Lappeenranta University of Technology
School of Industrial Engineering and Management
Degree Program in Computer Science

Bachelor's thesis

Joni Herttuainen

STATISTICAL SEGMENTATION METHODS AND COLOR VARIANCE ANALYSIS OF RETINAL IMAGES

Examiner(s): Professor Lasse Lensu

Supervisor(s): Professor Lasse Lensu

ABSTRACT

Lappeenranta University of Technology
School of Industrial Engineering and Management
Degree Program in Computer Science

Joni Herttuainen

Statistical segmentation methods and color variance analysis of retinal images

Bachelor's thesis

2014

33 pages, 9 figures, 2 tables, 6 appendices

Examiner(s): Professor Lasse Lensu

Keywords: color variation analysis, statistical classification, Naïve Bayes classifier, Gaussian mixture model, medical segmentation, retinal images.

Because of the demand of semi-automatic medical image segmentation tools, various of methods have been studied as possible candidates for the implementation. In this research, the effectiveness of Naïve Bayes and Gaussian Mixture Models classifiers on segmenting exudates in retinal images is studied and the results are evaluated with metrics commonly used in medical imaging. Also, because there are a number of methods (including methods used in the research) based solely on color information of the images, a color variation analysis of retinal images is carried out to find how effectively can retinal images be segmented using only the color information of the pixels.

FOREWORD

I would like to thank the Machine Vision and Pattern Recognition laboratory of the Lappeenranta University of Technology for offering me both the opportunity and the means to work in their laboratory and to carry out my research.

I would also like to thank the staff for their support and the kindness to help me with the problems I stumbled upon. Special thanks goes to Dr. Lasse Lensu for mentoring me and giving me the knowledge I needed to carry out this research.

I would also like to thank the University of Bristol and Bristol Eye Hospital for letting me use their non-public retinal image database in my research.

Last, but certainly not least, I would like to thank my family and friends for giving me all the support I needed.

TABLE OF CONTENTS

APPENDIX C: Example of a well segmented image

APPENDIX D: Example of a poorly segmented image

APPENDIX E: Example of a typical segmented image

APPENDIX F: Images of high Mahalanobis' distance

TABLE OF SYMBOLS

BG	Background
BristolDB	used database (provided by the University of Bristol and Bristol Eye Hospital)
DSC	Dice Similarity Coefficient
EM	Expectation Maximization (algorithm)
FG	Foreground
FJ	Figuero-Jain (algorithm)
GMALL	Gaussian Mixture models classifier classifying ALL pixels in given region
GMEXP	Gaussian Mixture models classifier classifying pixels by EXPanding representative points
GMM	Gaussian Mixture Model
IST	Interactive Segmentation Toolkit
JSC	Jaccard Similarity Coefficient / Jaccard index
LCT	Lightness Channel Thresholding (algorithm)
MICCAI	Medical Image Computing and Computer Assisted Intervention
NBALL	Naïve Bayes classifier classifying ALL pixels in given region
NBC	Naïve Bayes Classifier
NBEXP	Naïve Bayes classifier classifying pixels by EXPanding representative points
RAAD	Relative Absolute Area Difference
VISCERAL	Visual Concept Extraction Challenge in Radiology

1 INTRODUCTION

1.1 Background

Diabetic retinopathy is a very common symptom caused by diabetes and is the leading cause of blindness in the working population of western countries. The World Health organization expects the number of people suffering from diabetes to rise in the near future in both developed and underdeveloped countries [1]. Diabetic retinopathy is diagnosed based on eye fundus images (also called retinal images). If a retinal image contains certain types of lesions such as exudates, the patient is suffering from diabetic retinopathy. Since the current methods of detection and assessment of diabetic retinopathy are manual, expensive and require a trained ophthalmologist [2], an automated method would be a desired substituent.

Automatic detection of diabetic retinopathy started to gain attraction since the recognition of digital imaging as an accepted method to document the eye fundus [3]. Since then, numerous researches have been done on automatic detection of lesions in the eye fundus images. However, the so-called ground truth markings (i.e., exact markings in retinal images made by expert ophthalmologists) are required to develop accurate tools. Also, the superiority between the different methods of segmentation can not be truthfully addressed if they are not tested with the same data and measured with commonly used metrics [4].

Acquiring the ground truth currently requires the expert ophthalmologists to segment lesions manually. This is a time-consuming and expensive process, which could be easily speeded up, if there were proper tools. Therefore, there is a need for interactive medical tools that would help to segment lesions or structures accurately in retinal images.

Researchers in the Machine Vision and Pattern Recognition laboratory of Lappeenranta University of Technology have developed an image annotation tool for the ground truth acquisition process. A new version for the tool is currently under development. In the earlier version of the tool, the user could only mark a representative point of the lesion and the surroundings (background, the pigment epithelium of the retina) of the lesion. Various methods used in segmenting the medical images are being analysed as candidate implementations for the tool, including the statistical tools evaluated in this research.

1.2 Goals and limitations

In this research, gaussian mixture models and naïve Bayesian classifier was implemented to segment the retinal images in order to find the exudates within manual segmentations. The effectiveness of these statistical methods is studied and the results are evaluated by common metrics.

The purpose of the research is to study methods to be used in an interactive tool to help

ophthalmologists to mark lesions in images, and therefore, the results were evaluated with metrics appropriate for that purpose. The metrics used in the analysis of the methods that automatically detect diabetic retinopathy may vary from the ones used in this research.

Also, as some of the medical image segmentation tools such as the statistical classifiers used in this research concentrate solely on the colors of different objects, a variance analysis of the colors in pigment epithelium and exudates in retinal images was carried out and the results are presented in this thesis.

1.3 Structure

This report consist of three sections. In Section 2, there is a review on interactive medical tools. The section concentrates on finding out which algorithms have been used in those applications. The segmentation methods used in this research are also presented in Section 2, which includes also the methods of analyzing the variances and the metrics used for evaluating the results of segmentation.

The results for both the segmentation and color variance are presented in Section 3. The section contains the results plotted in visually demonstrative charts and a literary analysis of the results.

In the last section (Section 4), there is a summary of the research as well as some thoughts on the research. There is also a brief analysis on some questions arisen during the research in Section 4.

2 SEGMENTATION APPLICATIONS AND METRICS

2.1 Prior interactive segmentation applications and algorithms

There are several interactive segmentation applications when it comes to medical imaging. However, they typically use the same kind of methods to segment the images. They either use information about both the background and foreground, or just the foreground. The foreground signifies the object to be segmented and the background signifies the area surrounding that object.

In [5], there are several open-source programs listed as candidates to be implemented to VISCERAL (Visual Concept Extraction Challenge in Radiology) annotation tool. There are a number of evaluation criteria for the algorithms, but the main criterion was that every framework had to have a points-of-interest annotation tool for medical images and a semi-automatic segmentation tool, which should reduce the time of making the manual annotations of the 3D structures.

There are also a few interactive segmentation algorithms evaluated and presented in [6]. These algorithms are implemented in a segmentation application called IST (Interactive Segmentation Toolkit). IST is not actually a medical imaging software, even though it may also be used in medical imaging and it uses some of the same algorithms as in the applications mentioned in [5], as well as a couple of other algorithms.

The applications mentioned above are presented in Table 1 with the algorithms implemented in them.

Tool	**Algorithm(s)**
GeoS [7]	Generalized geodesic distance transforms (presented in [8]) and energy minimization
MITK [9]	Several thresholding methods (incl. Otsu's method), seeded region growing
3DSlicer [10]	Region growing, intensity based image segmentation
ITK-SNAP [11]	Snake evolution
ImageJ [12]	Robust Automatic Threshold Selection
MeVisLab [13]	Graph cuts
IST [6]	Seeded region growing, graph cuts, simple interactive object extraction, binary partition trees

Table 1: Segmentation applications. The name of the application is in the left column and the algorithms used in the application are presented in the right column.

2.2 Retinal image database

The retinal images and spatially accurate foreground data was provided by the University of Bristol and Bristol Eye Hospital. The database will be referred to as BristolDB. It consists of 107 retinal images and their ground truth masks, 90 of which have exudates. The resolution of the images is 540x536.

2.2.1 Generating inaccurate data

The segmentation algorithms need spatially inaccurate data as the input. Therefore, the accurate ground truth masks had to be made inaccurate. This was achieved by simply expanding each marking iteratively three times. In practice, this was carried out by expanding each boundary pixel of each marking from one to three (randomly selected) pixels. When this was done for each of the boundary pixels, the resulting markings were expanded again twice. This process can be seen iteration by iteration in Figure 1.

There was a total of 100409 ground truth pixels in data set. After the expansion, there were five times more pixels (a total of 628718 pixels) in the artificially generated inaccurate data.

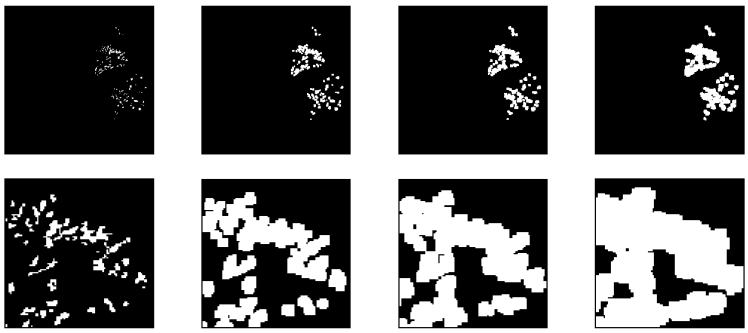


Figure 1: Production of artificially inaccurate masks. From left to right there are the ground truth markings, the results of the first, second and third iteration of expansion. On the top row, there are images of the whole mask and in the bottom are magnifications of the mask above.

