purple bright* is added.
Some of the tests with the previous labels are shown in Appendix E.

The earlier mentioned dark spots and brownish areas are to diffuse or in too few images to become classes.
The 20 labelled images are not chosen randomly but selected by two criteria; each image should contain several classes and the chosen images from a study should together represent as much different appearance and different classes as possible.
The eight classes are used as they are, but they are also merged so that the two healthy classes become the class *Healthy* and the two "certain" lesion classes (*Lesion white* and *Lesion yellow*) becomes *Lesion*. Merging more of the lesion classes will give too much variation in the class (the intensities of purple is near those of blue) and besides it is still interesting to show that the purple areas are in fact a lesion stage.
The reason for doing both is that the eight classes can have more specific definitions and can later be merged directly or weighted. Using the merged classes, overlapping healthy classes will

not appear as misclassifications nor in other ways turn up as poor results. Likewise with the lesion classes.

6.2 Scatterplots of Classes

The data set, the samples from the labelled images, are shown as 3D and three times 2D in Figure 6.2.

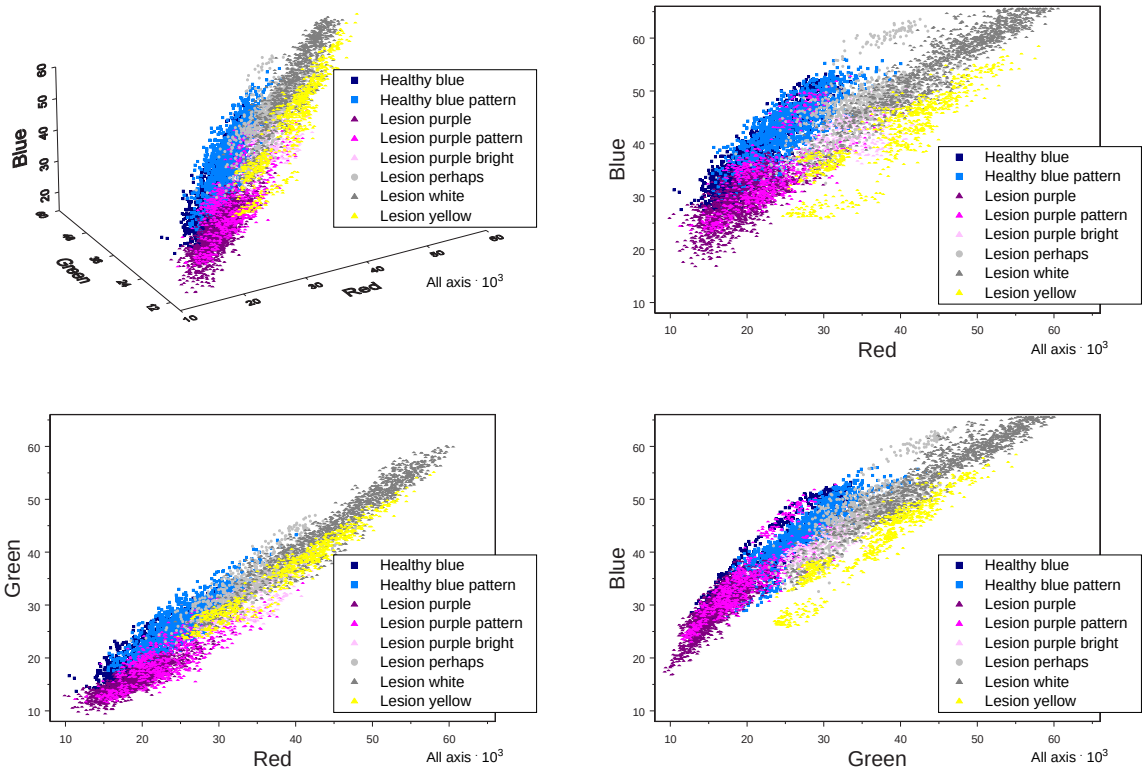


Figure 6.2: All the classes from all the labelled images. Note that S-Plus draws one class at a time hence concentrated points in one class can hide other classes.

The plots show that there are tendencies of grouping, but especially *Healthy blue pattern* and the non purple lesion classes have relatively large variations. The purple classes are mixed with the healthy classes, but are generally more red and less blue. The other lesion classes are generally brighter or at least have higher red and green intensities than the blue classes.
The slant oblong shape means that the colorbands are highly correlated, especially the red and green values which, further more, have alike values.
The plots also show that the blue colorband is saturated in at least one image.

6.3 Summary of the Classes

Summary statistics are shown in Table 6.1 for the data set. It shows that each class has different center locations when using at least two colorbands. The distances from the blue mean to the red and green mean are larger for the healthy classes than for the lesion classes, which gives the

		Red		Green		Blue	
Class	Pixels	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
Healthy blue	1488	20842	4253	21472	4987	37610	6412
Healthy blue pattern	1256	25399	4196	26926	4640	42918	5524
Lesion purple	1509	19189	3533	15869	2729	28772	4030
Lesion purple pattern	765	23191	4991	19791	4532	34450	6361
Lesion purple bright	370	33413	3535	30133	3461	42148	3139
Lesion perhaps	779	35177	4924	35371	4954	47858	6137
Lesion white	1268	45404	7526	45070	8030	54797	7410
Lesion yellow	756	39152	7323	36466	7086	41852	7561

Table 6.1: Summary statistics for each class.

blue appearance. For the purple classes the distance from the blue mean to the red is smaller than for the other classes. The purple classes are generally darker than the blue classes which again are darker than the rest of the lesion classes.
The pattern classes generally show less variation than their corresponding non-pattern classes. *Lesion white* and *Lesion yellow* together have the largest and second largest variance for all the bands. These observations are the opposite of what could be expected, after looking at the images of the tibia, but it was earlier observed that bright lesion shows a large variation. Adding areas together across images also increases the variance.

To test if the class centers are equal a two-sided Kruskal-Wallis rank sum test is carried out. The test values (Kruskal-Wallis chi-square) for the three bands are between 5200 and 6500 which results in p-values of zero (with S-Plus' precision) and are thus highly significant. At least one of the class means is therefore significantly different from the others, which is not surprising after examining the scatterplots and due to the amount of data.

6.4 Scatterplots of Classes in each Image

The classes are plotted for each image separately, to see if the different classes are more isolated by looking at one image at a time. Figure 6.3 shows two examples and all 20 plots can be seen in Appendix F. Figure 6.3 shows that the classes are less mixed than for the entire data set. For

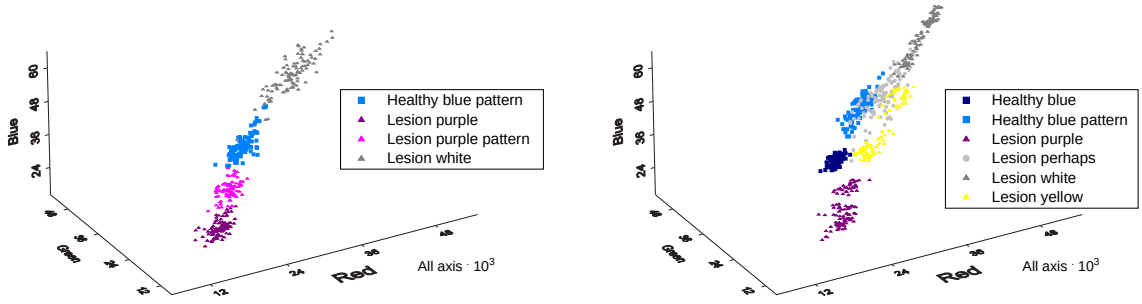


Figure 6.3: The classes shown for one image at a time. Left (STR1N-13-110): One of the best separable. Right (STR1N-15-102): One of the worst separable.

the STR1N-13-110 they are almost separable, but just next to each other. For STR1N-15-102

Lesion perhaps binds several classes together, which is expected for this class. Removing it results in almost isolated classes. The rest of the plots show that the two healthy classes often are totally mixed. Several of the classes are positioned at the same relative position to each other from image to image. This is an interesting observation and shows that for instance an intensity invariant measure or a color transformation should be used to separate the classes. *Healthy blue pattern* and *Lesion yellow* do not always follow this tendency.

6.5 Summary of each Class in each Image

It is shown above that the grouping tendencies are larger for a single image than for the merged ones. Therefore the standard deviation for each class in each image is calculated and then averaged for each class. This way it can be observed if the variation is smaller when looking at each image separately. Table 6.2 shows that the standard deviation is reduced to less than half for

		Red		Green		Blue	
Class	Pixels	Std($\sum$)	$\sum$(Std)	Std($\sum$)	$\sum$(Std)	Std($\sum$)	$\sum$(Std)
Healthy blue	1488	4253	2070	4987	1882	6412	2123
Healthy blue pattern	1256	4196	2597	4640	2307	5524	2251
Lesion purple	1509	3533	2221	2729	1701	4030	2565
Lesion purple pattern	765	4991	2645	4532	2160	6361	2515
Lesion purple bright	370	3535	2757	3461	2261	3139	1911
Lesion perhaps	779	4924	4029	4954	3561	6137	3036
Lesion white	1268	7526	3774	8030	3713	7410	3342
Lesion yellow	756	7323	2864	7086	2624	7561	2579

Table 6.2: Summary statistics for each class. The earlier shown standard deviations for each class ($Std(\sum)$) and the average of the standard deviation from the separate images ($\sum(Std)$) for each class.

some of the classes, hence merging a class across the images will result in larger variation. Now it is *Lesion perhaps* and *Lesion white* that have the highest standard deviations. The pattern classes now have a larger variance than their corresponding non-pattern classes, as expected.

6.6 Boxplots of the Classes

The classes are somewhat overlapping and to see the problem and the merged classes compared to the separate ones, boxplots of the data set are shown in Figure 6.4.
It shows that red and green separates the merged classes *Healthy* and *Lesion* better than blue, but they still overlap.
All bands supply information, but the red and green band look, again, very similar. At least blue and then red or green should be used to obtain a slight possibility of separating the classes. The challenge observed in the scatterplots is mainly to separate the healthy blue classes from *Lesion purple* and *Lesion purple pattern*. *Lesion purple* might be separated using the green and blue channel while *Lesion purple pattern* mix pretty much with *Healthy blue*.

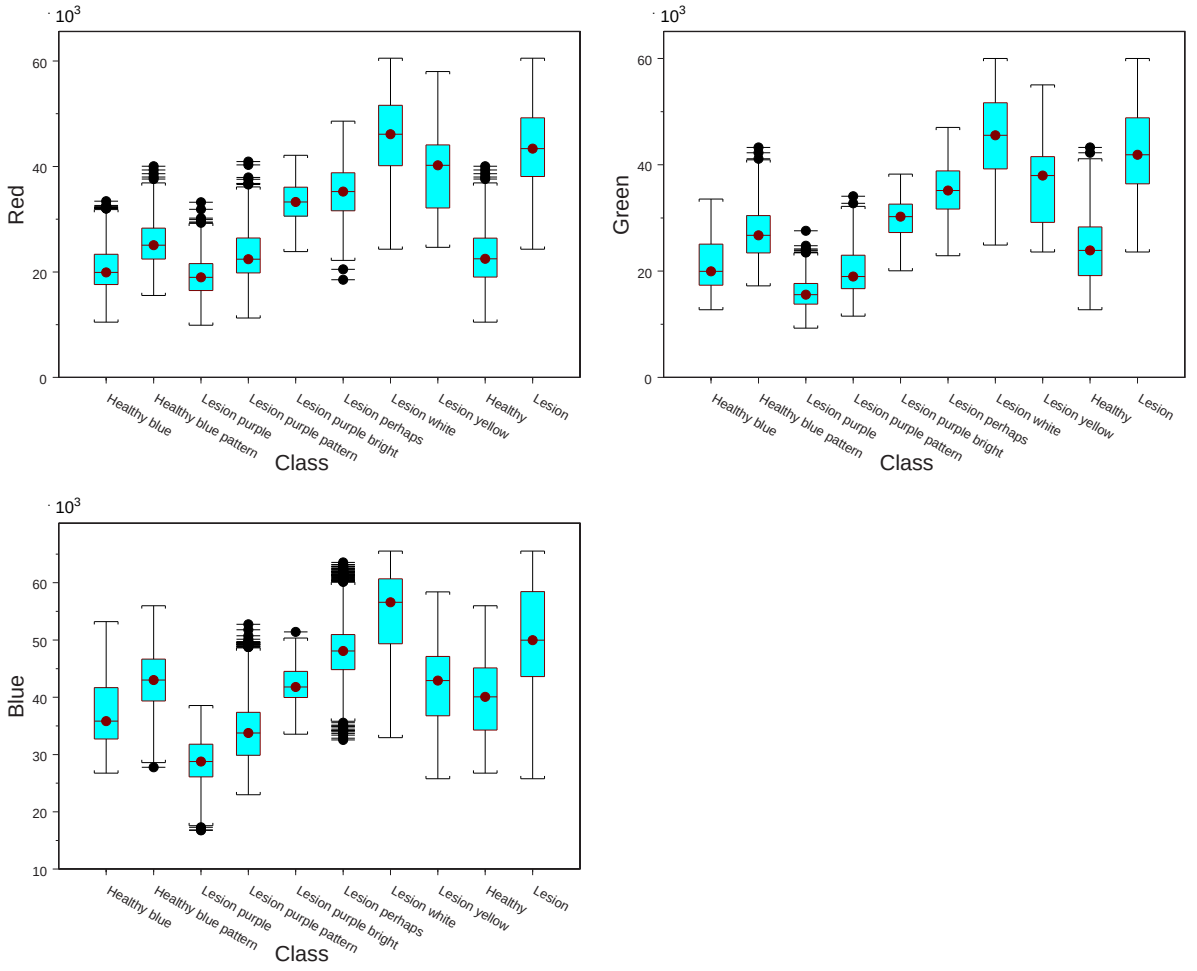


Figure 6.4: Boxplots of the classes. Note that the last two classes in each plot are the merged ones and hence redundant information.

6.7 Scatterplots of each Class

The difference in mean and variance from image to image is here studied further by plotting each class separately, see Figure 6.5 and 6.6. They show that the same class has different center positions in the different images and the variation between them is rather large. This has been observed before and is what could be expected due to different studies and biological variation. Another observation is that the variation and center positions move along a "straight line" with direction somewhat close to the direction of the diagonal of feature space (starting in 0,0,0), meaning that a large part of the variation could be due to variation in brightness in the images and / or studies (the relationship between the RGB values in each pixel is more or less retained for each class). *Lesion white* is a nice example of the positions on this "straight line".

At first glance it looks bad with large variance and overlapping classes, but if much of the variation is due to an intensity variation from image to image, then the classes in each image should still be separable to some degree, using some intensity invariant measure or by a color

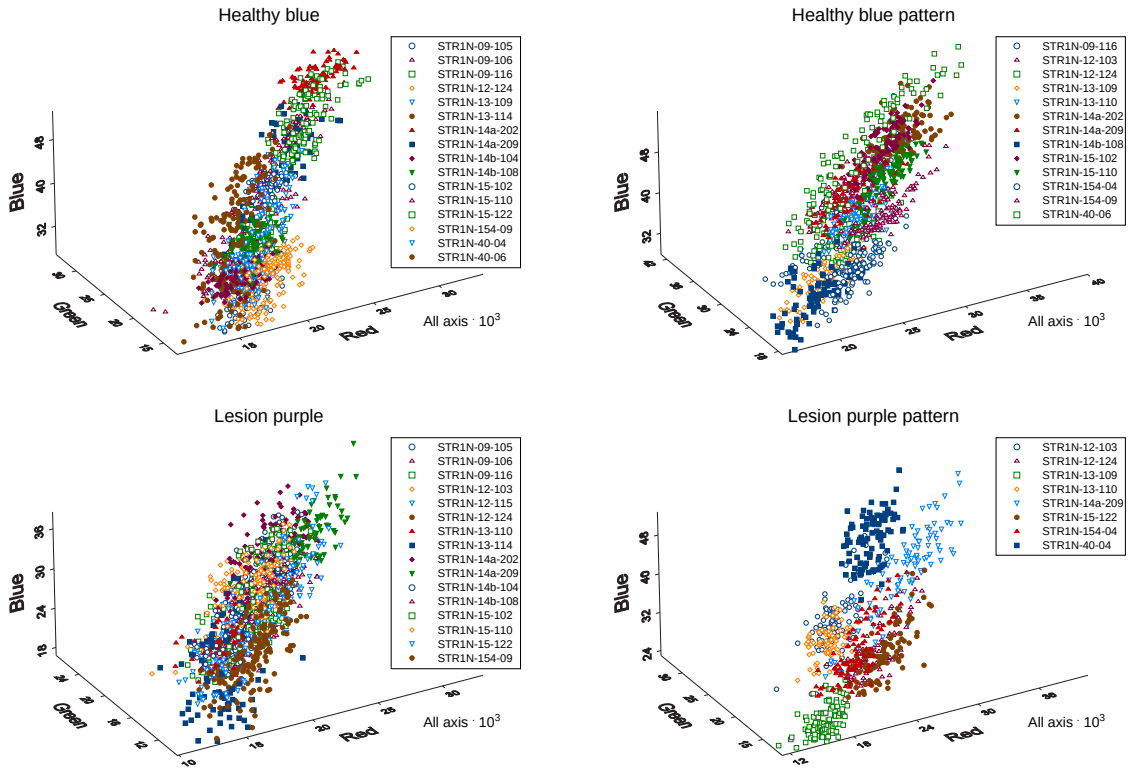


Figure 6.5: Classes shown separately and traceable to their respective image and study. Note that the axes from plot to plot are not equal.

transformation.

The observation leads to the formulation of the hypothesis "*the diverse appearances of the same type of area from image to image is mainly due to an intensity variation between the images*", which will be looked upon in the next section.

6.8 Intensity Variation between Classes and Images

The above formulated hypothesis "*the diverse appearances of the same type of area from image to image is mainly due to an intensity variation between the images*", is examined here. Two figures are generated showing the classes' mean values for each image. Figure 6.7 shows them connected by class and Figure 6.8 shows them connected by image, hence a possible variation structure should be visible.
Both plots in Figure 6.7 show that there is difference from image to image, but that a very large part is according to a general intensity variation. The not shown red vs. blue shows the same.
Figure 6.8 shows the same tendencies as above, and is even more convincing. These observations strongly indicate that the class centers move according to an intensity variation from image to image. This means that the class centers cannot be generalized as absolute values, but that separation might be possible using other approaches. Probably by using some intensity invariant measure or by color transformation.

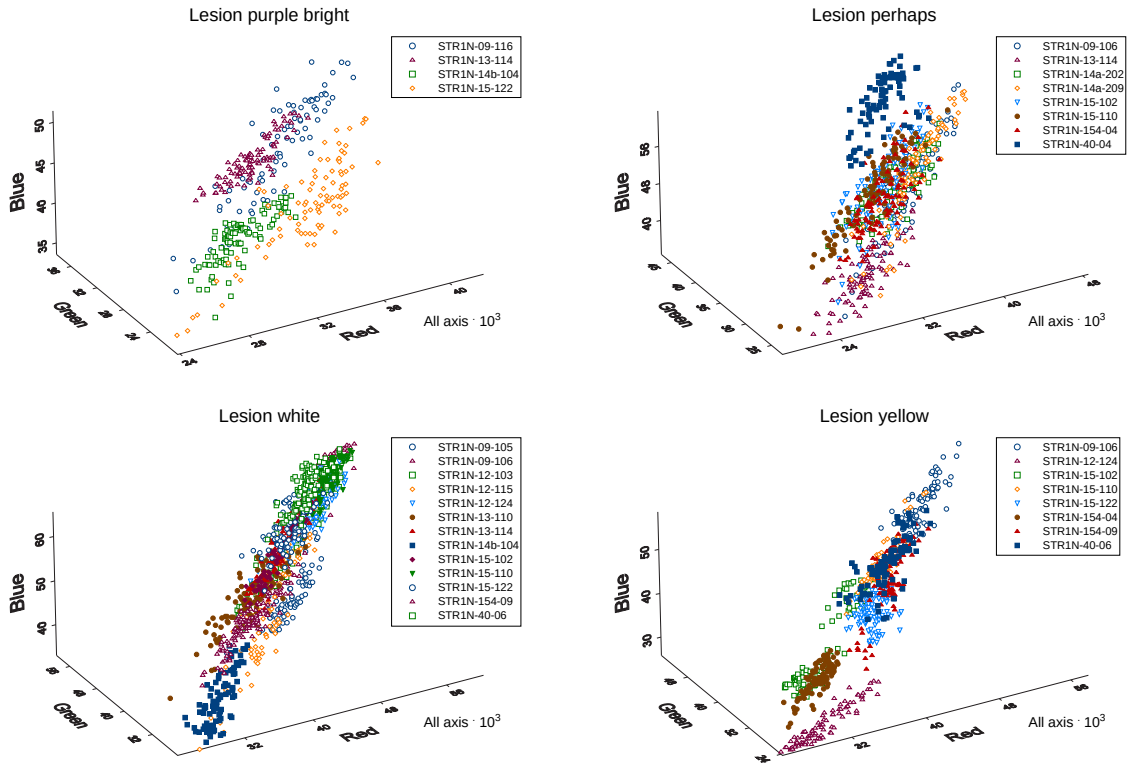


Figure 6.6: Classes shown separately and traceable to their respective study and image. Note again that the axes from image to image are not equal.

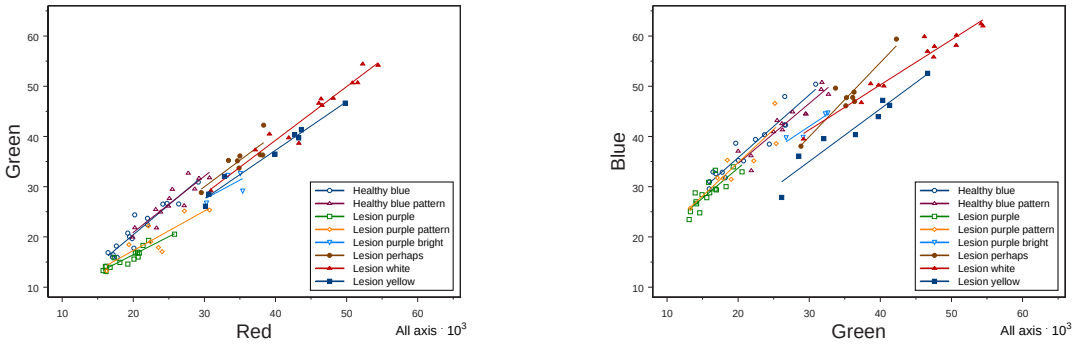


Figure 6.7: Relationship between the same class in different images. The class centers are used for a linear regression with class as the dependent variable, hence each line represents a class.

6.9 Conclusion

Eight classes in the images are defined and areas in 20 images are labelled according to them. Samples from these images constitute a data set used throughout the project to learn about the classes and images, and later to test the obtained partitioning results. The eight classes are found by visual inspection and carried out by an untrained person. There is not much physical background backing them up and they are changed twice during the project.

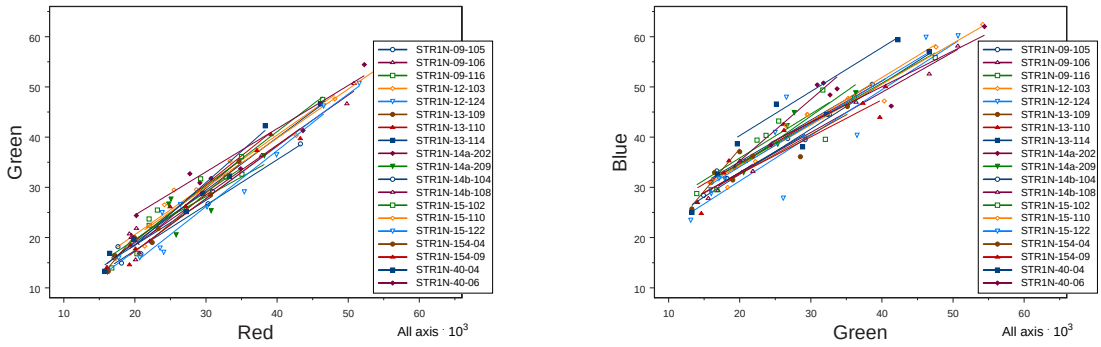


Figure 6.8: Relationship between classes in the same image. The class centers are used for a linear regression with image as the dependent variable, hence each line represents an image.

Some of the tests show that the red and the green bands have very similar values and are better than blue to separate the classes. At least two bands are acquired for a decent separation as this is not a one dimensional problem.

When looking at a single image the classes are more or less distinguishable, but it is not possible to generalize absolute values of the class centers.
Statistical tests show that the classes do not have equal mean values (at least one is different from the others, for each band). When using images, the amount of data is rather large and hence significance is easily obtained.

Visual inspection of the mean red, green and blue values for each class in all the images, and for all classes in each image, shows a general displacement of the classes. This indicates that the three bands' mean values are displaced alike, from image to image, hence a lot of the variation must be caused by a different intensity from image to image. The hypothesis "*the diverse appearances of the same type of area from image to image is mainly due to an intensity variation between the images*" is hence plausible.

A further approach could be intensity normalization, color transformation or studying the intensity variation between images and studies. The goal would be to understand the underlying structure, remove it and / or increase the distances between classes.

Chapter 7

Analysis of Color Transformations

This chapter deals with the reduction of the in-class variation between images caused by an intensity variation from image to image. Eight new color representation sets are generated, which are studied by boxplots, scatterplots etc. and an analysis of variance is carried out.
The purpose is, besides reducing the intensity variation, to find the best separating color representations.

7.1 Color Representations

Hands on experience shows that blue should be looked upon in comparison with red, green or both, but blue is not treated separately here, so the new color representations are generated for all bands. This results in calculating 22 new variables / features per pixel, with the purpose of normalizing the images or showing relationships between the bands according to the classes.

7.1.1 Trichromatic Colors

Trichromatic red, green and blue [6] are normalizations of the colors, in which each band is divided by the sum of the bands. Afterwards the sum of the bands hence equals 1 and each band contains its relative share of the intensity. See equation 7.1.

$$\begin{aligned} Tri.R &= \frac{Red}{Red + Green + Blue} \\ Tri.G &= \frac{Green}{Red + Green + Blue} \\ Tri.B &= \frac{Blue}{Red + Green + Blue} \end{aligned} \tag{7.1}$$

The trichromatic green and blue are the color representations that Visiopharm previously had the best results with. All three together are referred to as Tri.x.

7.1.2 Relative Distance between the Bands

Relative distance (or the ratio) between the bands is calculated by dividing one band by one of the other bands, see equation 7.2.

$$\begin{aligned} B.Div.R &= \frac{Blue}{Red} \\ B.Div.G &= \frac{Blue}{Green} \\ R.Div.G &= \frac{Red}{Green} \end{aligned} \tag{7.2}$$

Together they are referred to as x.Div.y.

7.1.3 Absolute Distance between the Bands

The absolute distance between the bands is defined as one band subtracted from another band, see equation 7.3.

$$\begin{aligned} B.Sub.R &= Blue - Red \\ B.Sub.G &= Blue - Green \\ R.Sub.G &= Red - Green \end{aligned} \tag{7.3}$$

Together they are referred to as x.Sub.y.

7.1.4 Distance from the Bands to the Mean Intensity

The distance from the bands to the intensity is defined as each band subtracted from the Intensity (the mean intensity), see equation 7.4.

$$\begin{aligned} R.Sub.I &= Red - Intensity \\ G.Sub.I &= Green - Intensity \\ B.Sub.I &= Blue - Intensity \end{aligned} \tag{7.4}$$

This transformation is known to efficiently remove highlight [16]. Together they are referred to as x.Sub.I.

7.1.5 IHS Color Representation

The IHS transformation [6] separates the intensity from the other parameters (Hue and Saturation) which can be a huge advantage in working with these images. The Intensity is defined as the mean intensity of the bands. Hue is the angle of the color that defines the "pure color".

Saturation is the inverse amount of black in the color, thus a high saturation results in the pure color, while a low saturation gives a dark color, see equation 7.5.

$$\begin{aligned}
Intensity &= \frac{Red + Green + Blue}{3} \\
Hue &= cos^{-1}\left(\frac{1}{2}\right) \qquad for\ Red = Green = Blue \qquad else \\
Hue &= cos^{-1}\left(\frac{1}{2}\frac{2 \cdot Red - Green - Blue}{((Red - Green)^2 + (Red - Blue)(Green - Blue))^{1/2}}\right) + \theta \\
Hue &= Hue - 2\pi \qquad for\ Hue > 2\pi \\
Hue &= 2\pi - Hue \qquad for\ Blue > Green \\
Saturation &= 0 \qquad for\ Intensity = 0 \qquad else \\
Saturation &= 1 - \frac{3 \cdot min(Red, Green, Blue)}{Red + Green + Blue}
\end{aligned} \tag{7.5}$$

where θ is the angle of the discontinuity in Hue. According to A. K. Poulsen [3], an angle of $\theta = \frac{\pi}{3}$ is the best for the present images, but it creates a discontinuity in the new color space. This is also the case when using $\theta = 0$. $\theta = -\frac{\pi}{3}$ does not result in a discontinuity and hence this angle is chosen. An exact angle is not needed as long as there is no discontinuity (resulting values below 0 or above 2π). The angle just shifts the values.

7.1.6 Standard Deviation

The standard deviation between the bands is calculated using equation 7.6.

$$Std.Dev = \sqrt{(Red - Intens.)^2 + (Green - Intens.)^2 + (Blue - Intens.)^2} \tag{7.6}$$

in which Intens. is short for Intensity. A low Std.Dev. means that the bands are alike and hence near a grayscale intensity.

7.1.7 Gradient

The gradient of each colorband is calculated as described in Chapter 4.6, p. 29. To avoid noise and local variation, the gradients are filtered by a 40 × 40 mean filter before the values are extracted. It gives the three variables: *Red.Grad*, *Green.Grad*, *Blue.Grad*. Together they are referred to as x.Grad. The gradients of each band are normalized between 0 and 255. The normalization factor for each band is the same from image to image.

7.1.8 Variance

The variance of each colorband is calculated as described in Chapter 4.7, p. 30. These are also filtered by a 40 × 40 mean filter before the values are extracted. It gives the following three variables: $Red.Var$, $Green.Var$, $Blue.Var$. Together they are referred to as x.Var. They are

normalized independently in the same way as above.

The trichromatic colors and IHS are known color representations while the rest are more or less invented for the purpose of reducing the intensity variation and / or enhancing the interesting inter-band relations. The gradient and variance are not color transformations but are included because they seem to differ for the bright lesion and maybe for the other classes as well. Thus, they might contribute to the description of the different classes.

7.1.9 Collinearity

When doing these color transformations the possibility of collinearity should be considered. If three or more points are located on a straight line, they are said to be collinear. This is fulfilled either when two color representations have a high correlation with each other or when one of the color representations can be predicted by one of the others or by a combination of others. E.g. with two of the trichromatic colors is it possible to predict the third. Hence there is redundant information using all three color transformations. The actual dimension of their feature space is only two.

Collinearity between variables is a problem when carrying out a regression analysis. It is normally assumed that the variables are independent when used in a regression analysis and if they are highly correlated, the resulting coefficients can be unstable. The inverse covariance matrix (used in a classifier) cannot be calculated if the variables are totally correlated and hence the problem is discovered. If the variables are "only" highly correlated the inverse covariance matrix can be calculated but might be unprecise.

Visual inspection can be used in order to find the problem. A check of the point cloud reveals that the color representations are collinear if a 2D feature space results in a line or if a 3D feature space results in a plane or a line. There are so many variable combinations that it is not feasible to check them visually.

SAS can check for collinearity when carrying out a regression analysis. In this project the dependent variable (class) is nominal hence the ordinary regression analysis cannot be used and the collinear test cannot be used either.

If several variables together can be represented by fewer dimensions, then correlation does not reveal the problem. Here a principal component analysis (PCA) can reveal the problem. If an eigenvalue of the principal components is close to 0 then the corresponding variable does not contribute and is either constant or predicted by the other variables. This test is used in the next chapter.

Collinearity can result in features wrongly chosen by the regression analysis, hence other methods are used in this chapter to examine the variables for their possible separation of the classes.

7.2 Summary

The mean values of all the color representations are shown in Table 7.1 and their standard deviation is shown in Table 7.2.

Class	Red	Green	Blue	Tri.R	Tri.G	Tri.B	B.Div.R	B.Div.G	R.Div.G
Healthy blue	20842	21472	37610	0.261	0.267	0.473	1.82	1.78	0.98
Healthy blue pattern	25399	26926	42918	0.266	0.282	0.452	1.71	1.61	0.95
Lesion purple	19189	15869	28772	0.300	0.248	0.452	1.52	1.83	1.21
Lesion purple pattern	23191	19791	34450	0.299	0.254	0.447	1.51	1.77	1.19
Lesion purple bright	33413	30133	42148	0.316	0.284	0.400	1.27	1.41	1.11
Lesion perhaps	35177	35371	47858	0.297	0.298	0.405	1.37	1.36	1.00
Lesion white	45404	45070	54797	0.312	0.309	0.379	1.22	1.23	1.01
Lesion yellow	39152	36466	41852	0.333	0.310	0.357	1.07	1.15	1.08

Class	B.Sub.R	B.Sub.G	R.Sub.G	R.Sub.I	G.Sub.I	B.Sub.I	Intens.	Hue	Satur.
Healthy blue	16768	16138	−630	−5799	−5169	10969	26642	3.11	0.241
Healthy blue pattern	17519	15992	−1527	−6349	−4821	11170	31747	3.07	0.211
Lesion purple	9583	12903	3320	−2087	−5408	7495	21277	3.41	0.257
Lesion purple pattern	11259	14659	3400	−2620	−6020	8639	25813	3.39	0.242
Lesion purple bright	8735	12015	3280	−1818	−5099	6917	35237	3.44	0.147
Lesion perhaps	12681	12486	−194	−4292	−4097	8389	39467	3.15	0.131
Lesion white	9394	9727	334	−3020	−3354	6374	48425	3.18	0.085
Lesion yellow	2701	5386	2685	−5	−2690	2696	39154	3.79	0.071

Class	Std.Dev	Red.Grad	Green.Grad	Blue.Grad	Red.Var	Green.Var	Blue.Var
Healthy blue	135	106	33	28	56	32	33
Healthy blue pattern	138	117	56	37	81	60	43
Lesion purple	96	119	42	38	76	48	52
Lesion purple pattern	111	126	57	43	91	65	56
Lesion purple bright	90	118	66	62	133	110	103
Lesion perhaps	104	130	80	49	123	101	69
Lesion white	79	101	56	41	96	82	63
Lesion yellow	44	95	50	40	88	74	61

Table 7.1: The mean value of each class for each color representation.

Table 7.1 shows that when two classes have the same mean value using one color transformation, then there are always other transformations when this is not the case. However, it is impossible to see if and how much they overlap with other classes.
By looking at the standard deviation, Table 7.2, it is seen that each class has the minimum standard deviation for at least one color transformation. It should therefore be possible for all the classes to be represented in a concentrated form. Possible mispartitionings are thus due to outliers, close class means or large standard deviations in the other classes.

It is not reasonable to choose between the color representations from these tables. If anything, they can be used to deselect color representations with several mixed classes. This way the x.Grad and x.Var might be left out.

7.3 Correlation of the Features with Age and with Each Other

The color representations from the sample data are correlated with age to see if there is a relationship and they are correlated with each other to check for collinearity. This is carried out both pixel-wise and image-wise.

Class	Red	Green	Blue	Tri.R	Tri.G	Tri.B	B.Div.R	B.Div.G	R.Div.G
Healthy blue	4253	4987	6412	0.016	0.016	0.018	0.17	0.16	0.10
Healthy blue pattern	4196	4640	5524	0.017	0.013	0.020	0.18	0.13	0.08
Lesion purple	3533	2729	4030	0.023	0.011	0.025	0.20	0.16	0.11
Lesion purple pattern	4991	4532	6361	0.027	0.017	0.027	0.22	0.18	0.15
Lesion purple bright	3535	3461	3139	0.017	0.010	0.017	0.11	0.10	0.08
Lesion perhaps	4924	4954	6137	0.017	0.010	0.020	0.15	0.10	0.06
Lesion white	7526	8030	7410	0.010	0.010	0.014	0.08	0.08	0.05
Lesion yellow	7323	7086	7561	0.016	0.007	0.019	0.10	0.08	0.05

Class	B.Sub.R	B.Sub.G	R.Sub.G	R.Sub.I	G.Sub.I	B.Sub.I	Intens.	Hue	Satur.
Healthy blue	3086	2354	1990	1543	1026	1705	5090	4.39	18.350
Healthy blue pattern	3311	2318	2116	1683	985	1770	4575	4.27	21.691
Lesion purple	2880	2188	1709	1400	890	1607	3209	6.48	16.972
Lesion purple pattern	4440	3227	2713	2205	1326	2424	4950	8.42	23.274
Lesion purple bright	3259	1829	2245	1763	827	1595	3056	10.11	15.634
Lesion perhaps	4534	3184	2318	2153	1078	2495	4977	6.39	19.690
Lesion white	2611	2276	1998	1351	1132	1491	7546	7.48	14.602
Lesion yellow	3422	2322	1544	1591	653	1880	7173	21.52	10.297

Class	Std.Dev	Red.Grad	Green.Grad	Blue.Grad	Red.Var	Green.Var	Blue.Var
Healthy blue	21	16	6	4	7	10	8
Healthy blue pattern	22	15	15	6	15	14	9
Lesion purple	18	15	14	10	23	26	22
Lesion purple pattern	27	11	19	10	27	30	18
Lesion purple bright	14	9	15	8	22	20	14
Lesion perhaps	31	30	26	12	29	25	20
Lesion white	18	17	16	10	23	23	21
Lesion yellow	14	19	18	11	31	28	22

Table 7.2: The standard deviation of each class for each color representation.

Correlation with Age

The pixel-wise correlation coefficient of red, green and blue with age is 0.01, -0.04 and -0.14, respectively. The correlation with age is surprisingly low and could be due to the intensity variation between images and the variation when looking at single pixels. It supports the fact that this is not a one-dimensional problem and the use of e.g. classes.

The new color representations generally correlate better with age than the RGB values, and is as high as -0.28 for B.Sub.R. Seven others have correlations above 0.20. For x.Sub.y and x.Sub.I five (out of six) are above 0.20.

The image-wise correlation coefficient of red, green and blue with age is 0.03, -0.09 and -0.33, respectively. This is higher than for the pixel-wise correlation. The correlation is less than found when examining the intensities (Table 4.2, p. 29), but they were based on the whole ROI from all images, compared to the image-wise correlation in which only parts of 20 images are used.

The new color representations generally correlate better with age, than the RGB values and is as high as 0.61 for R.Sub.I. Ten other color representations have a correlation above 0.33. For x.Sub.y and x.Sub.I five (out of six) are above 0.42.

Tri.R correlates with age with 0.59 with a level of significance at 0.005 (only 20 images are used). The Std.Dev correlates with age with -0.57, hence the Std.Dev between the bands is reduced with age. This results in more grayish colors and indicates that the *Lesion white* is increased with age.

Correlation between the Features
The pixel-wise correlation between the features is 0.97 for red with green, 0.86 for red with blue and 0.93 for green with blue. The high correlation is expected, especially the correlation between red and green which has already been observed from the scatterplots, etc.
The correlation of the color representations, shows that Intensity with red and green gives 0.98 and 0.99 respectively, and Std.Dev. with B.Sub.I gives 0.99. Collinearity could be a problem if these color representations are used simultaneously. The rest of the correlations are between 0 and 0.97 (the highest level of correlation between the original bands).

The image-wise correlations between the bands are 0.96 for red with green, 0.81 for red with blue and 0.89 for green with blue. These values are lower than for the pixel-wise correlations.
The correlation of the color representations, shows that Intensity with red and green gives 0.96 and 0.99, respectively and Std.Dev. with B.Sub.G and B.Sub.I both give 0.99 and B.Sub.I with B.Sub.R gives 0.98. The rest of the correlations are between 0 and 0.96.

7.4 Boxplots of the new Features

To see if one or more of the color transformations can be used for separation, a boxplot of each new feature is generated. The boxplots also include *Healthy* and *Lesion* besides the eight classes. *Healthy* is the merged healthy classes and *Lesion* is a merge of *Lesion white* and *Lesion yellow*. Figure 7.1 shows two of them. The rest can be observed in Appendix G together with the original colorbands for comparison. The main observations are described below.

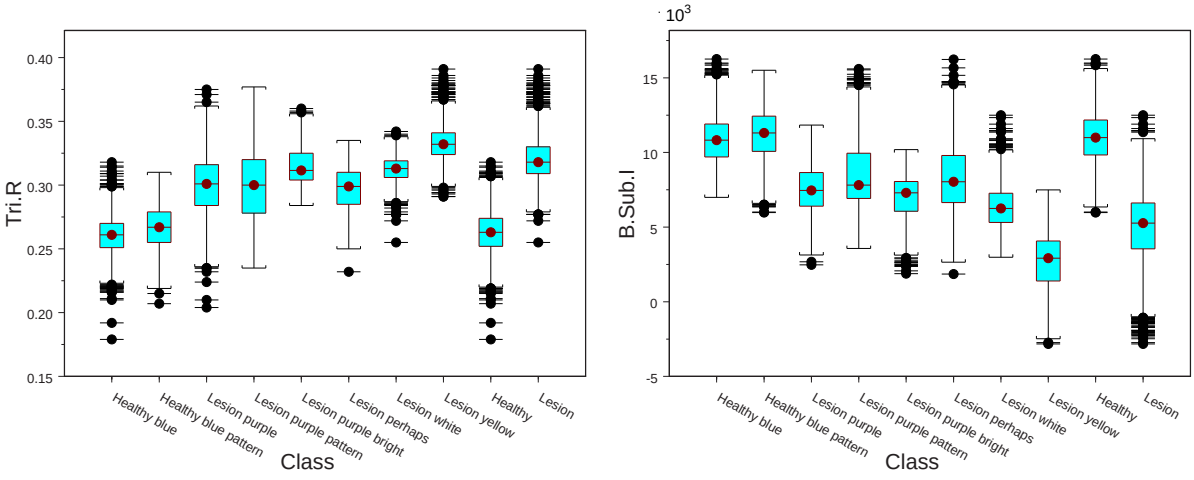


Figure 7.1: Boxplots of two new color representations. Note that the last two classes are the merged classes and hence redundant information. Note that the y-axis' are not the same.

The trichromatic transformation results in promising separation between *Healthy* and *Lesion*. The healthy classes have lower intensities for trichromatic green and higher intensities for trichromatic red than *Lesion purple* and *Lesion purple pattern*. Tri.B can totally separate *Healthy blue* from *Lesion yellow*.

The bands, divided by one of the other bands, show that B.Div.R separates *Lesion purple*

bright from *Healthy blue*. B.Div.G separates *Lesion yellow* from *Healthy blue* and almost also from *Healthy blue pattern*, *Lesion purple* and *Lesion purple pattern*. R.Div.G shows lower intensities for the healthy classes than for the purple classes. B.Div.R and B.Div.G can almost separate *Healthy* and *Lesion*.

The absolute distances between the bands, show that B.Sub.R and R.Sub.G respectively have higher and lower intensities for the healthy classes than for the purple classes. B.Sub.G almost separates *Healthy* and *Lesion*.

The bands, subtracted by the intensity, show that R.Sub.I and B.Sub.I respectively have lower and higher intensities for the healthy classes than for the purple classes. B.Sub.I almost separates *Healthy* from *Lesion yellow*. R.Sub.I is similar to R.Sub.G and B.Sub.I is similar to B.Sub.R.

IHS shows that Intensity is like the green band. The Hue is lower for the healthy classes than for the purple classes, but they all have a lot of outliers. Saturation separates the healthy classes from *Lesion yellow* and also *Lesion purple* from *Lesion yellow*.

Std.Dev is very similar to B.Sub.I.

Healthy blue is totally separated from *Lesion purple bright* using Blue.Grad and *Healthy blue* has a relatively low variation for Green.Grad.

The x.Var all totally separate *Healthy blue* from *Lesion purple bright* and the Blue.Var also separates *Healthy blue pattern* from *Lesion purple bright*.

Summary
There is no dominating feature that separates all the classes. There is no convincing separation of the healthy classes but several representations have different mean values (but are overlapping) so combining two or more of them might do the job.
The most promising variables are Tri.R, Tri.B, B.Div.G, R.Div.G and Hue, but only if two or more are combined.

7.5 Scatterplots of the new Features

Looking at the features in 1D reveals some details, but combining the color representations gives better probability of separation. The color representations are therefore shown in 3D scatterplots in Figure 7.2 and 7.3. The plots are rotated independently and shown from an angle which more or less reveals the best separation of the classes.
The scatterplots in Figure 7.2 show that the classes, using the color transformations, are still overlapping, but are a little different than the original ones. Tri.x and x.Div.y group the blue classes, the purple classes and the rest in three similar lumps, which could be an advantage in e.g. a classifier (if trying to separate only the different main groups of classes). *Lesion perhaps*, *Lesion white* and *Lesion yellow* are more concentrated than for the RGB colors, but they are still mixed together.
X.Sub.y shows an elongated shape and *Lesion purple* and *Lesion purple pattern* are more scattered than the original scatterplot. The center of these classes are moved away from the blue

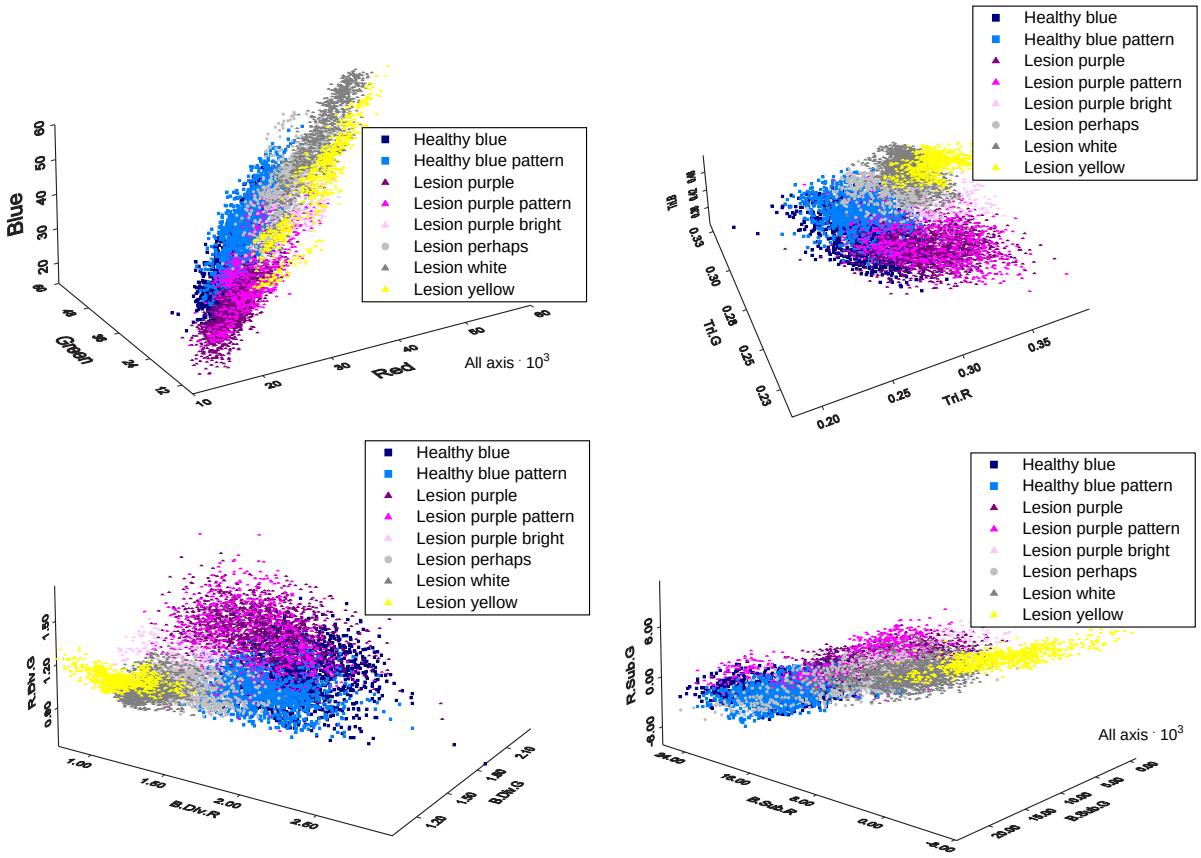


Figure 7.2: 3D Scatterplots of the color representations. Top: The original colorbands (RGB) and the trichromatic colors (Tri.x). Bottom: The relative colors (x.Div.y) and the absolute differences (x.Sub.y). Note that the axes and the rotation differs from plot to plot.

classes and closer to the other lesion classes. This means that the healthy classes are in one end of the feature space and the lesion stage more or less grows the further away an observation is from this end. This tendency is not perfect but might be interesting to examine further. Figure 7.3 shows that x.Sub.I is like the x.Sub.y. The IHS colors show large variation of the non-purple lesion classes. The classes are not overlapping more than some of the other color representations, but the large variance of especially *Lesion yellow* could create problems with a classifier if it uses the covariance.

The gradient and variance features show large variance for all the classes and they are mixed more than the other color representations. The usefulness of these is hard to see.

7.6 ANOVA

Analysis of variance (ANOVA) is here used to calculate the variance according to each class and according to the different images. The purpose here is also to examine the color representations' separation effect and to select the best separating ones.

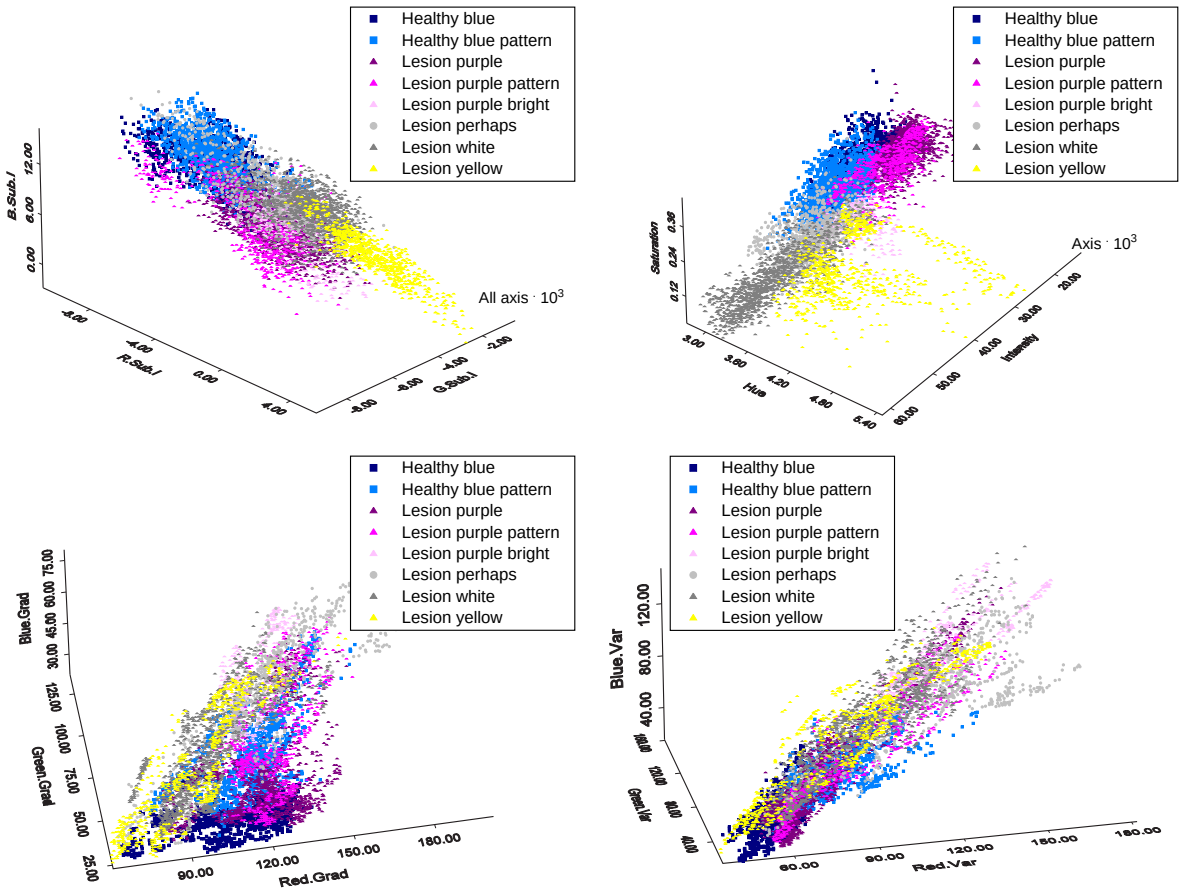


Figure 7.3: 3D Scatterplots of the color representations. Top: The bands subtracted the mean intensity (x.Sub.I) and the IHS colors. Bottom: The gradients (x.Grad) and the variance (x.Var). Note again that the axes and rotation differs from plot to plot.

7.6.1 Interpretation of ANOVA

Calculating ANOVA gives an F-value that is a measure of the ratio between the model and the error. The Pr(F) is the probability of significance of the F-value and due to the large amount of data, working with images, this is often 0 ($< 10^{-10}$). The larger the F-value, the better is the separation of some of the classes (but maybe not all of them).
When having several classes and several images there are two independent variables; the variance between the classes (F_{Class}) and the variance between the images (F_{Image}).
The larger the ratio (F_{Class}/F_{Image}) between the two variables, the more variance / information comes from the classes and not from the intensity variation from image to image. Hence a good color representation has both a high F-value and a large F-ratio.

7.6.2 Tests

ANOVA is carried out for all the color representations, including the original RGB. For the labelled images, Class and Image are used as independent variables and the 25 color represen-

tations, each as the dependent variable.

Two single images are also tested. It is STR1N-13-110 and STR1N-15-102, chosen as the worst and best separable, respectively (shown in page 52). Here the only independent variable is Class. The results can be seen in Table 7.3.

	Red	Green	Blue	Tri.R	Tri.G	Tri.B	B.Div.R	B.Div.G	R.Div.G
F_{Class}	5244	5958	3061	2268	5276	4963	3005	5521	1726
F_{Image}	446	578	423	297	657	549	343	663	263
Ratio	11.8	10.3	7.2	7.6	8.0	9.0	8.8	8.3	6.6
$F_{STR1N-13-110}$	1632	3342	1988	303	690	158	362	2131	210
$F_{STR1N-15-102}$	874	1096	964	708	694	375	720	1230	500

	B.Sub.R	B.Sub.G	R.Sub.G	R.Sub.I	G.Sub.I	B.Sub.I	Intens.	Hue	Satur.
F_{Class}	2554	2446	1277	2142	1213	2696	5090	1193	4897
F_{Image}	414	504	289	351	411	468	483	213	366
Ratio	6.2	4.9	4.4	6.1	3.0	5.8	10.5	5.6	13.4
$F_{STR1N-13-110}$	170	2235	1126	186	581	465	2529	145	854
$F_{STR1N-15-102}$	515	1207	956	610	443	737	1008	263	1040

Class	Std.Dev	Red.Grad	Green.Grad	Blue.Grad	Red.Var	Green.Var	Blue.Var
F_{Class}	10162	89	1758	1835	3363	4932	2928
F_{Image}	328	211	232	149	231	226	133
Ratio	31	0.4	5.4	12.3	14.6	21.8	22.0
$F_{STR1N-13-110}$	442	838	113	96	140	318	219
$F_{STR1N-15-102}$	776	577	381	350	782	518	326

Table 7.3: F-value from ANOVA of each color representation for all the images and for two single images.

Labelled Images
By looking at the RGB bands, it is observed that blue has smaller F-values than the other bands and also has the smallest Ratio. The color transformations have F-values which are generally lower than for the RGB. Std.Dev. has the largest F_{class} and Trichromatic green and B.Div.G have the largest F_{Image} while Blue.Grad and Blue.Var have the lowest.
Saturation, Std.dev, Blue.Grad and x.Var all have a larger Ratio than the largest from RGB (11.8 for red).

Single Images
For the original bands, the image with the best separable classes has F-values that are two or three times as high as for the image with less separable classes. This is only the case for six of the new color representations while the rest have larger F-values for STR1N-15-102. B.Div.G, B.Sub.G and Intensity have larger F-values for both images than the two lowest of the original RGB.

Choosing Promising Variables
Many of the new color representations do not seem to be better than the original ones. Combining them might contribute with new information. The ones chosen, based on the F-values and ratio, are: Tri.G, B.Div.G, Intensity, Saturation, Std.dev and x.Var. Tests of the single images result in choosing B.Div.G, B.Sub.G and Intensity. R.Sub.G and Saturation might also contribute.

7.7 Conclusion

Twenty-two new color representations (8 sets) are calculated from the images and included in the data set. Trichromatic and IHS are known color transformations, while the rest are more or less invented, based on the experience with the images. The gradient and variance are also included to see if they can describe the different classes.

Collinearity can be a problem when using the color transformations. Correlation reveals the problem between two variables, and thus which variable should not be used simultaneously. Some features have correlations with each other with up to 0.99.
The collinearity between three variables are not found by correlation. Here principal component analysis can reveal the problem, if an eigenvalue is close to 0. This will be the test for the tried variable combinations for the classification in the next chapter.

Based on the summary and scatterplots, the classes using x.Grad and x.Var are found overlapping more than the other variables.

Correlating the variables (averaged by image) to age shows that red and green have no significant correlations with age, while for blue it is -0.33. Half of the new color representations (11 out of 22) have higher correlations than the (-)0.33 and top at 0.61 for R.Sub.I. For x.Sub.y and x.Sub.I five (out of six) are between 0.42 and 0.61 and hence interesting in relation to further use.

The boxplots reveal that a few classes can be totally separated using different variables, one at a time. The blue classes overlap and so do the purple classes, and they also mix with each other. Trichromatic red looks like the best one for separating the blue classes from the purple ones. The most promising variables for separation are Tri.R, Tri.B, B.Div.G, R.Div.G and Hue, where at least two should be combined.

Scatterplots show that none of the color representation sets separate the classes notably better than RGB. The trichromatic colors and x.Div.y represent blue, purple and the rest of the classes in three equal sizes. They still overlap but look more separable than using RGB if only three main classes are to be separated.

x.Sub.y and x.Sub.I organize the classes somewhat according to the assumed lesion stages. At one end of the point cloud, in feature space, are the blue classes, then the purple classes and finally the rest of the classes. They are not perfectly sorted, but might be usable.

ANOVA points out B.Div.G, Intensity, Saturation as variables that might separate the classes. Tri.G, B.Sub.G, R.Sub.G, Std.dev and x.Var might also contribute.

None of the tried color transformations result in a convincing separation of the classes. The intensity variation is thus not as general as expected or the matching color transformation is not found.

Chapter 8

Classification of the Different Types of Areas

In this chapter the hypothesis "*the relative amount of the lesion area increases with age and the relative amount of healthy area decreases with age*" is tried rendered probable. The approach is classification of the images based on the defined classes and the original and new color representations. The aim is to obtain a high classification percentage of the classes and correlate the relative amount of the found areas to age, biomarker and histology. From the relative areas, an OA measure is generated and tested.
A Bayes classifier is defined and used with several input and output combinations and a classification evaluation measure (TDCM) is defined. The classes are merged to the three basis classes and used for classification besides trying the eight original classes.
Finally another classifier is tested. It is based on manually specified decision boundaries.

8.1 Classification of the Eight Classes

For the purpose of testing whether or not the different classes reduce the result for the rest of the classes, the classifier is run several times with different class combinations as input. First all the classes are used, then *Lesion perhaps* is left out, etc. The last combination consists of only *Healthy blue* and *Lesion white*. The images are classified for each of these combinations, using the color representations separately and in various combinations.
Besides using the classified classes as they are, they are also merged to a healthy class and to a lesion class. Due to the previous the interpretation of the purple areas etc. the merge is carried out in multiple ways. The merge combinations of healthy and lesion can be seen in Table 8.1.
The four healthy combinations and the five lesion combinations are combined so that both merged classes cannot contain the same class in the same test. This gives 16 combination possibilities. The classification percentages of these classes are higher than that of the single classes because pixels from a healthy class are considered correctly classified if they are classified as any of the healthy classes. This applies also to the lesion classes. These results are referred to as "post merged".

The classifier is implemented in C++ in order to control the process and because the amount of manual work in SAS will be considerable. The mean and covariance are calculated for all the color representations with respect to the classes. Mahalanobis distance is used so the covariance matrices are not pooled but separate for each combination of two color representations with

	Healthy combinations
1	*Healthy blue* *Healthy blue pattern* *Lesion purple* *Lesion purple pattern*
2	*Healthy blue* *Lesion purple*
3	*Healthy blue* *Healthy blue pattern*
4	*Healthy blue*

	Lesion combinations
1	*Lesion purple* *Lesion purple pattern* *Lesion purple bright* *Lesion perhaps* *Lesion white* *Lesion yellow*
2	*Lesion purple* *Lesion purple pattern* *Lesion purple bright* *Lesion white* *Lesion yellow*

	Lesion combinations
3	*Lesion purple bright* *Lesion perhaps* *Lesion white* *Lesion yellow*
4	*Lesion purple bright* *Lesion white* *Lesion yellow*
5	*Lesion white*

Table 8.1: Merge combinations for the healthy class and the lesion class.

respect to the classes.
A Bayes classifier, using Mahalanobis distance, results in decision boundaries that are quadratic surfaces (ellipsoids, paraboloids or hyperboloids). The classifier does not use the prior distribution of the classes. The information is not available and the amount of labelled classes would make a poor estimate because the entire images are not labelled, merely areas in them. The classes are hence assumed to have equal probabilities. The loss is likewise considered uniform for all class combinations. The classes are equally weighted when merged. A short explanation of the classifier can be seen in the Theory chapter, p. 106.

Besides evaluating the confusion matrices from the classifier, the classes are correlated with age. It is the relative amount of pixels found for each class in each image that are correlated with age. The post merged classes are also used to calculate confusion matrices and correlate their relative areas with age.

No filtering nor other postprocessing are carried out to smoothen the classified images.

8.1.1 Finding the Best Variable Combinations

All of the color representations are used, one at a time, as input for the classifier and afterwards the color representation sets (red, green and blue, then trichromatic red, green and blue, etc.) are tried. Combining the single color representations that each give a good classification result are then tried, besides the promising color representations chosen by exploring the boxplots and ANOVA.
Finally, a regression analysis is used in order to find the best variable combinations which are carried out by stepwise regression, forward selection and backwards elimination. The dependent variable (Class) is a nominal variable, hence a normal regression is not possible, but PROC Stepdisc (in SAS) can do the job. Table 8.2 shows the obtained results of the regression analysis. The stepwise regression and forward selection result in the same variable combination, while backwards elimination gives a different result. All four results are tried in the classifier.

Regression	Two variables	Three variables
Stepwise regression	*Saturation and B.Sub.R*	*Saturation, B.Sub.R and Red.Var*
Forward selection	*Saturation and B.Sub.R*	*Saturation, B.Sub.R and Red.Var*
Backwards elimination	*Tri.G and Tri.B*	*Tri.G, Tri.B and Hue*

Table 8.2: Results of regression analysis with different parameters and number of output variables.

8.1.2 Test of the Input

The input to the classifier should be normally distributed and the samples from a single class in a single image, generally fulfills this. This is also the case for the different color transformations of the samples. For the bright lesion classes there are exceptions.
When the samples are merged across the images, with respect to the classes, the resulting classes are not always normally distributed. Figure 8.1 shows the distribution of *Lesion purple* for the RGB and the trichromatic variables.

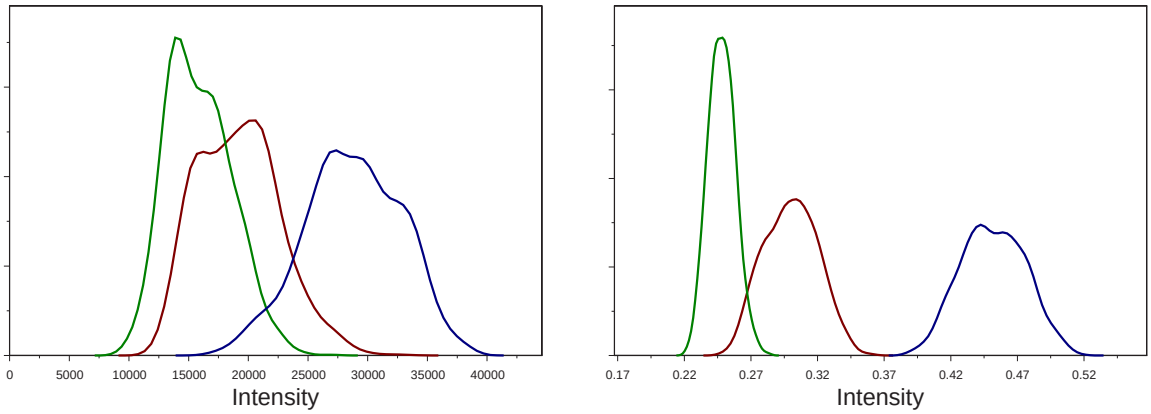


Figure 8.1: Distribution of *Lesion purple* when merged across the images. Left: Distributions of the original bands. Right: Distributions of the trichromatic colors. The plots are smoothed a bit by S-Plus. The distributions are colored with their respective band color.

Figure 8.1 shows that while the original bands are not normally distributed, the trichromatic colors approximately are and for trichromatic green it is normally distributed. The data is hence not always normally distributed, which it ought to be. This is also the case for the linear classifier, hence choosing it instead does not solve the problem.
The choice is still the quadratic classifier in order to make the most of the individual variation of each class.

8.1.3 Results

One of the classified images (STR1N-09-101) is shown with results from different input combinations and with different color representations, see Figure 8.2.
Figure 8.2 shows that the results differ depending on which color representations are used as input. The bright blue areas (*Healthy blue pattern*) are the main problem for this image because much of these areas are classified as *Lesion white* or as *Lesion perhaps*.

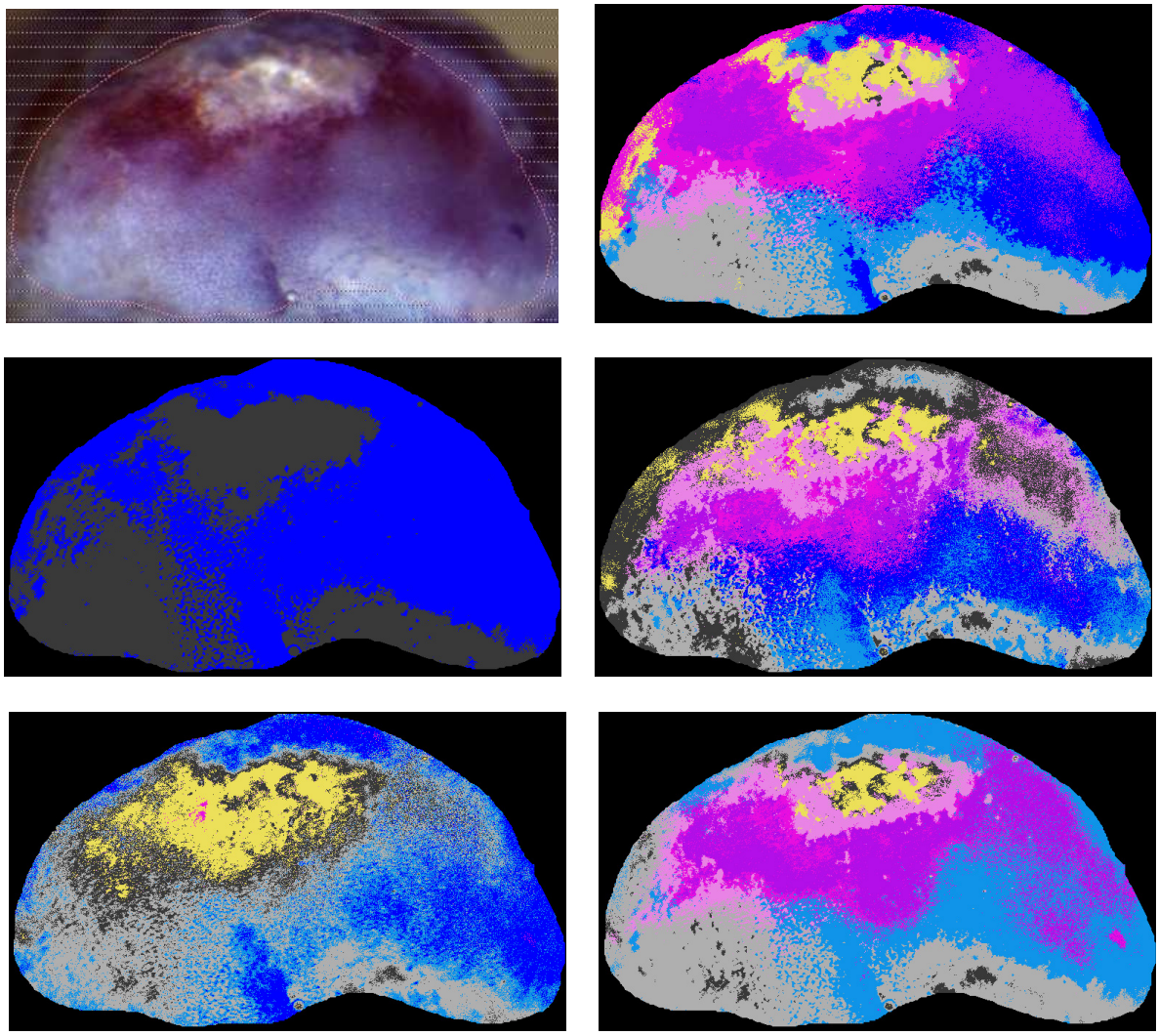


Figure 8.2: Classification examples of STR1N-09-101. The used pseudo colors are the same as in the scatterplots (dark gray represents *Lesion white* and light gray *Lesion perhaps*). Top row: The original image and the classified result with RGB as input. Middle row: RGB as input only for the classes *Healthy blue* and *Lesion white*. Right: X.Sub.I as input. Bottom row: Trichromatic red as input and all three trichromatic bands as input.

The classified image, based on RGB, is relatively well classified. In the bright lesion area, more *Lesion white* and less *Lesion purple bright* is expected and for the blue areas in the bottom of the ROI, less *Lesion perhaps* is expected.
The result of classifying with only two classes, using RGB, is that the bright lesion is correctly found but includes too large an area around it. The bright blue area is classified as *Lesion white* which is an example of why *Lesion perhaps* can be a useful class.
The classification using x.Sub.I results in many errors; their are too much of the lesion classes and they overlap one another. The area near the left border of the tibia is relatively dark and this also introduces misclassifications.
The classification based on trichromatic red is surprisingly bad. It finds none of the purple areas

and both *Lesion perhaps* and *Lesion yellow* are found in too many places. This is surprising. By looking at the mean intensities of the classes, for trichromatic red, it is found that the other lesion classes have mean values which are very close to and on both sides of the purple ones. Values slightly of the purple mean values are therefore not classified as purple but as another bright lesion class.
The classification image, based on all the trichromatic bands, is reasonable but *Lesion white* is found in much of the blue areas and there is hardly any *Healthy blue*.

Figure 8.3 shows the classification of the data set. The classes found have large variations but the class centers look correct and hence the classifier seems correctly implemented. The classes found using RGB look more like the original sample than the ones obtained using the trichromatic colors.

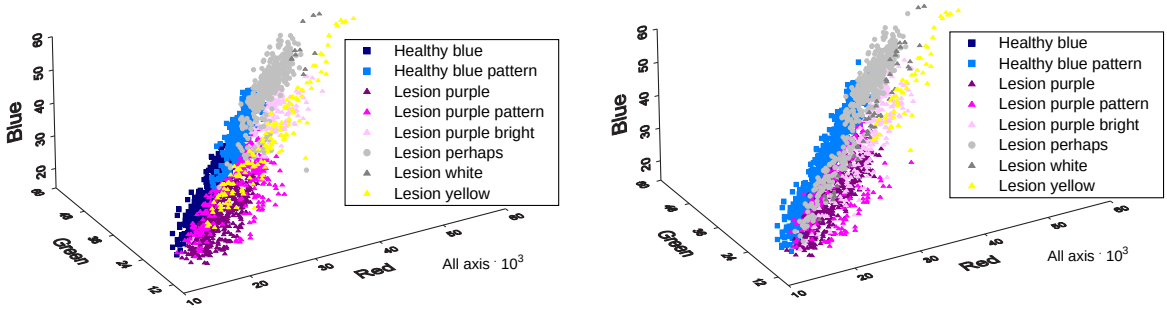


Figure 8.3: Classification examples as scatterplots of image STR1N-09-101. Left: Classification based on the RGB values. Right: Classification based on trichromatic values.

Confusion matrix
Some of the classification results are shown in Table 8.3. It is the diagonal of the confusion matrix that is shown, in which all the classes are used as input and both the eight classes and the post merged ones are used as output. A confusion matrix shows what pixels from each class are classified as. Each row represents an original class and each column represents a resulting / classified class. The pixels in the diagonal are those that are classified correctly, while the rest are misclassifications. It is normally the relative amount, in percentage, that is shown in the confusion matrix, so that each row sums up to 100 %. It is the diagonal, in percentage, that is used to summarize the results in Table 8.3.
To obtain a reliable correlation with age, most of the pixels in a class should be classified correctly and at the same time it should not find too many pixels from other classes. Only the first amount is truly revealed by the confusion matrix when looking at percentages. A combined measure multiplying "the relative amount of the class that is classified correctly" by "the relative amount of the classified pixels which are actually from the current class" is thus used (not shown). A high score thus demands not stealing many pixels from other classes and not letting other classes steal ones pixels. The measure is named True Diagonal of the Confusion Matrix (TDCM) and the formula can be seen in the Theory chapter, p. 107.

The highest correlation with age is obtained for trichromatic red with *Healthy blue*, *Lesion purple* and *Lesion white* as the input classes. The correlation of the relative area of *Lesion white* with age is 0.62. It is based on low classification percentages, where the TDCM measure is 0.49 and hence not reliable. Using two or three parameters, with or without Tri.R, has not improved the correlation with age.

	Healthy		Lesion						Post merged	
	blue		purple							
Variable combination		patt.		patt.	bright	perhaps	white	yellow	Healthy	Lesion
Single variable input										
Red	7	49	83	0	62	24	54	21	44	82
Green	6	22	91	32	52	54	47	9	25	94
Blue	9	1	90	32	70	52	44	0	10	96
Tri.R	67	30	0	0	1	41	66	77	89	71
Tri.G	19	8	74	4	49	44	10	86	23	87
Tri.B	68	56	0	5	49	22	63	75	92	43
B.Div.R	64	35	18	0	29	33	66	79	86	64
B.Div.G	14	66	68	0	53	34	46	78	58	78
R.Div.G	1	58	48	4	21	8	51	67	50	92
B.Sub.R	51	53	12	0	10	17	51	83	83	82
B.Sub.G	31	41	23	11	52	0	42	83	77	75
R.Sub.G	24	58	31	8	1	0	29	53	67	84
R.Sub.I	46	55	35	0	0	4	38	73	81	82
G.Sub.I	0	7	29	36	35	28	8	87	7	94
B.Sub.I	55	48	40	5	10	0	55	86	84	80
Intensity	6	25	90	33	59	49	52	0	25	94
Hue	42	58	60	0	6	12	1	38	84	70
Saturation	26	35	63	2	61	16	28	90	50	75
Std.Dev	54	46	24	6	53	0	37	89	80	82
Red.Grad	16	0	0	25	55	23	12	82	15	93
Green.Grad	98	23	16	0	38	37	0	0	74	34
Blue.Grad	90	64	0	6	76	20	0	0	95	25
Red.Var	95	65	0	0	80	4	18	0	91	23
Green.Var	97	67	0	0	77	5	25	0	94	24
Blue.Var	85	45	0	7	95	27	0	0	97	32
Multiple variable input										
RGB	55	63	82	27	83	63	59	91	84	88
Tri.x	0	80	77	16	74	51	56	82	67	92
x.Div.y	48	68	64	8	72	44	70	76	85	81
x.Sub.y	Collinear									
x.Sub.I	47	52	33	24	47	26	42	89	81	86
IHS	51	69	83	24	75	65	61	83	83	87
x.Grad	86	61	30	19	97	35	9	55	80	53
x.Var	94	68	32	14	97	40	33	29	90	50
Selected variable combinations										
Saturation,B.Sub.R,Red.Var	94	70	74	14	74	56	66	87	94	76
B.Div.G,Tri.B,Tri.G	42	68	78	3	68	52	51	87	84	82
Tri.G,Tri.B,Hue	60	63	64	6	67	51	62	74	92	72
Tri.R,Tri.B	53	71	64	9	73	51	56	82	90	77
Saturation,R.Sub.I	49	66	76	8	69	54	60	87	87	82
Tri.G,B.Div.G,Intensity	50	69	81	23	78	64	59	91	97	96

Table 8.3: Classification results represented by the diagonal vector of the confusion matrix for each test. These tests use all the classes. All results are in percentage correctly classified.

Using the TDCM measure the results are not that good when looking at all eight classes. Correlation with age based on the classified classes can therefore be doubtful.
Lesion white and *Lesion yellow* and in most cases also *Healthy blue*, have a high TDCM score. Other classes, mostly *Lesion purple pattern* and *Lesion purple bright* are below 0.1 and are sometimes 0.
The classification, based on three inputs from the same color representation set, generally gives better results than using only a single band.

The best variable combination is "Saturation, B.Sub.R and Red.Var" found by the regression

analysis. It is the best at classifying the blue classes and *Lesion white*, and also have relatively high classification percentages of the rest of the classes. "Red, Green and Blue", "Tri.G, B.Div.G and Intensity" and "Tri.R, Tri.G and Tri.B" also give reasonable results and will be examined further.

The post merged classes obtain high classification percentages but the TDCM measures are not that high.

Test for Collinearity
The collinearity between the variable combinations is tested by Principal Component Analysis (PCA) for combinations of two and three variables. The "proportion of variance" for the principal components is calculated. If the proportion is close to 0 then the corresponding principal component does not contribute to the variation and should be removed.
For the combination of variables in each color representation set (RGB, all the trichromatic colors, etc.), the proportion of the first principal component is between 0.75 and 0.95, for the second between 0.04 and 0.20. The third component is 0.05 or less, meaning that the third component contributes with 5 %, or less, of the total variation. In the case of trichromatic colors, where the problem is known to exist, the proportion of the third component is $2.49 \cdot 10^{-5}$. Values below this will result in an unreliable inverse covariance matrix. The fact that it is not lower, is due to rounded variables.
The classification using "B.Sub.R, B.Sub.G and R.Sub.G" resulted, as the only combination, in a direct calculation problem of the inverse covariance matrix. Its third component is $9.75 \cdot 10^{-17}$, so the problem is not surprising.
For the best separating variable combinations found above, the third component is 0.00523, 0.00714 and 0.000873 for "Red, Green and Blue", "Saturation, B.Sub.R and Red.Var" and "Tri.G, B.Div.G and Intensity", respectively. These are low values but not necessarily collinear. The classifier is executed again and each three variable combination is split in three combinations of two variables.
Features that are collinear and not rounded (like x.Sub.y) result in exactly the same classification percentages when combining the variables two by two as when all three are used simultaneously. If they are rounded, they differ from each other with a few percentage points. The original bands combined two by two give different and lower results, but they did not have that low third component. Even though the red and green bands earlier were found to correlate at 0.97, they both supply information for the classification and are known to be distinct in the purple areas. It is assumed not to be collinear.