2.2.2 Training data

The BristolDB includes the ground truth masks and the actual images, but no representative points. Therefore, also the training data had to be artificially generated since it is essential for the Bayesian segmentation algorithms.

Foreground trainining data consists of the representative points for the exudates. The data was generated by automatically selecting a representative point for each one of the exudates. In practice, the training data generating algorithm would take one region in the ground truth mask, take all the marked pixels of that region from the image, and take the pixel with the median distance to the origin in the RGB color space. This was done to all of the regions in all of the ground truth masks. Those median pixels were expanded to their 8-connectivity later on because the training data was not sufficient for all of the segmentation algorithms.

Background data for training was simply acquired by taking all the boundary pixels of all the solid areas in the artificially generated inaccurate (expanded) masks. This way there are notably more background pixels than there are exudate pixels, which simulates

the real situation in interactive segmenting tools.

It is worth mentioning that there would have been other ways to determine the representative point, such as taking the centroid of the region or the means of the RGB channel values to be used as representative pixel or a real pixel closest to the mean pixel. The problems with these methods are that in some cases the centroid of the region would not even represent the exudate, but the background and that the artificial mean pixel does not represent any real pixel. The problem with using a pixel closest to the mean or using the selected method of taking the pixel with the median distance is that there are many pixels with an equal distance to some point but they represent different colors. For example, three pixels with RGB values of (255,0,0), (0,255,0) and (0,0,255) are all of equal distance to origin, but represent very differently colored pixels (red, green and blue respectively).

2.3 Segmenting algorithms

2.3.1 Gaussian mixture models methods

The GMMBayes [14] was a project of the machine vision and pattern recognition laboratory of the Lappeenranta University of Technology. The main goal of the project was to study Bayesian classifier, Gaussian mixture probability density function models and maximum likelihood parameter estimation methods and to implement these methods to different classification tasks such as letter image recognition.

The outcome of the project was a new toolbox (GMMBayes Matlab Toolbox) for Matlab. The toolbox contains all the functions and methods used and developed in the GMMBayes project, as well as the documentation for the functions. In this research, the toolbox was implemented to classify the pixels to segment exudates in retinal images.

The toolbox has three different probability estimation algorithms, which are: basic Expectation Maximization (EM), Figuero-Jain (FJ) and greedy EM algorithms. The basic EM and FJ algorithms were used in this research. The main difference between these algorithms is that the FJ algorithm estimates the number of Gaussian components (Gaussian probability density functions) while the EM algorithm uses a fixed number of the components. Also, the maximum number of components must be defined as an input parameter for the FJ algorithm. There is a detailed description of these algorithms in [14].

2.3.2 Naïve Bayes Classifier

Naïve Bayes Classifier (NBC) [15] is a probabilistic classifier based on Bayes' theorem of conditional probabilities. NBC naïvely assumes that the features of a class are independent of each other. Hence the name, *naïve* Bayes classifier.

In practice, the NBC calculates the posterior probabilities for a sample belonging to each of the alternative classes based on its features. The sample is then classified to the class with the largest posterior probability. In the application of pixel classification, the

classes are foreground (FG) and background (BG), and the features are the color values. The posterior probabilities for given classes are calculated as follows:

$$p(FG) = \frac{P(FG)p(\overline{\mathbf{x}}|FG)}{P(FG)p(\overline{\mathbf{x}}|FG) + P(BG)p(\overline{\mathbf{x}}|BG)} \quad (1)$$

in which $p(FG)$ is the posterior probability for class FG (foreground), $p(\overline{\mathbf{x}}|FG)$ is the conditional probability of color channel x_i having the value as in the sample pixel if it was of class FG and $P(FG)$ is the prior probability for class FG.

2.3.3 Implementations of classifiers

For both the NBC and GMM classifiers two implementations were made. The first implementation uses the all the representative exudate pixels and background (the edge of expanded area) pixels in an image to teach the classifier. Then the classifier classifies all the pixels inside the expanded areas to either background or foreground (exudates). The first implemention is referred to as NBALL the second one is referred to as NBEXP. The classification was based on green and blue channels of the images and the red channel was left out because it brought no significant improvements.

The second implementation takes one expanded area, teaches the classifier with its representative points for background and foreground, takes the 8-connectivity of each representative areas' edge pixels and classifies those pixels. Then it takes the pixels in the 8-connectivity of the newly classified expand pixels and classifies those pixels and continues this iteration until no new foreground pixels are found. In other words, this implementation expands the representative exudate pixels as long as new exudate pixels are found. This implemention is described in algorithm 1. The first implemention is referred to as GMALL and the second one is referred to as GMEXP.

For the GMALL, the FJ algorithm (with a maximum of five components) was used, and for the GMEXP, the EM algorithm with one component was used because the representative points did not provide enough training data for FJ algorithm. Another reason was that using EM algorithm with more than one component did not improve the results compared to the one-component case, was much slower and tended to stop running occasionally.

Algorithm 1 The second method for segmentation

INPUT: *expanded_mask*, *image*;
OUTPUT: *segmented_mask*;
$rpps \leftarrow get\_representative\_points()$;
for each *expanded_area* **do**
  $resulted\_mask \leftarrow rpps\_in\_current\_expanded\_area$:
  while $new\_foreground\_pixels\_found$ **do**
    $neighbors \leftarrow get\_8\_nearest\_neighbors(resulted\_mask)$;
    $classified\_as\_exudates \leftarrow classify(neighbors)$;
    $resulted\_mask \leftarrow add\_pixels(classified\_as\_exudates, resulted\_mask)$;
  end while
  $segmented\_mask \leftarrow combine\_masks(segmented\_mask, resulted\_mask)$;
end for

2.4 Evaluation metrics

2.4.1 Dice's Similarity Coefficient

The Dice Similarity Coefficient (DSC) [16] is a way of measuring set agreement and is determined as follows:

$$D(A, B) = \frac{2|A \cap B|}{|A| + |B|}, \tag{2}$$

where A and B are the sets, $|A|$ denotes the pixels of set A and $D \in [0, 1]$.

The value of the DSC, as seen in the formula, varies between zero and one. The DSC value of zero means that there is no overlap between the sets, and the value one means there is a perfect overlap (i.e. the sets are exactly the same). In application of comparing segmentation algorithms' results to the ground truth, higher DSC indicates more truthful segmentation.

2.4.2 Jaccard index

The Jaccard index [17] (also known as Jaccard Similarity Coefficient) (JSC) is also a way of measuring set agreement and is determined as follows:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{3}$$

where A and B are the sets, $|A|$ denotes the pixels of set A and $J \in [0, 1]$.

The main difference between JSC and DSC is that while DSC is dependent of true positive markings and the areas of the segments being compared, JSC is dependent of true positive, false positive and false negative markings and it tends to be more critical than DSC (i.e., usually the JSC value is usually notably smaller than the DSC value for the same sets).

2.4.3 Relative absolute area difference

In MICCAI's (Medical Image Computing and Computer Assisted Intervention) Prostate MR Image Segmentation -challenge in 2012, one evaluation criteria of the competing algorithms was relative absolute volume difference (i.e., the percentage of the absolute difference between two volumes). The relative absolute volume difference is a neat way of evaluating how similar in size are the two regions being compared. This kind of a metric is a good indicator of how well region growing algorithms work. Thus, the relative absolute volume difference is modified to Relative Absolute *Area* Difference (RAAD) in this research and it will be used as an evaluation metric for the results. RAAD is determined as follows:

$$RAAD(A, B) = \left| \frac{|A|}{|B|} - 1 \right| \times 100\%, \tag{4}$$

where $|X|$ denotes the area of set X, B is the ground truth and RAAD $\in [0\%, \infty[$.

If the value of RAAD is 0 %, it means the two segments are of exactly the same size. If the value of RAAD is larger than 0 %, the areas of two segments differ. However, if the RAAD is 0 % it does not necessarily mean that the two segments represent the same exact region as shown in figure 2. Therefore, RAAD can not be used as the only evaluation metric.

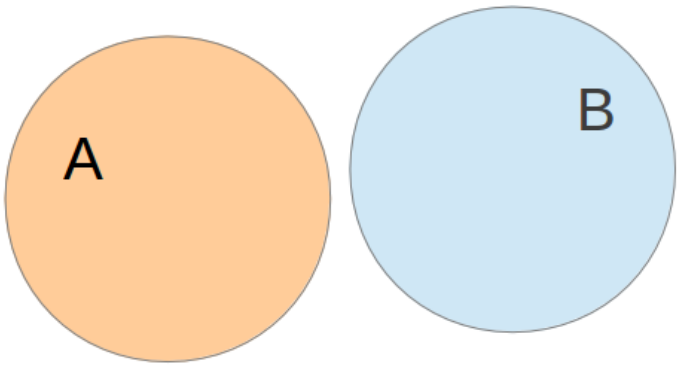


Figure 2: In this particular case, RAAD is 0 %, even though the segment A does not represent the region B at all (DSC = 0 and JSC = 0).