OA Measure
An OA measure is calculated using the relative areas of the classified images. The measure is defined as the predicted age.
The OA measure could also be defined as the correct age subtracted from the predicted age hence an OA value of 0 would mean that the osteoarthritis stage of the mouse is as if it were untreated. If the measure was positive the mouse would have an OA stage above what was expected and if the OA measure was negative the OA stage would be less than expected and hence it indicates that the used medication works. If age is used in the OA measure then when correlating this with age, they would not be as independent as they should be. Therefore the predicted age is used as the OA measure throughout the report.

A linear regression analysis is carried out, with the relative areas found by classification, for the best classifying variable combinations. The significant areas are included in the OA measure together with their respective parameter estimates.
The regression analysis of the classified areas by "Saturation, B.Sub.R and Red.Var" is shown below. Only the significant classes are left. The model with intercept and all the relative areas

do not have full rank because the relative areas sum to 1 for each image. The intercept is the least significant coefficient and thus it is removed first. "RA_..." is short for the "Relative Area of".

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
RA_HealthyBlue	1	0.09138	0.01410	6.48	<.0001
RA_LesionPurple	1	0.08893	0.01465	6.07	<.0001
RA_LesionPurplePattern	1	0.24689	0.05350	4.61	<.0001
RA_LesionPerhaps	1	0.09766	0.05727	1.71	0.0908
RA_LesionWhite	1	0.36097	0.04637	7.78	<.0001
RA_LesionYellow	1	0.27279	0.11410	2.39	0.0184

The OA measure thus becomes

$$\begin{aligned} \text{Predicted Age} \;=\; & 0 + 0.0914 \cdot \text{RA\_HealthyBlue} + 0.0889 \cdot \text{RA\_LesionPurple} \\ & +0.247 \cdot \text{RA\_LesionPurplePatt.} + 0.0977 \cdot \text{RA\_LesionPerh.} \\ & +0.361 \cdot \text{RA\_LesionWhite} + 0.273 \cdot \text{RA\_LesionYellow} \end{aligned} \tag{8.1}$$

The age of each mouse is predicted and their correlation with age is 0.56. Figure 8.4 shows the residual vs. the predicted age. When the dependent variable (Age) is a discrete variable, the residuals are expected to look like this. The regression analysis is therefore believed to be correct. The predicted age does not exceed 20.5 weeks and has residuals up to 11.5 weeks. This will be looked upon in the next section.

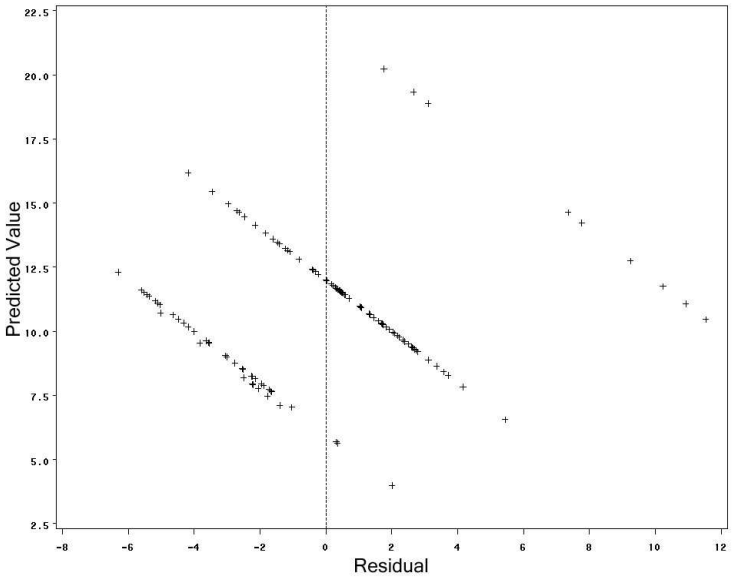


Figure 8.4: Residual plot.

For the most promising variable combinations the procedure is repeated and the results of the

	Age		Histology		Biomarker	
Variable combination	Corr	Significance	Corr	Significance	Corr	Significance
Saturation, B.Sub.R, Red.Var	0.56	$8.14 \cdot 10^{-12}$	0.29	$4.80 \cdot 10^{-2}$	0.05	0.82
Red, Green, Blue	0.57	$2.13 \cdot 10^{-12}$	0.40	$5.34 \cdot 10^{-3}$	0.13	0.55
Tri.G,B.Div.G,Intensity	0.58	$5.16 \cdot 10^{-13}$	0.41	$4.21 \cdot 10^{-3}$	0.03	0.89
Tri.R,Tri.G,Tri.B	0.56	$8.14 \cdot 10^{-12}$	0.33	$2.35 \cdot 10^{-2}$	0.06	0.79

Table 8.4: Correlation of the predicted age to age, histology and biomarker.

predicted age are shown in Table 8.4.
Table 8.4 shows that the OA measure from the four variable combinations has highly significant correlations with age at 0.56 - 0.58. Their correlation with histology differs between 0.29 and 0.41 and is either significant or highly significant. There is no significant correlation with biomarker. Their residual plots look like the earlier shown and hence the regression analyses are believed to be correct.
For red, green and blue as input, the OA measure does not include *Lesion white* because it was not significant in the regression analysis. This is not expected.

OA Measure with Logarithm, Quadratic and Exponential Input
The predicted age of the OA measure never reaches 22 weeks (the age of the oldest mice) and even though the residuals look correct, the regression analysis is tried improved.
A hypothesis is that "*the degeneration speed decreases with age*" and hence the logarithm of the relative areas is used as input together with the original relative areas. With red, green and blue as input, the result is that some of the new parameters are significant. One predicted age is now just above 22 weeks and the OA measure's correlation with age increases with 0.01 (to 0.58).
The exponential and the squared relative areas are also tried as input, one "set" at a time, together with the original relative areas. Some of these inputs are also significant but do not improve the result.
For the variable combination Saturation, B.Sub.R and Red.Var, the improvement using the squared values is 0.01 (to 0.57).
The different additional inputs give only small improvements and the residual is almost the same hence a better model is not found. The above hypothesis, "*the degeneration speed decreases with age*", is not rendered probable.

OA Measure of the Post Merged Classes
The regression analysis of the relative areas for "Saturation, B.Sub.R and Red.Var" removes RA_Lesion because the model does not have full rank when an intercept is used. RA_Healthy and RA_Lesion sum to 1 for each image.
The model with intercept and RA_Healthy and the model with both relative areas give almost the same predicted age. In both cases, it has a significant correlation with age at 0.31. The other variable combinations match or have lower correlations with age.

Effect of the Input and Output Combinations
The different combinations of input and output do not contribute positively to the classification and correlation at the same time. Using fewer classes results in better classification percentages of each class but the relative amount of each class has a lower correlation with age. For

instance, a lot of the purple areas are included in *Healthy blue* when purple is not defined as a class. Mixing two types of areas where one increases and the other one decreases with age gives together a low correlation with age. The unused types of areas / classes thus result in noise.
The idea of removing *Lesion perhaps* from the classification will therefore reduce the results. Leaving it out afterwards, so that the relative areas are independent of it, is expected to improve the classification percentages. It is like a garbage bin, so instead of misclassifying pixels, they end up here and are not used. The results are generally not improved.
The result is that all the classes should be included in the calculation and they should all be used after the classification, to establish the OA measure. They can be merged to fewer classes, but leaving them out will reduce the result.
The promising variables found by ANOVA, boxplots, mean and standard deviations are combined in multiple ways but none of them are better than those presented in Table 8.3. Maybe they improve the classification of a single class or maximum two, but the total amount of correct classified pixels is not improved. *Lesion purple pattern* generally has the lowest classification percentage and none of the tried combinations improves it, compared to the variables found by the regression analysis.
With several color presentations and multiple input and output combinations, but without TDCM measures near 1, tendencies are what can be learned from the results.
Going through the results, looking for indications of purple's effect on correlation shows that both *Lesion purple* and *Lesion purple pattern* generally have positive correlations with age. *Lesion purple pattern* has the highest correlation. For a few variables, they can be negative but that is only the case when the input consists of a single variable and resulting low TDCM measures.

8.2 Classification of Merged Classes

Instead of merging the classes after classification, they are merged prior to this. Three classes are generated; *Blue* from the two blue classes, *Purple* from *Lesion purple* and *Lesion purple pattern*, and *White* from the rest of the classes. In this way, the new classes are each described with the joint mean and variation for their respective original classes.
Purple and *White* are not joint for three reasons; firstly, *Purple* appears more similar to *Blue* than to *White*, hence a purple-white class will have large variance and overlap with blue. Secondly, it is still interesting to see *Purple*'s correlation with age (if it is positive, it is again rendered probable to be a lesion stage). Thirdly, to obtain an OA measure *Purple* and *White* should probably not be equally weighted due to different lesion stages.

There are much less output from this test such as the classification percentage for each class, the correlation of the relative areas with both age, biomarker and histology, and then the OA measure and its correlation with age, biomarker and histology.

8.2.1 Finding the Best Variable Combinations

A regression analysis is carried out as above in order to find the best variable combinations, see Table 8.5. It shows the same result using two variables. Stepwise regression and forward selection also result in the same combination for three variables, while backwards elimination includes another variable. All three results are, among others, tried in the classifier.

Variable combination	Two variables	Three variables
Stepwise regression	*Saturation and R.Div.G*	*Saturation, R.Div.G and Blue.Var*
Forward selection	*Saturation and R.Div.G*	*Saturation, R.Div.G and Blue.Var*
Backwards elimination	*Saturation and R.Div.G*	*Saturation, R.Div.G and Tri.B*

Table 8.5: Results of regression analysis with different parameters and number of output variables.

8.2.2 Results

The best results can be observed in Table 8.6. Here, the best is defined as high TDCM values for all bands. It shows that the classification gives reasonable results and that the TDCM measures

	Blue		Purple		White	
Variable combination	TDCM	Corr	TDCM	Corr	TDCM	Corr
Green	0.35	0.00	0.65	−0.06	0.68	0.14
Red, Green, Blue	0.67	−0.40	0.75	0.13	0.85	0.44
Tri.R,Tri.G,Tri.B	0.64	−0.46	0.68	0.10	0.81	0.47
Saturation,B.Sub.R,Red.Var	0.67	−0.34	0.71	0.05	0.80	0.42
Tri.B,R.Sub.I	0.64	−0.42	0.72	0.13	0.82	0.44
Tri.G,B.Div.G,Intensity	0.65	−0.46	0.71	0.18	0.85	0.46
Saturation and R.Div.G	0.61	−0.46	0.67	0.06	0.76	0.48
Saturation and R.Div.G and Blue.Var	0.64	−0.32	0.63	0.01	0.81	0.43
Saturation and R.Div.G and Tri.B	0.59	−0.35	0.69	0.00	0.79	0.43

Table 8.6: Classification results for the merged classes. First, the best result using a single band, then the best using two or three bands and then the variable combinations found by regression analysis.

for different classes are alike when using three variables as input. The variable combinations found when using regression analysis do not perform as well as the other combinations. The best classification is obtained by RGB and results in a correlation with age at -0.40, 0.13 and 0.44 for *Blue*, *Purple* and *White*, respectively. The correlations above 0.14 are significant, hence *Purple* is not significant for this variable combination.

Test for Collinearity

The variable combinations which are the same as in the previous section result in almost the same proportion of variance as earlier. The three new combinations give $1.03 \cdot 10^{-5}$ for "R.Div.G and Saturation", $6.70 \cdot 10^{-6}$ for "R.Div.G, Saturation and Blue.Var" and $1.28 \cdot 10^{-7}$ for "R.Div.G, Saturation and Tri.B". These values are all lower than for the other variable combinations and not used due to suspicion of collinearity.

OA Measure

The three best and most reliable variable combinations from Table 8.6 are used in a regression analysis to see which of the relative areas are significant and thereafter to estimate the coefficients for the OA measure (the Age prediction).

When using an intercept all the variables in the three tests are insignificant. Intercept is furthest away from significance. Removing the intercept changed the situation so that all three relative areas become significant in the tests. The two classifications with the best TDCM scores are shown below. Here with "Red, Green and Blue" as input:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
RA_Blue	1	0.05016	0.00883	5.68	<.0001
RA_Purple	1	0.11897	0.00842	14.12	<.0001
RA_White	1	0.22742	0.01933	11.77	<.0001

All three relative areas are highly significant and should be included in the age prediction which can be seen in equation 8.2.

$$\begin{aligned} \text{Predicted age} &= 0 + 0.0502 \cdot \text{RA\_Blue} + 0.119 \cdot \text{RA\_Purple} \\ &\quad +0.227 \cdot \text{RA\_White} \end{aligned} \tag{8.2}$$

The correlation of the predicted age to age, histology and biomarker can be seen in Table 8.7. Figure 8.5 shows the development of the relative areas according to age and also the predicted age. *Blue* and *White* shows convincing correlation with age based on the relative areas, while *Purple* shows less correlation and the relative areas for the old mice do not match perfectly to the regression line. It is also seen that the maximum of the Predicted age is below 20 weeks, but there is a convincing correlation with age.

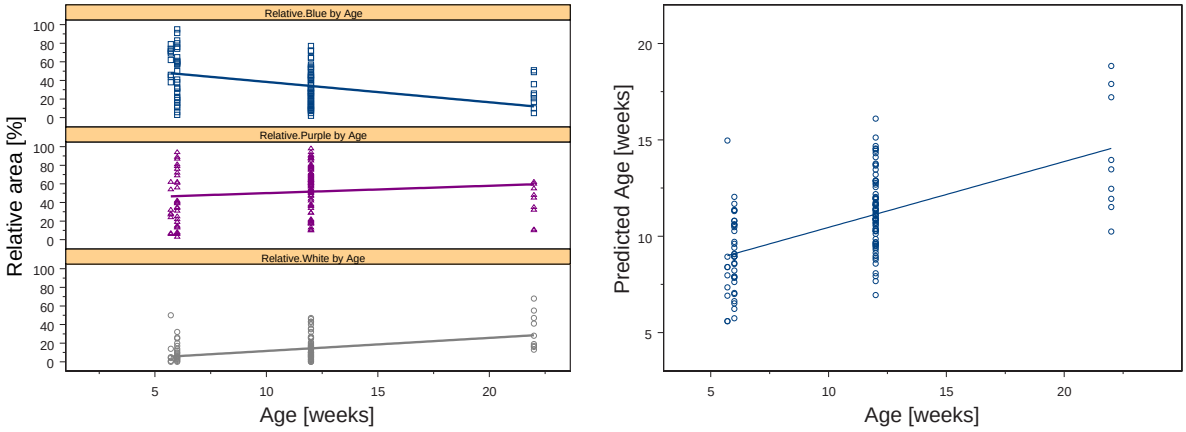


Figure 8.5: Classification results of the merged classes using RGB as input. Left: The relative area of each class with respect to age. Right: The predicted age with respect to age.

For the variable combination "Tri.G, B.Div.G and Intensity", the OA measure becomes

$$\begin{aligned} \text{Predicted age} &= 0 + 0.0486 \cdot \text{RA\_Blue} + 0.115 \cdot \text{RA\_Purple} \\ &\quad +0.244 \cdot \text{RA\_White} \end{aligned} \tag{8.3}$$

The parameters for the OA measures obtained with different variable combinations are alike, which indicates a good performance. It can be assumed that relative areas are alike and this indicates that most of the pixels are classified the same way.
The residual plots look like the one shown earlier and hence the regression analysis is believed to be correct. The obtained results are reasonable and match the obtained results by using the eight classes separately.

	Age		Histology		Biomarker	
Variable combination	Corr	Significance	Corr	Significance	Corr	Significance
Red, Green and Blue	0.57	$2.13 \cdot 10^{-12}$	0.40	$5.34 \cdot 10^{-3}$	0.13	0.15
Tri.G, B.Div.G and Intensity	0.58	$5.16 \cdot 10^{-13}$	0.41	$4.21 \cdot 10^{-3}$	0.03	0.74

Table 8.7: Correlation of the predicted age to age, histology and biomarker.

8.3 Classification by Manual Decision Boundaries

The results from the Bayes classifier are not optimal. The intensity variation is not removed nor circumvented sufficiently. The result using RGB as input is close to the best obtained, hence there must still be potential of increasing the classification percentages. With the current data set, the results are fair but for further use, the TDCM values should be in the order of 0.85 - 0.90 or higher. This is only just obtained for class *Lesion*.

A simple classifier based on decision rules is tried. It is not trained and the rules are based on human observation of the images, scatterplots, etc. Only the original RGB is used as input. It is the hope that general class info can be used. This might be less dependant of the intensity variation. The purpose is to separate the point cloud in feature space in three classes which are the same as the above (*Blue*, *Purple* and *Lesion*).
A pixel is defined as *Blue* if the distance from the mean of the red and green bands to the blue band is larger than, say, 10,000. A pixel is defined as *Purple* if the distance red - green is larger than, say, 0. A pixel is defined as *Lesion*, if the distance from the mean of the red and green bands to the blue band is smaller than, say, 10,000 and at the same time the sum of the bands is larger than, say, 80,000.
The results of the decisions are merged, and if the criterions are met simultaneously for more than one of the classes, then *Lesion* has first priority and *Purple* has second priority.
If none of the criterions are met, the pixel is unclassified and is left out of the following area calculations. The decision rules are listed in equation 8.4.

$$\begin{aligned} Blue &= Blue - \frac{1}{2}(Red + Green) > T_{Blue} \\ Purple &= Red - Green > T_{Purple} \\ Lesion &= (Blue - \frac{1}{2}(Red + Green) < T_{Lesion_1}) \cdot (Red + Green + Blue > T_{Lesion_2}) \\ Pixel &= 3 \cdot Lesion + 2 \cdot Purple \cdot (1 - Lesion) + Blue \cdot (1 - Purple) \cdot (1 - Lesion) \end{aligned} \tag{8.4}$$

where T_x are the thresholds. *Pixel* contains the class number, where 0 - 3 represents *Unclassified*, *Blue*, *Purple* and *Lesion*, respectively.
The thresholds are adjusted to optimize the TDCM values. This is carried out by brute force, using four for-loops within each other and trying a few values for each threshold. The basis of the trial values are those shown above.

8.3.1 Results

The optimal thresholds are found to be

$$T_{Blue} = 7,000 \ , \ T_{Purple} = 1,000 \ , \ T_{Lesion_1} = 15,000 \ , \ T_{Lesion_2} = 90,000 \tag{8.5}$$

Two of the classified images can be seen in Figure 8.6.

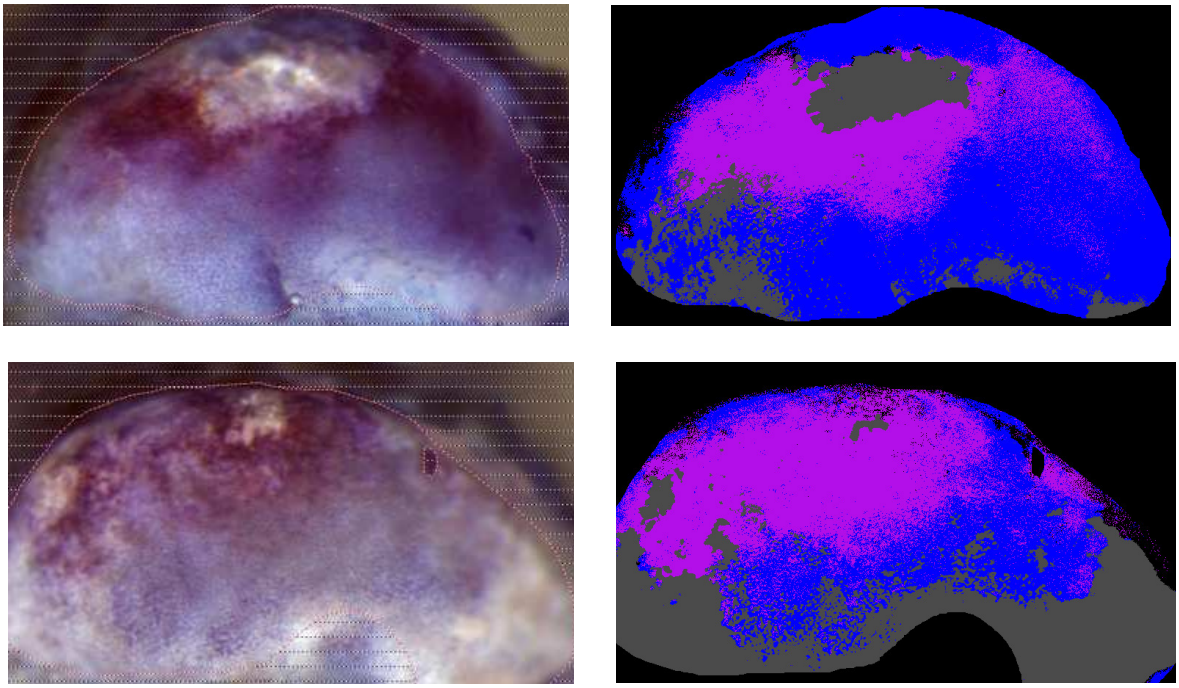


Figure 8.6: Original images and their classification results using decision boundaries. Top: Image STR1N-09-101. Bottom: Image STR1N-154-10.

Figure 8.6 shows that the classifications are pretty good. Class *Healthy blue* can be classified as *Lesion white* but not to the same degree as seen for the eight classes. It finds too little purple in image STR1N-09-101 and too much in STR1N-154-10.
The decision boundaries seem too simple for *Lesion white* because it finds pixels in the bright blue areas.

OA Measure
Table 8.8 shows the results for the optimized thresholds.

	Blue	*Purple*	*Lesion*	OA
TDCM	0.61	0.68	0.79	–
Corr to age	−0.41	0.17	0.37	0.44
Corr to Biom09	−0.16	0.05	0.18	0.02
Corr to Histo09	−0.02	−0.14	0.30	0.17
Corr to Histo12	−0.15	0.17	−0.09	−0.23

Table 8.8: The optimized results of manual decision boundaries.

The classification percentages of the three classes are lower than those found using the Bayes

classifier. Correlation of the found areas and of the OA measure with age are the only significant correlations. The correlations with biomarker and histology data are not significant due to the amount of data. The OA measure correlates a little better with age than the classes *Blue* and *Lesion* separately.
The residuals check out fine and the results are therefore assumed to be correct.

Further trials would be data fitting and therefore no more optimizations are tried.

8.4 Conclusion

A Bayes classifier is implemented and tried with different input combinations and different merging of the output classes.
The best classification result is obtained by combining Saturation, B.Sub.R and Red.Var and use all the classes as input and not merging them after classification. The variable combination is not found to be collinear and the predicated age has a correlation with age of 0.56 which is highly significant. With histology the significant correlation is 0.29, while there is no correlation with biomarker. RGB as input gives an OA measure which has a correlation with age of 0.57 and 0.40 with histology, both are highly significant. RGB has a little lower classification percentage and is not collinear.
Other promising variable combinations are removed due to collinearity, found by PCA. The variable combination Tri.G, B.Div.G and Intensity gives an OA measure that correlates with age at 0.58 but the combination is close to collinearity and thus the result is doubtful.
The predicted age does not reach the age of the oldest mice. Adding the logarithm, the squared or the exponential of the input does not improve the model.
The obtained results are not perfect, but reasonable. The correlation of the OA measure to age is a little better than obtained by Visiopharm [1]. The used studies are not entirely the same, so it would be fair to say that they match.

The classifier is tried with only three merged classes (*Blue*, *Purple* and *Lesion*). The best variable combinations found by regression analysis do not perform as well as other combinations and are collinear. RGB is not collinear and gives the best classification percentages. The resulting OA measure has highly significant correlation with age of 0.57 and to histology of 0.40.

A second classifier is implemented and based on manual decision boundaries. It describes *Blue*, *Purple* and *Lesion* by a few rules. The threshold of these rules are optimized and result in reasonable TDCM measures that are below the ones earlier obtained. The OA measure has a high correlation with age of 0.44 but does not correlate with histology nor biomarker.

The classification percentages for the three ways of classifying are alike, with the manual decision boundaries as the worst. A low correlation is not the same as a poor classification. The relative areas or the OA measure cannot have a relationship with age that correlates 100 % due to biological variation and intensity variation.

It is surprising that the classification with the original RGB is so close to the optimal obtained classification result. If the intensity variation can be removed, it must be possible to improve the obtained result.

Chapter 9

Clustering

This chapter describes the second solution approach for the hypothesis "*the relative amount of the lesion area increases with age and the relative amount of healthy area decreases with age*". When the class centers are not located at the same positions from image to image, clustering could be the solution, because it groups pixels together in "natural" groups, without a priori knowledge nor training. The solution is therefore unlabelled groups that might have to be identified afterwards, depending on the use. This identification can still be a challenge.
The crossing between two clusters should be where there is a change in pixel concentration or no pixels at all. As mentioned in the Problem Analysis, a literature search is conducted on clustering. The outcome is compiled to get an overview of the different types of clustering. The survey is located in Appendix K, where also distance measures (metrics) are found, besides a more thorough description of the popular clustering routines; k-means and ISODATA.
Articles about identification of clusters are found; the first article [14] uses clustering after classification to adjust the classes to more correct positions. The other article [15] obtains class centers by manual labelled data and after clustering, it tries to match the resulting clusters with the obtained class centers. For both articles, the purpose is to obtain more precise classes with respect to each image.

9.1 Initial Clustering Trials

S-Plus has functions for k-means, PAM (Partitioning Around Medioids) and Fuzzy clustering, among others. For these clustering routines, the number of clusters must be manually specified and is here tried with two, three and eight clusters, for each clustering routine on different images (a random sample from each image). The k-means results, using eight clusters, can be seen in Figure 9.1.
It shows that for all four images, the clusters almost split the pixel cloud in the feature space in equal sizes. Especially the two last images were expected to be clustered differently. They both have more concentrated points in the middle of the point cloud than in the ends of it, so they seem to have three natural clusters, but the clustering does not match this. The two samples are therefore also shown with three cluster trials, but again the distribution of the points is not matched by the clustering, see Figure 9.2. The middle cluster in each image should be larger, if each of the clusters should cover a "homogeneous" area.

Using PAM, Fuzzy clustering and / or just two clusters give similar results and do not seem

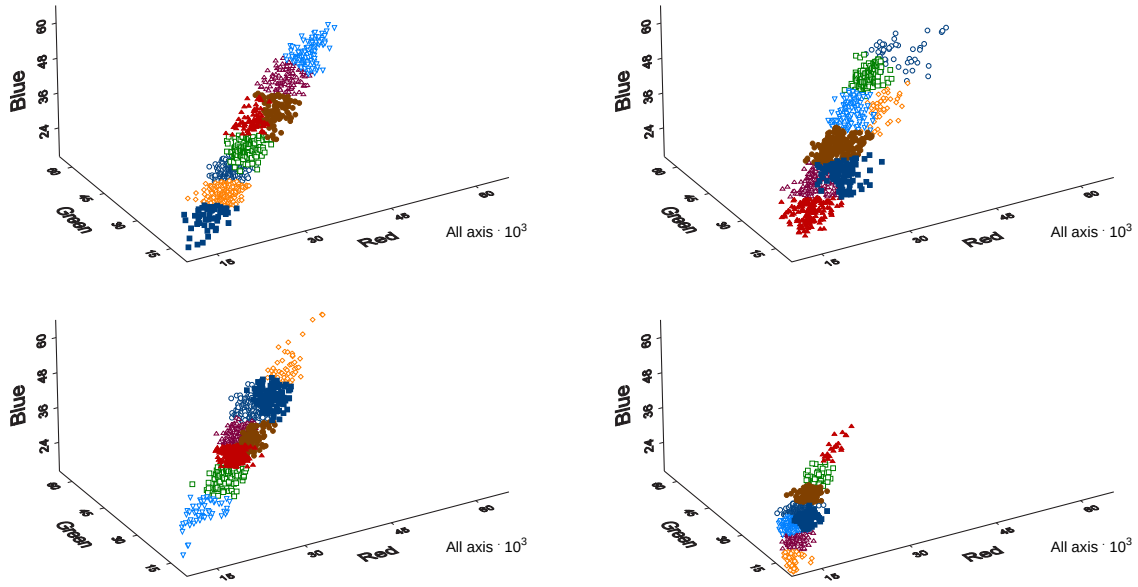


Figure 9.1: Examples of k-means clustering with eight clusters. Top: Samples from the ROI of STR1N-154-09 and STR1N-09-101. Bottom: Samples from the ROI of STR1N-15-102 and STR1N-40-10.

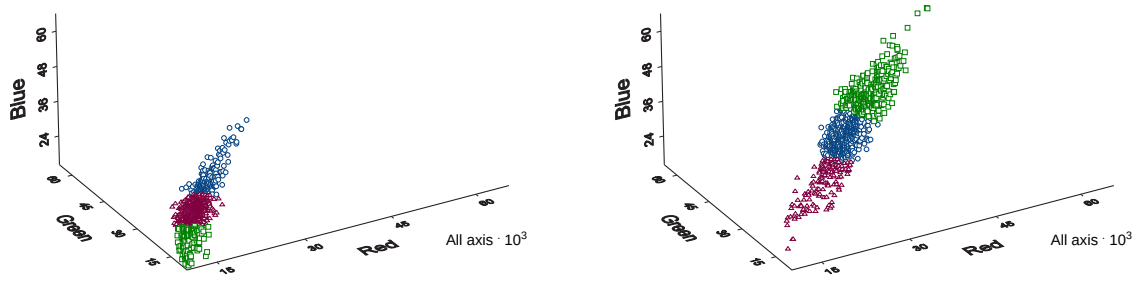


Figure 9.2: Examples of k-means clustering with three clusters. Left: Sample from the ROI of STR1N-40-10. Right: Sample from the ROI of STR1N-15-102.

directly usable. The tried clustering routines work best at separated or almost separated clouds.

A usable clustering routine should look in the neighborhood of the feature space for a given pixel to see how many pixels are present within a certain diameter in order to decide if the pixel in focus is related to them strongly enough, to be in the same cluster. These clustering types are not implemented in S-Plus nor in SAS.

Manual trials of splitting up the feature space into natural groups (splitting when there seems to be a change in the concentration) does not give usable results. The main groups are not found reasonable due to too much overlap. If there is a reasonable border separating the types of areas, then it is not visible in the scatterplots. Hence it is not believed that other clustering routines in Appendix K will result in usable clusters.

9.2 The Use of Non-natural Clusters

Visiopharm has suggested a clustering approach, which should be tried. The solution suggestion is to cluster each image with, say, eight clusters and then calculate the difference in distance between the clusters. If some of the differences are large enough, then the respective clusters must represent different types of areas (some represent parts of the healthy area while others must be lesion clusters). If two clusters are not that different, they should be merged. The idea is then to identify the clusters by using universal classes e.g. the class means found by manual labelling.

The approach does not necessarily demand natural and separable clusters, but merging and especially identifying them afterwards will give problems due to the intensity variation in the images and thus the universal classes are hard to define precisely.

A modified solution night be to use a simpler identification of the clusters without universal classes; the blueish clusters (low intensities) must be healthy, while the bright / white clusters (high intensities) must be lesions. Purple is a problem though, because after the suggestion was presented, it is found to be an early lesion stage. It overlaps with blue and is positioned parallel to blue in the elongated point cloud in feature space. Clusters that are too near / alike all the other clusters might overlap or it might be doubtful which type of area they represent and these should therefore not be included.
The OA measure could be defined as the ratio between the amount of healthy vs. lesion area, or as the percentage of lesion area compared to the entire ROI area, see equation 9.1.

$$OA\ measure = \frac{N_{lesion}}{N_{healthy}} \qquad or \qquad OA\ measure = \frac{N_{lesion}}{N_{ROI}} \tag{9.1}$$

where N_{lesion}, $N_{healthy}$ and N_{ROI} is the amount (in pixels) of lesions, healthy and the ROI, respectively.
The suggestion is somewhat more complicated because blue and purple areas will end up in the same clusters. However, purple and blue are defined earlier, hence the respective clusters can be separated in two groups using these definitions. The original and new solution is outlined in Figure 9.3.
The clusters are used by the following hypothesis: "*The clusters which are positioned at a certain distance from one another, must represent different types of areas. The bright areas must be lesions. The clusters with low intensities can be split in purple, which is believed to be the first lesion stage, and the rest must be healthy*".

9.3 Results

The K-means clustering is implemented in C++ and the images are clustered with $k = 2,\ 3 \cdots 8$. For five or more clusters the brightest cluster is a bright lesion or an exceptional bright area. For a constant number of clusters, it differs how many of the subsequent brightest clusters represent bright lesion. A clustering example is shown in Figure 9.4. The clusters are sorted so that the further away from origin (0,0,0) a cluster center is, the brighter it appears.
Figure 9.4 shows that the center of the bright lesion is represented by the brightest cluster and in order to represent the entire lesion, the three brightest clusters have to be used. The bright and fibrillated blue areas in the bottom of the image are clustered up to the second brightest

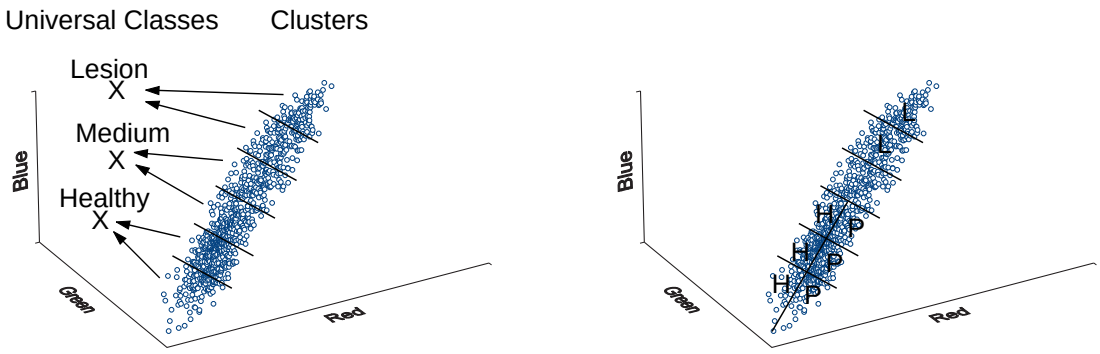


Figure 9.3: Sketches of the two suggested solutions. Left: The original suggestion by Visiopharm. Right: The simplified solution approach. H, P and L represents Healthy, Purple and Lesion, respectively.

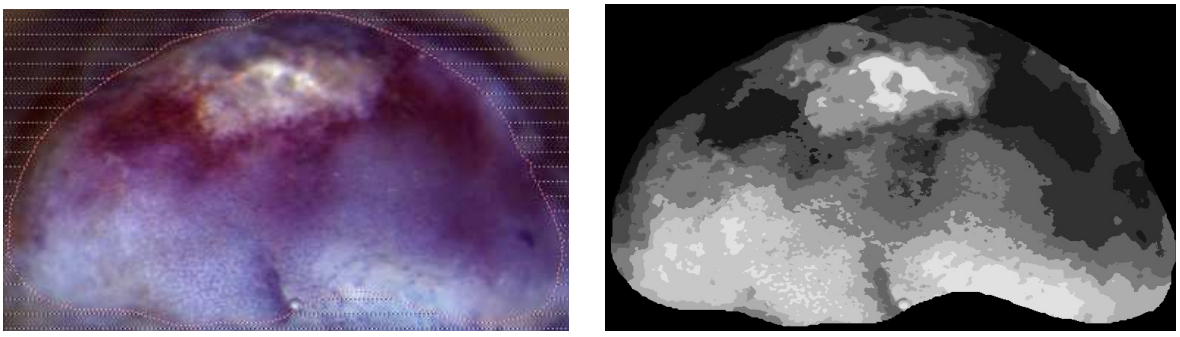


Figure 9.4: Example of clustering using eight clusters on image STR1N-09-101.

cluster. Here is thus an overlap problem, but it might be solved with the manual decision of the main groups. In the classifier these areas are also mixed and relative areas still show significant correlation to age. Thus no action is taken which keeps the objectivity of the clustering. The two darkest clusters approx. represent purple and the next two clusters approx. represent healthy blue.

The Euclidean distance is calculated between each combination of two clusters. The maximum distance (between the first and last cluster) is shown in Figure 9.5, in which also the distances from the first cluster to each of the other clusters are plotted, according to age. It shows that the distance between the cluster with the darkest and brightest intensities has no consistency to age. Some images do not use much of the dynamic range while others are spread over approx. 2/3 of the length of the diagonal in feature space. The distance from the darkest cluster to each of the other clusters grows a little with age, but does not seem convincing either.

The absolute value of the clusters' position is also checked and examples are shown in Figure 9.6. For the darkest cluster it shows that the red and blue intensities are increased with age, hence the darkest cluster of the older mice appears more bright and purple than for the younger mice. For the brightest cluster in each image, the red and green band increase with age

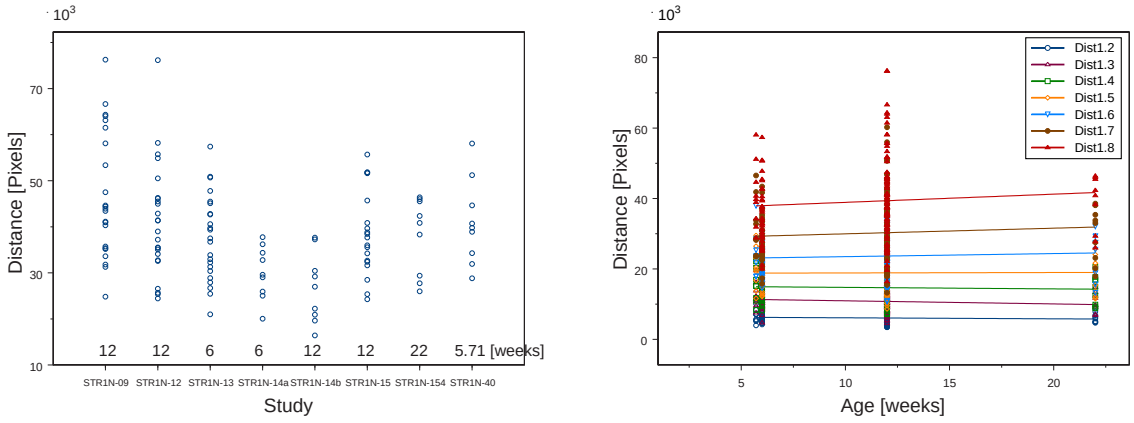


Figure 9.5: Distances between the clusters. Left: The maximum distance between two clusters in each image. The age of each study is shown as guidance. Right: The distance from the first cluster to each of the other clusters. A linear regression line is added.

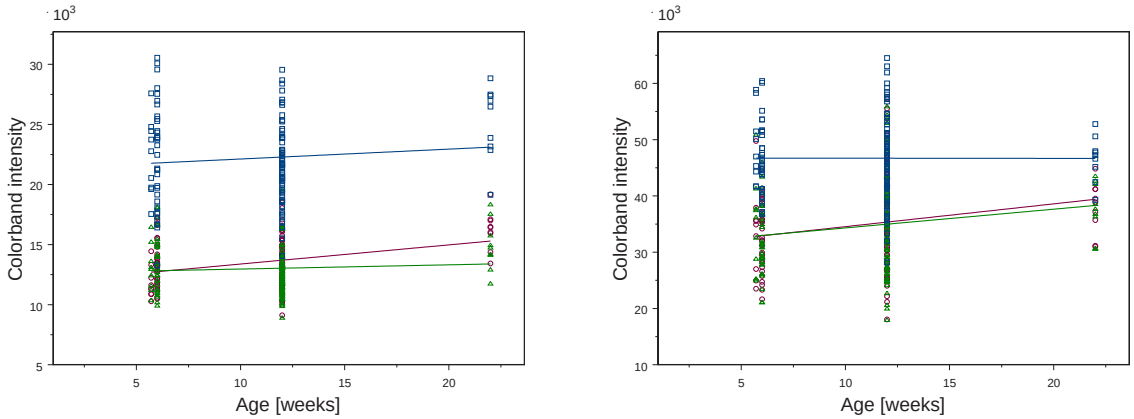


Figure 9.6: Intensity of each colorband for the cluster centers. Left: The darkest cluster in each image. Right: The brightest cluster in each image.

and hence represent areas that are brighter and more white. The tendencies are pointing in the right direction but larger differences are expected.