2.5 Usefulness of statistical methods

One of the main goals of the research was to find out, how effective are the statistical methods in exudate segmentation. The exudates are the most distinguishable kind of lesions in retinal images because they are light in color and provide high contrast to the background. Therefore, if statistical methods worked poorly in segmenting exudates, they would probably work even worse in segmenting lesions that are not as distinguishable (e.g. haemorrhages).

The resulting segments of these algorithms were evaluated with given metrics and the results are then compared with each other to find out the superiority between the algorithms. The results are also compared with the results of a simple Lightness Channel Thresholding (LCT) algorithm, which was found to be superior to K means clustering [18] and Otsu's method thresholding [19] algorithm with the same dataset. A brief description of the LCT algorithm and the comparison to K means clustering and Otsu's method thresholding can be seen in the Appendix A.

2.6 Color variance analysis

To study the variance of the color channel values, the variance was calculated for each channel and image individually. Also, the total variations in the background and the foreground were calculated. The results were plotted into demonstrating graphs, which are presented in Section 3.2 with a literary analysis of the results.

2.6.1 Coefficient of variation

The coefficient of variation [20] is a unitless metric used in statistics. It is a normalized measure of data dispersion and is used in this research to compare variation: the smaller the coefficient, the smaller the dispersion in the data. The coefficient of variation is determined as follows

$$c_v = \frac{\sigma}{\mu}, \tag{5}$$

where σ is the standard deviation and μ is the mean.

2.6.2 Mahalanobis' distance

Mahalanobis' Distance (MD) is also a unitless metric of calculating data dispersion. MD is used in determing how similar one sample set is to a known one. It takes into account that the variances of variables are different from each other and it also accounts for the covariance between those variables. A highly demonstrative explanation of MD can be seen in [21].

3 RESULTS

3.1 Segmentation

The whole Bristol database was used in the segmentation areas for the statistical segmentation algorithms. The results of the statistical algorithms were compared to the results of the LCT algorithm with the same dataset as stated in Section 2.5.

Each one of the statistical algorithms clearly outperformed the LCT algorithm on segmenting the exudates in the retinal images. The results can be seen in Figure 3. While the GMALL and NBALL algorithms seem to have only a bit better results compared to the LCT algorithm and measured with DSC and JSC, the RAAD of these two algorithms was clearly smaller than that of the LCT algorithm.

As expected, the algorithms expanding the representative points (GMEXP, NBEXP) outperformed those algorithms that classified all the pixels inside the expanded masks (GMALL, NBALL, LCT). By expanding the representative points, there will be no isolated pixels falsely classified as foreground in the resulting mask. What was surprizing, is that the NBEXP algorithm outperformed the GMEXP algorithm, even though there were couple of images the NBEXP algorithm could not segment at all. For those situations, the algorithms would score 0 for DSC and JSC and 100 for RAAD.

The gained results were tested with paired-sample T test to see if the differences were statistically significant. The difference between and statistical algorithm and LCT algorithm was significant ($p < .001$). Also, the differences between the implementations (NBEXP vs NBALL, GMALL vs GMEXP) were statistically significant ($p < .01$). Surprisingly, there was no statistically significant difference between GMMBayes and NBC algorithms.

It is worth mentioning that the images without lesions were outcluded from this evaluation. The plots considering all of the images are presented in Appendix B for better comparison to the plots of algorithms in Appendix A.

In Appendices C, D and E, there are parts of the best, the worst and most typical of the resulted resegementation masks shown respectively. Of course, the masks without lesions to begin with have not been taken into consideration in this comparison. As it can be seen in these images, the NBALL and NBEXP resulted very similar masks. The clear differences are that the NBALL algorithm tends to result a few more areas (in number) marked as exudates and that the NBEXP worked better with more isolated areas (see the most left exudate in Appendix E and compare GMALL and NBEXP to each other).

There were more differences between the GMALL and GMEXP algorithms than between the NB algorithms. This may be due to the fact that the two GM algorithms use different algorithms of calculations. The GMALL uses Figuero-Jain algorithm to evaluate the right number of components in the mixture models while the GMEXP uses the EM algorithm for which the number of components is fixed.

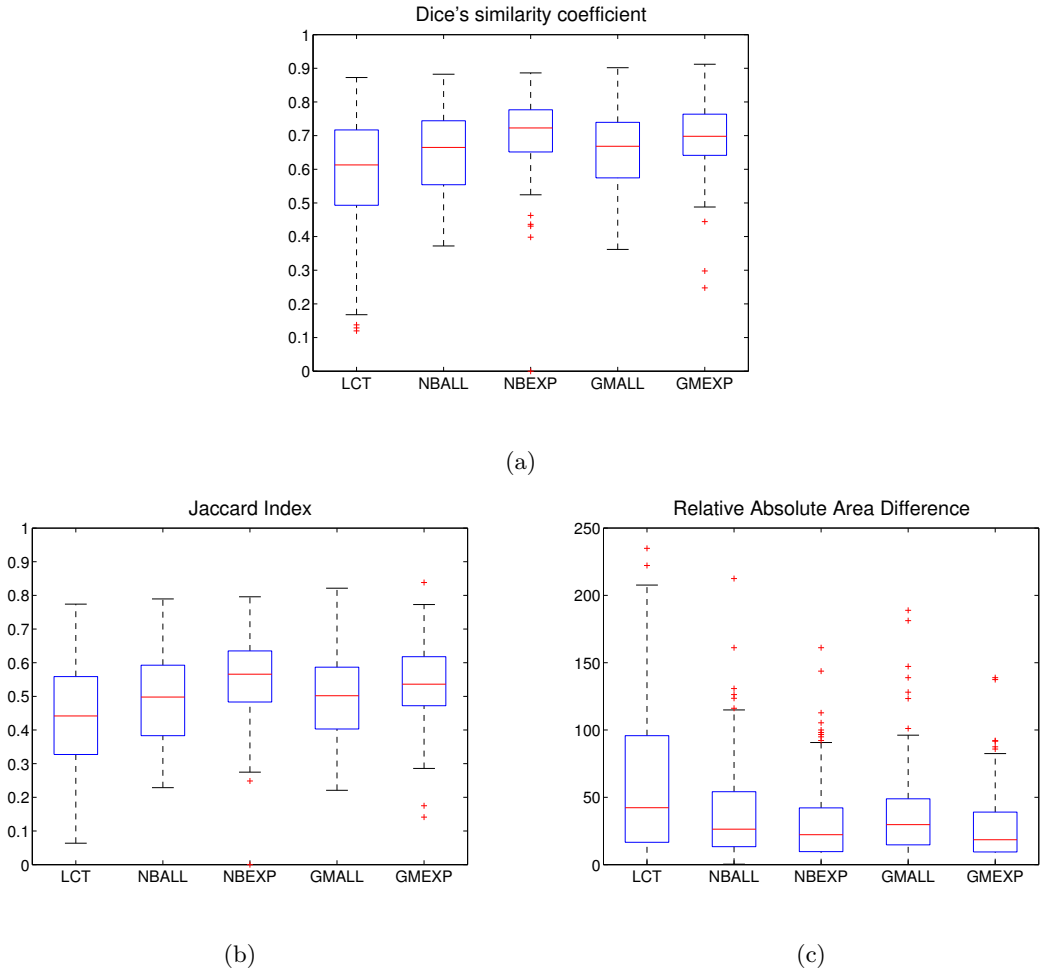


Figure 3: Results of evaluation. a) Dice Similarity Coefficient, b) Jaccard Index, c) Relative Absolute Area Difference. Note that in each of the figures the red line is the median, the edges of the boxes are 25th and 75th percentiles and the whiskers extends to the most extreme values not considered as outliers. Outliers are plotted individually (crosses).

3.1.1 Addition of red channels

As mentioned in Section 2.3.3, the red channel was left out because it did not improve the algorithms significantly. To be assured that the red channel did not improve the algorithms, the NBEXP algorithm was executed with all channels and the results were evaluated with the same metrics. The median DSC improved only by 0.0062. In addition, 45 of the 107 images were little more accurately segmented with NBEXP using all three channels and 32 were little more accurately segmented with NBEXP using only green and blue channels (i.e., for 30 of the 107 images the addition of red channel brought no improvement). The DSC results of the test can be seen in Figure 4.

The results were tested using paired-sample T test to verify that the improvement is not statistically significant. Null hypothesis was that the results are from the same distribution with equal means and with equal but unkown variances, and the alternative

hypothesis was that they are from distributions with unequal means. The null hypothesis was failed to reject with significance level of 0.05. The paired-sample T test was also done for both the JSC and RAAD. The test yielded similar results. To be exact, the p-value for all of the tests was $p > 0.50$. So, the addition of the red channel had no significant effect on the results, but it made the algorithm slightly slower since the classifiers had 50% more data to deal with.

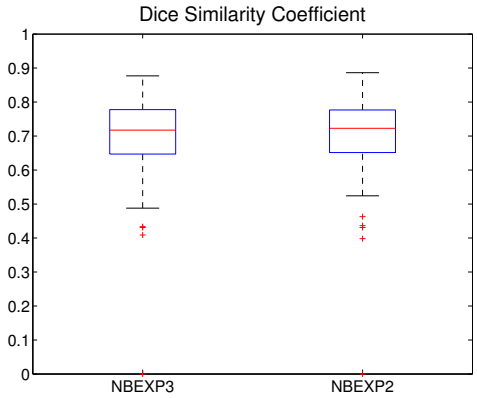


Figure 4: Results of adding the red channel. The 3-channel NBEXP (on the left) uses all three channels and the 2-channel NBEXP (on the right) uses only green and blue channels of the image. Note the similarity of the results. In the figure, the red line is the median, the edges of the boxes are 25th and 75th percentiles and the whiskers extends to the most extreme values not considered as outliers. Outliers are plotted individually (crosses).