The intensity and the amount of pixels in each cluster, in each study, are shown in Figure 9.7. The intensities of the red and blue band, according to cluster number and study, reveal that some of the studies with 12 week old mice and the study with 22 week old mice increase more than average as the cluster number increases. The lowest increase is seen for STR1N-14b (12 week old mice). The grid is smoothed a bit, so the difference is even larger than shown.
The amount of pixels in the clusters shows that STR1N-14a (study number 3 with 6 weeks old mice) have the largest amount of pixels for the brightest clusters but they have low intensities. This probably makes it possible for the routine not to generate too high an OA measure for

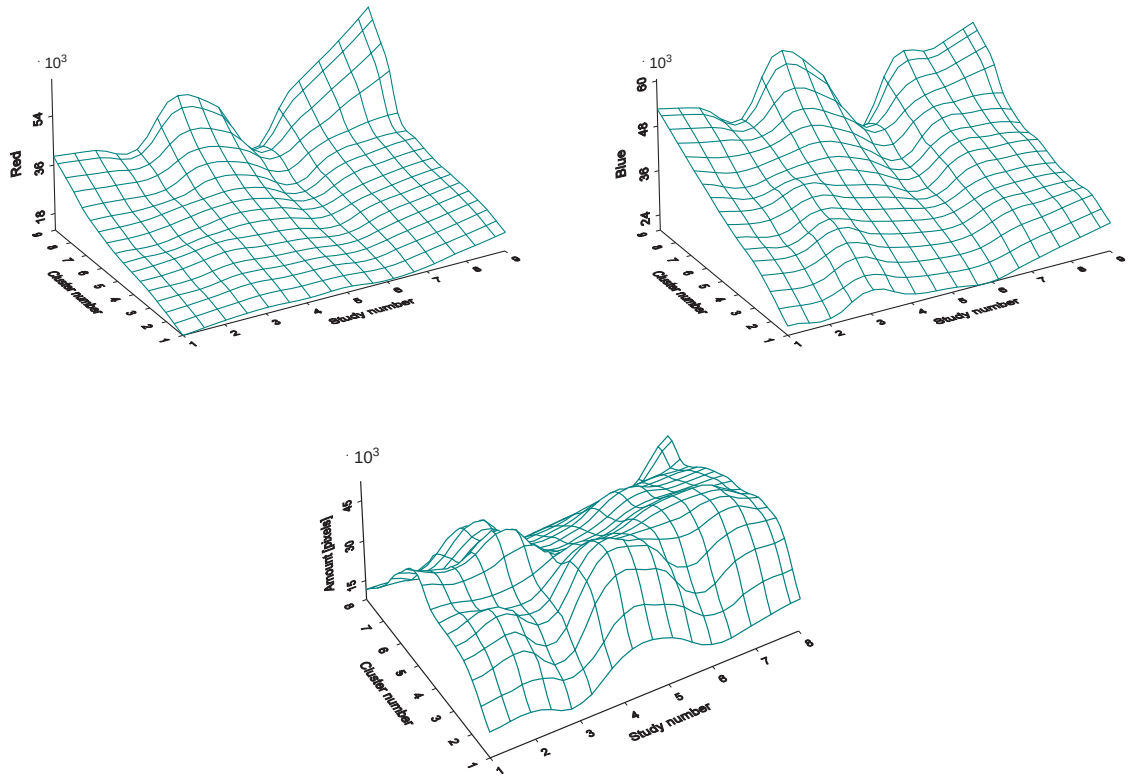


Figure 9.7: Intensities and amount of pixels for the clusters in each study. Top: The red and blue intensities. Bottom: The size of each cluster. Note that the mesh is smoothed a bit. Note also that the studies are sorted by age hence Study number from 1 to 8 represents STR1N-40 (5.71 weeks), STR1N-13 and -14a (6 weeks), STR1N-09, -12, -14b, -15 (12 weeks) and STR1N-154 (22 weeks), respectively.

this study's images. STR1N-14b (study number 6 with 12 weeks old mice) has the clusters with lowest intensities and a low amount of pixels in the brightest clusters. This will give the lowest OA measure of the studies which is not correct.
The above shown information is now shown gathered in one plot, see Figure 9.8. It is not easy to see the 3D structure, but the "2D" image hopefully helps in understanding the shape. The point cloud forms a "bridge" starting with low intensities and a low amount of pixels in the darkest clusters. Both measures increase with the cluster number and for the largest cluster numbers the amount of pixels decrease. The difference between the studies is negligible.

Clusters with intensities 10,000, 20,000 $\cdots$ 70,000 above the darkest clusters intensity are now considered lesions. Their summed area of the ROI area is calculated and correlated with age. Figure 9.9 shows these lesion percentages for distance thresholds of 20,000 and 35,000. Their correlation with age is -0.19 and 0.10, respectively. Using a distance threshold of 50,000 gives a correlation with age at -0.02. It is a relatively simple threshold but a larger connection was hoped for.

As shown above, it is difficult to find any relationship between the studies of the same age

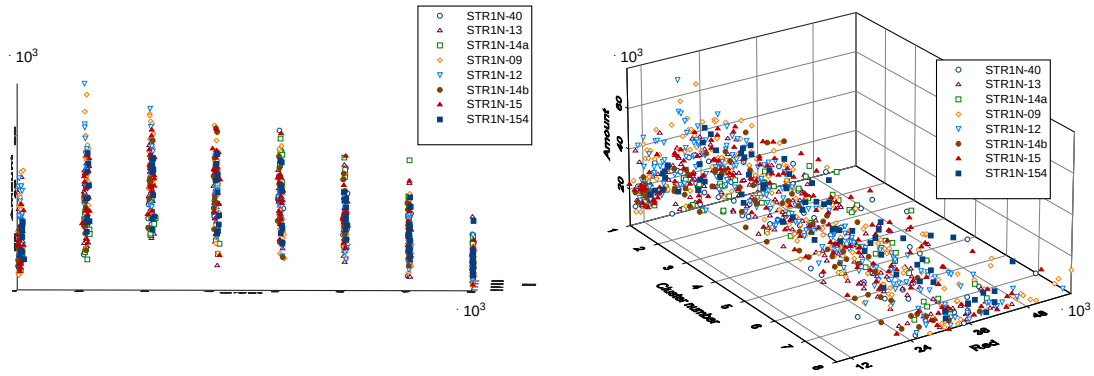


Figure 9.8: The intensity and amount of each cluster in each study. Left: The figure shown from the side with increasing cluster number on the x-axis and amount on the y-axis. Right: The same figure rotated.

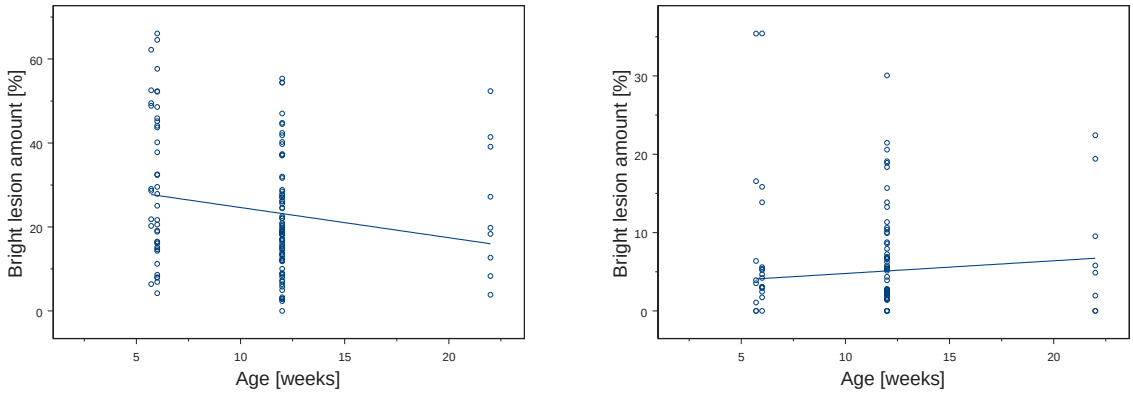


Figure 9.9: Lesion percentages based on distance from the darkest cluster. Left: The result using a distance threshold of 20,000. Right: The result using a distance threshold of 35,000.

or according to age for these clusters.

9.4 Conclusion

A clustering survey is conducted. The feature space of an image consists of one point cloud and the tried routines cluster it in equally sized parts. This is the result of K-means, PAM and fuzzy k-means clustering, tried with different numbers of clusters ($2 \cdots 8$), therefore the clusters represent no reasonable classes and are not directly usable.
Visiopharm has supplied a solution suggestion which is modified a bit due to the intensity variation and the fact that purple is a lesion stage. The resulting approach is to use the non-natural clusters. The ones containing pixels with a high intensity must represent lesions, while clusters containing pixels with a low intensity must represent blue or purple areas. Again, the relative areas of each group should be the basis for the OA measure.
After the images are clustered with the implemented k-means clustering, the results are exam-

ined. These show that there is not much consistency between age, intensity and the amount of pixels in the clusters. If this shows any tendency then it is not consistent for studies with mice of the same age.
Clustering is therefore believed not to be a suitable approach for this project's images.

Chapter 10

Further Classification

Due to promising, but not optimal results, it is in this chapter tried to improve the classification. The images are smoothed in order to reduce noise and two new classification evaluation measures are implemented.
Aventis has verified the higher probability of the bright lesions close to the upper border of the tibia. The area in which the bright lesions can emerge is somewhat sausage shaped. Aventis informs that for the younger mice bright lesions can only appear close to the upper border of the tibia, and only in severe cases bright lesions are found near the middle of the tibia.
It is further informed that fibrillated areas are early lesion stages. These types of areas are developing OA but there is still some proteoglycan left, which gives this fibrillated appearance. The pattern classes and *Lesion perhaps* fit this description.
With the purpose of using this new information, the classification is based on all the classes. The merged ones are not tried because they would merge the pattern classes with the non-pattern classes.

10.1 Improvements

Five things are tried in order to improve the classification

- Noise reduction, by smoothing the images before use.
- The spatial position of an area should affect its probability of being a bright lesion.
- Better detection of the fibrillated areas.
- Sensitivity and specificity are added to evaluate the classifications.
- Collinearity removal for the problematic variable combinations.

These improvements are described below.

Noise Reduction
To improve the signal-to-noise ratio (SNR), the images are median filtered before use. Trials are carried out with filter sizes of 3×3, 5×5 and 7×7 pixels. The sharpness of the original images is not perfect and blurring them, with a window of more than 7×7 pixels, gives images which are too blurred and might affect the classes in a negative way; instead of reducing the variance in each class, the classes might be diluted. Especially classes that are spatially small, like *Lesion yellow*, could be influenced by neighbor areas.
The data set is updated using these filtered images. The scatterplots, the regression analysis

and the classification are repeated for each of the three filters.

Comments on the New Information
The manually labelled classes do not behave according to the new information from Aventis; the bright lesions can occur in other positions than near the upper border of the tibia and especially the fibrillated areas occur in the lower part of the tibia. Image STR1N-09-101 is a good example of this (shown in Section 8.1.3, p. 74). The brightness of study STR1N-14a (in Chapter 4.2, p. 26) is explained by the large amount of pattern areas which are without respect of the spatial position.

This new information combined with visual observations raise new questions; the early lesion stage (fibrillated areas) emerge in the young mice and is clearly visible after 6 weeks, and is independent of the spatial position. At the age of 12 weeks there are less fibrillated areas and not a matching increase of the bright lesions. What becomes of those fibrillated areas which do not evolve to bright lesions? And why do the bright lesions "only" occur near the upper tibia if an earlier stage can occur everywhere?
The fact that bright lesions "only" appear near the upper border are shown in this report and at the same time it sounds reasonable that the fibrillated areas are an early lesion stage. In order for this not to be a paradox, there must be a stage between the fibrillated areas and the bright lesion stage. It could be the purple appearance but that is pure guesswork. Another guess would be biological variance, but that would be too extreme.

In the following classification both types of information are used but independently; the bright lesions are more likely to appear in a sausage shaped area near the upper border of the tibia and the fibrillated areas are probably an early lesion stage and independent of its spatial position.

Using the Spatial Position as A Priori Knowledge for the Bright Lesions
The spatial position of a pixel should have influence on its class probability. The probability map could be used as a priori knowledge for this task but as mentioned earlier, it is based on relatively few images and is hence not precise.
A distance measure from the upper border of the tibia is therefore used, based on the ROI mask for each image. First, cracks and concave areas in the mask are reduced by closing (a morphological operation of dilation followed by erosion). The kernel is a circle with a diameter of 70 pixels which is a rather large size but for a few images it is necessary due to wide cracks in their masks.
For each vertical line the topmost pixel of the mask is identified. Lines without any masks are left out and likewise with the 20 most left and 20 most right vertical lines of the mask. From the resulting line that follows the upper border, the euclidian distance is calculated for all the positions in the image. The recursive euclidian distance transformation (SEDT) is used again. Now the distance of 0 - x pixels forms a sausage shaped area in the upper part of the tibia. Image 10.1 shows the area for a distance up to 100 pixels. Due to no exact information / definition of the size of the area with larger bright lesion probability, the implementation is designed, so the probability of a pixel being a bright lesion is reduced linearly with the distance from the upper tibia. The probability of a lesion near the upper border is higher than without the factor and gets lower the further away from the upper border.

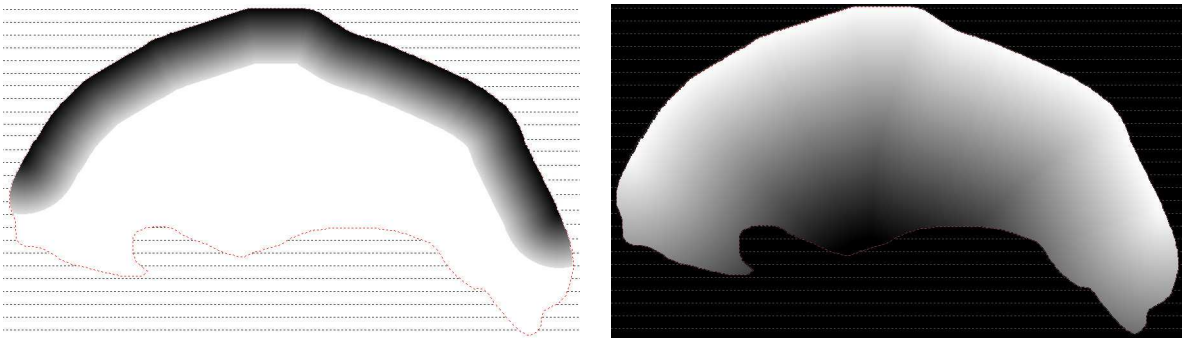


Figure 10.1: Implementation of the bright lesion information. Left: The sausage shaped distance measure shown for a distance of 100 pixels form the upper border. Right: The used probability factor for a pixel to be a lesion. The darker the lower probability of being a bright lesion.

Using the Fibrillation Information

The local variance is hoped to assist the separation of the fibrillated and non-fibrillated areas. The local variance, one for each colorband, is thus three new features. The fibrillation has a fine structure and hence the earlier included variance is based on too large an area. Here the variance is calculated for window sizes of 3×3, 5×5 and 7×7 pixels. It is followed by a mean filter where window sizes of 7×7, 11×11, 15×15, 21×21 and 25×25 pixels are tried. The variance is calculated for the median filtered images and for the original images. This results in 60 combinations and the most promising window combinations are selected based on boxplots of the labelled areas. The local variance is hereafter extracted like the other features were and included in the classifier.

Removing the Collinearity Problem

To avoid the collinearity problem when using x.Sub.y or other problematic combinations, it is the idea to use a mixture of the linear and quadratic classifier (uses the pooled covariance and separate covariances, respectively). The separate covariance matrices are pulled toward the common covariance matrix. This is known as Regularized Discriminant Analysis, proposed by Friedman [13]. The resulting covariance matrices become

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma} \tag{10.1}$$

where $\hat{\Sigma}$ is the pooled covariance matrix and $\alpha \in [0, 1]$. α can be determined by cross-validation and in this case it could be in the order of 0.9 - 0.99, hence keeping most of the separate covariance. This will result in moving the covariance matrix so that the inverse covariance matrix can be calculated correctly, but still keeping the advantages of separate covariances.

Sensitivity and Specificity

Instead of using the TDCM measure, there are two measures that when combined are more precise. They are called "sensitivity" and "specificity" and are normally used in a two class situation in the field of medicine, in which an experiment either gives an effect or no effect. Here is only one boundary separating the outcome (positive or negative) and hence increasing the classification percentage of one class will reduce it for the other. Sensitivity is a measure of how many pixels of the first class that are correctly classified (how many of the positive cases are

found positive). Specificity is a measure of how many pixels of the second class that are correctly classified (how many of the negative cases are found negative). Adjusting the boundary that separates the two classes gives an increased sensitivity and a decreased specificity or vice versa. A Receiving Operator Characteristic (ROC) curve can be plotted where this trade-off situation is visualized and a reasonable boundary can be selected. It is the sensitivity and against (1 - specificity) that are plotted.
The equations are as follows

$$\text{Sensitivity} = \frac{\text{Found positive}}{\text{All positive}} \qquad \text{Specificity} = \frac{\text{Found negative}}{\text{All negative}} \tag{10.2}$$

Both measures should be as close to 1 as possible. The sensitivity is the same measure as the diagonal of the confusion matrix but is not in percentage.
In this project there are eight classes so when calculating the sensitivity for one class, the rest of the classes are considered as one (merged) class and for this, the specificity is calculated. The classification results are now evaluated using these measures.

10.2 Initial Results

The suggested improvements are here tried before implementation.

10.2.1 Noise Reduction

The median filtering improves some of the classification percentages but the effect varies from feature to feature. Table 10.1 shows the result for the best variable combinations for the first classifier. For the original bands separately, the sensitivities are reduced for all filter sizes. For

	Healthy		Lesion					
	blue		purple					
Variable combination		patt.		patt.	bright	perhaps	white	yellow
RGB, original	0.55	0.63	0.82	0.27	0.83	0.63	0.59	0.91
RGB, 3x3 median	0.58	0.63	0.85	0.29	0.85	0.66	0.60	0.92
RGB, 5x5 median	0.50	0.67	0.84	0.17	0.69	0.39	0.70	0.72
RGB, 7x7 median	0.52	0.57	0.71	0.35	0.55	0.35	0.67	0.69
Saturation,B.Sub.R,Red.Var original	0.94	0.70	0.74	0.14	0.74	0.56	0.66	0.87
Saturation,B.Sub.R,Red.Var 3x3 median	0.94	0.71	0.79	0.14	0.76	0.60	0.67	0.88
Saturation,B.Sub.R,Red.Var 5x5 median	0.92	0.73	0.75	0.16	0.67	0.41	0.75	0.16
Saturation,B.Sub.R,Red.Var 7x7 median	0.64	0.65	0.78	0.25	0.51	0.35	0.76	0.15

Table 10.1: Classification sensitivity using different median filters.

the other features separately, there are improvements of single classes but generally the results are the same or reduced. It is the opposite situation for the combined variables, where most classes are improved for the 3 × 3 pixels window and some also for the 5 × 5 pixels window. For the window of 7 × 7 pixels most of the sensitivities are reduced.
The sensitivity and specificity measures are shown in Table 10.2 and are increased using the 3 × 3 pixels median filter for both variable combinations. Most of the sensitivity and specificity measures are increased and none of them are decreased. For these variable combinations the filtering is nothing but an improvement and the 3 × 3 pixels median filter will be used on the images before the classification.

Variable combination	Healthy blue		Lesion purple					
		patt.		patt.	bright	perhaps	white	yellow
RGB, original sensitivity	0.55	0.63	0.82	0.27	0.83	0.63	0.59	0.91
RGB, original specificity	0.88	0.90	0.92	0.95	0.98	0.97	0.99	0.99
RGB, 3x3 median sensitivity	0.58	0.63	0.85	0.29	0.85	0.66	0.60	0.92
RGB, 3x3 median specificity	0.90	0.90	0.93	0.95	0.98	0.97	0.99	0.99
Saturation,B.Sub.R,Red.Var, original sensitivity	0.94	0.70	0.74	0.14	0.74	0.56	0.66	0.87
Saturation,B.Sub.R,Red.Var, original specificity	0.85	0.95	0.96	0.97	0.98	0.98	0.97	0.99
Saturation,B.Sub.R,Red.Var, 3x3 median sensitivity	0.94	0.71	0.79	0.14	0.76	0.60	0.67	0.88
Saturation,B.Sub.R,Red.Var, 3x3 median specificity	0.87	0.95	0.96	0.97	0.98	0.98	0.98	0.99

Table 10.2: Sensitivity and specificity measures of the classification with and without median filtering.

When examining the 3D scatterplots of the different filter sizes (not shown), not much difference is found compared with the original one. The classes, or more precisely the parts (the sample from each image) that each class are merged of, are a bit more concentrated.

10.2.2 The Spatial Position of the Bright Lesions

The distance measure is initially calculated for the labelled classes, see Figure 10.2. The result

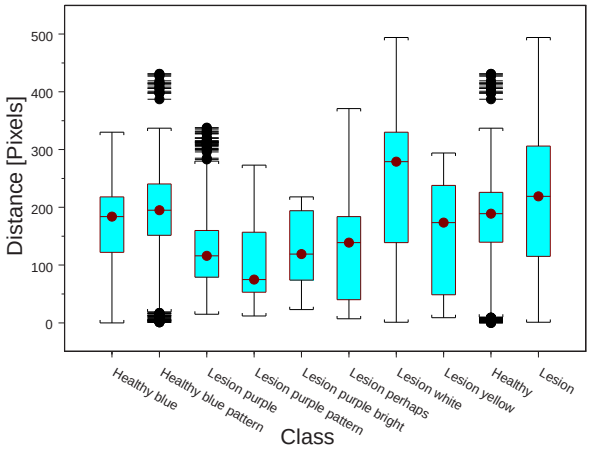


Figure 10.2: Boxplots of the first distance trial.

is large variations and overlapping classes. *Lesion white* has the largest distance between the upper border and its average position which is the opposite of what is expected.
The manually labelled classes do not follow the tendency of a larger possibility of bright lesions emerging near the upper border. Either the areas are marked badly or else the tendency is not correct for the used images (they are not the same as used for calculating the probability map in Chapter 5, and are subjectively selected). The distance of the manually labelled lesions cannot be included in the classification as a normal feature.
The previously shown distance factor is still implemented but as a factor on the bright lesion classes. The probability of a pixel being a bright lesion is increased if it is positioned near the upper border of the tibia and vice versa further away.

This still suggests a problem because the labelled lesions in the classified images are positioned away from the upper border of the tibia (they do not follow the tendency). The sensitivity of the bright lesions will thus be reduced, but the images might be classified better and could give a better correlation with age.

10.2.3 The Fibrillation Measure

The local variance of each band is calculated and extracted, see the boxplots in Figure 10.3. The pattern classes show larger variance than non-pattern classes for all the bands. The local

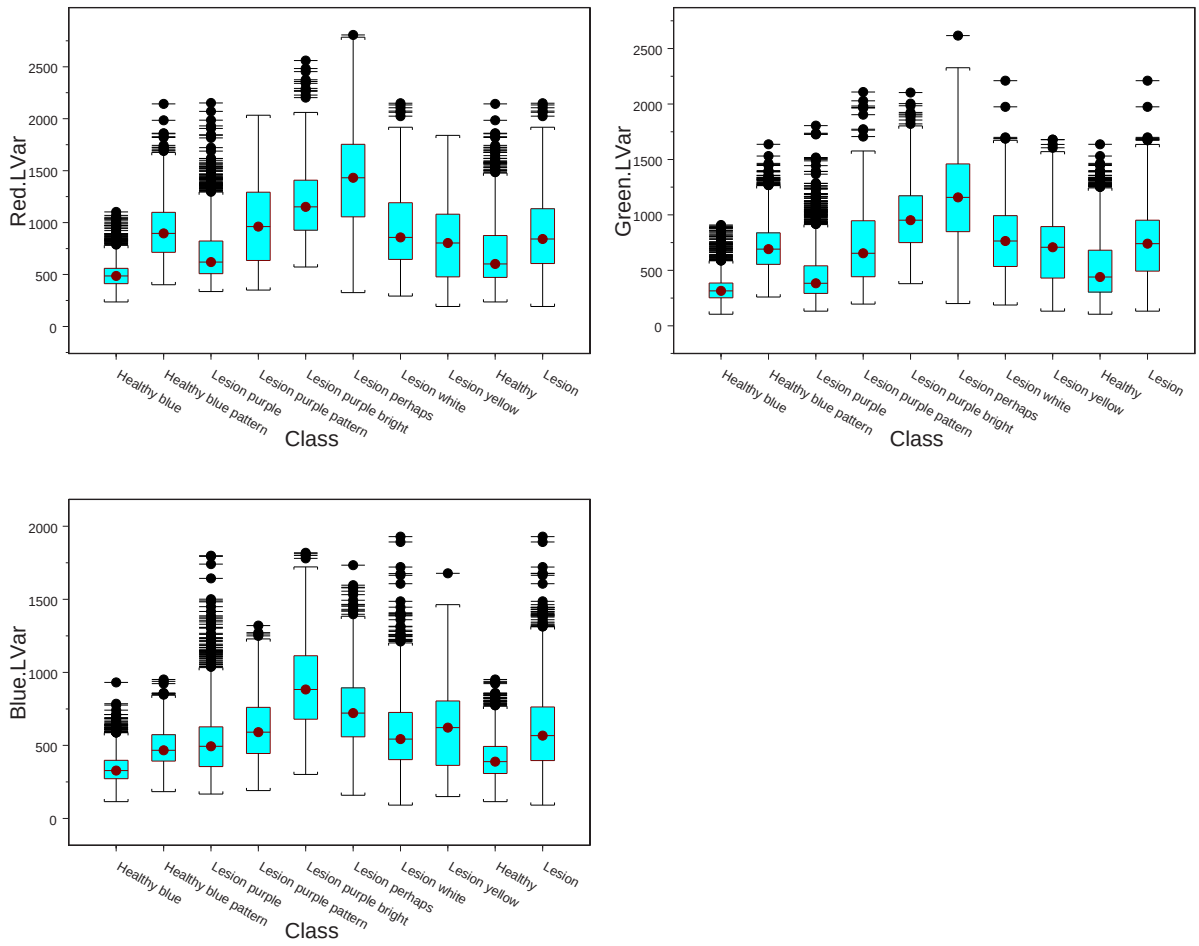


Figure 10.3: Boxplots of the new features and the first distance trial.

variance for the red and green bands gives the wanted difference even though when choosing between so many window sizes and combinations, a larger difference could be expected. Table 10.3 shows the classification results using the three features simultaneously. The classification using the local variance, one band at a time, does not give good results. For the three features combined, the sensitivity of the pattern classes is rather low. It is therefore not surprising that none of the local variance features are selected by regression analysis (neither by forward, backward nor stepwise regression).

	Healthy		Lesion					
	blue		purple					
Variable combination		patt.		patt.	bright	perhaps	white	yellow
Red.LVar,Green.LVar,Blue.LVar sensitivity	0.83	0.56	0.26	0.06	0.66	0.35	0.16	0.09
Red.LVar,Green.LVar,Blue.LVar specificity	0.56	0.91	0.94	0.98	0.97	0.97	0.96	0.97

Table 10.3: Sensitivity and specificity measures of the classification using the local variance features.

A manual trial, combining the local variance of the red band with any combination of two original bands, increases the sensitivity compared to using only two original bands. However, it does not match other classification results e.g. using red, green and blue simultaneously. The classification percentage of *Lesion purple bright* increases but it represents relatively few pixels, so the amount of correctly classified pixels is only increased a bit compared to the other reductions.

10.2.4 Regularized Discriminant Analysis

The common covariance is calculated and the Regularized Discriminant Analysis is implemented. For x.sub.y it does however not solve the problem nor for the other variable combinations with collinearity problems. The inverse of the common covariance matrices still gives problems. The variables are hence collinear for all the classes and the regularization is therefore without effect. This could be predicted for the trichromatic features but was not expected for some of the other variable combinations.

10.3 Results

The 3×3 pixels median filter is implemented and used throughout the chapter.

The sensitivity and specificity are good evaluation measures and more intuitive than the TDCM measure, generated in the first classification chapter. There are of course two measures to examine but it is quite informative.

10.3.1 OA Measure

The OA measures are updated due to the filtered images. With RGB as input, the measure becomes

$$\begin{aligned} \text{Predicted Age} \;=\; & 0 + 0.0672 \cdot \text{RA\_HealthyBlue} + 0.101 \cdot \text{RA\_LesionPurple} \\ & +0.165 \cdot \text{RA\_LesionPurplePat.} + 0.2 \cdot \text{RA\_LesionPurpleBright} \\ & +0.265 \cdot \text{RA\_LesionPerhaps} + 0.310 \cdot \text{RA\_LesionYellow} \end{aligned} \tag{10.3}$$

The regression analysis includes the same relative areas as earlier in the model. The result is a correlation between the predicted age and age at 0.60. This is the best result obtained yet.

For the input combination of Saturation, B.Sub.R and Red.Var the same relative areas are

included in the OA measure and with similar parameters.

$$\begin{aligned}\text{Predicted Age} \quad = \quad & 0 + 0.09 \cdot \text{RA\_HealthyBlue} + 0.0827 \cdot \text{RA\_LesionPurple} \\ & +0.266 \cdot \text{RA\_LesionPurplePat.} + 0.0988 \cdot \text{RA\_LesionPerh.} \\ & +0.344 \cdot \text{RA\_LesionWhite} + 0.273 \cdot \text{RA\_LesionYellow}\end{aligned} \tag{10.4}$$

The predicted age correlates with age at 0.56. This is the same result as without the filtering, hence the improved sensitivity and specificity do not improve the result. This is surprising and might indicate that the correlation, using classification, cannot get much better. Even better sensitivity and specificity should be obtained, before this assumption is confirmed.

10.3.2 Distance Measure

The distance measure is implemented and tested using RGB as input. The distance is converted to a factor which is multiplied on the classification distance (which is minimized in the classification routine).
From a center point, the factor increases with the distance away from the border and vice versa closer to the border. If the center is positioned at zero distance from the upper border, all pixels will be scaled with a factor above 1, which increases the classification distance and therefore reduces the pixels' probability to be classified as a bright lesion class. If the center of the factor is positioned in the middle of the tibia, the probability of it being a lesion is increased near the border.
The position of the center is tried at different distances from the upper border, and the regression analysis, OA measure and its correlation with age, are calculated each time.
First, all the lesion classes are multiplied by the factor, but the results are below the previously obtained ones. When it is applied to only *Lesion white* and *Lesion yellow*, it improves the result, see Figure 10.4. It shows that the optimal position of the center is approx. 500 pixels away from

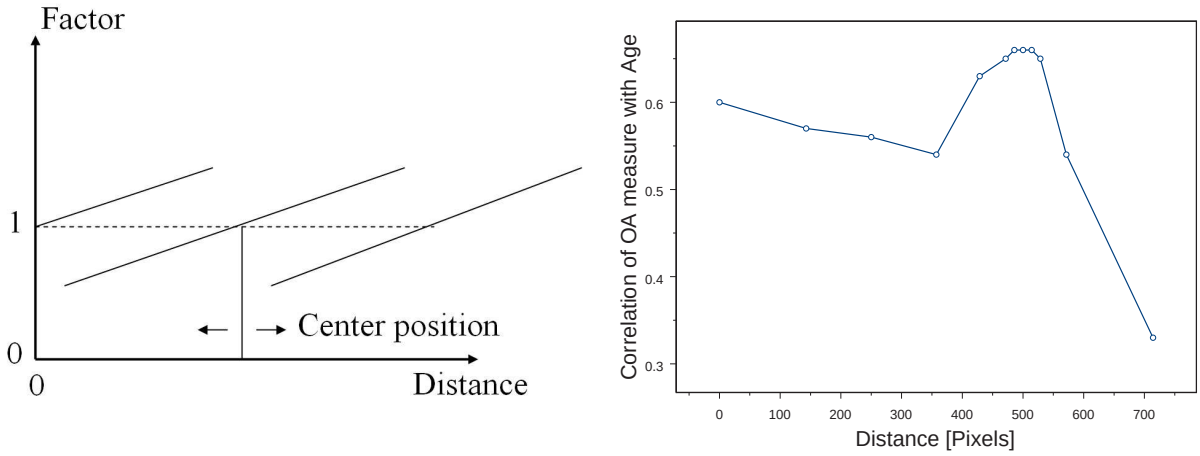


Figure 10.4: Effect of the distance factor on the bright lesion classes. Left: The distance factor tried at different positions. Right: The resulting optimization curve for the correlation of the OA measure with age.

the border. This means that all the pixels' probabilities of being bright lesions are increased.

The probability is approx. tripled at the upper border and is linearly decreased to a factor of one, 500 pixels away. The OA measure now correlates with age at 0.66 and includes the bright lesion classes

$$\begin{aligned} \text{Predicted Age} \;=\; & 0 - 0.21 \cdot \text{RA\_HealthyBlue} - 0.235 \cdot \text{RA\_HealthyBluePat.} \\ & -0.203 \cdot \text{RA\_LesionPurple} + 5.73 \cdot \text{RA\_LesionPurpleBright} \\ & +0.15 \cdot \text{RA\_LesionWhite} + 0.131 \cdot \text{RA\_LesionYellow} \end{aligned} \tag{10.5}$$

The sensitivity of these lesion classes is increased (to 0.86 and 0.96 for *Lesion white* and *Lesion yellow*, respectively) but they find too many of the other classes' pixels. The specificity for the bright lesion classes is therefore reduced.

The basis for the obtained result is not deeply founded. A trial with an OA measure using only the two bright lesion classes gives a low correlation with age. The plot of the residuals is similar to the one earlier shown and hence the regression analysis itself is found to be correct. The interpretation of the optimization is not clear and thus the result might be data fitting. The labelled lesions, using the eight classes, were earlier revealed not to be positioned near the upper boarder. Without "better" labelled classes and more background information, further optimization is not carried out.

10.4 Conclusion

Aventis has verified that there is a higher probability of the bright lesions being positioned near the upper border of the tibia than in the middle of it. Actually it is only in severe cases that lesions should be found near the middle of the tibia. Aventis also informs that the fibrillated areas are an early lesion stage. Five improvements for the classifier are suggested, partly based on this new information.

The noise reduction improves the classification sensitivity and specificity for up to a 7×7 pixels median filter. The best filter size is found to be 3×3 pixels and it is used in the further classification. It improves or matches all of the sensitivity and specificity measures for the RGB input and improves the correlation with age to 0.60. For the combination of Saturation, B.Sub.R and Red.Var all the sensitivity and specificity measures are also improved or matched, but the correlation of the resulting OA measure with age is the same (0.56).

The collinearity of single classes can be removed by Regularized Discriminant Analysis. The color transformations for this project result also in collinearity for the common covariance matrix and hence the problem is not solved.

The new evaluation measures (sensitivity and specificity) are quite informative concerning the trade-off between classes using various combinations of input variables.

The classification of the fibrillated areas are tried improved using local variance. For the red and green bands, there is difference between the pattern and non-pattern classes. A regression analysis does not include any of these new measures in the two or three variable combinations suggested. Replacing any of the red, green or blue bands with either the local variance of red

or green, reduces the classification result.

The a priori knowledge of the position of the bright lesions is implemented. The distance to the labelled classes shows that *Lesion white* has the mean position that is furthest away from the upper border. This is the opposite of what was found earlier and in contradiction to the verification by Aventis. The distance is therefore not used as an input variable but used as a factor in the classification. It adjusts the pixels' probability of being *Lesion white* or *Lesion yellow*. The probability is linearly decreased with the distance from the upper border. The optimum is found by increasing the probability of being a bright lesion for all the pixels and results in a correlation of the OA measure and age at 0.66. There is basis for further optimization but also for validation of the result because the interpretation of the optimization is uncertain.

Chapter 11

Theory

This chapter contains theory on some of the used functions and models. Some of the information might be *common knowledge* and is explained with the purpose of a more explicit formulation of the approach used.

11.1 Correlation and Significance

The correlation routine [8], used in this project, is the sample estimation of the true Pearson product-moment correlation

$$r = \frac{\sum(x_i - \overline{x}) \cdot \sum(y_i - \overline{y})}{\sqrt{(\sum(x_i - \overline{x})^2 \cdot \sum(y_i - \overline{y})^2)}} \tag{11.1}$$

where r is the normalized correlation coefficient, x and y are two variable vectors and i is the counter for variable vectors going from 1 to N, where N is the length of each variable vector.

The p-value of the correlation is found by calculating the test value

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{11.2}$$

where t is the test value and r and N are the same as above.

The test value is matched by table lookup of critical values for the t distribution (with parameters t and df = N-2). The found level of significance is halved in order to obtain the two-sided level of significance.

11.2 Modelling of the Intensity by Position, Area and Age

Analysis of variance is, among others, used to examine the effect of age, position etc. on each colorband of the aligned images.
The analysis of variance model [8] includes the following variables and effects:
There are four age groups: Age.
For each Age, there is one or more studies: Study(Age)
For each Study, there are some images: Image(Study·Age)

There are two types of areas (middle and border): Area
For each Area, there are some sample positions: Position(Area)

The input effects become: Age(i), Study(j), Image(k), Area(l), Position(m)

The used model is

$$\begin{aligned} Color\ rep \quad = \quad & \mu + Age_i + Study(Age)_{j(i)} + Image(Study \cdot Age)_{k(ji)} + Area_l \\ & + Area(Age)_{l(i)} + Area \cdot Study(Age)_{lj(i)} + Area \cdot Image(Study \cdot Age)_{lk(ji)} \\ & + Position(Area)_{m(l)} + Position \cdot Age(Area)_{mi(l)} \\ & + Position \cdot Study(Age \cdot Area)_{mj(il)} \\ & + Position \cdot Image(Study \cdot Age \cdot Area)_{mk(jil)} + z \end{aligned} \tag{11.3}$$

where μ is the mean value and z is the error or noise.

The components of the equation are sorted in systematic / fixed effects (represented with small letters) and in random effects (represented with capital letters).

$$\begin{aligned} Color\ rep \quad = \quad & \mu + age_i + area_l + area(age)_{l(i)} + position(area)_{m(l)} + position \cdot age(area)_{mi(l)} \\ & + Study(Age)_{j(i)} + Image(Study \cdot Age)_{k(ji)} + Area \cdot Study(Age)_{lj(i)} \\ & + Area \cdot Image(Study \cdot Age)_{lk(ji)} + Position \cdot Study(Age \cdot Area)_{mj(il)} \\ & + Position \cdot Image(Study \cdot Age \cdot Area)_{mk(jil)} + z \end{aligned} \tag{11.4}$$

This equation is examined using PROC MIXED in SAS, in which the first line of the equation are the mixed inputs and the rest are the random effects.

For the red band the SAS call is as follows

```
PROC MIXED data = WORK.CanDiskTot;
class Age Area Position Study Image;
model Red = Age Area Area(Age) Position(Area) Position·Age(Area);
random Study(Age) Image(Study·Age) Area·Study(Age) Area·Image(Study·Age) ···
    Position·Study(Age·Area) Position·Image(Study·Age·Area);
lsmeans Position(Area) /pdiff=all;
```

Reduction of the model is carried out by removing the least significant systematic effect from the model and then executing it again. When only significant effects are left, the process is stopped and the rest of the effects constitute the model.

The last line of the call (lsmeans Position(Area) /pdiff=all;) tells SAS to estimate the mean values of the bands for each Position and also the difference between them.

11.3 Bayes Classifier

The used classifier is a Bayes classifier. Each pixel is classified according to which class it has the largest probability of belonging to. The probability is based on the mean and covariances.