3.1.2 Weigths and Components of the FJ algorithm

Even though the FJ tended to use more than just one or two components for the boundary pixels while building the classifier, mostly just one or two of the components had a significant weight associated with it. In Figure 5, there is an illustration of the weights of the components for the image. Note how low is the weight of the 3rd, 4th and 5th component of the classifier. Because of the fact that almost with all of the images, the FJ algorithm tended to use two or more components, the segmentation of the images using GMEXP with EM algorithm using two fixed components instead of just one was carried out.

The results of GMEXP using EM with two components were similar to the results with EM using only one component. The difference between the results was similar to the case of adding the red channel to the classifying. According to the paired-sample T test, the slight difference was not statistically significant. Also, using two components was more time-consuming.

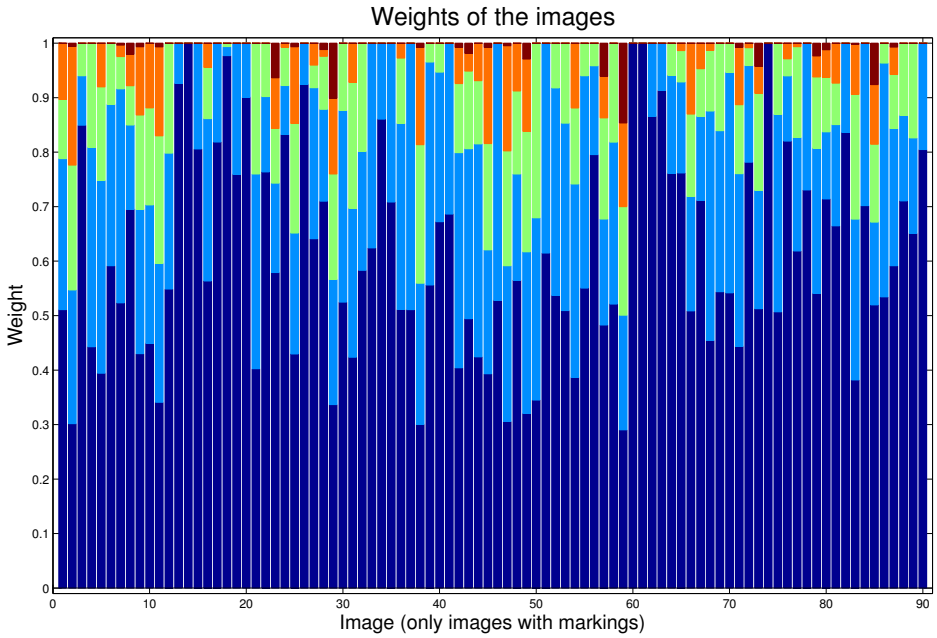


Figure 5: Stacked weights of the components per image. In this figure it is shown how many of the 5 components (number of bars) the classifier used on each image individually. The weight of the component can also be seen (size of the bar of that color). Note how most of the images require basically two components.

3.2 Color variance analysis

This part of the report concentrates on the color variation of the retinal images. If the distribution of the ground truth pixels' color overlaps too much that of the background (pigment epithelium) pixels, global statistical classification of pixels in terms of segmenting would be problematic. Correspondingly, if there is only a slight overlap, using the stastical classification for segmentation should work very well. This applies to most of the segmentation algorithms that uses only the color information.

3.2.1 Using all pixels

There were several evaluations done for the variance analysis. In the first one, all of the ground truth pixels, segmented masks' pixels and expanded masks' (both with and without ground truth pixels in them) pixels were plotted into a boxplot chart to get a larger view of the color variances. The results were unexpected. Intuitively, the variation of the expanded masks' pixels should be huge because of them being both exudate and background pixels and because the color of the background of the individuals *seemed* to vary a lot in the dataset.

In Figure 6, there are 4 charts having three boxplots having 3 boxes each. The boxes visualize the variation of the pixels for all of the three color channels: red, green and blue. There are charts for the expanded masks', ground truth masks' and segmented masks'

pixels. The fourth chart is a visualization of the the expanded masks' pixels without the ground truth pixels.

Note how in Figures 6a and 6b there is a slight visual difference between the variation. That may be because of the fact that the expanded masks had over five times more pixels than the ground truth masks, as stated in Section 2.2.1. Since there are multiple times more background pixels than the ground truth pixels in expanded masks, the ground truth pixels' values weights less than those of background pixels'. Anyhow, the most significant finding is that the variation between patients' pigment epithelium (i.e., the background) is relatively small to that of the exudate (ground truth) pixels.

The difference of the variation between the ground truth pixels and the segmented masks' pixels (Figures 6c and 6d respectively) seems trivial but the two-sample T test reveals that the difference is significant ($p < 0.001$) for all of the three channels. Therefore, the variation of the segmented masks' pixels is smaller than that of the ground truth masks' pixels. This finding was somewhat expected since the classifiers were trained with the representative pixels that were pixels of median distance to the (0,0,0) in the RGB space. These findings are also supported by the coefficients of variations which are represented in Table 2.

Mask (pixels') type / Color channel	**Red**	**Green**	**Blue**
Expanded mask with ground truth	0.2204	0.2951	0.3982
Expanded mask without ground truth	0.2238	<u>0.2495</u>	<u>0.3744</u>
Ground truth masks	0.1848	0.2984	0.3910
Segmented masks	<u>0.1818</u>	0.2905	0.3883

Table 2: Coefficients of variation of pixels. The smallest coefficient for each channel is <u>underlined</u>.

What is more, by looking at the red channels' boxes of all of the images, one may notice how in each one of the images, the whiskers extend to the maximum value of 255. From this it may be induced that there is some information lost in the red channels due to the fact that the pixels ca not exceed that value. In photography, images having similar white pixels (i.e., pixels having value of (R,G,B)=(255,255,255) that should represent something else than plain white) are called overexposured pixels. When this overexposure happens to just one channel, the channel is called a clipped channel. This may occur if the parameters of the camera are not properly set prior to the photographing.

In the Bristol database, the red channels are clearly clipped in many of the images. The images have a mean of 1,0% red channel pixels with value 255 in them. For some images, this number was as big as 4,4%. If we calculated the percentage of pixels having red channel value over 250, (areas having these values are not useful since many of the

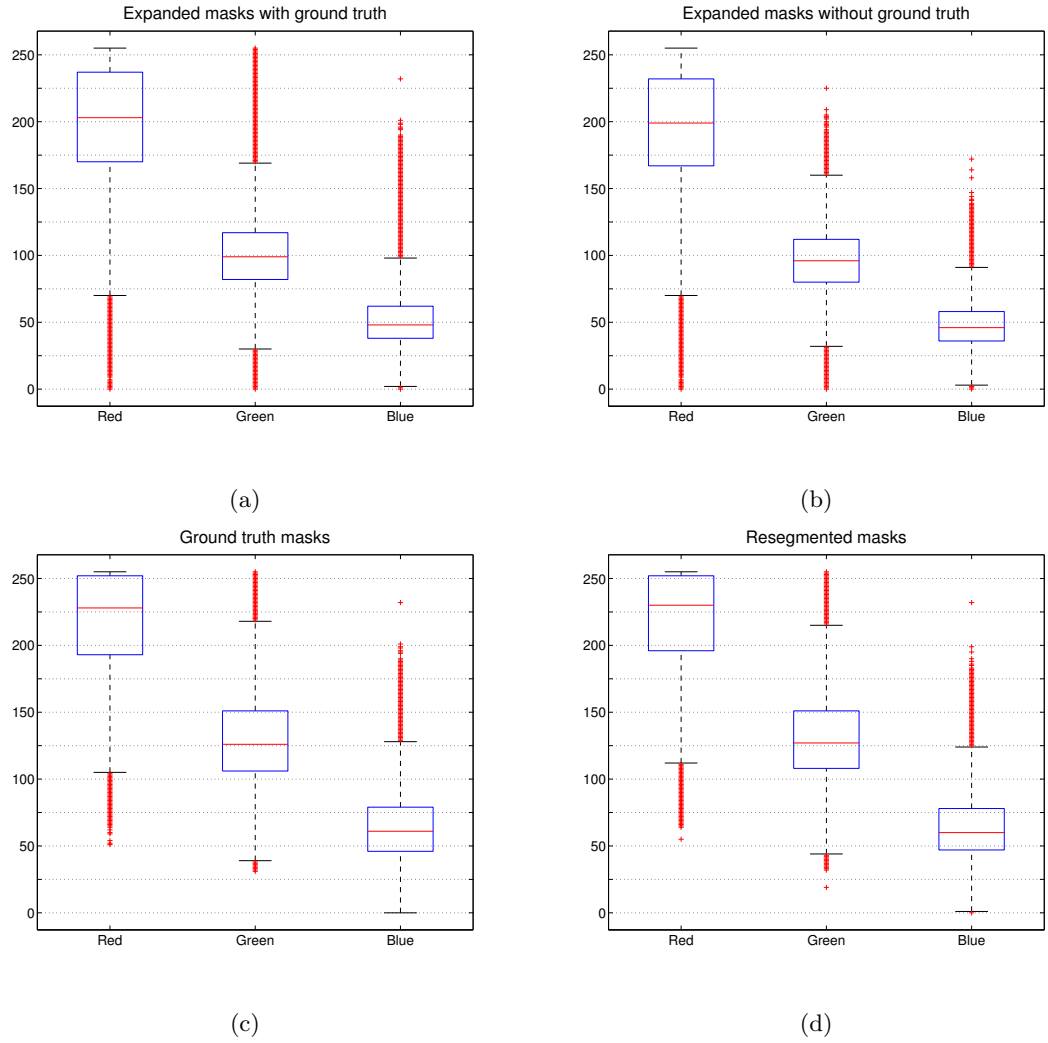


Figure 6: Variation of pixels by channels. a) Expanded masks with ground truth, b) Expanded masks without ground truth, c) Ground truth masks, d) Segmented masks. The y axis number is the value of the pixel for the channel (max: 255). Note that in each of the figures the red line is the median, the edges of the boxes are 25th and 75th percentiles and the whiskers extends to the most extreme values not considered as outliers. Outliers are plotted individually (crosses).

exudates have this red channel value) the mean would be 10,9% and the worst case would be 54,9%. The black background areas around the round area representing retina in the images was excluded from the calculations since they provide no information about the patients' retina.

There is a part of an image representing bad red channel clipping in Figure 7. In Figure 7d you can see a large area of clipped pixels (pixels with red channel values over 250) represented by blue pixels. These areas are called 'blown out' areas. Note how in the red channel (Figure 7b), there are no visible highlights in that area. If the exudates were in that area (as they are in some images) there would be no way to detect them directly from the red channel of the image (i.e., there is no profit in using the red channel for segmenting these images).

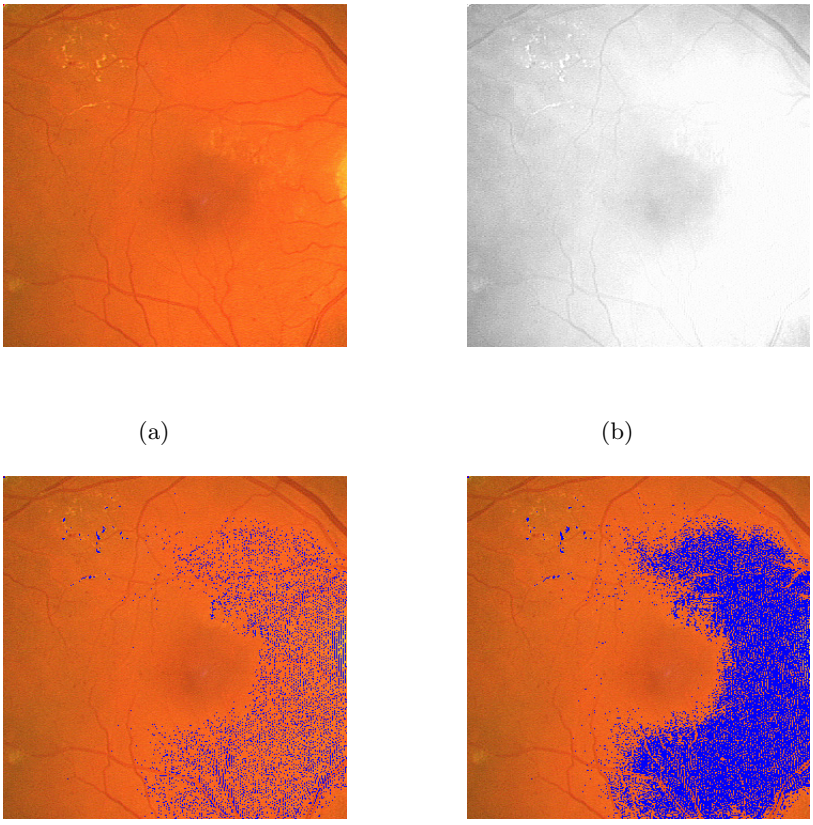


Figure 7: Clipping of the red channel of an image. a) Part of a clipped image, b) Red channel of the partial image, c) Clipped pixels (having the red channel value of 255, marked in blue), d) pixels exceeding red channel value of 250 (may be considered as clipped).

3.2.2 Individual images

To evaluate pixels' variation image by image, the variations for the pixels marked in the masks (same mask types as in Section 3.2.1) were calculated for each image individually. At this point, the red channels were excluded from the analysis. The results can be seen in Figure 8. In the chart are shown both green channels' variation and blue channels' variation as stacked bars. The x axes represent the number of an image and the y axes represent variance. The axes are scaled similarly in each of the images for better comparison.

The findings were similar to the findings of all of the pixels. Almost for every segmented mask, the variation was smaller than the variation of the ground truth mask of the image, and almost for every expanded mask, the variation was bigger than that of the expanded masks without ground truth pixels to them. Also, the expanded masks' variation tends to be smaller than the segmented masks variation - a finding already discovered in the evaluation of all of the pixels.

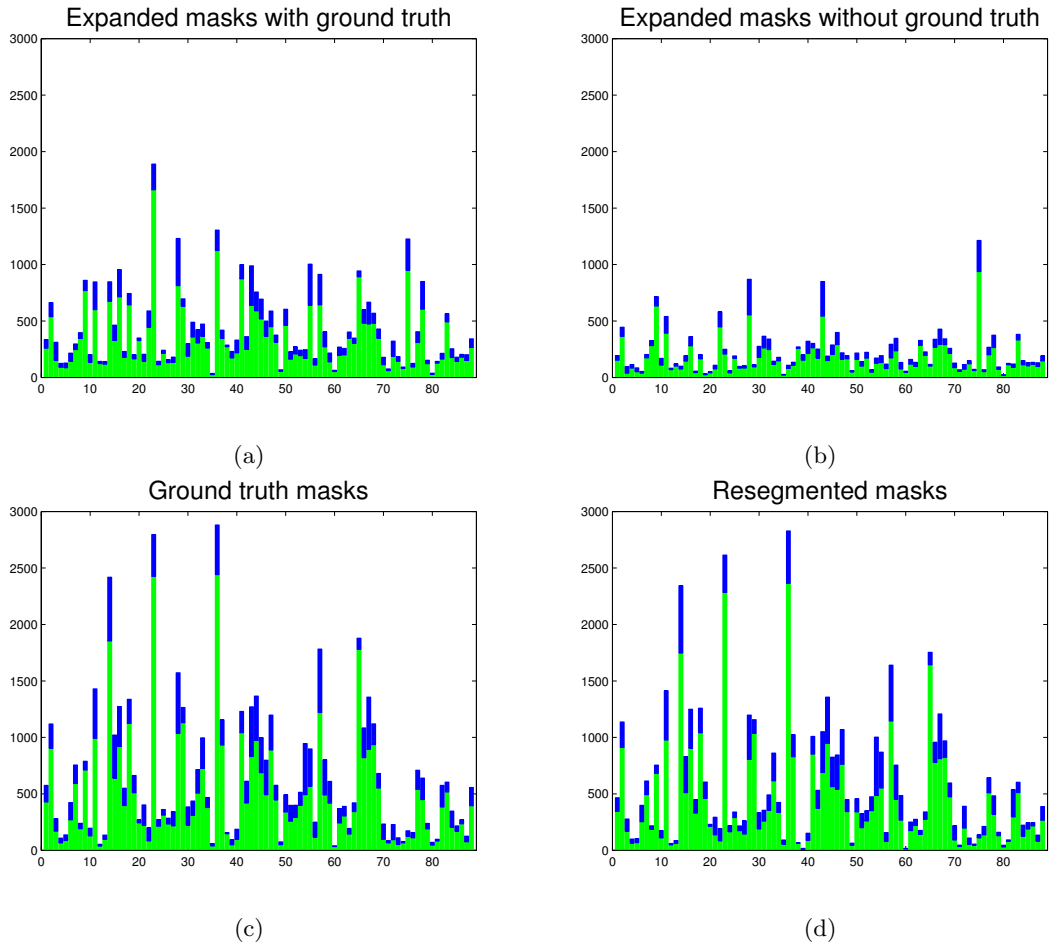


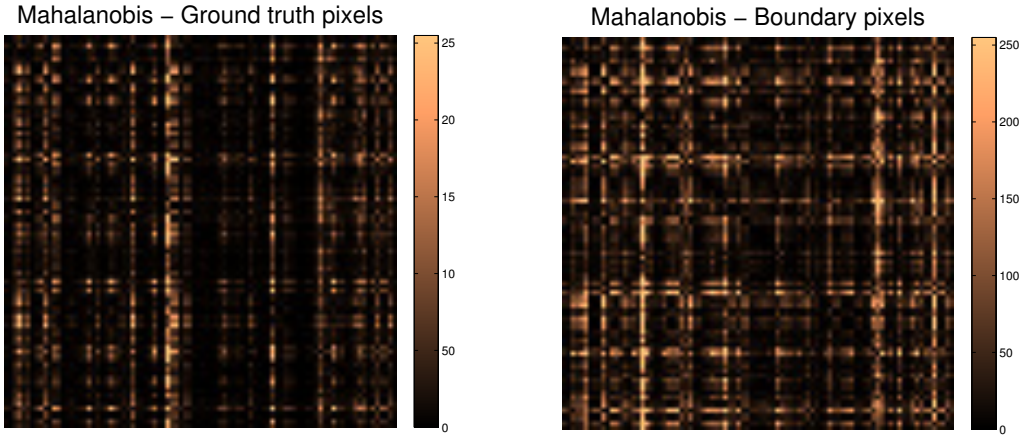
Figure 8: Variation of pixels by image. a) Expanded masks with ground truth, b) Expanded masks without ground truth, c) Ground truth masks, d) Segmented masks. In this image, the variations of green and blue channels are shown as stacked bars. The x axis is the image number and the y axis is the variance.