Loss functions and a priori knowledge can also be included in the probabilities. Often a common (pooled) covariance matrix is used, but here the variances differ so a separate covariance matrix is used for each class. The distance from class a to b is not the same as from class b to a, due to the different variation of the classes. This is known as Mahalanobis distance and results in a quadratic classifier where the planes that separate the classes are not flat but quadratic surfaces (ellipsoids, paraboloids or hyperboloids).

SAS uses cross-validation (or leave-one-out) where the pixel in question is left out of the calculation of the mean and covariance matrix. The effect of this is minimal in cases without extreme outliers, and is left out here. The distance measure to the discriminant function can be seen in equation 11.5.

$$D_j{}^2 = (x - \overline{x_j})' {cov_j}^{-1} (x - \overline{x_j}) + Log|cov_j| \tag{11.5}$$

where $D_j{}^2$ is the distance from class j to the pixel, x is the vector with the pixel intensities, $\overline{x_j}$ is the mean vector of class j and cov_j is the covariance matrix of class j.
In the 1-dim. case cov_j is just the variance of the variable and in the n-dim. case, it has the size $n \cdot n$. Likewise is it for $\overline{x_j}$; for 1-dim. it is a scalar and for n-dim. it is a vector of length n. The classifier is implemented to handle with 1 - 3 dimensions.

The classifier is tested against the results obtained in SAS. They are not totally the same, but that is due to the fact that SAS was trained and tested on the random sample, while the implemented classifier is tested on all the labelled data.

11.4 True Diagonal of the Confusion Matrix

The classification percentage of a class is normally defined by how many pixels from the class that are correctly identified. This is not a good measure because a lot of other classes' pixels can be included in the found amount of pixels and this is not revealed by the above measure. Here another measure is defined which includes both "the relative amount of the class that is classified correctly" and "the relative amount of the classified pixels which are actually from the current class". This is named the True Diagonal of the Confusion Matrix (TDCM) and can be seen in equation 11.6 and 11.7.
For class 1 the calculations are as follows

$$\begin{aligned} \text{TDCM(Class 1)} \quad = \quad & \frac{\text{Number of correct classified class 1 pixels}}{\text{Total number of class 1 pixels}} \\ & \cdot \frac{\text{Number of correct classified class 1 pixels}}{\text{Total number of pixels classified as class 1}} \end{aligned} \tag{11.6}$$

which for all the classes can be written as

$$TDCM(i) = \frac{{x_{ii}}^2}{\sum_j x_{ij} \cdot \sum_j x_{ji}} \tag{11.7}$$

where i is the class number and $j = 1, \cdots$ the number of classes. The TDCM(i) is the squared value in the diagonal divided by the sum of the row and by the sum of the column which it is a part of. The TDCM(i) goes from $0 - 1$ where 0 is "unusable" and the closer it is to 1, the higher the quality the found pixels have and therefore the more reliable is the following correlation with age etc.

Chapter 12

Discussion

A high correlation to age of the OA measure is a nice goal, but it is the classification behind which should be optimized. If the sensitivity and specificity for all the classes is close to one, a very reliable foundation is obtained for the establishment of the OA measure. This will most likely not correlate with age at, say, a degree of 0.95 or above. With a good classification result, it is possible to reveal the nature of the relationship of each type of area to age and hence to the OA stage.

A sensitivity and specificity of, say, 0.95 or above is rarely obtained and especially not when working with biological images. The results are thus reasonable for *Lesion yellow* with a sensitivity at 0.92 and a specificity at 0.99. The rest of the classes have sensitivities at 0.58 - 0.85 and specificities at 0.90 - 0.99. *Lesion purple pattern* is an exception with a sensitivity at 0.27.

The obtained lesion information (purple is most likely a lesion stage and bright lesions are more likely to emerge near the upper border of the tibia) could probably be included in Visiopharm's present approach to improve it. E.g. the histograms could be obtained for different areas of the tibia (near the border and in the middle of the tibia) and measure these separately.

The manually marked bright lesions in the aligned images have a correlation with age at 0.27. This indicates one of three things; the areas cannot be simplified to just a healthy and a lesion class, or the OA stage cannot be based on the amount of bright areas, or the author is not qualified for the task of marking bright lesions. Due to the belief that purple represents an early lesion stage and the parameter estimates differ within different types of areas of the three main classes in the OA measure, it seems likely that first is the explanation. The two-class approach (healthy or lesion) is too simple and accounts for the low correlation with age.

The advantages of classifier approach compared to Visiopharm's approach using histograms is that the classifier uses more specific information (the classes interpretation and their mean and covariance). Here there is a foundation for a deeper explanation of how the areas develop during the disease and thus how and why the OA measure works. E.g. if the OA measure have to be certified this approach must be a huge advantage.

12.1 Future Work

12.1.1 External Factors that Could be Improved

A calibration of the image acquisition would most probably result in less variation between the images and hence improve the consistency of the class centers.
The tibia curves and are found to be darker near the upper border of the tibia. More diffuse light could remove this problem e.g. by a second, and larger light ring. The ROI could also be drawn a little smaller (the present ROIs can be reduced by e.g. opening), but that would be a waste of information.
An equal amount of mice of the different ages and also an equal age distance could improve the knowledge regarding the development of the OA appearance. The mice are currently a part of a series of tests and hence changing the circumstances is probably not possible.

12.1.2 Internal Factors that Could be Improved or Tried

The collinearity problem could be examined further. Variable combinations chosen by e.g. stepwise regression which performed a little better than using RGB are excluded due to collinearity. There is not a sharp border for where the problem arise and further tests could be used to validate the different variable combinations.

If the clustering approach is to be tested further it could be in the feature space of x.Sub.y or X.Sub.I, but only in 2D due to collinearity using all three variables in these color representation sets. The classes using these sets are more or less sorted in the point cloud according to the lesion stages they represent. The obtained clusters could therefore have a more simple relationship to age and OA stage. R.Sub.I should be one of two x.Sub.I variables used. Alone it has an image-wise correlation with age of 0.61.

When even better classification results are obtained the robustness should be tested. The leave-one-out approach is suggested for this task; the classification is repeated each time leaving out the labelled images for one of the studies when estimating the mean and covariance of the classes. Each time, all studies are classified. The result obtained for each excluded study is averaged, the median is found or the smallest result is used to validate the robustness of the approach.

The labelling might be unprecise, particularly for the bright lesions. The different types of areas could be explained, defined and labelled by the staff that work with this project at Aventis. Doctors in this field may have some input that can supply information to the definitions.

It should be feasible to classify the fibrillated areas better. Multiple trials are here conducted using local variance. It is not expected that this approach can be improved further. The other types of areas are not completely homogeneous but there must be usable information in fibrillated areas. They may fibrillate within a certain frequency span and thus the Fourier Transform could be tried. If a new measure is still insignificant (e.g. using stepwise regression), it can be applied after a normal classification to separate the areas in the blue and purple classes in their respective pattern and non-pattern classes.

Chapter 13

Conclusion

The age of the tibias is predicted and used as a highly significant OA measure. The approach is a Bayes classifier, using mean and separate covariance from eight manually defined classes. It uses the median filtered RGB as input and obtains a correlation with age at 0.60. The underlying classification percentages are reasonable even though for the class *Lesion purple pattern* it should be improved.
There is an intensity variation between the images, keeping RGB from performing even better. Various combinations of color representations have shown promising reduction of the intensity variation, but the better performing ones are removed due to suspicion of collinearity.
The result, obtained by RGB, is tried optimized by scaling of the two bright lesion classes. The scaling depends on the spatial position of the pixel in focus. The optimized OA measure has a correlation with age at 0.66. The interpretation of the optimization is not clear and thus the result might be doubtful.
Several trials of improvement, and experiments with clustering, have not lead to better results. The classification of the areas can therefore still be improved, probably by further attempts to reduce the intensity variation.

The correlation of the non-optimized OA measure with age is higher than the corresponding result obtained by Visiopharm. Not all the image data is the same and therefore it would be fair to say that the classifier approach match their measure on the distributions of the colorbands.

The classification was improved by noise removal using a small median filter. Some images are found to be saturated. A few pixels in the red band were found to be below the normally used dynamic range and can be found explicit or removed by median filtering.

Biomarker and histology information are compared and histology is found to be the most reliable measure of the OA stage of these two, if any.

Several tests indicate that studies containing mice tibia of the same age do not behave alike. There is a large biological variation between the tibias and the appearance of their OA stage. Results averaged by each study was expected to show a larger tendency according to age.

The hypothesis, "*the diverse appearances of the same type of area from image to image is mainly due to an intensity variation between the images*", is rendered probable. It has not resulted in

a successful removal nor reduction though.

The hypothesis, "*the bright / white lesions are more likely to appear near the border of the tibia than in the middle of it*", is rendered probable. The result is based on manually aligned images and bright lesion markings in these. More precisely, the bright lesions are mostly found at or near the upper border of the tibia. This result is verified by Aventis.

The hypothesis, "*the purple areas are more likely to appear next to bright / white lesions than in the other positions of the tibia*", is rendered probable. The measure is based on two purple measures and a blue measure, which are defined by the authors perception of the colors.

The main hypothesis, "*the relative amount of the lesion area increases with age and the relative amount of healthy area decreases with age*", is rendered probable. During multiple classification trials, the blue areas have a negative correlation with age while the bright lesion areas have a positive correlation to age. The only found contradiction to the hypothesis is the fact that the class *Lesion white* is not included in the OA measure, based on the RGB values. Other of the bright lesion classes are included though. The purple areas are found to have a positive correlation with age hence, again, it is indicated that purple areas represent an early lesion stage.

The routines used in this project clearly show than automatic image analysis can be used to establish a reliable OA measure and that other approaches can be used than those previously developed by Visiopharm. This way knowledge of the purple lesions and of the position of the bright lesions is obtained. The precision of the OA measure obtained by the classifier approach match Visiopharm's and it is believed that there is potential for even further improvements.

Bibliography

[1] "Image Analysis in Models of Osteoarthritis", Visiopharm, 2002.

[2] "Correspondence with Aventis concerning new tests", Visiopharm, 2002.

[3] "Colour Space Conversion", A. K. Poulsen, Visiopharm, 2003.

[4] Medic-Media ApS (slidgigtlaboratoriet, Rigshospitalet) "www.slidgigt.dk".

[5] Arthropharm Pty. Ltd. "www.arthritis.au.com/htm/home.htm".

[6] "Digital Image Processing", R. C. Gonzalez and R. E. Woods, Addison-Wesley Publishing Company Inc., 1992.

[7] "Image analysis, Vision and Computer Graphics", J. M. Carstensen, Technical University of Denmark, 2001.

[8] "En Introduktion til Statistik", Knut Condradsen, Vol. 2, 5^{th} edition, IMM, 2001.

[9] "Advances in active appearance models", G. J. Edwards, T. F. Cootes and C. J. Taylor, Proc Int. Conf. on Computer Vision, p. 137 - 142, 1999.

[10] "Active appearance models", M. B. Stegmann", "www.imm.dtu.dk/∼aam", IMM, DTU, 2000.

[11] E. Forgy "Cluster analysis of multivariate data: efficiency versus interpretability of classifications", Biometry 21, 768, 1965.

[12] J. MacQueen, "Some methods for classification and analysis of multivariate observations", In Proc. 5^{th} Berkeley Symp. Math. Statist. Prob., 1967.

[13] J. Friedman, "Regularized discriminant analysis", Journal of the American Statistical Association vol. 84, p. 165 - 175, 1989.

[14] D. Gutfinger and J. Sklansky, "Robust Classifiers by Mixed Adaption", IEEE Trans. on Pattern Analysis and Machine Intelligence. Vol 13 (6), p. 552 - 567, 1991.

[15] M. Setnes and R. Babuska, "Fuzzy Relational Classifier TRaind by Fuzzy Clustering", IEEE Trans. on Systems, Man and Cybernetics, Vol 29 (5), p. 619 - 625, 1999.

[16] S. Wesolkowski, "Shading and Highlight Invariant Color Image Segmentation", University of Waterloo, Canada.

The references for the clustering survey are found in Appendix K.

Appendix A

Osteoarthritis

This appendix looks into what osteoarthritis is, its extent, who gets it and why. Information is provided in order to know how it is diagnosed and the possibility of treatment.

A.1 What is Osteoarthritis

Osteoarthritis (OA), also known in Europe as osteoarthrosis, is a degenerative joint disease. The cartilage in the joint between two bones is weakened and in more progressed cases almost or completely gone. The function of the cartilage is shock absorbtion (mechanical stress) and to some extent to guide the movements of the joint. Because of the reduced protection of the bones in the affected joints, the bones are exposed to more mechanical stress and hence they are modified to absorb the larger shocks; e.g. when a person is exercising, the bones grow stronger to handle the larger stress. Unfortunately, in the case of osteoarthritis the bones are only strengthened close to the joints and this can result in hardening of the bone (sclerosis), death of bone cells (necrosis) and the changing of the inner bone blood supply etc. When the cartilage is gone, the bones start to wear down because of the direct contact between them.

A.2 Who gets Osteoarthritis

The older a person gets, the larger the possibility of developing osteoarthritis. At an age of 50 years, around 50 % have osteoarthritis and people in the seventies have approximately 85 % chance of having this disease.

Besides a few exceptions, it is impossible to generalize who gets osteoarthritis. There is an equal amount of cases when looking at sex and race, but the osteoarthritis strikes differently within these groups. People who are overweight have a larger risk of getting osteoarthritis in the knees, but not in the ankles, while smokers have a smaller risk of getting osteoarthritis in the knees (might be due to lower weight than average).

The exceptions are

- Work related osteoarthritis can only be traced to farmers (hips), ballet dancers (ankles), football players (knees) and mine workers (spine).
- In some severe cases it is inherited, caused by a genetic error.

- An injury (like a bone fraction) will result in osteoarthritis after 10 to 30 years due to the overload of the joint at the moment of fraction. People who suffer from osteoporosis (weakening of the bone) have a smaller chance of getting osteoarthritis, probably due to softer bones and thus less mechanical stress in the joints.
- People with hypermobile joints have a higher risk of getting OA.

All animals with joints similar to human beings also have the possibility of getting osteoarthritis, including the dinosaurs. The weather has no influence on the risk of getting OA, but hot and dry weather can reduce the symptoms.

A.3 Cartilage

As mentioned above, the function of the cartilage is shock absorbtion and to some extent to guide the movement of the joint. The cartilage consists of collagen (fibrous proteins), proteoglycan (sugar molecule) and synovial fluid (water and hyaluronan - a sugar molecule). The water binds to the proteoglycans and is "pumped" in and out of the joint due to relaxation and load, respectively. The synovial fluid also works as a lubricator and the result is a very low friction between the bones. The friction between cartilage and bone is one tenth of the friction between ice and skate!
The cartilage is very good at containing water, in fact, a teaspoon of proteoglycans can contain the water from an entire bathtub!
Figure A.1 shows the knee and how OA affects it.

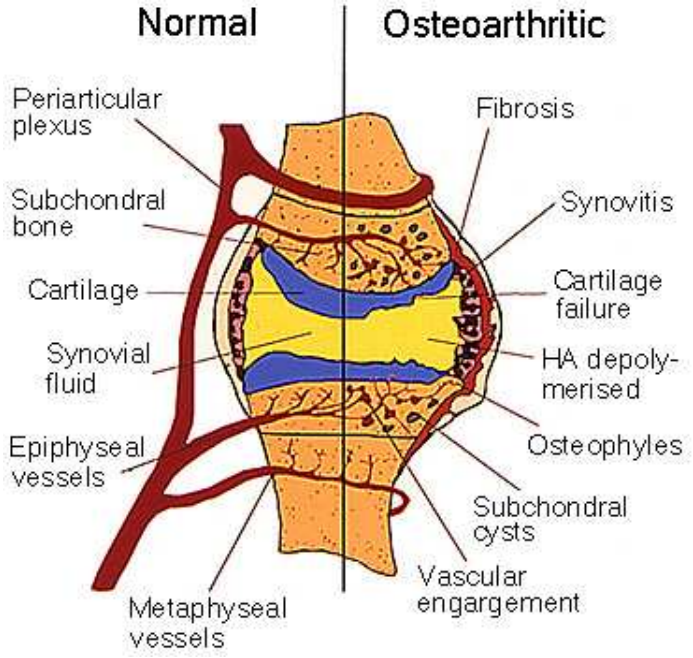


Figure A.1: A model of a knee with and without osteoarthritis.

When the cartilage degenerates it cannot contain as many proteoglycans and these bind more water and as a consequence, the joint swells. Later the proteoglycans degenerate to smaller

sugar molecules which can escape from the cartilage. The cartilage cannot contain the same amount of water as earlier and hence the joint decreases in size. Later again, the fight results in the death of the cartilage cells and the remaining cartilage is heavily reduced.
Normal cartilage is undergoing a constant "fight" during which the cartilage is torn down and rebuilt at the same time. These catabolic and anabolic processes result in a constant renewal of the cartilage, just like it happens in the bones. In OA the anabolic process is weakened compared to the catabolic.
There are no nerves in the cartilage and therefore the disease itself causes no pain. It is the nerves in the bone, muscles and other tissues around the cartilage, which send signals of pain due to changes in these parts. There can be situations where a joint in pain cannot physically be pointed out to be osteoarthritis.

A.4 The Stages of Osteoarthritis

Osteoarthritis is a disease in constant progress. The first symptoms are sore and stiff joints. These symptoms are a direct cause of overload and there can be long periods of time without an indication of the disease. In the middle stage, the pain is more constant but the mobility is still good. In the progressed case the pain is extended and also appears at night. As a consequence, wrong or bad positions of the joints are obtained to avoid it and after a while they become permanent.

A.5 Diagnosis

Several methods are used for diagnosing OA but they are either not perfect or too expensive

- X-rays normally show OA, but not in all cases. Sometimes the opposite happens; a patient with an OA diagnosis based on X-ray may not have any other symptoms (like pain).
- Blood test are currently not as a diagnostic method on humans but it probably will be, as will urine. The problem is that even if it is possible to diagnose OA, e.g. by measuring HP Creatine (HPC), it will only be a snapshot on the present degeneration and the "fight" of the cartilage (the degeneration speed), and not on how advanced the situation is. Hence it is possible to diagnose but not to measure the OA stage.
- Taking a physical sample is a bad idea because the cartilage cannot regenerate and hence a sample would essentially lead to OA.
- Magnetic Resonance Imaging (MRI) provides good images for the diagnose of OA because it provides an image containing both bone and soft tissue and hence a better overall representation of the joint. It is thereby possible to detect OA in its early phases.

Unfortunately MRI is too expensive for routine use, so the initial diagnose is often based on the patients symptoms, X-ray and only if in doubt, MRI is used.

A.6 Treatment

The treatments of today cannot regenerate cartilage but only slow or stop the degeneration, therefore the earlier OA is diagnosed the better.

Some treatment possibilities are

- Medicine. There are many OA drugs on the market, but they are far from perfect. Most of them are only dealing with the symptoms and are hence pain killers, mobility increasing and anti-inflammatory. Most of them have unwanted side effects.
- Natural medicine. In the natural end of the drug scale, ginger and hip is to be found, but they only reduce the inflammation and increase the mobility and can result in reduction of the amount needed of normal medicine.
- Surgery. Surgery is another alternative particularly given to those whom years of drug treatment does not help. The replacement of the hip is by far the most normal surgery related to OA. It accounts for approx. 6,000 out of the 12,000 OA operations in Denmark a year.
- Exercise. For a long time, exercise has been looked upon as bad for OA patients, but now even moderate exercise, the kind that does not overload the joints, is recommended. Swimming is e.g. a good way to exercise the whole body without worsening the OA.

A.7 Medicine Development

Medicine is a huge industry and OA is a large area. Due to increasing life span and the growing economy of the 3rd world, the target group is expanding rapidly. The developed medicine, as most medicine, is tested on animals. This is done to get fast results and later on, to test for side effects.

A.7.1 Osteoarthritis' Extent and Expense

In the USA 1 % of the BNP or 10 % of the health budget is used to treat osteoarthritis. In Denmark there are over 12,000 operations a year related to OA. It is mainly in the hips and knees but also the spine, thumb and the big toe are treated. This results in more than 150,000 "bed days" a year. The public expense is not only operations and medicine but also lost work, etc. and the problem is growing. That is why a great deal of money is spend on developing medicine against this disease.

A.7.2 Testing on Animals

The developed OA medicine is initially tested on mice and is the only test used in this project.

The mouse is genetically altered (named STR1N) which gives it predisposition for developing osteoarthritis. The correspondence between the age of the mouse and how advanced the disease is should thus be high. It is therefore relatively simple to test if the treatment is functional.
With the present method, the mice have to be between 6 and 12 weeks old before the treatment effect can be measured.

A.7.3 Test of Treatment

There are currently two indications of the stage of OA

- Biomarker. Urine samples can be collected from the mice. The degeneration of cartilage can be measured in the urine, but it is a measure of the present process (the speed of degeneration) and not the OA stage.

- Histology. A slice (or several) of the cartilage can be examined and the extent of cracks and dents in the cartilage is a measure of the osteoarthritis stage. This is the current gold standard for testing the treatment effect of medicine.

The histology method is not perfect

- The work concerning the test is quite time consuming (manual work). Actually it takes 6 weeks to estimate the cartilage condition from 75 mice.
- Due to human involvement the result can be error prone and not objective nor reproducible.

The span of time from a test is initiated to the results are at hand, is relatively long. Hence developing a method to speed this test up is time and work saving. The sooner the result of a test is present, the faster can a change or further testing be initiated and hence make medicine development cheeper.

Appendix B

Earlier Work

This chapter resumes the earlier reports from Visiopharm to Aventis.

B.1 Image Analysis in Models of Osteoarthritis (July 2002)

Some of the information under Osteoarthritis and Data and the Image Acquisition come from the reports from Visiopharm and will not be repeated here. The focus here is on the tests and results.
The data used here [1] is ADEP, STR1N-04, STR1N-09, STR1N-12, STR1N-13.
ADEP and STR1N-04 are of poor quality due to bleaching, caused by multiple image acquisitions submerged into water.
Study STR1N-05 was excluded due to bleaching and because it was not blinded. A joint degeneration index (OAIx/S) was developed as a measure of the gross-morphological changes. The index has highly significant correlation at 0.54 with age and 0.43 with biomarker (even though this relationship is not linear). The correlation with histology was only 0.08 and not at all significant.
The routines are tested on RGB, IHS, YIQ and trichromatic RGB and from these color representations Mean, Standard deviation, Skewness and Coefficient of variance are extracted.
The best result is defined by a large significant correlation with histological values. This is obtained using trichromatic green and then by calculating the standard deviation, entropy (uniformity) and Inverse Mode Density Level (IMDL).
The entropy's correlation with age (for trichromatic green) was found to 0.54 and is highly significant.
The observed bleaching results in a little more green than red (mean values), while the opposite case is without bleaching.
The color variation for the young animals is significantly smaller than for the older animals.

Image Quality
The imaging was carefully studied by Visiopharm and the following problems were found

- The noise is worse in the red and blue channel than in the green band.
- Lower spatial resolution in the red and green bands than in the blue band.
- The green channel is shifted horizontally up to 10 pixels, which is bad if looking for edges.
- In some images it is only a part of the green and blue bands that are shifted.
- Double vertical resolution in some lines in the red and blue channel.

- 5 images in the STR1N-05 study are only 8 bit and are therefore excluded.

Other problems were also found and fixed, so the present data is without them.

The speckle, and some of the other problems, have been reduced significantly due to image reduction (cubic interpolation to the half size in each direction).

B.2 Correspondence with Aventis Regarding new Tests (October 2002)

The second report [2] first concludes that the High treatment of STR1N-04 has aggravated the cartilage and hence a low biomarker (most of the cartilage is gone so the turnover decreases). This hypothesis is assisted by the low doses treatment where the biomarker is significantly higher than for the control group. The Histological data does not supply agreement nor disagreement with this hypothesis.
Two new studies with 40 and 154 day old mice were added with the former STR1N-05, STR1N-09 and STR1N-12 which contain 12 week old mice and STR1N-13 which contains 6 week old mice. ADEP and STR1N-04 are not included due to bleaching.
The trichromatic blue is found to be related to the amount of stained cartilage based on the relation to age. The staining result is purple (blue and red) rather than just blue. The entropy seems to increase as the staining intensity is decreasing, but is also high for minimal damage! The combination of the entropy of trichromatic green and trichromatic blue shows a U-shape making it possible to very accurately measure the level of degeneration. If only one of these should be used it would have to be the trichromatic blue (its average value).

The biomarker and histology are no good measures, but for the untreated animals (control groups) they are assumed usable.

B.3 Colour Space Conversion (June 2003)

The Bachelor thesis by A. K. Poulsen [3] describes a lot of different color transformations. He also tests the effect of the hue origin, which can be arbitrary chosen, when transforming from RGB to IHS colorspace.
The analysis is based on the studies STR1N-04, STR1N-09, STR1N-12, STR1N-13, STR1N-14, STR1N-40 and STR1N-154. The best correlation with age is obtained from the mean values of Saturation in HLS, Saturation in HSV, Value in HCV and Trichromatic red. Test of the Hue origin in IHS transformation shows a large dependency on the angle and the best (out of four) was found to be $\frac{\pi}{3}$, the opposite of the blue color dominating the tibia images. The resulting correlation is 0.24 with a level of significance at $6.12 \cdot 10^{-3}$ for the standard deviation.

Appendix C

Alignment of the Images

The images are aligned using the following approach, referring to Figure C.1.
The smooth contour of the right tibia, illustrated by the black line, is visually identified (not

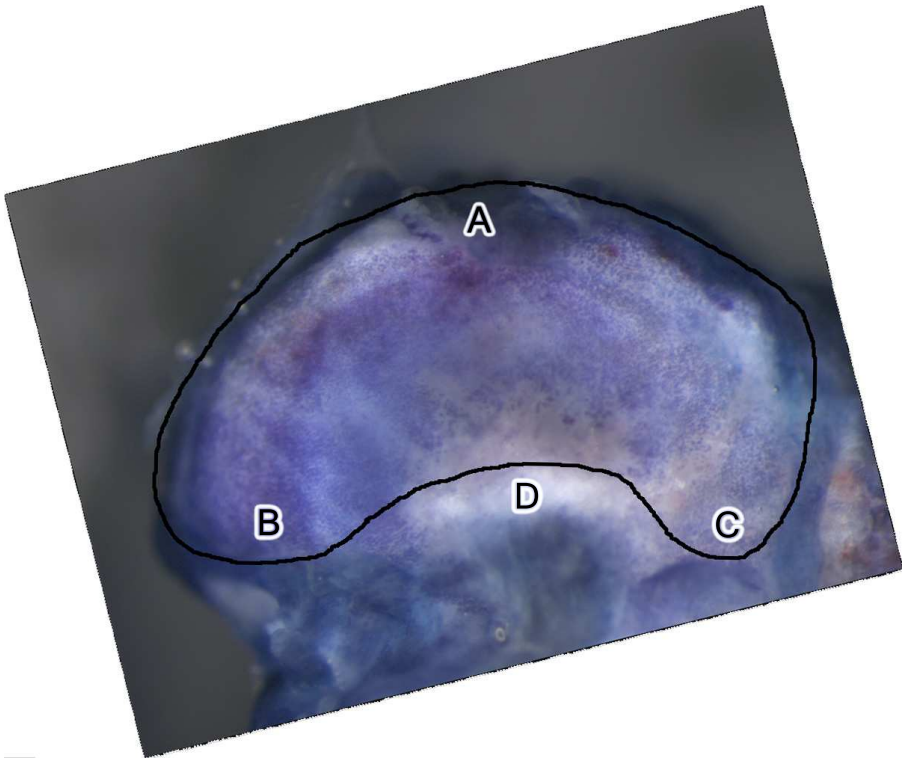


Figure C.1: Alignment example (STR1N-15-102). The images are rotated so the positions B and C are approx. horizontally aligned and then the images are scaled and moved so the positions (mainly A and D) match those from the other images.

drawn) and from this, the four positions denoted A - D. The smooth contour is for the top of the image in the out-of-focus area, but without the cartilage lumps. In the part next to the left side of the tibia, the smooth contour is between the in-focus and out-of-focus areas.
The image is rotated so that the positions B and C are horizontally aligned. The image is now scaled (retaining the x/y ratio) and moved iteratively so that mainly the positions A and D fit

the corresponding positions in the other images.
To move and scale the first image in each study and to make sure that they correspond to the other studies, the black line in Figure C.1 is drawn physically on the screen (on vita-wrap)! This is very useful to align the images and to run through them afterwards in order to check them.
The images are cut to a square of the size of the largest original row / column size, resulting in a size of 1044×1044 pixels.

Appendix D

Aligned Images

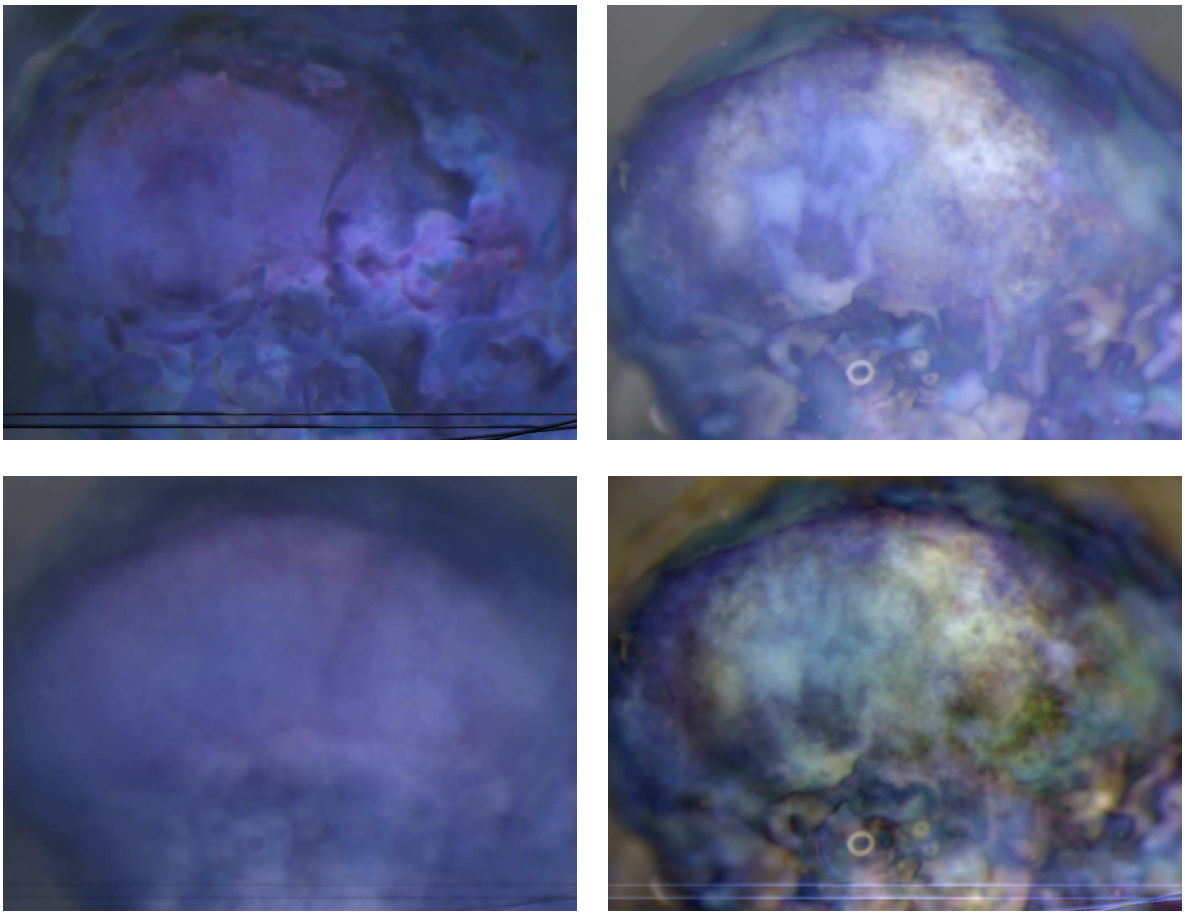


Figure D.1: Aligned images of the 5.71 week old mice. Top row: Minimum and Maximum image. Bottom row: Mean and variation image. Note that the images are brightened individually for a better perception and that the lines in the images are due to an artifact of the alignment.

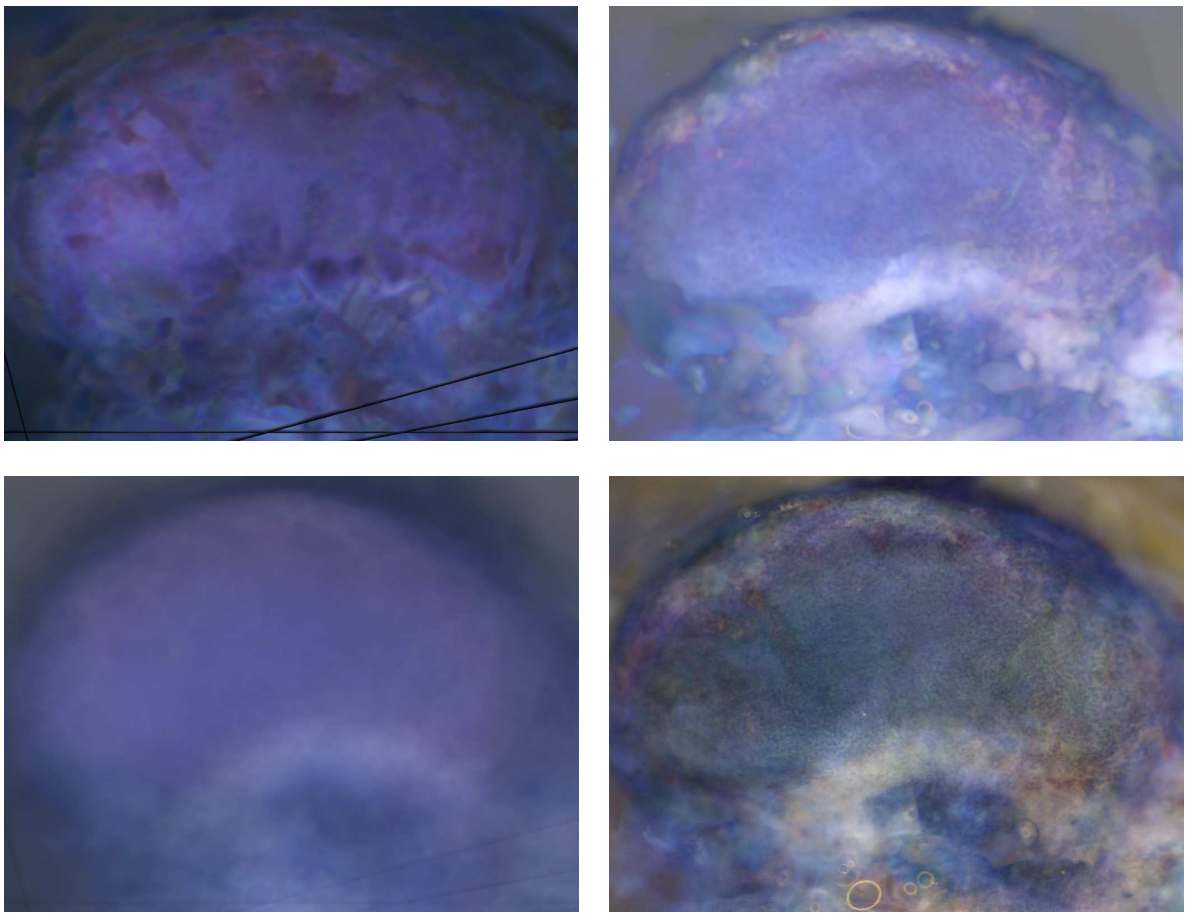


Figure D.2: Aligned images of the 6 week old mice. Top row: Minimum and Maximum image. Bottom row: Mean and Variation image. Note that the images are brightened individually for a better perception and that the lines in the images are due to an artifact of the alignment.

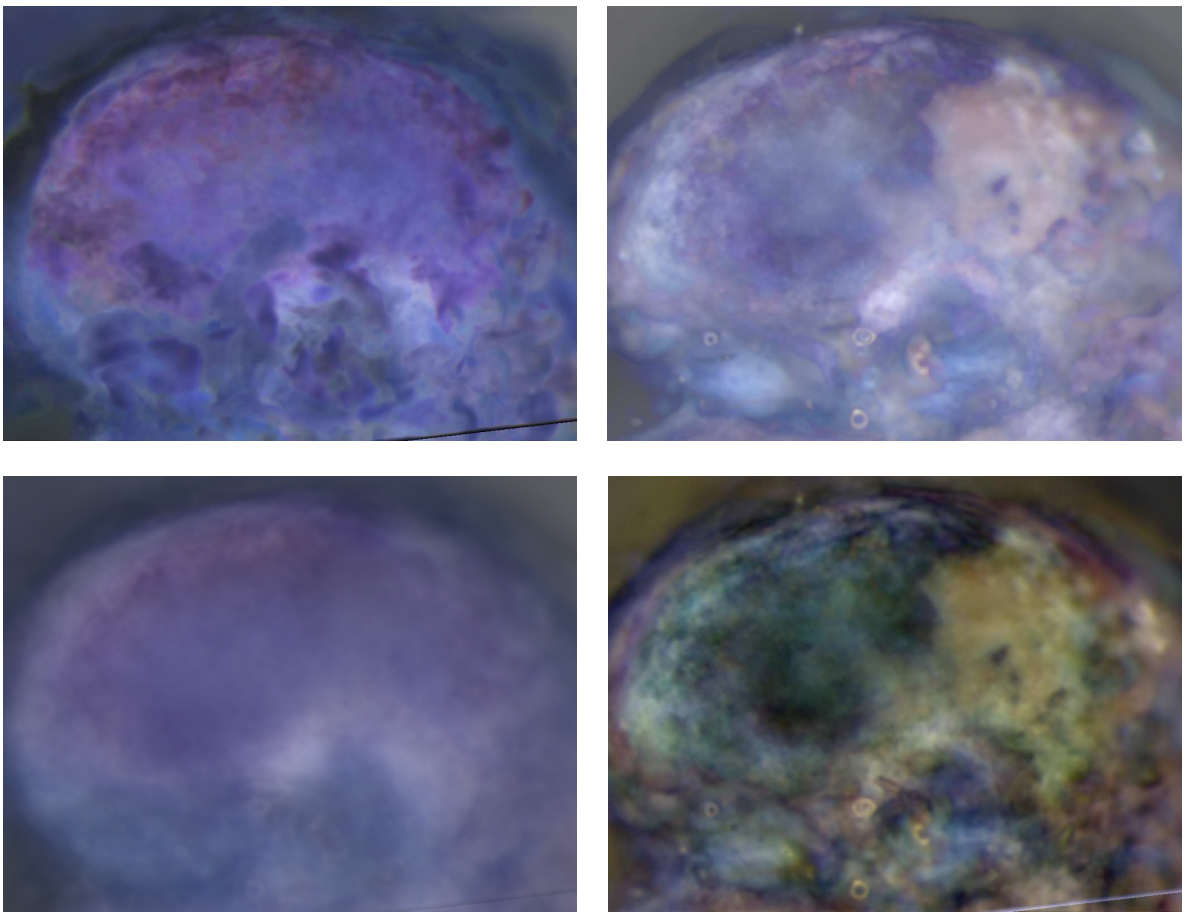


Figure D.3: Aligned images of the 22 week old mice. Top row: Minimum and Maximum image. Bottom row: Mean and variation image. Note that the images are brightened individually for a better perception and that the lines in the images are due to an artifact of the alignment.

Appendix E

Previous Labels and Their Results

This chapter shows some of the tests and plots from the first two labelling trials, that lead to changes in the class definitions and the class perceptions.

E.1 Results from the First Labels

The first version of the labels do not contain the class *Lesion purple bright*, but contain *Lesion pattern* which is varying bright areas. The purple classes are believed to be healthy.
The data set is based on 14 labelled images, where only "pure classes" are labelled.