3.2.3 Mahalanobis' distance

In Figure 9, there are two charts. The first one (9a) represents the median MDs of the exudate (marked in ground truth masks) pixels. In the chart each pixel represents the median distance from one images' exudate pixels to another ones' ground truth pixels. Each row of pixels represents the median MD of pixels of an image to pixels of another image represented by each column. The brighter the pixel is, the higher the MD is. In the second image (Figure 9b), the same is done for the boundary (background) pixels of the expanded masks.

By looking the scales of the two charts in Figure 9, it is clear that the median distances for the background pixels were notably higher than those of the exudate pixels. This contradicts with the findings made in sections 3.2.1 and 3.2.2 of variances of background being smaller than the variances of ground truth pixels. However, this supports the assumption that the variances of the background pixels were smaller due to the large number of the pixels and also supports the intuitive presumption of having a high variance in individuals' pigment epitheliums.

For some images, the MD was particularly high with all of the images (bright rows in Figure 9, some of these images are presented in Appendix F.



(a) (b)

Figure 9: Mahalanobis distances. a) median MDs between images' Ground truth pixels, b) median MDs between images' boundary pixels. Each row and each column represents an individual image and every pixel represents the MD.

4 DISCUSSION

In this research, the main goal was to study how effectively statistical methods do segment the exudates in the human retina. Both of the used methods, Naïve Bayes and Gaussian mixture model classifiers, were able to outperform the benchmark algorithm (Lightness channel thresholding) measured with all given metrics (Dice's similarity coefficient, Jaccard index, relative absolute area difference). Therefore, they also outperformed K means clustering and Otsu's method thresholding with the same retinal image database.

Both the NBC and GMM classifier had two implementations to evaluate. Intuitively, the more sophisticated GMM classifier should have outperformed the simple NBC, but based on the experiments, it was the other way around. However, this difference was statistically not significant. Out of two different statistical implementations, the algorithms expanding the representative points was found superior to the algorithms going through the whole region. Still, the question remains, is the implementation effectice enough to be used in a semi-automatic medical tool.

Another goal was to find out how effectively the exudates can be segmented using only the color information of the pixels. To answer this question, a color variance analysis of retinal images was carried out. In the analysis, there were contradicting findings discovered. The variance of background (pigment epithelium) pixels of the patients was notably smaller than that of the exudate pixels. Also the coefficients of variation supported this finding. But when the dispersion of the pigment epithelium pixels between individuals was measured with the Mahalanobis' distance, it showed remarkably higher dispersion than with the exudate pixels, which supports the fact that the pigment epitheliums of individuals varied considerably in the dataset. This may be explicable by the proportion of the background pixels, and therefore, by the amount of the pixels considered as outliers.

Despite of the contradicting results in the variance analysis, it is clear that there is a lot of variation in the color of individuals' pigment epithelium and exudates. Therefore, using color normalization or image processing methods to reduce the variances would be suitable for methods based solely on color information.

Based on the acquired results of statistical segmenting and color variance studies, the statistical tools may be used in segmentation of the exudates because of the high contrast between them and the pigment epithelium. However, because of the high amount of variance, it is questionable whether or not the statistical methods can be used to segment other types of lesions (e.g., haemorrhages) in retinal images.

REFERENCES

[1] Oliver Faust, Rajendra Acharya, EYK Ng, Kwan-Hoong Ng, and Jasjit S Suri. Algorithms for the Automated Detection of Diabetic Retinopathy Using Digital Fundus Images: a Review. *Journal of medical systems*, 36(1):145–157, 2012.

[2] Akara Sopharak, Bunyarit Uyyanonvara, and Sarah Barman. Automatic Exudate Detection from Non-dilated Diabetic Retinopathy Retinal Images Using Fuzzy C-means Clustering. *Sensors*, 9(3):2148–2161, 2009.

[3] Tomi Kauppi. *Eye Fundus Image Analysis for Automatic Detection of Diabetic Retinopathy*. Doctoral thesis, Acta Universitatis Lappeenrantaensis. Lappeenranta University of Technology, 2010.

[4] Tomi Kauppi, Valentina Kalesnykiene, Joni-Kristian Kamarainen, Lasse Lensu, Iiris Sorri, Asta Raninen, Raija Voutilainen, Hannu Uusitalo, Heikki Kälviäinen, and Juhani Pietilä. The DiaretDB1 Diabetic Retinopathy Database and Evaluation Protocol. In *BMVC*, pages 1–10, 2007.

[5] Oscar Jiménez and Henning Müller. Prototype of 3D Annotation Software Interface. Technical report, Visceral, 2013. `http://www.visceral.eu/`.

[6] Kevin McGuinness and Noel E O'Connor. A Comparative Evaluation of Interactive Segmentation Algorithms. *Pattern Recognition*, 43(2):434–444, 2010.

[7] Antonio Criminisi, Toby Sharp, and Andrew Blake. GeoS: Geodesic Image Segmentation. In *Proc. European Conference on Computer Vision (ECCV)*, volume 5302 of *Lecture Notes in Computer Science*, pages 99–112. Springer, 2008.

[8] Pekka J. Toivanen. New Geodesic Distance Transforms for Gray-scale Images. *Pattern Recogn. Lett.*, 17(5):437–450, May 1996.

[9] Ivo Wolf, Marcus Vetter, Ingmar Wegner, Marco Nolden, Thomas Bottger, Mark Hastenteufel, Max Schobinger, Tobias Kunert, and Hans-Peter Meinzer. The Medical Imaging Interaction Toolkit (MITK): a Toolkit Facilitating the Creation of Interactive Software by Extending VTK and ITK. In *Medical Imaging 2004*, pages 16–27. International Society for Optics and Photonics, 2004.

[10] Steve Pieper, Michael Halle, and Ron Kikinis. 3D Slicer. In *Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on*, pages 632–635. IEEE, 2004.

[11] Paul A. Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C. Gee, and Guido Gerig. User-Guided 3D Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability. *Neuroimage*, 31(3):1116–1128, 2006.

[12] Michael D Abràmoff, Paulo J Magalhães, and Sunanda J Ram. Image Processing With ImageJ. *Biophotonics international*, 11(7):36–42, 2004.

[13] F. Ritter, T. Boskamp, A. Homeyer, H. Laue, M. Schwier, F. Link, and H. O Peitgen. Medical Image Analysis. *Pulse, IEEE*, 2(6):60–70, 2011.

[14] Pekka Paalanen, Joni-Kristian Kamarainen, Jarmo Ilonen, and Heikki Kälviäinen. Feature Representation and Discrimination Based on Gaussian Mixture Model Probability Densities Practices and Algorithms. *Pattern Recognition*, 39(7):1346–1358, 2006.

[15] Kevin P Murphy. Naive Bayes Classifiers. *University of British Columbia*, 2006.

[16] Lee R Dice. Measures of The Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302, 1945.

[17] Paul Jaccard. Étude Comparative de la Distribution Florale dans une Portion des Alpes et du Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[18] Tapas Kanungo, D.M. Mount, N.S. Netanyahu, C.D. Piatko, R. Silverman, and A.Y. Wu. An Efficient K-means Clustering Algorithm: Analysis and Implementation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):881–892, Jul 2002.

[19] Nobuyuki Otsu. A Threshold Selection Method from Gray-Level Histograms. *Systems, Man and Cybernetics, IEEE Transactions on*, 9(1):62–66, Jan 1979.

[20] X. He and S.O. Oyadiji. Application of Coefficient of Variation in Reliability-Based Mechanical Design and Manufacture. *Journal of Materials Processing Technology*, 119(1–3):374 – 378, 2001.

[21] R. De Maesschalck, D. Jouan-Rimbaud, and D.L. Massart. The Mahalanobis Distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1 – 18, 2000.

APPENDIX A: Lightness Channel Thresholding

LCT algorithm is a simple thresholding algorithm, which takes RGB image as an input, converts it to an Lab image and applies a global threshold to the Lightness channel to determine which pixels are foreground and which pixels are background. The below results are acquired testing the K means clustering (K), Otsu's method thresholding (A), LCT (L) and two different singular value decomposition (S1, S2) based algorithms with BristolDB. In the table, the best value for each category is in **bold text**.

		Methods		
		LCT	Otsu's method	K means clustering
DSC	Median	**0.69**	0.65	0.63
	Mean	**0.68**	0.65	0.65
JSC	Median	**0.52**	0.48	0.46
	Mean	**0.55**	0.52	0.51
RAAD	Median	**26.34**	29.43	31.29
	Mean	73.41	77.09	**65.03**

Table 3: Comparison of methods. The best result in each category is underlined.