E.1.1 Scatterplots of Classes

The data set is shown as a 3D scatterplot and as three times 2D scatterplots in Figure E.1. Here the labels are represented with another set of colors than in the report itself.
The plots show that there are tendencies of grouping, but especially *Healthy blue pattern* and *Lesion white* seem to have large variations. The lower the RGB values are, the larger is the possibility of a healthy class. The classes are mixed, but the healthy classes almost separate from the lesion classes if *Lesion perhaps* is removed. This can be seen in the red vs. blue and green vs. blue plots. *Lesion white* might still overlap a bit with some of the healthy classes. The classes are mixed more in the red vs. green plot.

To test if the class centers are equal, the Kruskal-Wallis rank sum test is carried out.
The test values (Kruskal-Wallis chi-square) for the three bands are between 2200 and 2500 which results in p-values of zero and hence are highly significant. At least one of the class means is therefore significantly different from the others, which is not surprising after exploring the scatterplots and due to the amount of data.

E.1.2 Summary of the Classes

Summary statistics are shown in Table E.1 for the data set.
Table E.1 shows that the classes have different center locations when using at least two color-bands. The distances from the blue mean to the red and green mean are larger for the healthy classes than for the lesion classes, which gives the blue appearance. For the purple classes, the red mean is larger than the green mean and the blue mean is lower than for the other classes. The purple classes are darker than the blue classes which again are darker than the lesion classes.

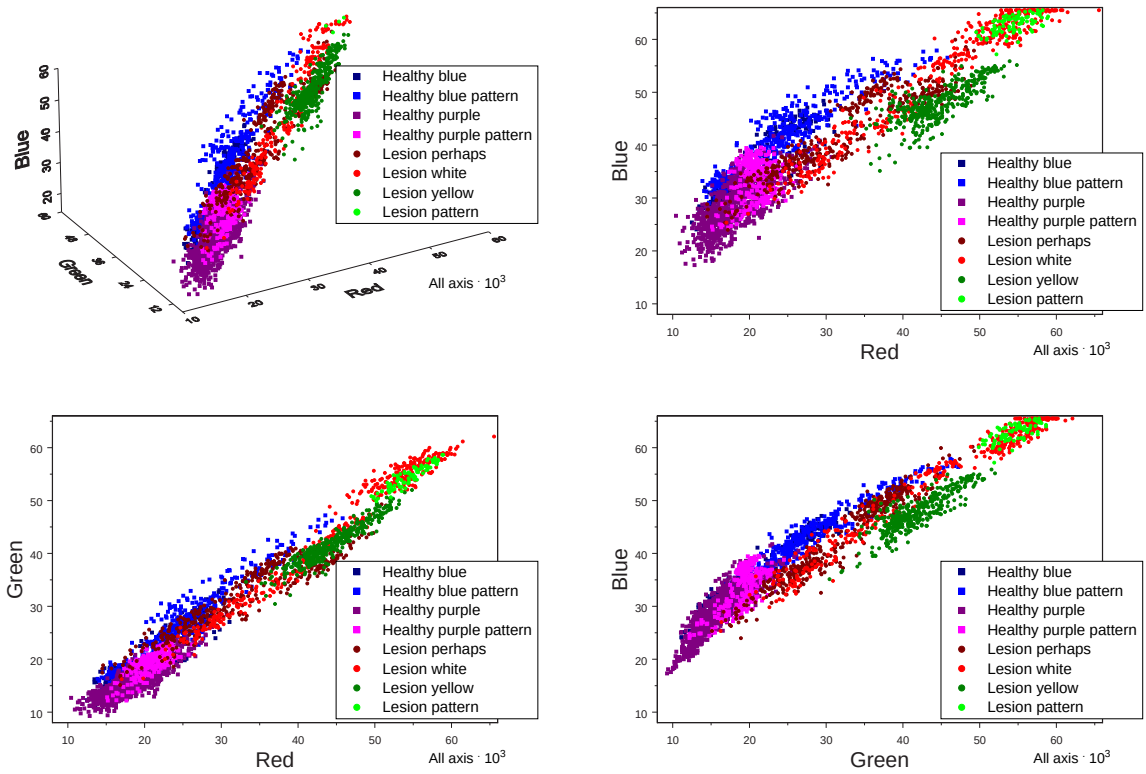


Figure E.1: All the classes from all the labelled images. Note that S-Plus draws one class at a time hence concentrated points in one class can hide other classes.

		Red		Green		Blue	
Class	Pixels	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
Healthy blue	277	20092	4415	20129	5329	34880	6197
Healthy blue pattern	409	24563	5939	26290	6591	41665	5982
Healthy purple	971	18170	3201	15537	2909	28618	4254
Healthy purple pattern	286	20666	1848	19165	2044	33678	3796
Lesion perhaps	398	32897	9100	31965	9588	42877	8854
Lesion white	375	44865	7783	44417	6439	53498	7134
Lesion yellow	383	44458	3929	41790	3786	47730	3630
Lesion pattern	86	54433	2373	54473	2184	62837	1640

Table E.1: Summary statistics for each class.

The pattern classes have less variance for all the bands than their respective non-pattern classes, which is surprising. This must be due to merging samples from different images.

E.1.3 Discriminant Analysis

A discriminant analysis is carried out in SAS for each band separately and with the bands simultaneously, hence doing one and three dimensional discriminant analysis. The covariances are not pooled which means that each class has its separate covariance matrix. The results are cross-validated which means that the observation in focus is left out of the used covariance matrix and mean value, also called leave-one-out.

No prior distribution is used hence equal probabilities of the classes are assumed.
The results can be seen in Table E.2 - E.4.

	Healthy				Lesion			
Colorband	blue	blue patt.	purple	purple patt.	perhaps	white	yellow	patt.
Red	0	40	61	83	3	42	82	95
Green	5	40	75	81	15	37	82	97
Blue	2	43	70	62	25	0	78	98

Table E.2: The diagonal vectors of the confusion matrixes from the 1-dim. discriminant analysis. It is a measure of how many pixels of a class that are actually classified as that class, in percentage.

The results in Table E.2 are not perfect. The best found class is *Lesion pattern* and it is probably because it is only defined in one image and therefore narrow defined. At the same time it is the brightest class, so there are no competing classes in the direction of higher intensities. *Healthy blue* finds very little of its own class.
It is a one-dimension discriminant analysis and with the mixed data poor results are expected.
The confusion matrix of the discriminant analysis using all three bands simultaneously can be seen in Table E.3.

	Classified as							
	Healthy				Lesion			
Correct class	blue	blue patt.	purple	purple patt.	perhaps	white	yellow	patt.
Healthy blue	**29**	31	25	0	14	0	1	0
Healthy blue pattern	17	**70**	2	0	6	3	3	0
Healthy purple	5	1	**73**	0	19	0	2	0
Healthy purple pattern	0	0	0	**99**	0	1	0	0
Lesion perhaps	7	0	16	0	**75**	0	1	0
Lesion white	0	1	0	32	0	**44**	20	2
Lesion yellow	2	7	1	0	2	7	**79**	3
Lesion pattern	0	0	0	0	0	4	3	**93**

Table E.3: Confusion matrix from the discriminant analysis with red, green and blue as simultaneous input. All values are in percentage.

The confusion matrix, in Table E.3, shows that combining the bands results in better classification. A lot of the missing pixels in the two blue classes are due to misclassification by the opposite blue class. The misclassification is not that bad because these classes are expected to have the same behavior with respect to age.

	Healthy				Lesion			
Colorband	blue	blue patt.	purple	purple patt.	perhaps	white	yellow	patt.
Red	−0.144	0.133	−0.167	0.046	0.037	0.177	0.200	0.032
Green	−0.006	−0.005	−0.083	0.053	0.010	0.098	0.152	0.011
Blue	−0.123	−0.128	0.049	0.157	−0.118	−0.166	−0.120	−0.045
RGB	−0.367	−0.330	0.019	0.102	0.233	0.318	0.349	0.039

Table E.4: Correlation of age with the relative area found for each class.

The correlation of the found classes with age, shown in Table E.4, is for the one dimensional

discriminant analysis', quite low (the largest is 0.20). Here all the images from the studies are classified and the area percentage of each class in each image is correlated with age, one class at a time to see how the relative area of that class evolves with age.
The better the classes are classified, the more reliable is the correlation with age.
The results from the bands used simultaneously show higher correlation than for the single bands, but the highest correlation (-0.37) is for the most mixed class (*Healthy blue*) so the result is useless. Better is *Healthy purple pattern* with 0.99 correctly found pixels, but it also finds a lot of *Lesion perhaps* which actually is 3/4 of the found pixels, hence it is not a reliable measure either. This information disappears when showing the confusion matrix in percentages.
The problem is the same for most of the other classes. The best ones are *Healthy blue pattern* and *Lesion yellow* which have correlation with age at -0.33 and 0.35, respectively. The amount of *Healthy blue pattern* is thus reduced with age and vise versa for *Lesion yellow*. These results are in accordance with what is expected.
All the correlations are highly significant, even the low ones, due to the amount of data.

The purple classes have a positive correlation with age for the areas found using all three bands and also for some found by a single band. The classification results are doubtful, but if they could be trusted, it would mean that the purple areas increase with age and hence is a lesion stage and not a staining failure in healthy cartilage.

E.1.4 Conclusion

The classes are overlapping and a single band is not enough to distinguish between them.
The classification results are not as good as hoped for, not even using all three bands simultaneously. The correlation results are therefore not that trustworthy.
The purple classes might increase with age, but this observation is based on low classification percentages.
More images should be labelled, this time with a broader definition of the classes.

E.2 Results from the Second Labels

The second version of the labels does not contain the class *Lesion purple bright* nor *Lesion pattern*, hence only seven classes are identified.
The data set is based on 22 labelled images, this time larger areas are labelled to represent more variation of the classes.

E.2.1 Scatterplots of Classes

The data set is shown as a 3D scatterplot and as three times 2D scatterplots in Figure E.2. The colors representing the classes are here the same as for the first set of labels.

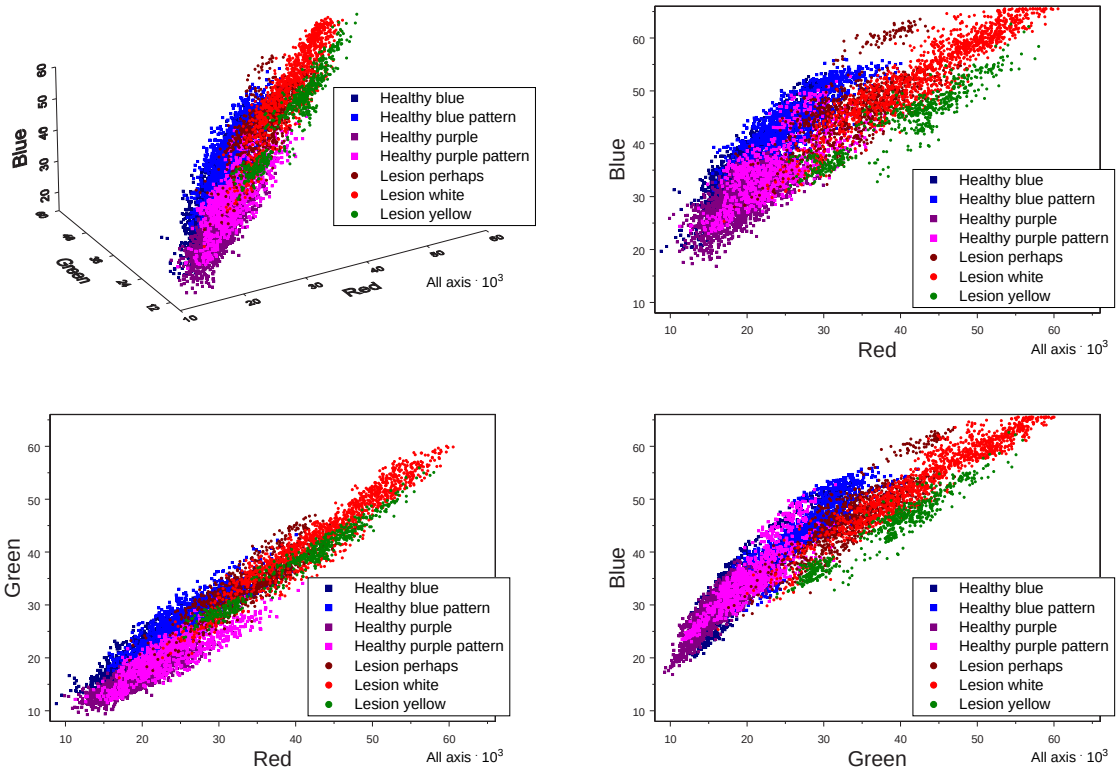


Figure E.2: All the classes from all the labelled images. Note that S-Plus draws one class at a time hence concentrated points in one class can hide other classes.

There are clearly more pixels now than for the first labels and the classes vary more. This results in more mixed classes, but also gives a more realistic impression of the problem.

To test if the class centers are equal, the Kruskal-Wallis rank sum test is carried out.
The test values (Kruskal-Wallis chi-square) for the three bands are between 4800 and 6100 which result in p-values of zero and hence highly significant. At least one of the class means is therefore significantly different from the others.

E.2.2 Summary of the Classes

Summary statistics are shown in Table E.5 for the second data set.

		Red		Green		Blue	
Class	Pixels	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.
Healthy blue	1596	20276	4321	20866	5015	36460	6863
Healthy blue pattern	1446	25903	4452	27257	4582	43588	5741
Healthy purple	1620	19264	3504	16082	2753	29253	4026
Healthy purple pattern	850	24742	4496	21581	3563	36635	5504
Lesion perhaps	858	34502	5176	34420	5257	47047	6243
Lesion white	1259	43584	7900	43362	8320	53489	7571
Lesion yellow	562	40702	7427	38327	6917	44464	6337

Table E.5: Summary statistics for each class.

Table E.5 shows that *Lesion white* and *Lesion yellow* are darker than before, while the rest of the classes are brighter. The distances between the class' means are hence decreased.
The classes *Healthy purple pattern* and *Lesion white* have larger variance than before, while the rest of the classes generally have less variance.

Healthy purple pattern generally shows more variation than its corresponding non-pattern class. *Lesion white* has the largest variance for all the bands.

E.2.3 Discriminant Analysis

The results of the analysis, carried out the same way as for the first data, can be seen in Table E.6.

	Healthy				Lesion		
Colorband	blue	blue patt.	purple	purple patt.	perhaps	white	yellow
Red	5	47	76	10	59	53	23
Green	0	50	83	45	52	54	32
Blue	0	44	83	41	39	60	0

Table E.6: The diagonal vectors of the confusion matrixes. It is a measure of how many pixels of a class that are actually classified as that class, in percentage.

Table E.6 shows that *Healthy blue pattern*, *Healthy purple* and *Lesion perhaps* are better classified, while the rest of the classes are worse classified than for the first data set.

	Classified as						
	Healthy				Lesion		
Correct class	blue	blue patt.	purple	purple patt.	perhaps	white	yellow
Healthy blue	**59**	22	16	2	0	0	0
Healthy blue pattern	22	**62**	0	8	8	0	1
Healthy purple	8	0	**79**	13	0	0	0
Healthy purple pattern	17	4	29	**46**	3	0	0
Lesion perhaps	1	12	0	4	**60**	14	9
Lesion white	0	0	0	2	24	**66**	8
Lesion yellow	0	0	0	0	3	4	**93**

Table E.7: Confusion matrix from the discriminant analysis with red, green and blue as simultaneous input. All values are in percentage.

By examining the confusion matrix using RGB as input, shows in Table E.7, that *Healthy blue*,

Healthy purple and *Lesion perhaps* are better found than with the first data set, *Lesion yellow* is the same and the rest are worse. Especially *Healthy blue* is much better while *Lesion purple pattern* is much worse.
The misclassification is mostly between the blue classes, between the purple classes and between the lesion classes. These misclassifications are of the best kind, because the classes within these groups are mostly divided for better classification and can be merged afterwards, to a healthy and a lesion class or to a blue, a purple, and a lesion class.

	Healthy				Lesion		
Colorband	blue	blue patt.	purple	purple patt.	perhaps	white	yellow
Red	−0.159	0.129	−0.151	0.137	0.185	0.124	0.224
Green	0.030	0.006	−0.064	0.016	0.131	0.073	0.161
Blue	−0.079	−0.124	0.108	0.021	−0.113	−0.122	−0.166
RGB	−0.435	−0.226	0.061	0.374	0.259	0.158	0.496

Table E.8: Correlation of age with the relative area found for each class.

Table E.8 shows that correlation with age differs from the first data set. The most trustworthy classes, with respect to the confusion matrix, are *Healthy purple* and *Lesion yellow*, which have correlations with age 0.06 and 0.50, respectively. The less trustworthy classes *Healthy blue* and *Healthy purple pattern* have correlations at -0.44 and 0.37.
These results indicate (again) that the purple classes are increasing with age (correlations at 0.06 and 0.37) and hence actually a lesion stage. This time the correlation is based on more reliable, but not perfect, classifications.
The rest of the classes behave as expected; the healthy blue classes have a negative correlation with age (decreases with age) while the rest have positive correlations with age.

E.2.4 Conclusion

The second set of labels results in a larger data set but only larger variation within two of the classes.
The correlation with age top at 0.50 for *Lesion yellow*. This and other results are more reliable this time and the purple classes show higher positive correlations with age (0.06 and 0.37).
Going through the images again with this new information (that the purple areas could very well be a lesion stage) often results often in finding purple only next to or even around lesion areas. Purple areas are therefore now believed to be a lesion stage which has altered the label definitions a bit and therefore also the labels. At the same time a new lesion type is found. It is *Lesion purple bright* which is clearly a bright lesion but it is so bright purple that it is extracted from *Lesion white*.

The positive, but somewhat uncertain, correlation with age and the discovery of purple next to the lesions led to the hypotheses and tests of the purple areas behavior in Chapter 5.

Appendix F

Class Samples of Labelled Images

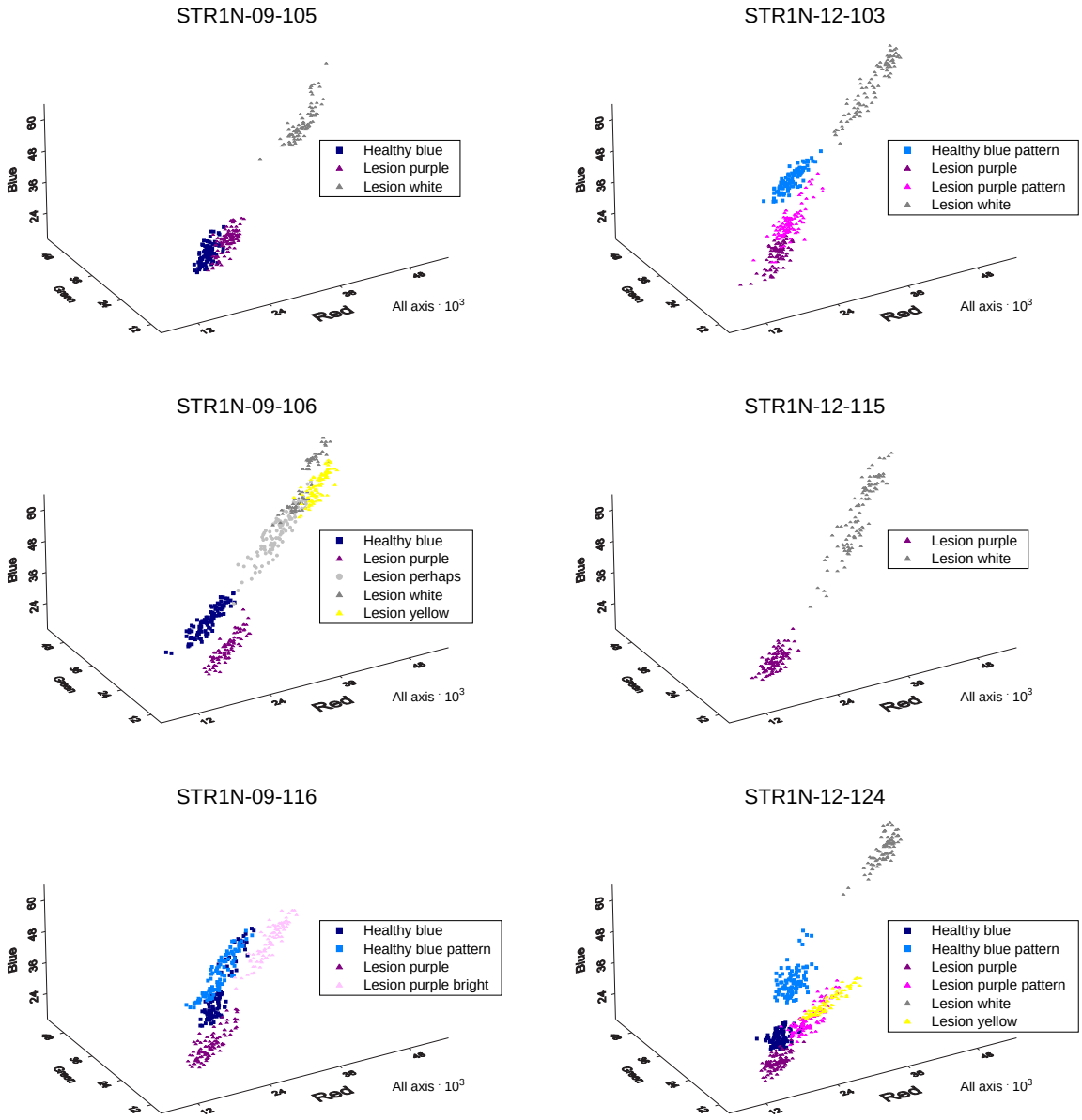


Figure F.1: Samples from the labelled images. Left column: Study STR1N-09. Right column: Study STR1N-12.

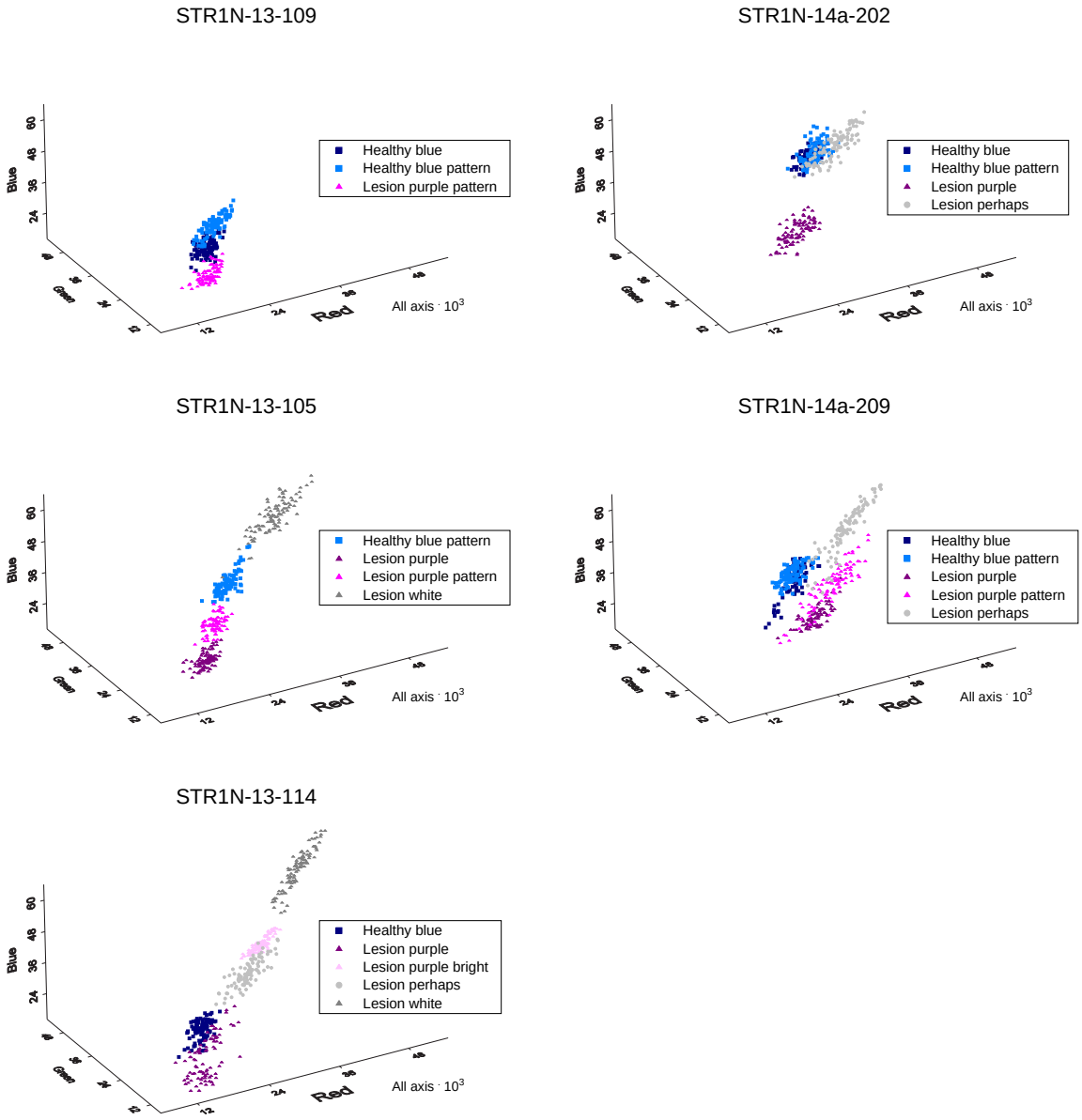


Figure F.2: Samples from the labelled images. Left column: Study STR1N-13. Right column: Study STR1N-14a.

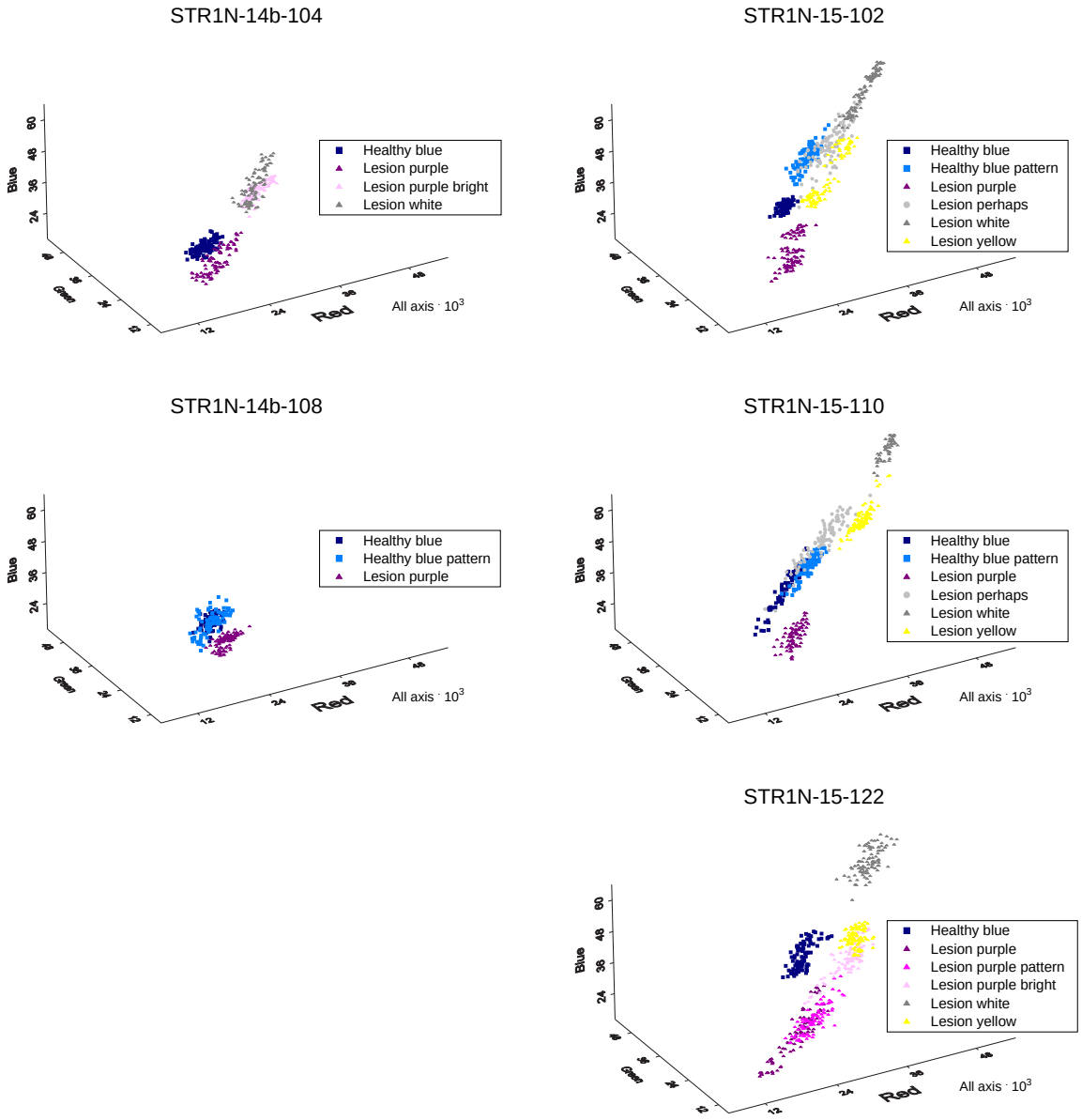


Figure F.3: Samples from the labelled images. Left column: Study STR1N-14b. Right column: Study STR1N-15.

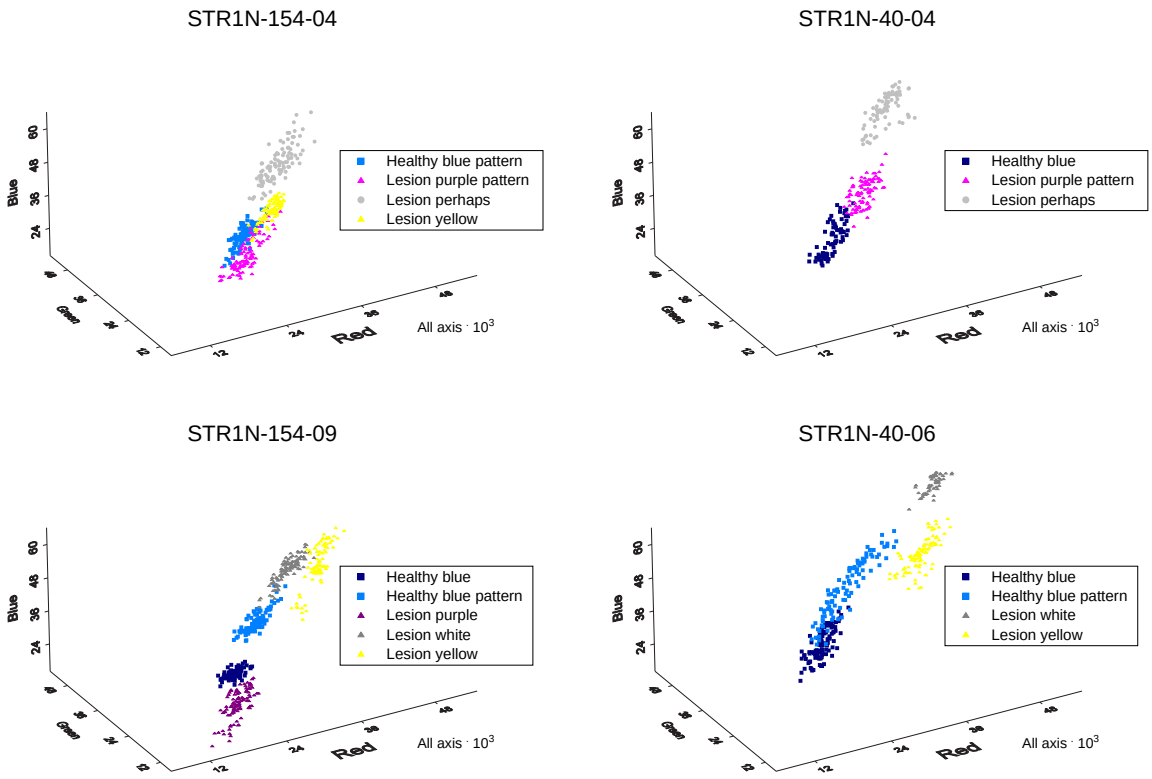


Figure F.4: Samples from the labelled images. Left column: Study STR1N-154. Right column: Study STR1N-40.

Appendix G

Boxplots of the Color Transformations

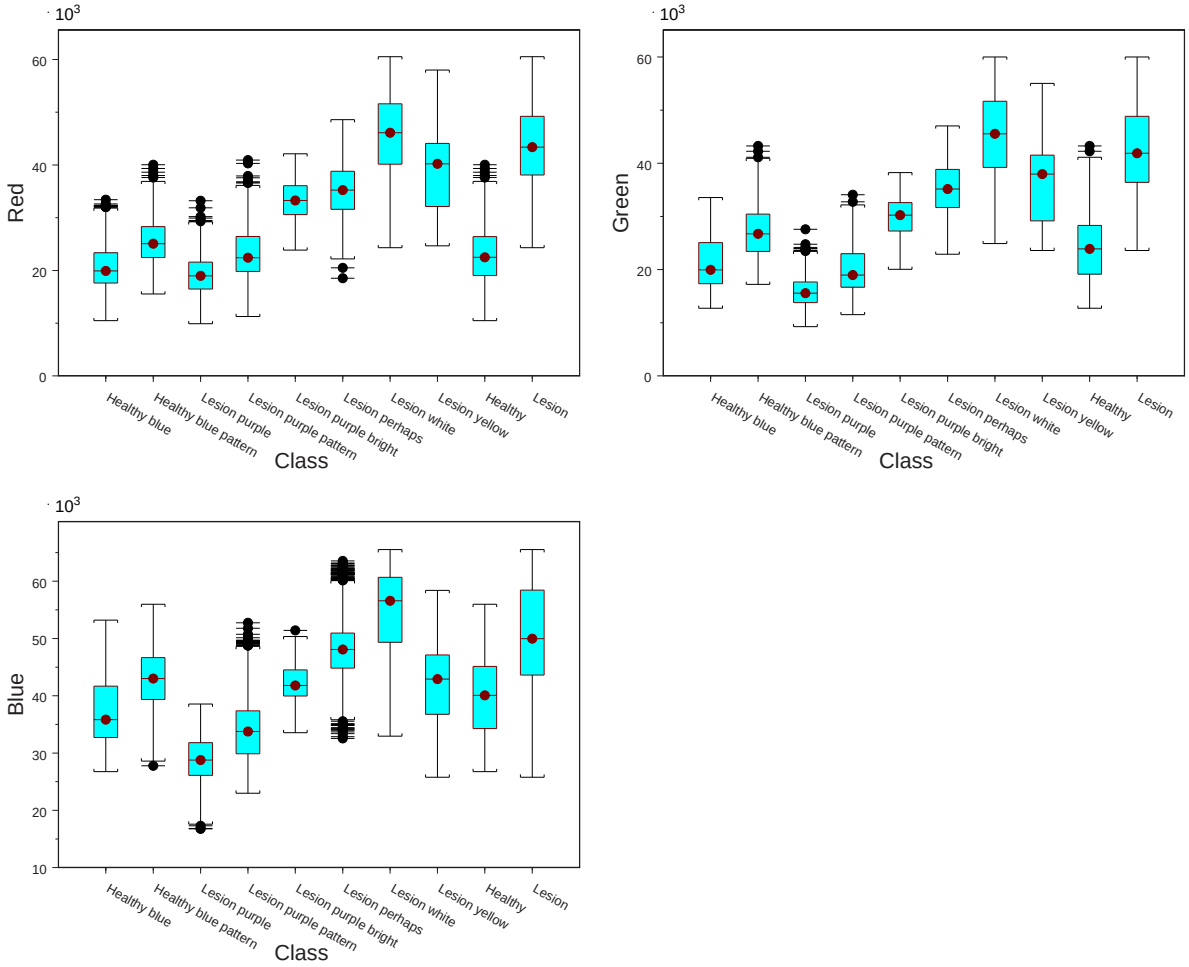


Figure G.1: Boxplots of the original colorbands.

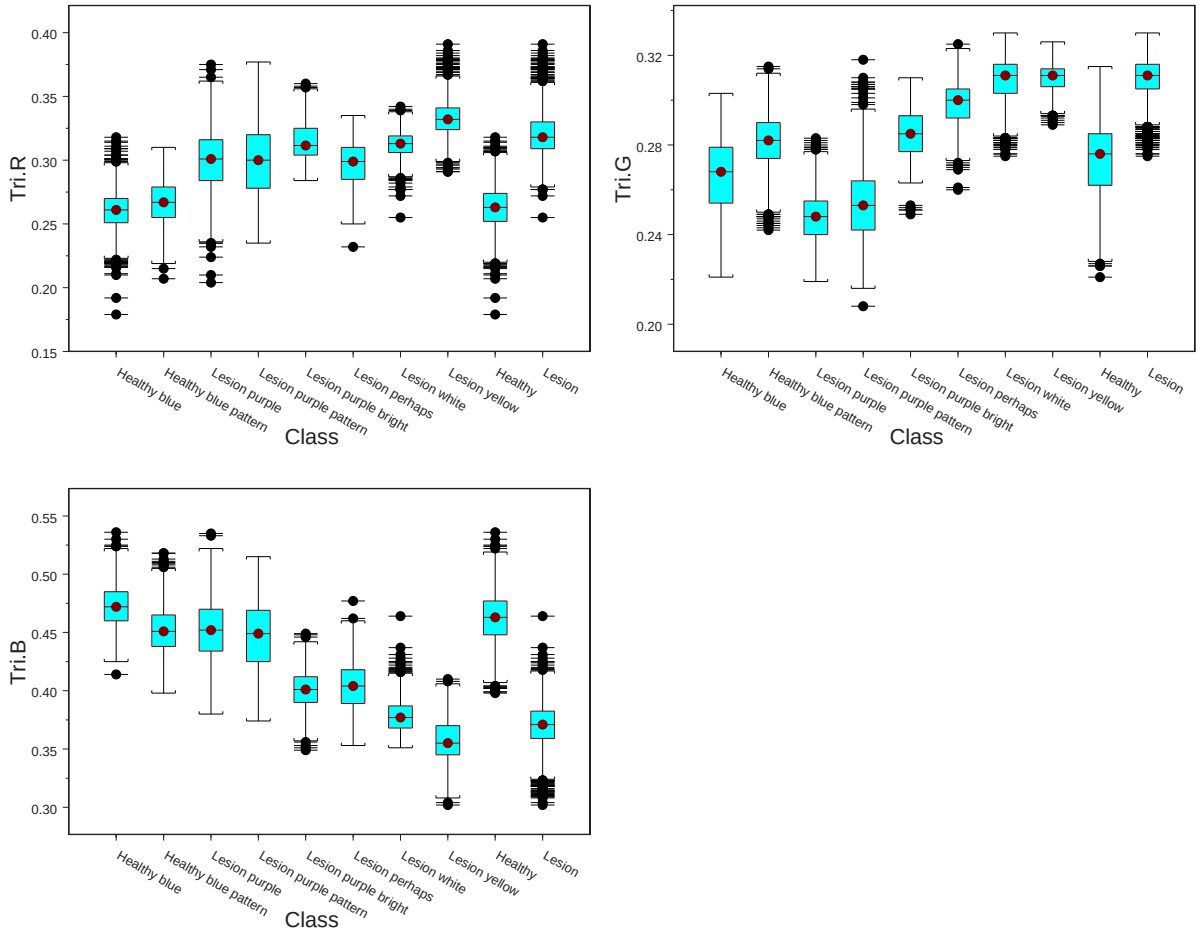


Figure G.2: Boxplots of the trichromatic colorbands.