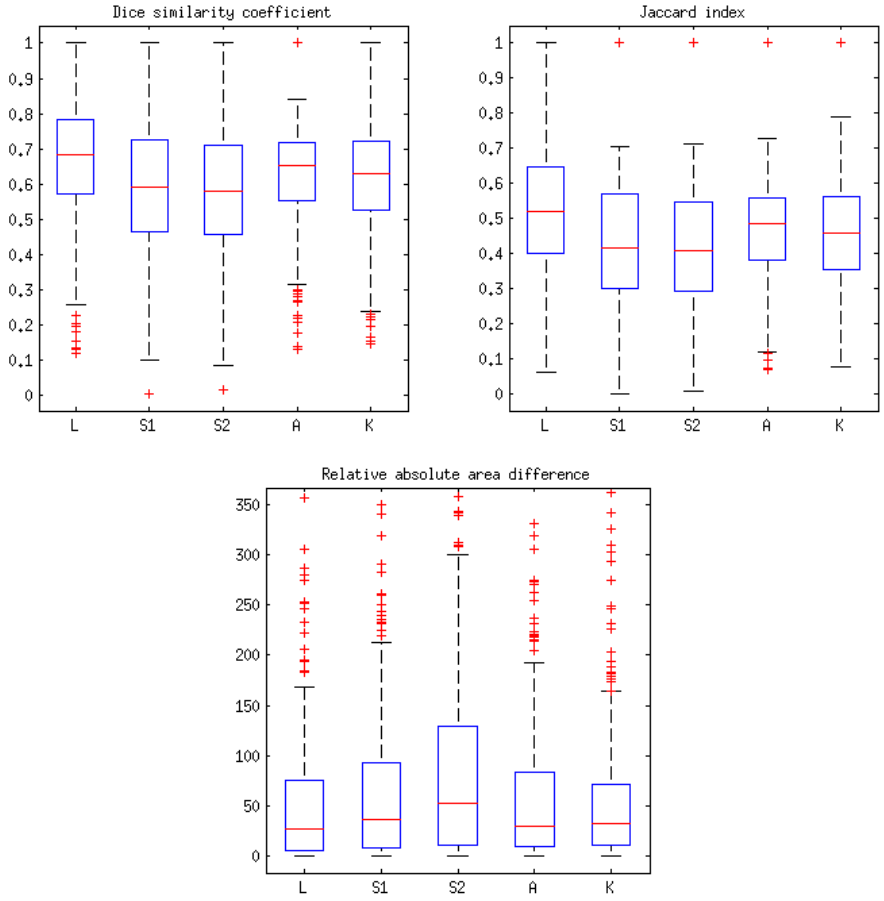
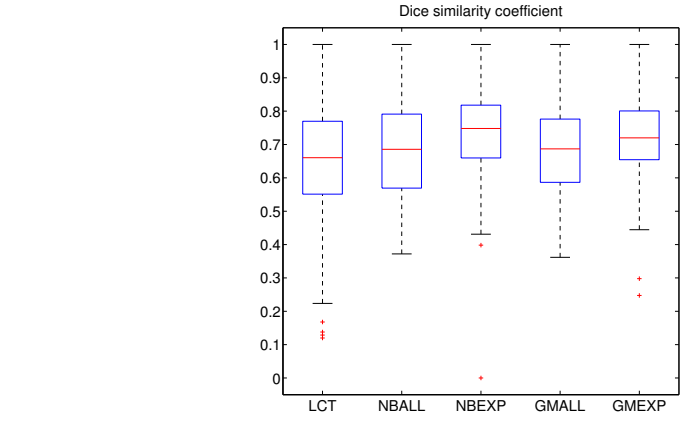


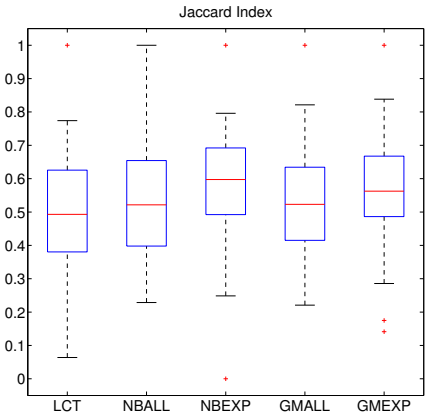
Figure 10: Comparison of methods. Note that in these results, the images with no lesions have not been outcluded from evaluation. These images shows as perfect results in the plots.

APPENDIX B: Plots with all of the images

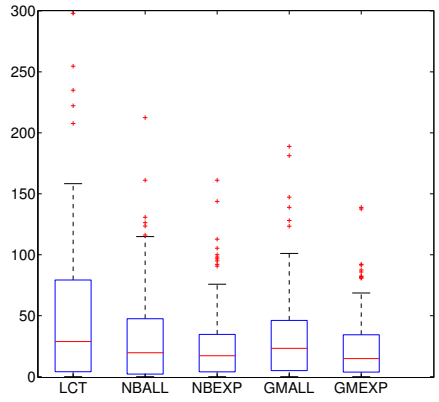
In these plots, all of the images (including those which had no lesions) are included in the evaluation. If the image had no lesions, it got a perfect result (full '1' measured with JSC and DSC, and a '0' measured in RAAD).



(a)



(b)



(c)

Figure 11: Results of evaluation. a) Dice Similarity Coefficient, b) Jaccard Index, c) Relative Absolute Area Difference. Note that in each of the figures the red line is the median, the edges of the boxes are 25th and 75th percentiles and the whiskers extends to the most extreme values not considered as outliers. Outliers are plotted individually (crosses).