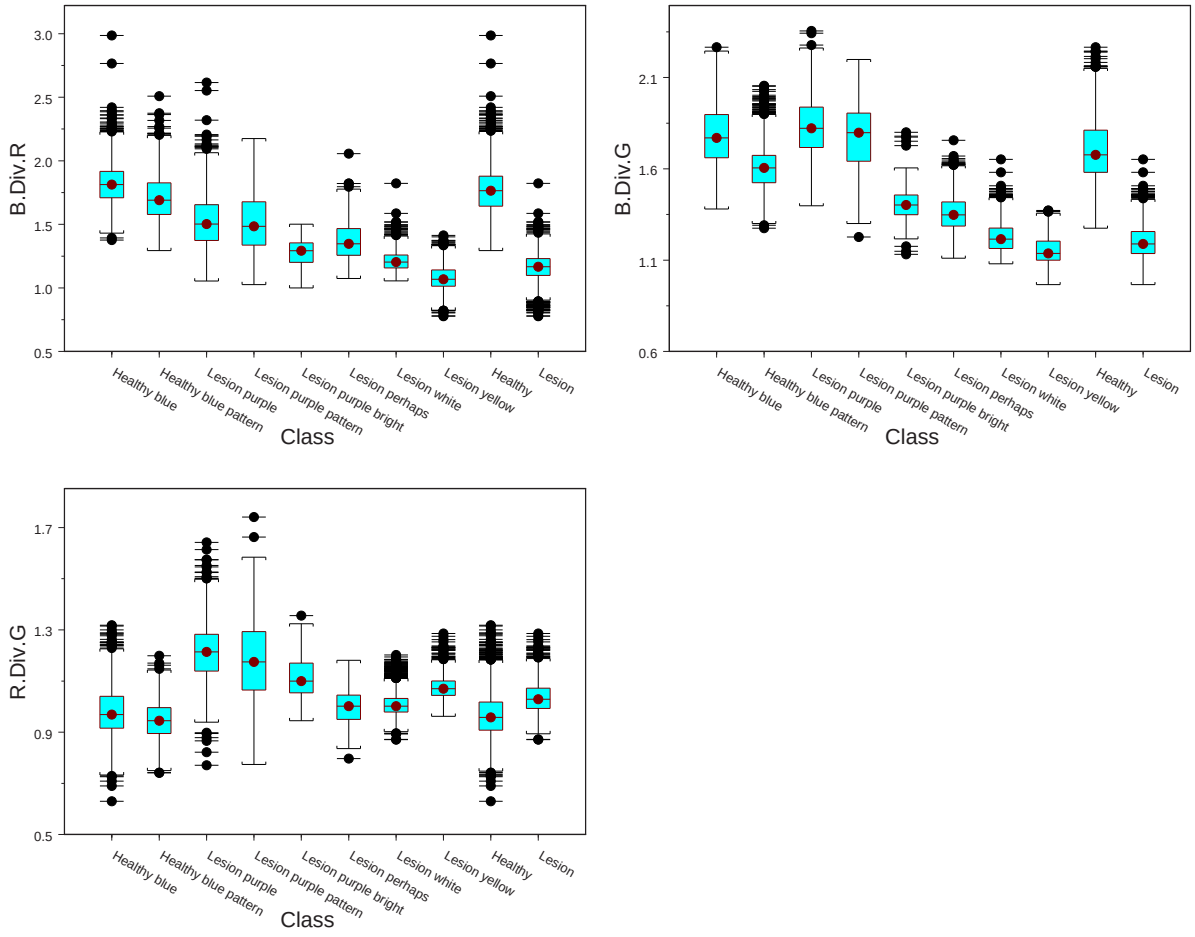


Figure G.3: Boxplots of the bands divided by another band.

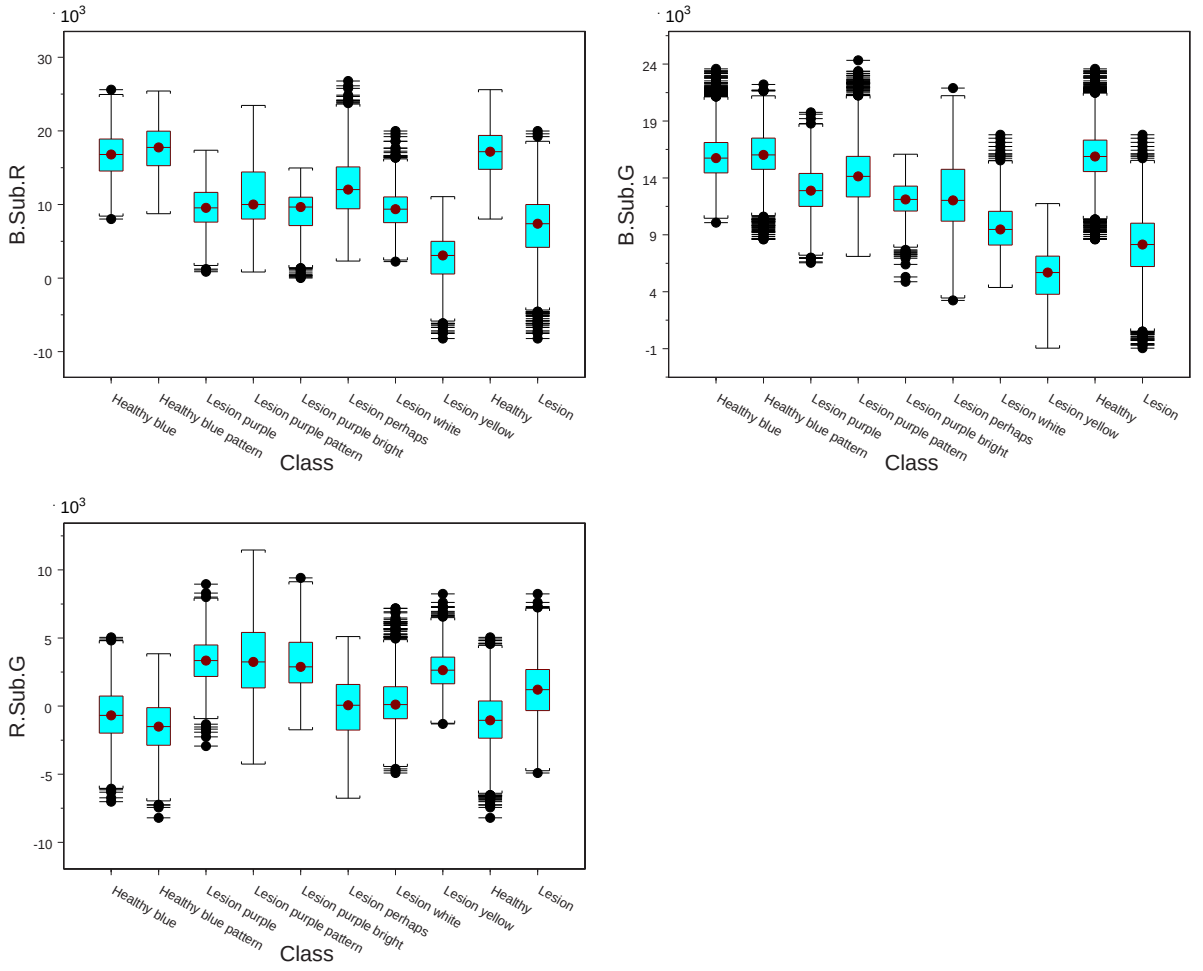


Figure G.4: Boxplots of the bands subtracted by another band.

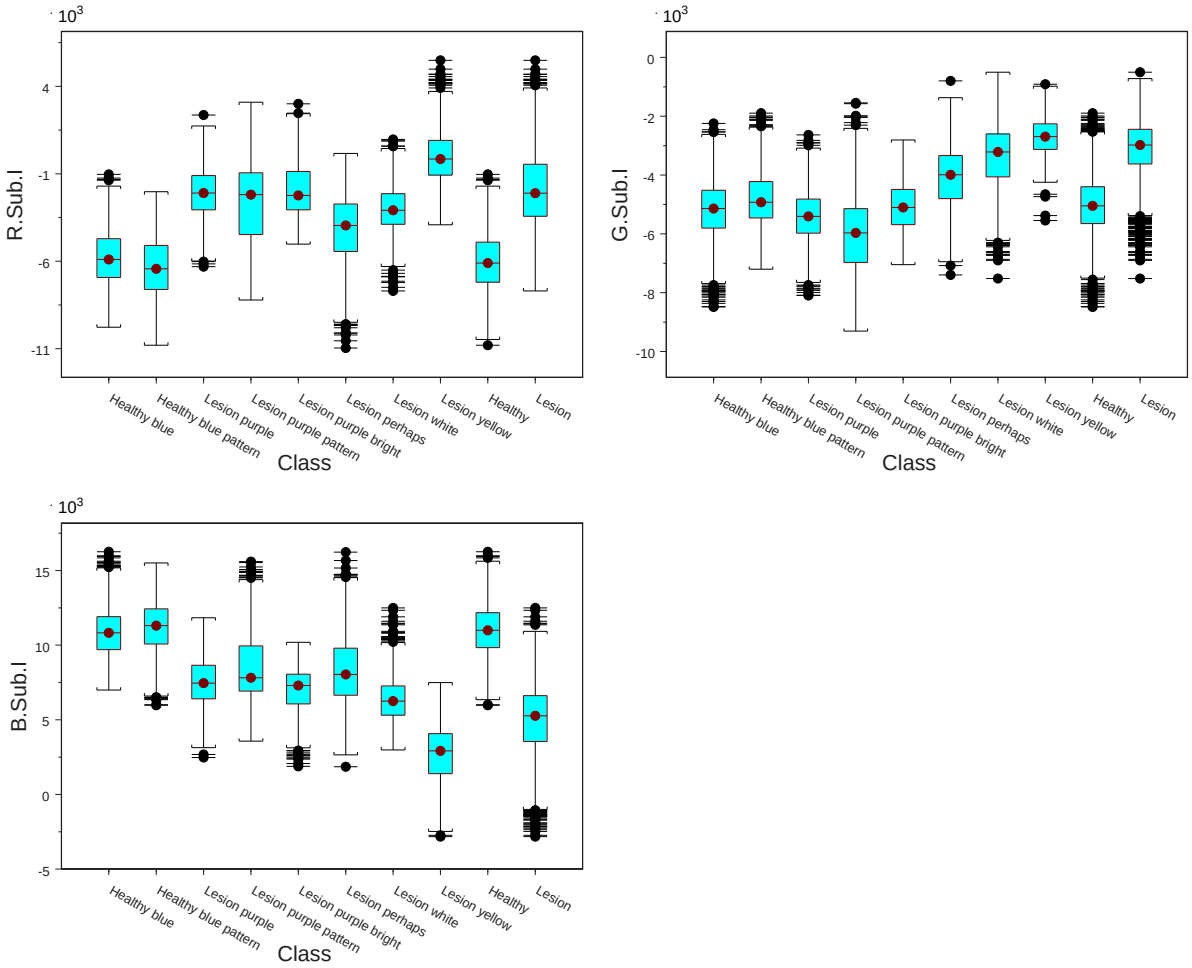


Figure G.5: Boxplots of the bands subtracted by the Intensity.

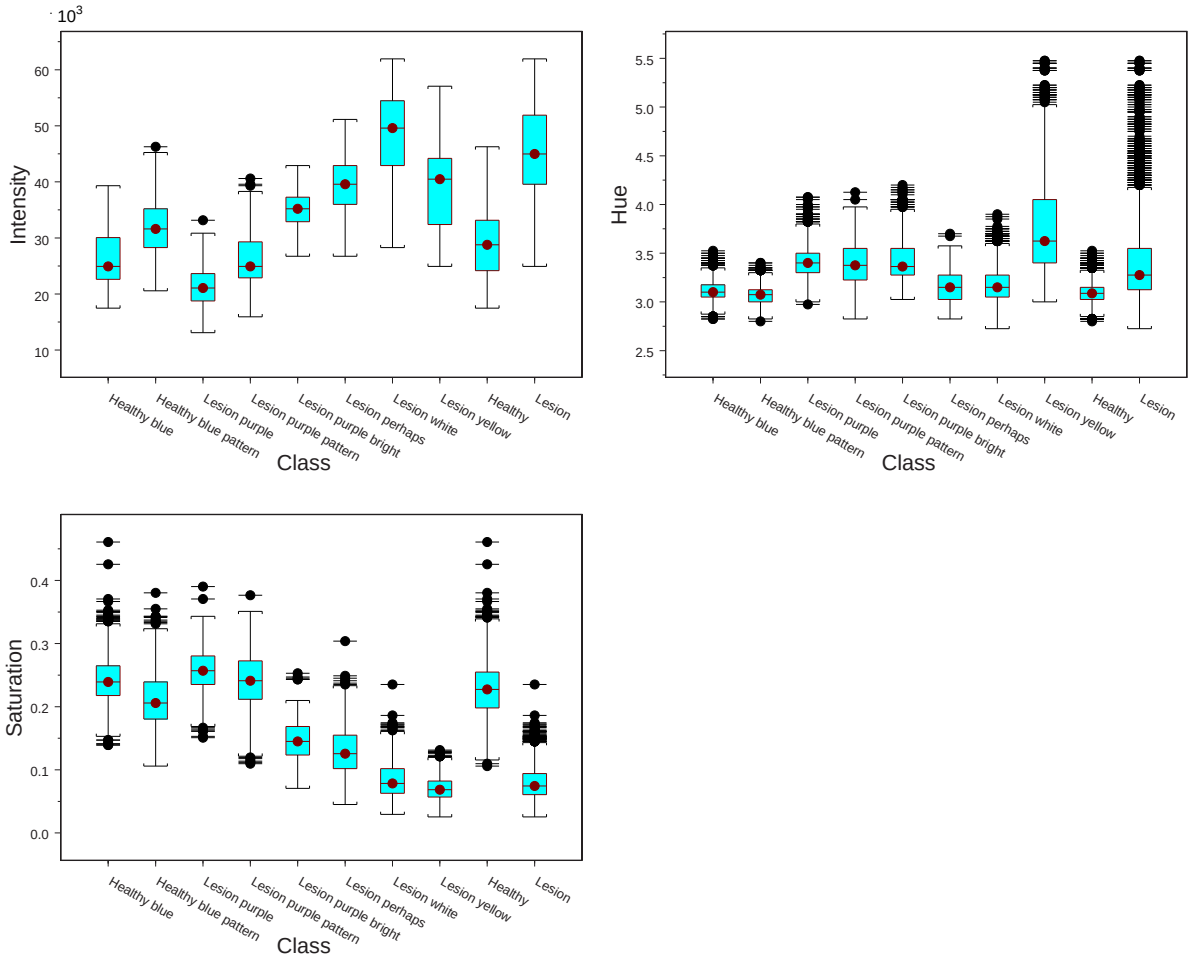


Figure G.6: Boxplots of the IHS colorbands.

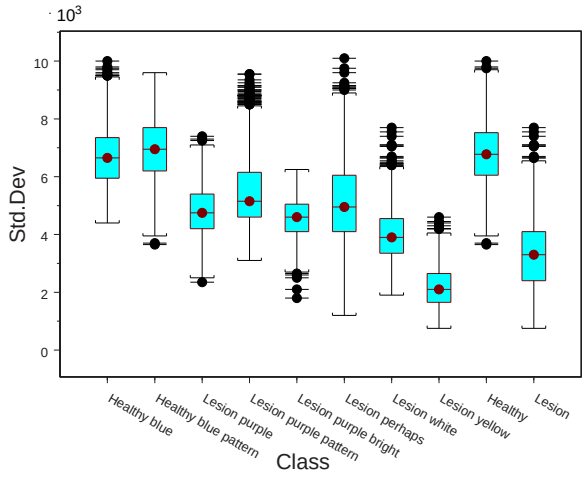


Figure G.7: Boxplot of the standard deviation between the bands.

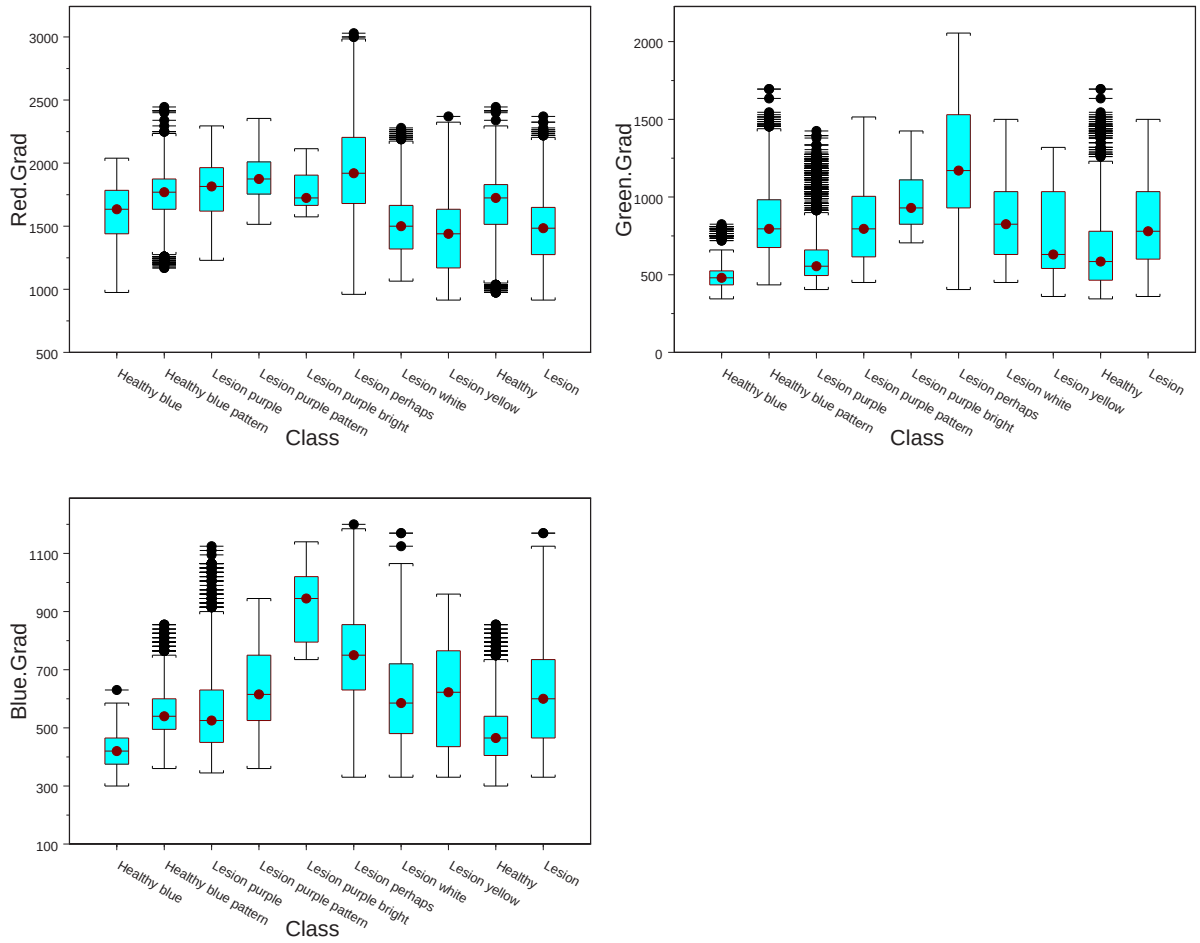


Figure G.8: Boxplots of the gradient of the bands.

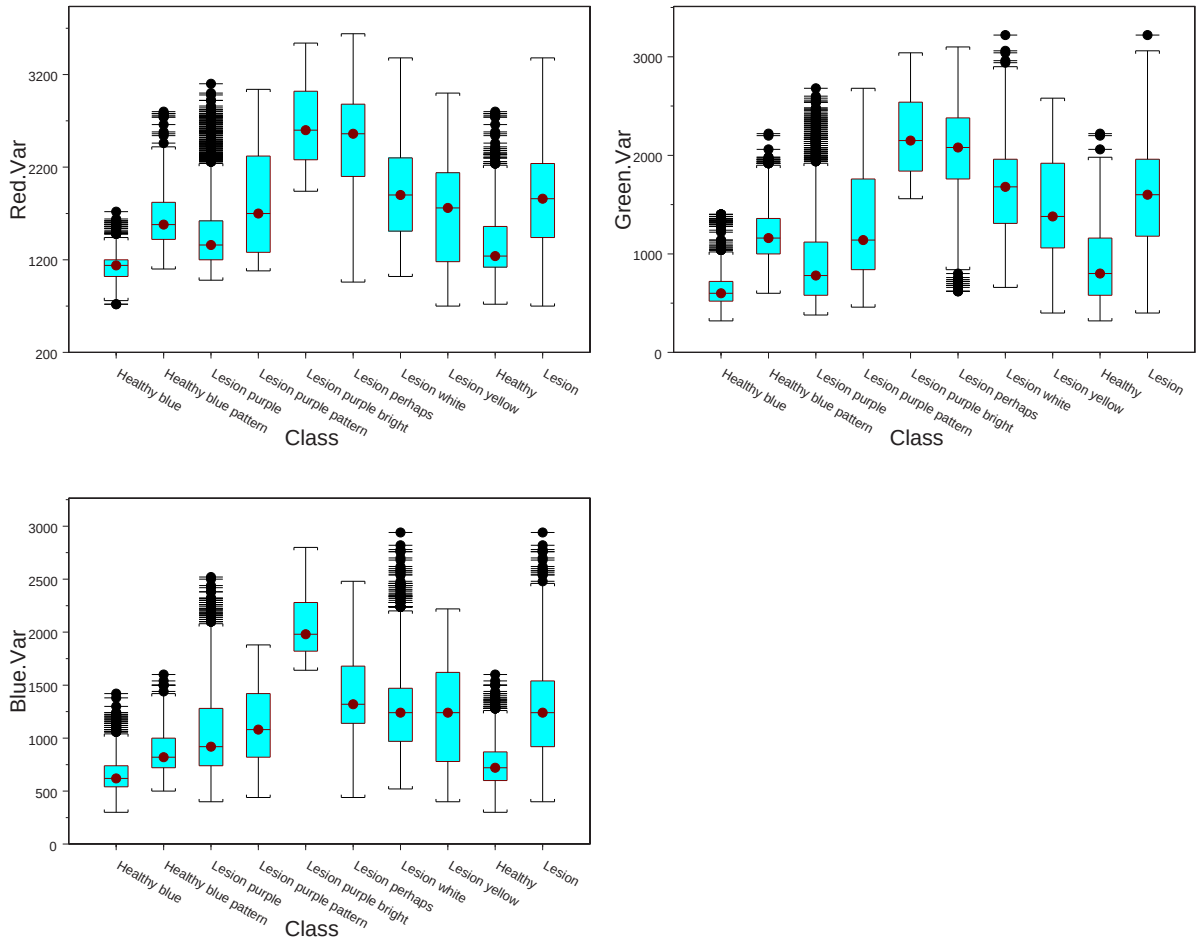


Figure G.9: Boxplots of the variance in the area.

Appendix H

Visual C++ Function List

Many of the C++ files have several functions and are stored with "01", "02" etc. after the filename. It is only the workspaces main function that are described here.

AreaOut - Loads image and masks and generates a RGB data file for each mask (group)

Bleached - Test the studies for bleaching, by the sum of R, G, B of the ROIs

CalcCov - Generates mean and covariance matrix for discriminant analysis

CalcMeanStd - Generates mean and std.dev. of the classes in latex file format

CanonDiscrim - Extracts pixels for CanDisk etc. from the warped images

Classify4 - Classifies the image with different class combinations and calls Testclassify4.cpp. For eight and three classes and for manual decision boundaries

ColorClass - Colors a single clustering / classifying result (image) to pseudo colors

ConvSTR1N15 - Finds the data (image and mask) in a sub directory and rename (and moves it)

DistanceFactor - Finds and smoothen the upper border of the tibia and calculates each pixels nearest distance to it

Gradient - Generates gradient images from the original RGB images

Histogram - Calculation of histograms for R, G and B and Mean and Std. dev.

IHS - Generates IHS images from the original RGB images

Kmeans - Perform kmeans clustering on the images, with different k and color transformations

LesionPurple - Uses the manual marked lesions to prove purple around lesion

LesionWhite - Uses the manual marked lesions to generate a probability map

LocalVar - Trials with local variance to find the fibrillated areas

MergeWarped - Merge warped studies

Pixvar - Generates std.dev. images from the original RGB images

PostClust - Processes the found clusters to end up with an OA measure

ShowRois - Isolate ROI and saves it as 3 x 8 bit for all the files in a directory

Smooth - Median filtering of the images

TestClassify4 - Test classification by correlation of classes and OA measure to age, histology and biomarker and generates the confusion matrix

Trichromatic - Generates Trichromatic images from the original RGB images

Unblind - Find files (image and mask) and rename them to the unblinded number

Variance - Generates variance images from the original RGB images

Warpdata - Loads image sequence and calculates mean image (with respect to inactive pixels)

Appendix I

S-Plus Function List

Many of the S-Plus functions are split in two, and should be called separately, to avoid that S-Plus crashes. The second function is named the same as the first but with an "a" added at the end. Only the first of each function is listed below. The functions are sorted by chapter.

Biomarker and Histology

HistoPlots(x) - Plots histology vs. biomarker and Histology vs. Histology

Initial Analysis of the Images

PlotImage3D(FileIn) - 3D Scatter of an image without labels

Histogram(File) - Generates R,G,B Histogram

HistogramStudyAll() - Generates R,G,B Histogram joint for all studies

PlotBleachAbsAge(x) - Plots the mean values by image

PlotBleachTriAge(x) - Plots the relative mean values by image

PlotBleachMeanAbs(x) - Plots the mean values by study

PlotBleachMeanTri(x) - Plots the relative mean values by study

Examination of the Bright Lesions

ShowCanonAge(Band) - Draws the red, green or blue band against age

ShowCanonPoints(Band) - Draws the red, green or blue band against points

ShowLesionPurple1 - Plots purple, defined as red-green, as a function of distance from bright lesions, for each study

ShowLesionPurple2 - Plots purple, defined as trichromatic red > 0.27, as a function of distance from bright lesions, for each study

ShowLesionPurpleTriBlue - Plots blue, defined as trichromatic blue > 0.48, as a function of distance from bright lesions, for each study

ShowLesionPurpleActually1 - Plots mean purple, defined as red-green, as a function of distance from bright lesions, for each study

ShowLesionPurpleActually2 - Plots mean purple, defined as trichromatic red, as a function of distance from bright lesions, for each study

ShowLesionPurpleActuallyTriBlue - Plots mean blue, defined as trichromatic blue, as a function of distance from bright lesions, for each study

Defining, Labelling and Analysis of Classes

PlotLabelClass3D(x) - 3D Scatterplot for classes with 100 pixels, for all labelled images

PlotLabelClass2D(Col1, Col2, File) - 2D Scatterplot for classes with 100 pixels, for all labelled images

PlotLabelSingleClass(Grp, FileOut,count) - All images but only one Class

PlotLabelMean2D(Col1,Col2, FileOut) - Plots Mean data for two of R, B, G, Class as depended

PlotLabelMean2DImage(Col1,Col2, FileOut) - Plots Mean data for 2 of R, B, G, Image as depended

ShowClass100Boxplot2(Var,Factor,Scale) - Boxplot for each variable (and 8 Class)

PlotLabelImage3D(File, title) - 3D Scatterplot for Class with 100 pixels, for one image

Analysis of color transformations

Anova2(x) - ANOVA on classes to get the variances and significanse

Anova2Single(File) - ANOVA on a single image (1000 pixels classes) to get the variances and significanse

DataForSas1(x) - Makes class a factor with correct names and export it for SAS

DataForSas1a(x) - Makes class a factor with correct names and export it for SAS (only Healthy and Lesion)

DataForSas2(Image) - Makes class a factor with correct names and export it for SAS

CompareClasses(x) - Compare classes 2 by 2 using Hotelling's T-test

PlotColorRep3D - 3D Scatterplot for new color rep. with 100 pixels, for all images

Classification of the Different Types of Areas

PlotClassification3D - 3D Scatter plot of an image with found classes

PlotBluePurpleWhiteAge - Plots the amount of Blue, Purple and White against Age

Clustering

ClusteringPlotStudyAgeVs - Generates 3D plot of clusters means vs. age

ClusteringPlotStudyAgeVsMaxDist - Plots the maximum distance between two clusters

ClusteringPlotFirstClust - Plots the first cluster in each image according to age

ClusteringScatterPlot - Scatterplot colored according to the found clusters

PlotImage3DClust - 3D Scatterplot an image without labels

PlotImageBandStudy - 3D mesh of study, cluster and a color band

PlotImageBandStudyAmount - 3D mesh of study, cluster and amount

PlotImageBandStudyAmountCluster - 4D scatter plot of study, cluster and amount and a colorband

GenerateLabelData.ssc - Contains all the data import and preparations

Run-scripts.scc - Contains all the calls for all the functions

Appendix J

SAS Function List

Discrim.sas - Discriminant analysis of the data set, for each color representation and combinations

DiscrimSingleImage.sas - Discriminant analysis of a single image, for each color representation and combinations

Regression.SAS - "Regression analysis" using PROC Stepdisk to find the best separating color representations

RegressionOA.SAS - Regression analysis using PROC Reg to find the significant relative areas for the OA measure

LesionPosition.SAS - "Regression analysis with two dependent variables.". Modelling the lesion position according to age using PROC Mixed

Appendix K

Clustering Survey

This appendix describes multiple clustering routines. It starts by explaining some general approaches and terms, then a lot of clustering routines are explained. The several distance measures are defined and a thorough explanation of k-means clustering and ISODATA is given. At the end the literature list is shown.
The process of partitioning a data set into unlabelled groups with similar characteristics, where the groups are sufficiently different from other groups, is called clustering. A more mathematical formulation is: The process of minimizing the intra-variance in groups and maximize the inter-variance between groups.

Since this survey is only a small part of the project, the following areas have less focus

- This survey is mainly focused on numerical data, while binary, nominal, interval-scaled variables etc. are of less importance.
- Large data sets and high dimensionality are not big issues.
- Processing time is of less importance due to no real time demands and relatively small data sets.
- Neural networks is a large subject and can be designed and trained for classification. It is not a search topic here, but some of the found clustering methods use similar approaches.

Besides the references from each algorithm, two clustering surveys [38, 39] are found.

Clustering vs. Classification
Classification (supervised) requires training data where clustering (unsupervised) does not.
Some of the reasons for using unsupervised clustering are

- Manual labelling large data sets are costly.
- Eliminating human precision and objectivity.
- Class labels may not be known prior to grouping.
- Exact center position may not be known.

K.1 Main Clustering Methods

There are several ways of categorizing clustering algorithms, the following is just one of them.

Single-pass Methods
The clustering is carried out in only one iteration, which in some cases makes the clusters dependent upon the order in which the data points are processed. The main group is Hierarchical clustering.

Relocation Methods
After initialization of the cluster centers, the data points are iteratively reassigned (interchanged / switched) between the clusters. This approach is prone to reach a local optimum rather than a global optimum. It is generally not possible to determine if the global optimum solution has been reached.

Other Methods
The following methods do not fit in the above definitions, are extensions or a is mix of them

- Fuzzy clustering.
- Nearest neighbor methods (assign data points to the same cluster as an user-defined number of their nearest neighbors).
- Density based (the number of data points in the neighborhood defines the quality of the data point).
- Model-based (a model is hypothesized for each of the clusters and the idea is to find the best fit of that model to the other models).
- Artificial neural networks (ANN).
- Decision Rules (rules for sorting pixels into classes).
- Hybrid methods (either a mix of 2 or more clustering algorithms or if clusters are initialized by training data - guided clustering).

Designing a Clustering Algorithm
There are three global decisions that have to be made when designing a clustering routine:

- Define a measure of similarity. The most widely used is the Euclidean distance.
- Define a criterion function to measure how well each data point is represented by its cluster center, e.g. the mean square error (MSE).
- Define an algorithm / process to minimize the criterion function. This is the part of the clustering process which has most variants.

Clustering Challenges

- Finding the optimal number of clusters, k.
- Assess the validity of a given clustering.
- Permitting the classes natural shape rather than forcing spherical shapes.
- Prevent the order of the data points to affect the clustering.
- Prevent the order of the splitting and merging to affect the clustering.

K.2 Main Clustering Terms

The set of cluster centers is called a codebook. A single cluster center is sometimes referred to as a prototype.

Parametric vs. Non-parametric
Parametric (mixture modelling) means that underlying class-conditional densities (mean and standard deviation) are assumed and hence estimated. Non-parametric clustering makes no such assumption, but separates the data point into natural dissimilar clusters.

On-line vs. Off-line
On-line learns / is adjusted continuously, for every new data point, and hence avoids storing the complete data set. The learning rate is normally decreased with time. Off-line routines only update after each iteration.

Hierarchical vs. Non-hierarchical
There are two major groups of clustering; Hierarchical and Non-hierarchical clustering.
Hierarchical (or sequential) clustering either starts with each data point as a cluster and then merges them one by one and ends up with just one cluster, or the other way around; starting with one large cluster and then splitting it until each cluster only contains one data point. The process is logged so every merge or split can be traced.
Non-hierarchical (or flat) clustering on the other hand starts with k cluster centers and assign all data point to the closest one and reassign them iteratively as the center locations are updated. Here there is no history of the process but the adjustment possibilities and results are far better. Non-hierarchical clustering can be used with large data sets where Hierarchical is computational prohibitive.

Hard vs. Fuzzy
Hard clustering (crisp clustering) means that a data point either or not belongs to a cluster. Fuzzy clustering on the other hand means that a data point can be a member of several clusters with a degree of membership summing to 1.
Fuzzy clustering is better at avoiding local minima, but not guaranteed.
It is normally used when the cluster boundaries are not clearly defined (overlapping).
The term Hard is normally left out of the name, so Hard clustering is assumed if Fuzzy is not specified.

Deterministic vs. Stochastic
If the clustering gives the same result at multiple runs, it is deterministic otherwise it is said to be stochastic.

Pattern Matrix vs. Proximity Matrix
When measured features are used directly as input they are ordered in a pattern matrix. When differences (likeness, affinity or association) are used they are ordered in a proximity matrix, which can contain either similarities or dissimilarities [87].

Polythetic vs. Monothetic
Polythetic algorithms uses all dimensions / features simultaneously while monothetic uses the features one at a time. The monothetic approach is faster, but tends to give poor results. Far most algorithms are polythetic.

K.3 Initialization

Non-hierarchical clustering suffers from the fact that the cluster number, k, has to be specified and likewise with the k initial cluster centers. Several solutions have been proposed through time (especially to deal with the last challenge):

- To find the optimal cluster number the algorithm should be run with 2 to K_{max} clusters selecting the best (minima / maxima) according to RMS Standard Deviation (a homogeneity measure) [78], to Bayesian information criterion (BIC) [64, 65, 66] or using "intra cluster distance" divided with "minimal inter cluster distance" [63].
- Use the first k data points.
- Use every m/k data point, where m is the total number of data points.
- Run the clustering several times each time with new random cluster centers among the data points and keep the best result.
- Run a hierarchical clustering first and make a cutting level according to the desired number of k clusters. The advantage of this approach over random initialization grows with the number of clusters.
- Initial clusters are spaced according to standard deviation distance, away from the central mean.
- Equally spread. e.g. four in each dimension will result in 4·4·4 = 64 clusters for a 3 dimensional data set.
- Starting with J subsamples, each being a set of initial cluster centers for the k-means clustering of the sub data set consisting of all the J subsamples. The initial clusters centers that result in minimal distortion are chosen as the initial clusters centers for the entire data set [67].
- Some algorithms have it built in like the Monte Carlo Cross-validation [38, 77, 87] which is a method for estimating an unknown distribution.
- See also [6] that tries to estimate the true number of clusters.
- See [86] for more initialization routines.

K.4 Non-Hierarchical Clustering

Here the data points are initially assigned to cluster centers and iteratively switching between them as they are updated. The result obtained by non-hierarchical clustering is normally a local minimum.

RGB Clustering
Partition of the data set into clusters on a grid using its multidimensional histogram. The peaks are used as cluster centers.

K-means Clustering
K-means is probably the most used clustering routine and was developed in 1967 by J. MacQueen [75]. In its original form, it first assigns every data point to their nearest clusters, then it recalculates the cluster centers as a mean of their data points. The data points are reassigned, and so on, until convergence.
The entire process and variations etc. are explained in Section K.14.

Isodata / Migrating Mean clustering Algorithm
ISODATA [41, 42] is one of the most used k-means variants. It is not bound by a specific number of clusters but can merge and split them to obtain better clusters.
The entire process and variations etc. are explained in Section K.15.

Forgy's K-means Clustering (1965)
The parent of K-means but here all cluster means are updated even if there are no changes [1]. Known to converge [2].

Sammon's Nonlinear Mapping
Sammon's Nonlinear Mapping [51] merges single data points and here the cluster centers are updated using gradient descent (minimizing the error).
The main usability is projection to visualize data structures in a lower dimension.

Incremental Clustering
Incremental Clustering (or adaptive clustering or sequential leader clustering) is a single pass algorithm:

1. First data point is a cluster.
2. If the next data point is closer to a cluster than a specified threshold, add it and update the cluster center or else make it a new cluster.
3. Repeat step 2 for all data points.

Variants

- An extra threshold defining the minimum distance allowed for a new cluster. If the distance is larger than allowed to merge but smaller than allowed to generate a new cluster, the data point is unclustered and first assigned to the closest cluster after going through all data points.
- After every M data point, all clusters with distances below a merge threshold are merged.
- See also Sequential Leader [15], Shortest Spanning Path (SSP) [16] and COBWEB [17].

Batchelor & Wilkins' Algorithm
Batchelor & Wilkins' Algorithm [76] (or maximum distance algorithm).

Specify f: Average distance between clusters.

1. Start with a random data point.
2. The next data point is selected as the one furthest away from the other clusters (by maximum or minimum distances).
3. If the distance, from the data point to the set of cluster centers, is above the specified f (the average interset distance), create a new cluster or else add it to the nearest cluster.
4. Repeat step 2 - 3 until convergence.

Clustering LARge Applications (CLARA)
A sampling based method specialized in dealing with large amounts of data [49].
The basic routine is k-medoids (k-means with cluster medians) but instead of using every data point it selects a random sample to cluster. The idea is that a representative sample will converge to the same result as the whole data set. It can be run several times and return the best result.

Clustering Large Applications based upon RANdomized Search (CLARANS)
CLARANS [33] is similar to CLARA, but the random sample is updated / renewed in each iteration resulting in faster convergence.
See also COD-CLARANS [36].

K.5 Hierarchical Clustering

Hierarchical clustering only merges two clusters or splits one cluster at a time. A step (merge or split) cannot be undone. It generates a tree structure showing all the merging and splitting and needs only one iteration. It can use a lot of different distance measures which can be seen in Section K.13.
A diagram of the process is called a dendrogram and gives a good visual idea of the data. A cutting level can be added so the present clusters, at this level, are kept. Results can also be shown using notes or Shepard plot.

K.5.1 Agglomerative (Bottom up, Joining or Pairwise Clustering)

Agglomerative hierarchical clustering (or AGglomerative NESting (AGNES)) only merges clusters. The main agglomerative approach is:

1. Start with one data point in each cluster.
2. Find the two clusters that are closest and merge them.
3. Repeat step 2 until the wanted k clusters are reached or till only one cluster is left.

There are two major groups of agglomerative clustering; the matrix method and the graph method.

Matrix Method

Here the similarity matrix is used and it is updated each time clusters are merged.
The different matrix methods are similar and mostly differ in the way they define and update the similarity matrix. This is why a general equation exits which can describe most of the variants by only changing a few constants.

Single Link (-age) or (Pairwise) Nearest-neighbor
Single linkage [4, 86] defines that clusters that are to be merged, are the ones with the minimum of the minimum distances between the data points in them.
The larger a cluster is, the easier it is to join and hence it is sensitive to noisy data. It contracts the space of relations among the data points in the nearby of the clusters.
It tends to find straggly, elongated, chained clusters.
Johnson's algorithm [5] uses the same approach.

Complete Link (-age) or Furthest-neighbor
Complete linkage [5, 86] is close related to single linkage, but clusters to be merged are the ones with the minimum of the maximum distances between any two data points in them.
The larger a cluster, the harder it is to join. It expands the space in the nearby of the clusters i.e. it moves apart from other clusters.