APPENDIX C: Example of a well segmented image

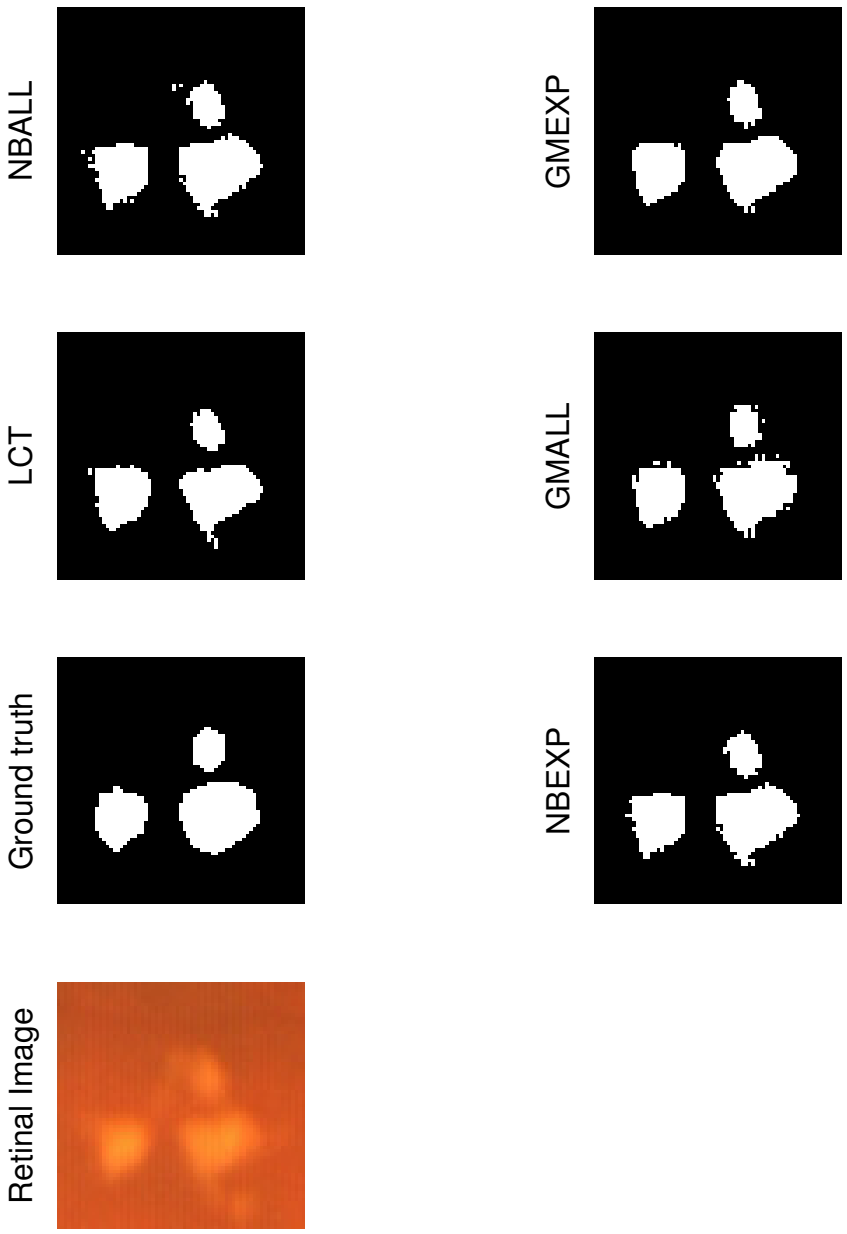


Figure 12: Example of a well segmented image

APPENDIX D: Example of a poorly segmented image

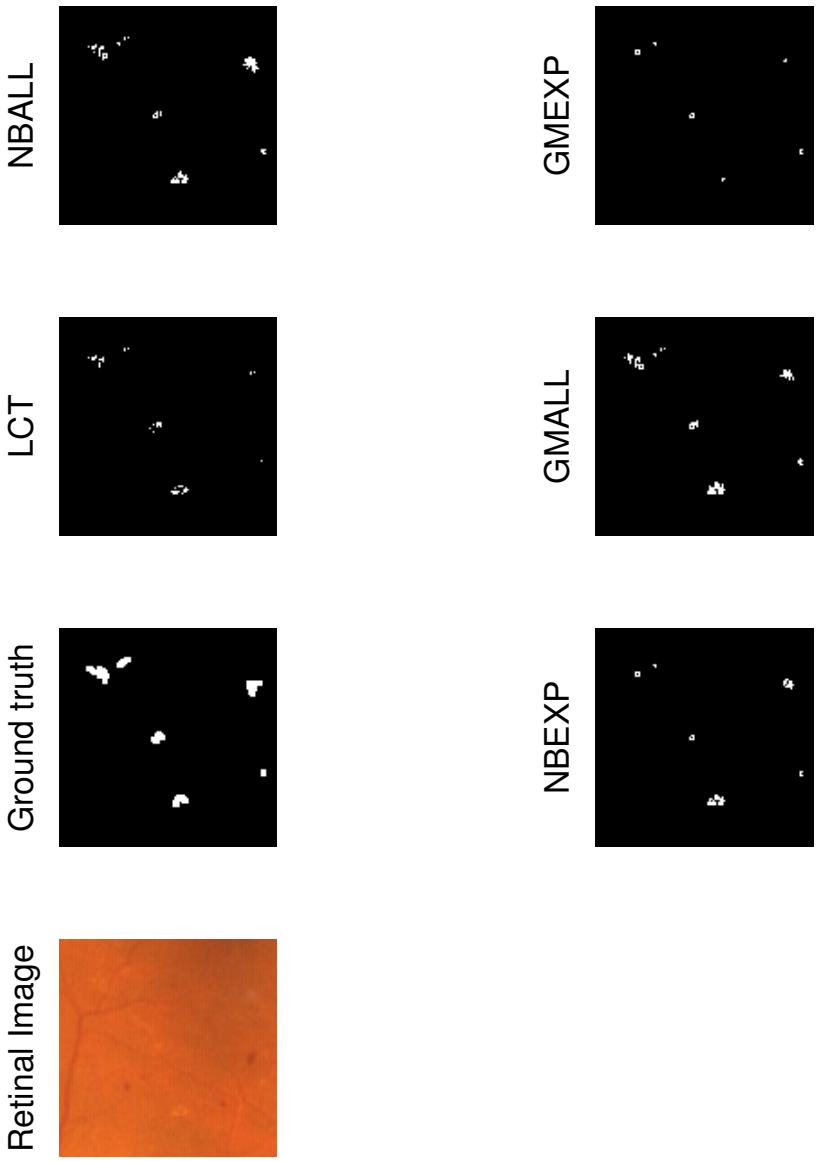


Figure 13: Example of a poorly segmented image

APPENDIX E: Example of a typical segmented image

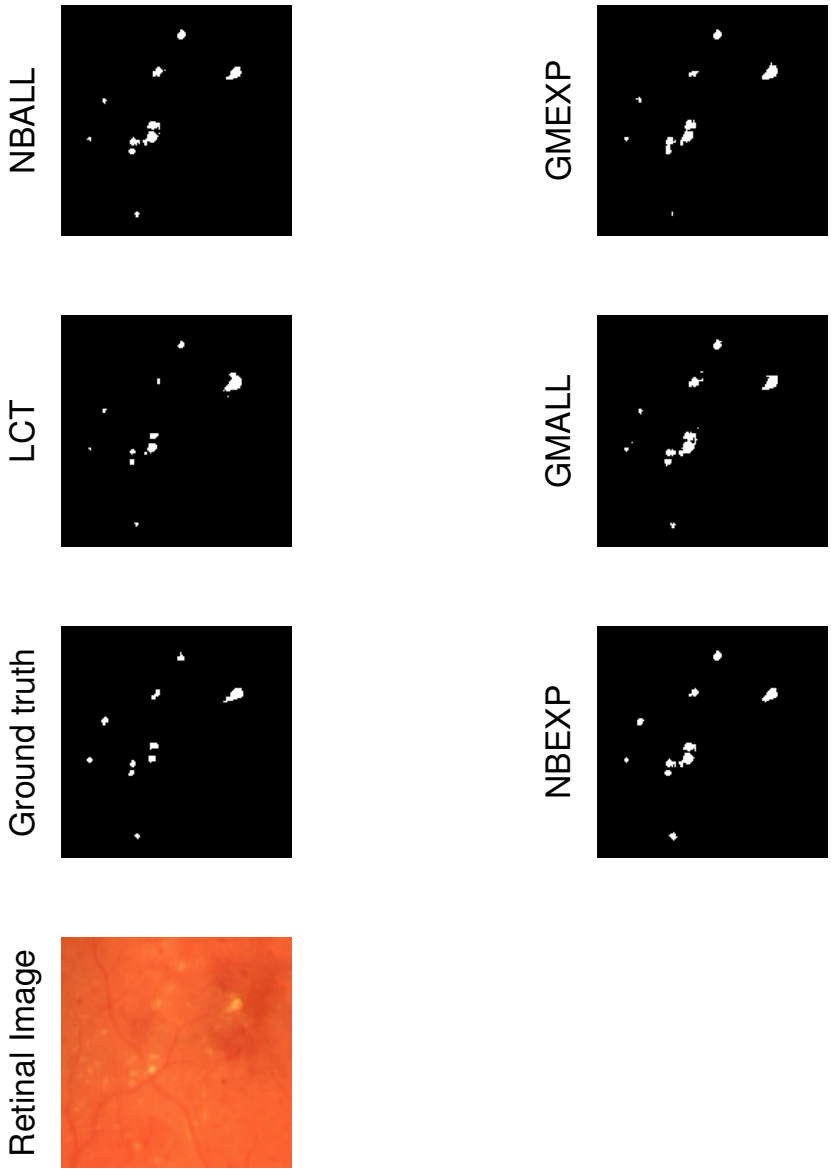
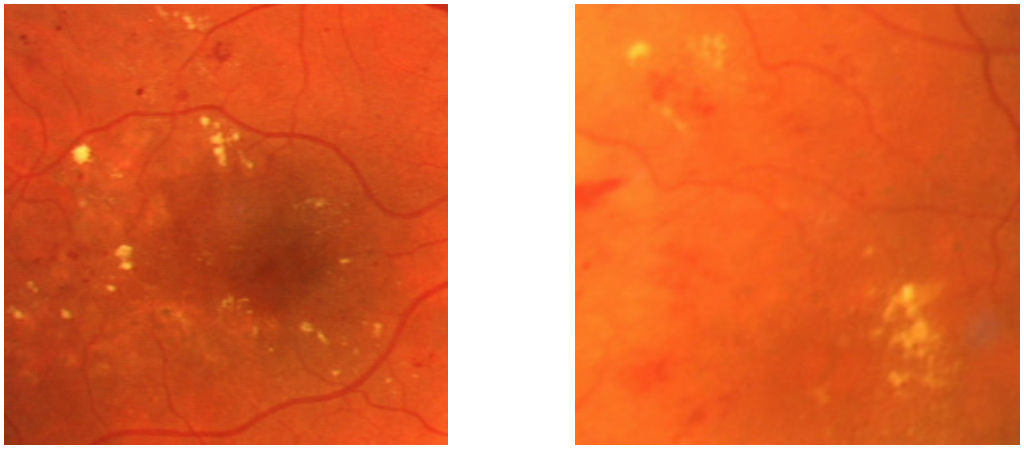


Figure 14: Example of a typical segmented image

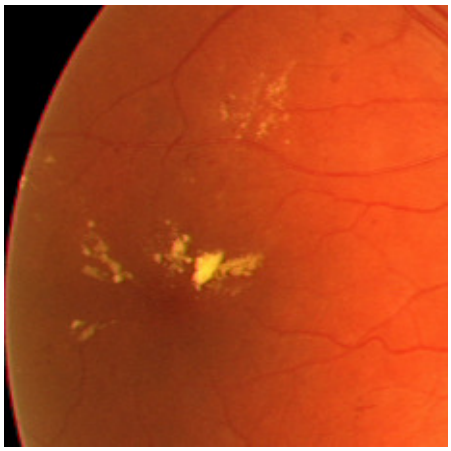
APPENDIX F: Images of high Mahalanobis' distance

These images represent images in the BristolDB for which the Mahalanobis' distances were notably high in the dataset.

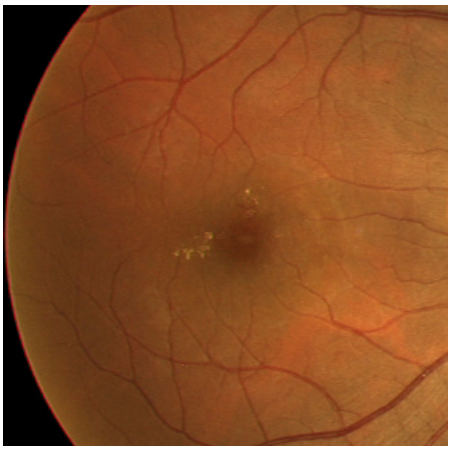


(a)

(b)



(c)



(d)

Figure 15: Images of high Mahalanobis' distance. a) high MD for both exudates and background, b) high MD for exudates, c) high MD for background, d) very high MD for background.