It tends to find extremely compact clusters.

Unweighed Pair-Group Method using Averages (UPGMA)
UPGA (or unweighed arithmetic average clustering) is as Complete linkage but clusters are compared using cluster averages. It calculates the Root Mean Square distance between all pairs of data within two different clusters.
It tends to find globular clusters.
A variant exists in using centroids instead of the average, called UPGMC. Here the centroids are dependent on weights, so large clusters do not dominate when merged.

Weighed Pair-Group Method using Averages (WPGMA)
WPGMA (or weighed arithmetic average clustering) is used when different groups are unequally sampled. Instead of reducing information, large groups are given a lower weight. It results in increased separation between groups in comparison compared to UPGMA.
A variant exists in using centroids instead of the average, called WPGMC. Here the centroids are dependent on weights so that large clusters do not dominate when merged.

Ward's Hierarchical Clustering Method
Ward (or Minimum variance) [62] suggests a within group variance approach where the total sum of variance (W) is minimized.

Maximum-Cut, Minimum-Cut, Median-Cut, Optimal-Cut
Maximum-Cut separates a graph in two and maximizes the sum of weights of the edges between the subsets, hence obtaining the minimum mean distance for each cluster center.

Minimum-Cut calculates the minimum number of links that must fail before at least one pair of nodes cannot communicate.

Median-Cut [83] is used for quantization of colors. It represents colors by a synthesized color map, each color with an equal number of pixels. Repeatedly, it subdivides the colorspace into smaller rectangular boxes. It uses adaptive partitioning to decide which way to split the box.

See also Optimal-Cut [84] (or Variance Based Clustering).

Iterative Shrinking Method
Iterative shrinking method [80] does not merge the selected clusters but their data points are relocated to the nearest clusters.

General Equation
Several of the above algorithms can be calculated from the same general equation, by altering some parameters:

$$\begin{aligned} Sim(C_{new}, C_{old}) &= a_1 \cdot Sim(C_{new1}, C_{old}) + a_2 \cdot Sim(C_{new2}, C_{old}) \\ &\quad + b \cdot Sim(C_{new1}, C_{new2}) + c \cdot |Sim(C_{new1}, C_{old}) - Sim(C_{new2}, C_{old})| \end{aligned} \tag{K.1}$$

where Sim(x,y) is the similarity measure between the clusters x and y and a_1, a_2, b, and c are constants defining the approach.

For single linkage, the parameters are:
$a_1 = a_2 = 1/2\ ,\ b = 0\ ,\ c = -1/2$
For complete linkage, the parameters are:
$a_1 = a_2 = 1/2\ ,\ b = 0\ ,\ c = 1/2$
For Weighed Pair-Group Method using Averages, the parameters are:
$a_1 = a_2 = 1/2\ ,\ b = c = 0$

Graph Methods

These methods are derived from graph theory. The distance / line between data points is called edges and the data points for vertices or nodes.

Spanning Tree

A spanning tree contains all vertices on the graph. If the edges are weighted by dissimilarities, then the Minimal Spanning Tree (MST) is the spanning tree with the minimal sum of weights among all possible spanning trees. The use of this theory in clustering [11] results in MST edges that are deleted if they have the longest length. The approach is similar to single-linkage.

Prim's Algorithm

Builds upon a single partial minimum spanning tree. At each step adding an edge connecting the vertex nearest to, but not already in, the current partial minimum spanning tree.

Kruskal's Algorithm

Maintains a set of partial minimum spanning trees (called forest), and repeatedly adds the shortest edge in the graph whose vertices are in different partial minimum spanning trees.
See also AUTOCLUST+ [37].

K.5.2 Divisive (Top Down or Splitting)

DIvisive ANAlysis (DIANA) starts with all the data points joined in one cluster which is splitted up again and again. It needs more computation than agglomerative, because it is more complicated to calculate which cluster to split and how to do it.
The main divisive approach is as follows:

1. Start with one cluster.
2. Find the "worst" cluster and split it.
3. Repeat step 2 until the wanted k clusters are reached.

The worst cluster can be defined as the largest one, the one with the largest variance or with the largest ESS (defined below), etc.
Split could be: Mean-median in one feature direction, perpendicular to the direction of the largest variance.

Minimum and maximum distance are extremely sensitive to outliers. Average and mean (the fastest) distance are more robust to outliers.

Variants

- Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) [27] which clusters

using tree structures - Clustering Features (CF).
- Clustering Using REpresentatives (CURE) [20] which has several representatives for each cluster and shrink them towards the center. Does not favor spherical shapes and are more robust to outliers. It is one of the most successful clustering methods for clustering data of any shape.
- ROCK [21] is used for categorical data, looking at inter-connectivity (the number of common neighbors) between clusters.
- CHAMELEON [23] defines similarity between clusters by both inter-connectivity and minimum distances.

K.6 Nearest Neighbor Clustering

The objective here is to look at each data point's k nearest neighbors.

Jarvis-Patrick method
This algorithm [3] needs only a single pass, but is not a Hierarchical even though the result is similar to those produced by single-linkage.
Specify: The k nearest neighbors needed to be considered and K_{min}: The necessary level of similarity between nearest neighbors.
Two data points are assigned to the same cluster if they both are one of each others k nearest neighbors and have at least k_{min} common k nearest neighbors.

K.7 Fuzzy Clustering

Fuzzy clustering uses degree of membership [54] which means that a data point can be a member of several clusters with a total degree of membership summing to 1.
The main usage is when clusters are not well defined (overlapping) such as images covering geographical areas and vegetation.
Many of the described clustering algorithms can be altered to a fuzzy version and k-means is one of the most popular ones. The fuzzy version is called Fuzzy c-means (FCM) where the c also refers to the number of clusters.

Fuzzy ISODATA
Fuzzy ISODATA [53] is a directly changed version of the hard ISODATA.

Fuzzy C-shell Algorithm
It is, with an adaptive variant, used for detecting circular and elliptical boundaries [52].

Possibilistic C-Means (PCM)
When using PCM [61] the degree of membership (probabilistic) is changed to possibilistic, an absolute membership (no need for summing to 1).

Fuzzy K-Nearest Neighbor (FKNN)
FKNN [55] is a diffuse variant of the hard version. The k nearest data points of the one being processed are ordered in greater or lesser resemblance. The data point in question is assigned mostly to the first on the list, less to the next and so on.

Modified Fuzzy C-means (MFCM)
This algorithm [56] consists of a hard and a fuzzy part:
Hard: The histogram of each colorband is smoothed and convoluted with different Gaussian functions hereby obtaining a set of thresholds that isolates the ranges of color intensity. It is used to generate independent "pre-clusters" / initial clusters. Further processing results in fewer and isolated cluster centers.
Fuzzy: The data points not yet assigned (those between pre-clusters), are assigned. The pre-clusters are NOT updated during this process (the centers are fixed).

Fuzzy Weights
Instead of using mean or median as cluster center, weighted average is used by Weighted Fuzzy Expected Values (WFEV) [58]. It is primarily used to avoid outlier influence. The weight could be the inverse of the distance from the data points to its cluster center, hence the further away the less influence.

Spatial Models
As with k-means, spatial information can be included in the algorithms. By adding an extra term to the membership function, a spatial penalty can represent e.g. edge information [60].

K.8 Density-based

This approach is not directly distance based and is efficient to discover clusters of arbitrary shape and also to avoid outliers. It continues growing a cluster as long as the density (number of data points) in the neighborhood exceeds some threshold, i.e. each data point in the cluster should contain at least a minimum number of data points within a specified radius.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
DBSCAN [19] defines a data point as a core object, if there are more data points than a threshold in its ϵ-neighborhood defined by the radius ϵ. A cluster is defined as a maximum set of density-connected points, i.e. the connection of multiple number of core object results in arbitrary shape. If the core object has another core object within its radius, they are directly density-reachable. If they only have mutual data points, they are density-reachable and if they are only connected through other core objects, they are density-connected. Due to this definition of core objects, outliers will be ignored.

Generalized DBSCAN (GDBSCAN)
GDBSCAN [19] is generalized so it also can take spatial dimensions (like gradient information).

Ordering Points To Identify the Clustering Structure (OPTICS)
To avoid the parameters in DBSCAN (and all other clustering routines) OPTICS [18] can be used. Here the data points are not clustered but an augmented cluster ordering is computed. Core-distance is a data point's smallest distance to make it a core object. Reachability-distance is the maximum value of the core-distance and the Euclidean distance between a core object and a data point.
It is used for automatic or interactive clustering analysis.

DENsity-based CLUstEring (DENCLUE)
DENCLUE [50] uses density distribution functions. Each data point's influence is modelled by an influence function, describing the data points impact on its neighborhood. The overall density is the sum of all influence functions. Clusters are now defined where there are local maxima (density attractors) of the overall density function, σ. Densities smaller than the noise threshold ξ are considered outliers. This approach is similar to "shape matching" where Gaussian shapes are fitted on the data set. It has the advantages of having a solid mathematical foundation, generalizes "all" other clustering methods, resistant to large amounts of noise, arbitrary clustering shapes, but strongly dependent on the density parameter σ and the noise threshold ξ.
See [81] for an iterative approach to find σ.
See also TURN [34] and DBCluC [35].

Cohesion-based Self-merging Algorithm (CSM)
CSM [79] compares clusters not only on the data points but also on their distributions. It defines the radius of a cluster and the joinability.

K.9 Model-based

In this clustering type, a model is hypothesized for each of the clusters and the idea is to find the best fit of that model to the other models.

CLASSY
CLASSY is an adoptive Maximum Likelihood clustering algorithm and is based on the assumption of a multivariate normal mixture model for the data. It attempts to estimate the number of components of the mixture via a sequence of hypothesis tests using a likelihood criterion. It uses a fuzzy approach where data points can belong to several clusters.
It starts with one large cluster and keeps splitting while cluster centers, weights and variance-covariance matrices are adjusted. Convergence is obtained when the weights change less than a threshold.
Clusters can also be merged and the process is logged creating a tree of clusters. Splitting occurs when skew and kurtosis of the variance-covariance matrices exceed a threshold. Merging occurs when cluster centers, weights and variance-covariance matrices are alike.

COBWEB
COBWEB creates a hierarchical clustering in the form of a classification tree.

K.10 Artificial Neural Network (ANN)

ANN [12 - 14, 88 - 90] uses quantitative features only. It may adaptively learn its interconnection weights. It can act as a feature normalizer and feature selector by appropriate selection of weights. It has a simple architecture and is single layered. Data points are presented at the input and associated with the output nodes. The weights between the input and output nodes are iteratively changed (called learning).
It is difficult to adjust the parameters.

Competitive Learning
Here the nearest cluster to a new data point is called the winner.
Winner-takes-all; several clusters compete about the data point. The closest one wins and moves towards the data point. Non-winning clusters do not move and have not learned, they play no role, and are underutilized.

Variants
- The Krishnamurthy Implementation of the Conscience Principle [45]: The same cluster should not be allowed to win all the time and hence a distance penalty grows each time it wins. Clusters near the winner also move in the direction of the data point.
- Dog-Rabbit Algorithm or The McKenzie-Alder Learning Vector Quantization Method [44]: The dogs (clusters) at a certain distance from the rabbit (new data point) moves the most against the rabbit, while dogs closer moves less, and dogs far away moves very little.
- Fuzzy Competitive Learning, which is the basic alteration to a fuzzy version [57].

Self Organizing Maps (SOM)
SOM [24] is defined so that each cluster has a weight and only the weighted winner becomes active and moves towards to the data point. The more a cluster wins the higher, its weight gets. Only one data point is active in each cluster.

The approach is as follows:
1. Chose the dimension and size of the map (Neurons in each dimension).
2. For each new data point calculate the distance to all cluster centers.
3. Recompute all cluster centers vectors with the new vector, using both a distance radius on the map and a learning rate (the moving distance) that decreases with time.

Vector Quantization (VQ)
Here the goal is to minimize the average (squared) quantization error. It is mostly used in signal coding where multidimensional vectors are quantized by finding a number of multidimensional prototypes (cluster centers). In other words; lossy data compression by reducing data to a small set of representatives.

Learning Vector Quantization (LVQ)
Kohonen [46] altered the basic algorithm by the following:
- The winner is determined by the greatest correlation, hence the dot product should be calculated.
- All clusters in the neighborhood move towards the winner (are reinforced) and those further away move away from the winner (a step towards extinction).

Neural Gas
Here the clusters are sorted according to their distances to the new data point. Then all cluster centers are updated using the new data point, in this order.

Growing Neural Gas
Growing Neural Gas [47] is like Neural Gas, but connects the clusters for a period of time (a number of interactions). For each new data point, the two nearest clusters are connected and

the age, set to 0. All ages are then increased by 1. Connections older than a threshold are deleted. Delete the connection between the two clusters with the highest accumulated error and split them into three clusters and connect the new one with the two old ones.

Growing Cell
Growing Cell [47] connects the initial clusters as a chain. For each new data point, delete the connection between the two clusters with the highest accumulated error and split them into three clusters and connect the new one with the two old ones.

Growing Grid
Growing Grid is like Self Organizing Maps, but here for each new data point, both a distance radius and a decreasing learning rate, is observed for all clusters. The data point is added by observing the maximum accumulated error as with Growing Neural Gas.

LBG (Linde, Buzo & Gray) or Generalized Lloyd
Search each cluster for the closest data points in the data set (called a Voronoi set). Update each cluster center using the mean of those data points. Repeat until none of the cluster centers change.

K.11 Evolutionary Approaches for Clustering

This approach is motivated by natural evolution. Cluster centers are encoded as chromosomes. Selection, recombination and mutation transforms one or more input chromosomes into one or more output chromosomes. A fitness function determines the likelihood of surviving into the next generation for each chromosome. The fitness function is inversely proportional to the squared error value. Selection, recombination, mutation and fitness updates repeat until convergence.
Generic algorithm [28], evolution strategies [29] and evolutionary programming [30] are algorithms using this approach.

Greedy Randomized Adaptive Search Procedure (GRASP)
The Greedy algorithm [85] contains a Restricted Candidate List (RCL), a sample of the data (e.g. 5 % of the data points), which are better candidates for a solution.
The first cluster center is arbitrarily chosen among the data points in the RCL.
Each new cluster is the data point furthest away from the existing cluster centers.
After each interaction, a solution construction phase and post processing are carried out to optimize the found solution.
Local search technique: If the solution is not a local minimum, it is replaced with one of its neighbors.
Extended GRASP: Uses mutation; with a probability p, each of the cluster centers can be altered, making a good search space exploration.

K.12 Other Clustering Algorithms

Simulated Annealing
Simulated annealing [31] is a stochastic search technique designed to avoid solutions which cor-

respond to local optimum, using stochastic relaxation. It accepts, with some probability, a new solution of lower quality. It is statistically guaranteed to find the global optimal solution.

SYNERgistic Automatic Clustering Technique (SYNERACT)
Combines hierarchical divisive clustering and k-means clustering in order to avoid limitations in ISODATA (a priori knowledge needed to specify parameters).

Grid-based Methods
Grid-based clustering uses a multiresolution grid data structure, obtained by quantizing the space into a finite number of cells. The clustering is performed on this grid which typically makes the routine time independent of the number of data points.
STatistical INformation Grid approach (STING) [26] looks at the statistical information in the grid.
WaveCluster [25] is clustering using wavelet transformation.
CLIQUE [32] is a grid- and density-based approach for high dimensional data.

Ant Based Approach
Here [82] the angle is copied from the stochastic principles of an ant colony (or any swarm) combined with the deterministic principles of the k-means algorithm.
The ants-like agents move randomly on a 2D grid and are allowed to move (pick up and later drop) data points and thereby classifying them.

The Following Clustering Approaches have not been Studied
- Mixture decomposition schemes
- Vacuum shell detection
- AMOEBA
- Ford-Fulkerson Labelling Algorithm
- Optimal-Cut or Variance Based
- Murtagh's Reciprocal Nearest Neighbor algorithm
- Tabu search
- Function optimization clustering methods (using e.g. branch-and-bound)
- Voronoi Based Adaptive K-means Clustering
- Pose clustering
- Focusing techniques
- Ejcluster
- MONA
- FANNY

K.13 Distance Measures

There are several different distance measures (metrics) which favors different shapes / distributions of the clusters. Below some of them are showed.

City-block Distance a.k.a. Manhattan Distance a.k.a. L_1 Norm

$$d_1(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{n} |x_{i,k} - x_{j,k}| = ||\mathbf{x}_i, \mathbf{x}_j|| \tag{K.2}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the clusters and n is the dimension of feature space.

Euclidean Distance a.k.a. L_2 Norm

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{n} |x_{i,k} - x_{j,k}|^2} = ||\mathbf{x}_i, \mathbf{x}_j||_2 \tag{K.3}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the clusters (feature vectors) and n is the dimension of feature space.

Minkowski Distance a.k.a. L_k Norm

$$d_p(\mathbf{x}_i, \mathbf{x}_j) = \sqrt[k]{\sum_{k=1}^{n} |x_{i,k} - x_{j,k}|^k} = ||\mathbf{x}_i, \mathbf{x}_j||_k \tag{K.4}$$

where k is the distance dimension (a positive integer), $\mathbf{x}_i$ and $\mathbf{x}_j$ are the clusters and n is the dimension of feature space. For $p = 1$ it is equal to the City-block distance and for $p = 2$ it is equal to the Euclidean distance.

Mahalanobis Distance

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i, \mathbf{x}_j)\Sigma^{-1}(\mathbf{x}_i, \mathbf{x}_j)^T \tag{K.5}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the clusters and Σ is the covariance matrix of the clusters.

Tanimoto Distance
The Tanimoto distance is commonly used for binary features

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\mathbf{x}_i^T \mathbf{x}_i + \mathbf{x}_j^T \mathbf{x}_j - \mathbf{x}_i^T \mathbf{x}_j} \tag{K.6}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the clusters.

Mutual Neighbor Distance (MND)

$$MND(\mathbf{x}_i, \mathbf{x}_j) = NN(\mathbf{x}_i, \mathbf{x}_j) + NN(\mathbf{x}_j, \mathbf{x}_i) \tag{K.7}$$

where $NN(\mathbf{x}_i, \mathbf{x}_j)$ is the neighbor number of $\mathbf{x}_j$ with respect to $\mathbf{x}_i$ (1 is the closest neighbor, 2 is the second closest, etc.).

Minimum Distance a.k.a. Nearest Neighbor

$$d_{min}(\mathbf{x}_i, \mathbf{x}_j) = \min_{\substack{x \in \mathbf{x}_i \\ y \in \mathbf{x}_j}} ||x - y|| \tag{K.8}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are clusters, x and y are the data points in the respective clusters.

Maximum Distance a.k.a. Furthest Neighbor a.k.a. Chebychev a.k.a. L_∞ Norm

$$d_{max}(\mathbf{x}_i, \mathbf{x}_j) = \max_{\substack{x \in \mathbf{x}_i \\ y \in \mathbf{x}_j}} ||x - y|| \tag{K.9}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are clusters, x and y are the data points in the respective clusters [86].

Average Distance
d_{avg} is the mean of all inter-distances:

$$d_{avg}(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{N_i N_j} \sum_{x \in \mathbf{x}_i} \sum_{y \in \mathbf{x}_j} ||x - y|| \tag{K.10}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are clusters, x and y are the data points in the respective clusters and N_i and N_j are the number of data points in the respective clusters.

Mean Distance

$$d_{mean}(\mathbf{x}_i, \mathbf{x}_j) = ||\mu_i - \mu_j|| \tag{K.11}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are clusters and μ_i and μ_j are the cluster centers.

K.14 K-means

K-means is described by MacQueen in 1967 [75] and is probably the most used clustering routine. It is based on minimizing the sum of squared distances from the data points to their respective clusters by interchanging / switching them from cluster to cluster.
For each cluster the Error Sum of Squared distances (ESS) is calculated as a measure of dissimilarity. The Total Error Sum of Squares (TESS) is the sum of each cluster's ESS and is reduced during the iterations hence when it stops decreasing, a local minimum is reached. It is known to converge [2].

Input
k: Wanted number of clusters.

Optional
The initial cluster positions.
Maximum number of iterations.
Convergence threshold (default is 0).

Process
1. If initial cluster positions are not specified, choose them randomly or those that are as far apart as possible (see Section K.3 on this issue).
2. Calculate the Euclidean distance from each data point to each centroid (cluster center).
3. Assign each data point to its closest centroid.
4. Recalculate the positions of the centroids.
5. Continue step 2 - 4 until convergence or the specified number of iterations is reached.

Could be run / tested with different values of k to find a suitable one.
It is possible to merge clusters manually after the routine has stopped hence it is better to select k too large than too small.

Formulas
The Error Sum of Squares (ESS) and the Total Error Sum of Squares (TESS) are calculated as follows

$$\begin{aligned} ESS(i) &= \sum_{j=1}^{N_i} ||x_{ij} - \mu_i||^2 \\ TESS &= \sum_{i=1}^{C} \sum_{j=1}^{N_i} ||x_{ij} - \mu_i||^2 \end{aligned} \tag{K.12}$$

x_{ij} is the data point j in cluster i, μ_i the mean of cluster i, N_i is the number of data points in cluster i and C is the number of clusters.

Variations
- The centroid can be changed to medoids, the median of the data points (called Partitioning Around Mediods (PAM) [49]), or to the average. Changing from centroids results in a more dynamic clustering algorithm.
- Handles categorical data (called k-modes [9, 10, 22]) by replacing centroids with modes and distances with dissimilarity measures or a mixture (called k-prototype).
- Fuzzy: Expectation Maximization (EM) where the partial membership results in cluster means based on weighted measures.
- Leave-one-out: When calculating the distance from a data point to its cluster center, the data point has been extracted from the cluster thus it has no effect on the location of the center. This is only a minor validation mechanism which normally does not change the result.
- Continuous: The initial cluster centers are based on the distribution of the data points. Hence it assigns more cluster centroids where there are more data points. It only updates a random sample of the data set, resulting in a much faster convergence.
- Edge-adaptive k-means [74]: Here edge information is incorporated by adding an additional term to the objective function of the K-means algorithm. The use of a priori knowledge results in efficient reconstruction of boundaries in the image.
- COP-k-means [73]: Another a priori knowledge is pairwise constraints like; Must-link and Cannot-link for two data points must be in the same cluster or cannot be in the same cluster, respectively.
- Morphological operations [68] use a priori anatomical information of the region of interest.
- A k-means like algorithm has been analytically derived using Local Search, called LKM [72]. It differs in the way that it looks at the after effect of moving a data point before moving it. It is extended to a variant, for large data sets with monotone convergence property, called A-KLM.
- K-means is not perfect, resulting in a large verity of variants, but not all problems can be overcome. [71] discusses 9 problems (like a priori knowledge of the number of clusters and non-invariant to scale transformations) and solutions, for some of them.

- To speed up the algorithm, [70] suggests organizing / sorting the patterns in a k-dim. tree structure resulting in faster finding the closest cluster center for a given data point.
- K-means can also be used for reducing the dimensions of the data set. This can be used for compression or visual presentation of multidimensional data [69].
- See [7] for more variations.

Strength
Simple and efficient.
Will converge to a local minimum.
Independent of the order of the data by only updating centroids after each iteration.

Weakness
The manual choice of k clusters and possibly other parameters.
No steps against noisy data or outliers for which it is sensitive.

K.15 ISODATA

Iterative self-organizing data analysis technique (ISODATA) [8, 41 - 43] - the last "A" is for pronunciation! - is based on k-means clustering. The difference is that ISODATA is capable of varying the number of clusters by splitting and merging them hence obtaining a much more flexible clustering algorithm.

Input
Maximum number of clusters allowed.
or / and
Maximum cluster variance (for splitting).
Maximum distance separation (for merging).
or / and
Minimum number of data points per cluster.

Optional
Initial number of clusters.
The initial cluster positions.
Maximum number of iterations.
Convergence threshold (default is 0).
Maximum number of mergers per iteration.

Process
1. If initial cluster positions are not specified, choose them randomly or as far a part as possible.
2. Perform k-means clustering.
3. Split clusters that contain too dissimilar data points according to the specified threshold.
4. Merge clusters that are closer than the specified threshold.
5. Continue step 2 - 4 until convergence or the specified number of iterations is reached.

Split and merge only if in agreement with the possible specified "Maximum number of clusters

allowed" and "Minimum number of data points per cluster".

Variations
Distributed ISODATA (D-ISODATA) [42].
Fuzzy version [40].

Strength
Self-organizing capabilities and hence very flexible.
Successful at finding spectral clusters.
Depending on which options are specified, the possibility of human error is minimized.
Is usable as input or modified input to a supervised classifier.

Weakness
Little human control over the clusters hence making inter-comparison difficult.
The possibility of many iterations before convergence.
Data must be linear separable. Long, narrow or curved clusters are not handled properly.
Initial guesses on some of the options are difficult.
Performance is highly dependent of the choice of these options.
Equal covariance is assumed.

K.16 References

[1] S. Z. Selim and M .A. Ismail, "K-means type algorithms: a generalized convergence theorem and characterization of local optimality", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 6, P. 81 - 87, 1994.
[2] E. Forgy "Cluster analysis of multivariate data: efficiency versus interpretability of classifications", Biometry 21, 768, 1965.
[3] R.A. Jarvis, E.A. Patrick "Clustering Using a Similarity Measure Based on Shared Near Neighbors", IEEE Transactions on Computers, C22, P. 1025 - 1034, 1973.
[4] P. H. Sneath and R. R. Sokal, Numerical Taxonomy, Freeman, London, UK, 1973.
[5] B. King, "Step-wise clustering procedures", J. Am. Stat. Assoc. 69, P. 86 - 101, 1967.
[6] R. C. Dubes, "How many clusters are best? - an experiment" Pattern Recogn. 20, 6, p. 645 - 663, 1987.
[7] M. R. Anderberg, "Cluster Analysis for Applications", Academic Press, Inc. New York, NY, 1973.
[8] G. H. Ball and D. J. Hall, "ISODATA, a novel method of data analysis and classification", Tech. Rep. Standford University, Standford, CA, 1965.
[9] E. Diday, "The dynamic cluster method in non-hierarchical clustering", J. Comput. Inf. Sci. 2, p. 61 - 88, 1973.
[10] M. J. Symon, "Clustering criterion and multi-variate normal mixture", Biometrics 77, p. 35 - 43, 1977.
[11] C. T. Zahn, "Graph-theoretical methods for detecting and describing gestalt clusters" IEEE Trans. Comput. C-20, p. 68 - 86, 1971.
[12] J. K. Hertz and R. G. Palmer, "Introduction to the Theory of Neural Computation", Addision-Wesley Longmann Pucl. Co., Inc.Reading, MA, 1991.
[13] A. K. Jain and J. Mao, "Neural Networks and pattern recognition". Comp.: Imitating Life,

p. 194 - 212, 1994.
[14] I. Sethi and A. K. Jain, "Artificial Neural Networks and pattern recognition: Old and New Connections", Elsevier Sci. Inc., New York, NY, 1991.
[15] J. A. Hartigan, "Clustering Algorithms", John Wiley and Sons, Inc., New York, NY, 1975.
[16] J. R. Slagle, C. L. Chang, S. R. Heller, "A clustering and data-reorganizing algorithm", IEEE Trans. Syst. Man Cybern. 5, p. 125 - 128, 1975.
[17] D. Fisher, "Knowledge acquisition via incremental conceptual clustering", Mach. Learn. 2, p. 139 - 172, 1987.
[18] M. Ankerst, M. Breunig, H. P. Kriegel and J. Sander, "Optics: Ordering points to identify the clustering structure", Proc. 1999 ACM-SIGMOD Int. Conf. Manage. of Data, p. 49 - 60, 1999.
[19] M. Ester, H. P. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases", Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining, p. 226 - 231, 1996.
[20] S. Guha, R. Rastogi and K. Shim, "Cure: An efficient clustering algorithm for large databases", Proc 1998 ACM-SIGMOD Int. Conf. Manage. of Data, p. 73 - 84, 1998.
[21] S. Guha, R. Rastogi and K. Shim, "Rock: A robust clustering algorithm for categorical attributes", Proc 1999 Int. Conf. Data Engineering p. 512 - 521, 1999.
[22] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values", Data Mining and Knowledge Discovery (2), p. 283 - 304, 1998.
[23] G. Karypis, E. H. Han ad V. Kumar, "CHAMELEON: A hierarchical clustering algorithm using dynamic modelling", COMPUTER (32), p. 68 - 75, 1999.
[24] T. Kohnen, "Self-organized formation of topologically correct feature maps", Biological Cybernetics (43), p. 59 - 69, 1982.
[25] G. Sheikholeslami, S. Chetterjee and Z. Zhang, "WaveCluser: A multi-resolution clustering approach for very large spatial databases", Proc. 998 Int. Conf. Very Large Data Bases, p. 428 - 439, 1998.
[26] W. Wnag, J. Yang, R. Muntz, "STING: A statistical information grid approach to spatial data mining", Proc. 1997 Int. Conf. Very Large Data Bases, p. 186 - 195, 1997.
[27] T. Zhang, R. Ramakrishnan and M. Livny, "BIRCH: an efficient data clustering method for very large databases", Proc 1996 ACM-SIGMOD Int. Conf. Manage. of Data, p. 103 - 114, 1996.
[28] J. H. Holland, "Adaption in Natural and Artificial systems", University of Michigan Press, Ann Arbor, MI, 1975.
[29] H. P. Schwefel, "Numerical Optimization of Computer Models", John Wiley and Sons, Inc., New York, NY, 1981.
[30] L. J. Fogel, A. J. Owens and M. j. Walsh, "Artificial Intelligence Through Simulated Evolution", John Wiley and Sons, Inc., New York, NY, 1965.
[31] S. Kirkpatrick, C. D. Gelatt Jr. and M. P. Vecchi, "Optimization by simulated annealing", Science 220, p. 671 - 680, 1983.
[32] R. Agrawal, J. Gehrke, D. Gunopulos, P.Raghaven, "Automatic subspace clustering of high dimensional data for data mining applications", Proc 1998 ACM-SIGMOD Int. Conf. Manage. of Data, p. 94 - 105, 1998.
[33] R. Ng and J. Han, "Efficient and effective clustering method for spatial data mining", Proc. 1994 Int. Conf. Very Large Data Bases, p. 144 - 155, 1994.
[34] A. Foss and O. R. Zaïane, "TURN* unsupervised clustering of spatial data", ACM-SIGKDD

Int. Conf. on Knowledge Discovery and Data Mining, 2002.
[35] O. R. Zaïane and C. H. Lee, "Clustering spatial data in the presence of obstacles and crossings: a density-based approach", IDEAS Int. Database Engineering and Applications Symposium, 2002.
[36] A. K. H. Tung, J. Hou, J. Han, "Spatial clustering in the presence of obstacles", Proc. ICDE Int. Conf. On Data Engineering, 2001.
[37] V. Estivill-Castro and I. Lee, "Autoclust+: Automatic clustering of point-data sets in the presence of obstacles", Int. Workshop on Temporal and spatio-Temporal Data Mining, p. 133 - 146, 2000.
[38] O. R. Zaïan, A. Foss, C. H. Lee and W. Wang, "On Data Clustering Analysis: Scalability, Constraints and Validation", PAKDD Adv. Knowledge Discovery and Data Mining, p. 28 - 39, 2002.
[39] A. K. Jain, M. N. Murty, P. J. Flynn, "Data Clustering: A Review", ACM Comp. Surveys, Vol. 31(3), p. 264 - 323, 1999.
[40] W. Pedrycz and J. Waletzky, "Fuzzy clustering in software reusability", Software-Practice Experience 27 (3), p. 245 - 270, 1997.
[41] g. H. Ball and D. J. Hall, "A clustering technique for summarizing multivariate data", Behav. Sci. 12, p. 153 - 155, 1967.
[42] M. K. Dhodhi, J. A. Saghri, I. Ahmad, R. U. Mustafa, "D-ISODATA: A Distributed Algorithm for Unsupervised Classification of Remotely Sensed Data on Network of Workstations", Jour. of Parallel and Dist. Comp. 59, p. 280 - 301, 1999.
[43] W. Niblack, "An Introduction to Digital Image Processing", Strandberg, Birkerød, DK, 1985.
[44] P. McKenzie and M. Alder, "Unsupervised learning: the dog rabbit strategy", Proc. Ieee Int. Conf. Neural Networks (2), 1994.
[45] A. K. Krishnamurthy, S. C. Ahalt, D. E. Melton and P. Chen, "Neural networks for vector quantization of speech and images", IEEE J. on Selected Areas in Com. vol. 8, p. 1449 - 1457, 1990.
[46] T. Kohonen, "Learning Vector Quantization for Pattern Recognition", Technical Report TKK-F-A601, Helsinki University of Technology, 1986.
[47] B. Fritzke, "Fast learning with incremental RBF networks. Neural Proc. Letter 1(1), p. 1 - 5, 1994.
[48] H. Frigui and R. Krishnapuram, "A Robust Clustering Algorithm Based on Competitive Agglomeration and Soft Rejection of Outliers", IEEE Conf. on Comp. Vision and Pattern Recogn., p. 550 - 555, 1996.
[49] L. Kaufman and P. J. Rousseeuw, "Finding Groups in Data: An Introduction to Cluster Analysis", John Wiley and Sons, Inc., New York, NY, 1990.
[50] A. Hinneburg and D. A. Keim, "An Efficient Approach to Clustering in Large Multimedia Databases with noise", KDD'98, New York, 1998.
[51] J. W. Shanmmon, "A nonlinear mapping for data structure analysis", IEEE Trans. Comp. 18(5), p. 491 - 509, 1969.
[52] R. N. Dave, " Generalized fuzzy C-shells clustering and detection of circular and elliptic boundaries", Pattern Recogn. 25, p. 713 - 722, 1992.
[53] J. C. Bezdek, "Pattern Recognition With Fuzzy Objective Function Algorithms", Plenum Press, New York, Ny, 1981.
[54] C. T. Zahn, "Graph-theoretical methods for detecting and describing gestalt clusters", IEEE

Trans. Comput. C-20, p. 68 - 86, 1971.
[55] J. A. Givens Jr, M. R. Gray and J. M. Keller, "A fuzzy K-Nearest-Neighbour algorithm", IEEE Trans. Syst. MCypern, vol. SMC-15(4), p. 580 - 585, 1985.
[56] S. U. Lee and Y. M. Lim, "On the color image segmentation algorithm based on the thresholding and the Fuzzy c-Means techniques", Pattern Recogn. vol 23(9), 1990.
[57] F. L. Chung and T. Lee, "Fuzzy competitive Learning", Neural Networks, Vol 7(3)m P. 539 - 552, 1994.
[58] M. Schneider and M. Craig, "On the use of fuzzy sets in histogram equalization", Fuzzy Sets Syst., vol 45, p. 271 - 278, 1992.
[59] J. C. Bezdek, "A physical interpretation of fuzzy ISODATA", IEEE Trans. On Syst. Man and Cybernetics 6, p. 387 - 389, 1976.
[60] D. L. Pham, "Spatial Models for Fuzzy Clustering", Comp. Vision and Image Unders. 84, p. 285 - 297, 2001.
[61] R. Krishnapuram and J. M. Keller, "Possibilistic approach to clustering", IEEE Trans. on Fuzzy Syst. (1), p. 98 - 110, 1993.
[62] J. H. Ward Jr, "Hierarchical Grouping to Optimise an Objective Function" J. Amer. Statist. Assoc. 58, No 301, p. 236 - 244, 1963.
[63] S. Ray and R. H. Turi, "A new approach to clustering-based colour image segmentation", Proc. of Signal and Image Proces., p. 345 - 349, 1996.
[64] C. Fraley and A. E. Reftery, "How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis", Comp. J. (41), p. 578 - 588, 1998.
[65] A. Dasqupta and A. E. Reftery, "Detecting features in spatial point processes with clutter via model-based clustering", J. Amer. Stat. Assoc., 93, p. 294 - 302, 1998.
[66] S. Mukerjee, E. D. Feigelson, G. J. Buba, F. Murtagh, C. Fraley and A. E. Raftery, "Three types of gamma ray bursts", Astrophys. J., 508, p. 314 - 327, 1998.
[67] U. Fayyad, C. Reina and P. S. Bradley, "Initialization of Iterative Refinement Clustering Algorithms", Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining, p. 194 - 198, 1998.
[68] C. W. Chen, J. Luo and K. J. Parker, "Image Segmentation via Adaptive K-Mean Clustering and Knowledge-Based Morphological Operations with Biomedical Applications", IEEE Trans. On Image Proc. 7, No 12, p. 1673 - 1683, 1998.
[69] A. Morrison, G. Ross and M. Chalmers, "Combining and comparing clustering and layout algorithms", IEEE TCVG Sym. on Visualization, 2002.
[70] K. Alsabti, S. Ranka and V. Singh, "An efficient space-partitioning based algorithm for the k-means clustering", PAKDD Meth. for Knowledge Discovery and Data Mining, p. 355 - 359, 1999.
[71] I. Davidson, "Understanding K-Means Non-hierarchical Clustering", Albany Tech. Report: 02-2, 2002.
[72] B. Zhang, G. Kleyner and M. Hsu, "A Local Search Approach to K-Clustering", HP Labs Technical Report, 1999.
[73] K. Wagstaff, C. Cardie, S. Rogers and S. Schroedl, "Constrained K-means Clustering with Background Knowledge", Proc. of the 8th Int. Conf. on Mach. Learn. p. 577 - 584, 2001.
[74] D. L. Pham, "Edge-adaptive clustering for unsupervised image segmentation", IEEE Image Proc.(1), p. 816 - 819, 2000.
[75] J. MacQueen, "Some methods for classification and analysis of multivariate observations", In Proc. 5th Berkeley Symp. Math. Statist. Prob., 1967.

[76] B. Batchelor and B. Wilkins, "Method for location of clusters of patterns to initialize a learning machine", Electronics Letters Vol. 5(20), p. 481 - 483, 1969. [77] P. Smyth, "Clustering using Monte Carlo cross-validation", Proc. ACM-SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 1996.
[78] S. Sharma, "Applied Multivariate Techniques", John Wiley & Sons, 1996.
[79] C. R. Lin and M. S. Chen, "A Robust and Efficient Clustering Algorithm based on Cohesion Self-Merging", SIGKDD 02 Edmonton, Alberta, Canada, 2002.
[80] O. Virmajoki, P. Fränti and T. Kaukoranta, "Iterative shrinking method for generating clustering", IEEE Conf. Img. Proc., Vol.2, p. 685 - 688, 2002.
[81] A. Denton, Q. Ding and W. Perrizo, "Efficient Hierarchical Clustering of Large Data Sets Using P-trees", CAINE'02, San Diego, Nov. 2002.
[82] J. L. Deneuborg, "The dynamics of collective sorting, robot-like ants and ant-like robots", 1st Int. Conf. on Simul. of Adapt. Beh., MIT Press, p. 356 - 363, 1990.
[83] P. Heckbert, "Color image quantization for frame buffer display", Computer Graphics vol. 16(3), p. 297 - 307, 1982.
[84] S. J. Wan, S.K.M. Wong and P. Prusinkiewicz, "An algorithm for multidimensional data clustering", ACM Transactions on Mathematical Software vol. 14(2), p. 153 - 162, 1988.
[85] S. Sahni and T. Gonzales, "P-complete approximation problems," Journal of the ACM vol. 23, p. 555 - 565, 1976.
[86] M. R. Anderberg, "Cluster analysis for applications", Academic press, New York, 1973.
[87] E. Backer, "Computer-assisted Reasoning in Cluster Analysis", Prentice Hall, London, 1995.
[88] M. Sonka and J. M. Fitzpatrick, "Medical Imaging, vol. 2", SPIE, Bellingham, Washington, 2000.
[89] M. Sonka, V. Hlavac and R. Boyle, "Image Processing, Analysis, and Machine Vision", Brooks/Cole, California, USA, 1999.
[90] R. C. Gonzales and R. E. Woods, "Digital Image Processing", Addison-Wesley, USA, 1992.